Bayesian Integrated analysis of multiple types of rare variants to infer risk genes for schizophrenia and other neurodevelopmental disorders

Hoang T Nguyen[1], Douglas M Ruderfer[1,2], Gulio Genovese[3,4], Menachem Fromer[1,5], Pamela Sklar[1,3], Xin He[6], Patrick F Sullivan[7,8], Shaun M Purcell[1,3,9], and Eli A Stahl[*] [1,3]

[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, New York 10029, USA
[2]Vanderbilt
[3]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA.
[4]Department of Genetics, Harvard Medical School, Massachusetts, USA.
[5]Verily Life Sciences
[6]Department of Human Genetics, University of Chicago, Chicago, IL 60637, USA.
[7]Departments of Genetics and Psychiatry, University of North Carolina, Chapel Hill, North Carolina 27599-7264, USA.
[8]Karolinska Instituet
[9]Sleep Center, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, USA.

March 19, 2017

Abstract

Integrating rare variation from family and case/control studies has successfully implicated specific genes contributing to risk of autism spectrum disorder (ASD). In schizophrenia (SCZ), however, while sets of genes

[*]eli.stahl@mssm.edu

have been implicated through study of rare variation, very few individual risk genes have been identified. Here, we apply hierarchical Bayesian modeling of rare variation in schizophrenia and describe the proportion of risk genes and distribution of risk variant effect sizes across multiple variant annotation categories. Briefly, we developed a pipeline based on the previous work used in ASD studies to jointly estimate genetic parameters for one or multiple combined populations of any disease. We applied this method to the largest available collection for rare variants in schizophrenia (1,077 families, 6,699 cases and 13,028 controls). We defined five variant annotation categories: disruptive (nonsense, frameshift, essential splice site mutations), damaging (predicting damaging by seven algorithms), silentCFPK (silent mutations within frontal cortex-derived DHS) de novo mutations, and disruptive and damaging missense case/control singletons. We estimated that 8.01% of approximately 20,000 genes are risk genes (95% credible interval, CI, 4.59-12.9%), with mean effect sizes (95% CIs) of 12.25 (4.8- 22.22) for disruptive de novos, 1.44 (1-3.16) for missense damaging de novos, and 1.22 (1-2.16) for silentCFPK de novos. The mean effect sizes of damaging and disruptive singleton variants for three case-control populations were 2.09 (1.04-3.54), 2.44 (1.04, 5.73) and 1.04 (1-1.19) respectively. Our analysis identified only two known SCZ risk genes with FDR<0.05: SETD1A and TAF13; and two other genes with FDR < 0.1: RB1CC1 and PRRC2A. We further used FDR to analyze the enrichment of several candidate gene sets. Significant results are observed in gene sets previously implicated in schizophrenia (including in a subset of these data): FMRP targets, promoter targets of CHD8, splice targets of RBFOX, constrained genes, genes with de novo mutations in ASD and developmental disorder, synaptic genes (need to describe some 'abnormal' genes). Novel result was observed for essential genes which were found significantly with autism genes in a recent study. We also applied the pipeline to infer genetic parameters for a total of 10,792 families and 4,058 cases/controls of four other neurodevelopmental disorders: autism spectrum disorder (ASD), intellectual disorder (ID), developmental disorder (DD) and epilepsy (EPI). The predicted proportions of risk genes in these diseases were smaller than that in SCZ ($< 5\%$ for all diseases), and larger in ASD (5%) than in the other disorders ($< x\%$). We report 164 and 58 genes with FDR < 0.05 for DD and ID, respectively. Of these, 101 of 161 and 15 of 58 genes are not currently known DD and ID genes. Overall, our results in schizophrenia replicate those of previous studies, confirming the robustness of our approach. Our method is able to identify novel risk genes for SCZ as well as for other diseases. We conduct power analyses under our inferred model to quantify the improvement in power to detect risk genes as more data become available.

Contents

1 Introduction

Schizophrenia (SCZ) is a complex psychiatric disorder particularly characterized by psychosis, and by positive, negative and cognitive symptoms, with severe medical and social-functioning comorbidities and high public health costs. Despite high reduction of reproductive fecundity and a lifetime risk of 0.7%, a very high heritability of 60-80% has been observed for the disease (Lichtenstein et al., 2009; Sullivan et al., 2003). The genetic architecture of SCZ is highly polygenic with the contribution of common, rare and denovo variants (Purcell et al., 2014; Fromer et al., 2014; Singh et al., 2016; Stefansson et al., 2009; Purcell et al., 2009). With the production of high-quality next-generation sequencing data, the genetics of schizophrenia and other diseases can be increasingly better characterized, especially for rarer variants.

Rare variants in case/control samples and de novo mutations have been successfully leveraged to identify specific SCZ risk genes (Singh et al., 2016; Takata et al., 2016), or to implicate gene sets for this disease (Purcell et al., 2014; Fromer et al., 2014). However, the genetic architecture of SCZ for rare variants and de novo mutations has not been inferred. Such analyses could help gain further insight into this disease, for example by using the estimated number of risk genes to calibrate gene discovery false discovery rates, or by using the distribution of effect sizes to estimate power for rare variant association studies. A better understanding of our certainty in sets of risk genes for SCZ will result in a better picture of biological pathways specific for the disease.

Here, we aim to develop a pipeline for integrative analysis of case-control rare variants and de novo mutations in order to infer genetic architecture and identify risk genes for SCZ as well as other diseases. To do this, we extend a hierarchical model Bayesian analysis framework (TADA, Transmission And De novo Association) which was developed for autism spectrum disorder (ASD) (He et al., 2013). The new framework (`extTADA`, extended Transmission And De novo Association) can be used to analyze only de novo data, only case-control data or the combination of both. `extTADA` uses all variant classes to jointly estimate genetic parameters (therefore it assumes that all classes play important roles in the genetic architecture of the tested disease). In `extTADA`, a conditional model for case-control sample frequency allows rapid analysis without population frequency parameters (which are very poorly estimated for rare variants), facilitating estimation of parameters via Markov Chain Monte Carlo (MCMC). In addition, we designed `extTADA` for the analysis of data from multiple population samples.

In this study, we used `extTADA` to analyze the largest available exome-sequence dataset, including 19,727 (6,699+13,028) case+control samples and 1,077 trio/quad families for SCZ. We estimated mean relative risks (RRs) of different variant annotation categories as well the proportion of risk genes for disease. Based on these meta-analysis, the false discovery rate (FDR) information of all genes was used to calculate the enrichment of known gene sets and novel gene sets. Analysis of separate classes of variants/mutations in terms of

annotation and rarity helps provide a detailed picture of the disease's rare variant genetic architecture. Finally, we used available data for other neurodevelopmental diseases: intellectual disability (ID), autism spectrum disorder (ASD), epilepsy (EPI) and developmental disorder (DD), totaling 10,792 trios and 4058 cases/controls. We are able to identify additional new significant genes for ID and DD based on `extTADA` results.

The pipeline is publicly available at https://github.com/hoangtn/extTADA.

2 Results

The `extTADA` pipeline and its comparison with `TADA` is described in Figure S1. Figure S2 summarises the workflow of the current study. As presented in Figure S2, variants/mutations in this study were divided into categories: synonymous, missense, loss-of-function (LoF), missense damaging (MiD), silent mutations within frontal cortex-derived DHS (silentCFPK), and then three main categories were used in the analysis: MiD, loF and silentCFPK.

2.1 The `extTADA` pipeline

We used a Bayesian approach to integrate de novo (DN) and case control (CC) rare variant data to infer genetic architecture parameters and to identify risk genes under a model with additive to dominant deleterious risk alleles. The framework is extended from the Transmission and Disequilibrium Association (TADA) model proposed by He et al. (2013); De Rubeis et al. (2014), as shown in Figure S1. Primary extensions to the TADA model facilitate joint Bayesian inference of rare variant genetic architecture model parameters (including the risk gene mixture proportion π, which is fixed in `TADA`), and include a likelihood formulation in which all variant categories contribute to the inference, which also allows inference based on multiple samples. `extTADA` also uses an approximate expression for case-control data probability that eliminates population allele frequency parameters, and controls the proportion of protective variants by constraining effect size distribution scale parameters. We used the same symbols for parameters as those used in He et al. (2013); De Rubeis et al. (2014) in the following sections. For comparison, we also described in detail methods originally presented in the TADA papers (He et al., 2013; De Rubeis et al., 2014).

In summary, for a given gene, all variants of a category (e.g., LoF, MiD) were collapsed and considered as a single count. Let q, γ and μ be the population frequency of genotype (for case/control or transmitted/nontransmitted data), relative risk (RR) of variants associated with the disease, and mutation rates of de novo mutations respectively. At each gene, two hypotheses $H_0 : \gamma = 1$ and $H_1 : \gamma \neq 1$ were compared. A fraction of the genes π, assumed to be risk genes, were represented by the H_1 model. Under this model, relative risks (γ) were assumed to follow a probability distribution. The model H_0 described non-risk genes, and relative risks (γ) of genes were set to equal 1. As in He

et al. (2013), we modeled de novo (x_d) and case (x_{ca}) control (x_{cn}) data as Poisson distributions and their hyper parameters as Gamma distributions. In addition, we used Gamma distributions as priors for hyper parameters, a Beta distribution to constrain π to less than 0.5, and a nonlinear function to constrain mean RRs and their variances as described in Table 1.

$x_{dn} \sim P(2N_{dn}\mu\gamma_{dn})$	$\gamma_{dn} \sim Gamma(\bar{\gamma}_{dn} * \beta_{dn}, \beta_{dn})$	$\bar{\gamma}_{dn} \sim Gamma(\bar{\bar{\gamma}}_{dn}, \bar{\bar{\beta}}_{dn})$ $\beta_{dn} = e^{a*\bar{\gamma}_{dn}^{b}+c}$
$x_{ca} \sim P(N_1 q\gamma_{cc})$	$\gamma_{cc} \sim Gamma(\bar{\gamma}_{cc} * \beta_{cc}, \beta_{cc})$ $q \sim Gamma(\rho, \nu)$	$\bar{\gamma}_{cc} \sim Gamma(\bar{\bar{\gamma}}_{cc}, \bar{\bar{\beta}}_{cc})$ $\beta_{cc} = e^{a*\bar{\gamma}_{cc}^{b}+c}$ $\frac{\rho}{\nu} = mean(\sum(x_{cn} + x_{ca}))$ $\nu = 200$
$x_{cn} \sim P(N_0 q)$	$q \sim Gamma(\rho, \nu)$	$\frac{\rho}{\nu} = mean(\sum(x_{cn} + x_{ca}))$ $\nu = 200$
$\pi \sim \ Beta(1, 5)$		

Table 1: Parameter information used in all analyses. N_{dn}, N_1, N_0 are sample sizes of families, cases and controls respectively. $\bar{\gamma}$ is mean RRs and β controls the dispersion of $\bar{\gamma}$. $\bar{\bar{\gamma}}$ and $\bar{\bar{\beta}}$ are priors for $\bar{\gamma}$ and are set in advance (they are inferred from simulation data). β is inferred from the equation $e^{a*\bar{\gamma}^{b}+c}$ inside the estimation process with a = 6.83, b = -1.29 and c = -0.58.

At each gene, a Bayes Factor (BF) was calculated for each category to compare models H_1 and H_0 ($BF = P(data|H_1)/P(data|H_0)$). If there were multiple categories, then BF_{gene} was the product of BFs for all categories. Data could be from different populations; therefore, we extended the BF_{gene} for multiple populations as the product of BFs of all populations as in Equation 1. To infer significant genes, BFs were converted to false discovery rates (FDRs) using the approach in Newton et al. (2004).

$$BF_{gene} = \left[\prod_{h=1}^{Ndn_{pop}} (\prod_{k=1}^{Cdn} BF_{dn_{hk}})\right] \left[(\prod_{a=1}^{Ncc_{pop}} \prod_{b=1}^{Ccc} BF_{cc_{ab}})\right] \quad (1)$$

in which: Ndn_{pop}, Ncc_{pop} are the number of populations of DN and CC data respectively; C_{dn}, C_{cc} are the number of categories of DN and CC data in that order.

To calculate BFs in Equation 1, hyper parameters of different populations, and categories (assuming ϕ_1 and ϕ_0 for H_1 and H_0 respectively) in Table 1 were needed in advance. Therefore, they were estimated simultaneously based on a

mixture model of two hypotheses as in Equation 2.

$$P(x|\phi_1, \phi_0) = \prod_{i=1}^{Gene\ numbers} [\pi P_{1i} + (1-\pi)P_{0i}] \tag{2}$$

where P_{1i} and P_{0i} at the i^{th} gene were calculated across populations and categories as follows:

$$\begin{aligned}
P_{1i} &= P_{1i}(x_i|\phi_1) \\
&= \left[P_{1i(dn)}(x_{i(dn)}|\phi_{1(dn)})\right]\left[P_{1i(cc)}(x_{i(ca)}, x_{i(cn)}|\phi_{1(cc)})\right] \\
&= \left(\prod_{h=1}^{Ndn_{pop}} \prod_{k=1}^{Cdn} P_{1i(dn)_{hk}}(x_{i(dn)_{hk}}|\phi_{1(dn)_{hk}})\right)\left(\prod_{a=1}^{Ncc_{pop}} \prod_{b=1}^{Ccc} P_{1i(cc)_{ab}}(x_{i(ca)_{ab}}, x_{i(cn)_{ab}}|\phi_{1(cc)_{ab}})\right) \\
P_{0i} &= P_{0i}(x_i|\phi_0) \\
&= \left[P_{0i(dn)}(x_{i(dn)}|\phi_{0(dn)})\right]\left[P_{0i(cc)}(x_{i(ca)}, x_{i(cn)}|\phi_{0(cc)})\right] \\
&= \left(\prod_{h=1}^{Ndn_{pop}} \prod_{k=1}^{Cdn} P_{0i(dn)_{hk}}(x_{i(dn)_{hk}}|\phi_{0(dn)_{hk}})\right)\left(\prod_{a=1}^{Ncc_{pop}} \prod_{b=1}^{Ccc} P_{0i(cc)_{ab}}(x_{i(ca)_{ab}}, x_{i(cn)_{ab}}|\phi_{0(cc)_{ab}})\right)
\end{aligned}$$

To simplify the estimation process in Equation 2, we approximated the original model $P(x_{ca}, x_{cn}|H_j)$ using a new model in which case counts were conditioned on total counts: $P(x_{ca}|x_{ca} + x_{cn})$ (Figure S1). A Markov Chain Monte Carlo (MCMC) method was used to sample parameters. We extracted samples from the MCMC process and used their modes for all analyses. In addition, all credible intervals (CIs) of 5% and 95% were also used in the inferences of parameters.

2.2 Test `extTADA` on simulated data.

We simulated data for three situations: one CC category, two CC categories, and one CC category and one DN category. For CC data, the original TADA model was used to simulate data, and then the CC approximate model was used in the estimation process.

2.2.1 Test `extTADA` for only CC data

To test the approximate CC model for different parameter values, we simulated two situations: one CC class and two CC classes. We tested on different sample sizes in order to evaluate the performance of the approximate model.

Overall, high correlations (~ 1) were observed for both situations (Figure 1 and 2). For sample sizes > 3000 cases, good estimation was observed, but slight over estimation was observed for the sample size of 1092, especially for risk-gene proportions.

An additional analysis was carried out to assess the performance of specific simulated values. Correlations were calculated for each mean RR and π value. For one CC class, mean RRs were estimated well by the model with correlations ~ 1 (Figure S3). However, the proportion of risk genes was affected by

mean RRs. They were estimated well when mean RRs were between 1.5 and 3.5, but underestimated with smaller mean RRs and slightly overestimated with larger mean RRs (Figure S3). For two CC classes, high correlations (≥ 0.97) between simulated and estimated values were seen for all parameters. In addition, small mean RRs of a given class did not directly affect the estimated values of proportions of risk genes (Figure S4).

The issue of poor estimation for one class, but good estimation for $>$ one class was expected. This was an advantage of using multiple classes compared to using only one class in the estimation process when the clustering signal was not very strong. Small mean RRs could result in difficulties in the calculation process to differentiate between a risk gene (mean RR > 1) and a non-risk gene (mean RR ~ 1). If one class was used then many risk genes would be considered to be non-risk genes. If more than one class was used, such risk genes would be assigned as genuine risk genes due to the information available from other classes.

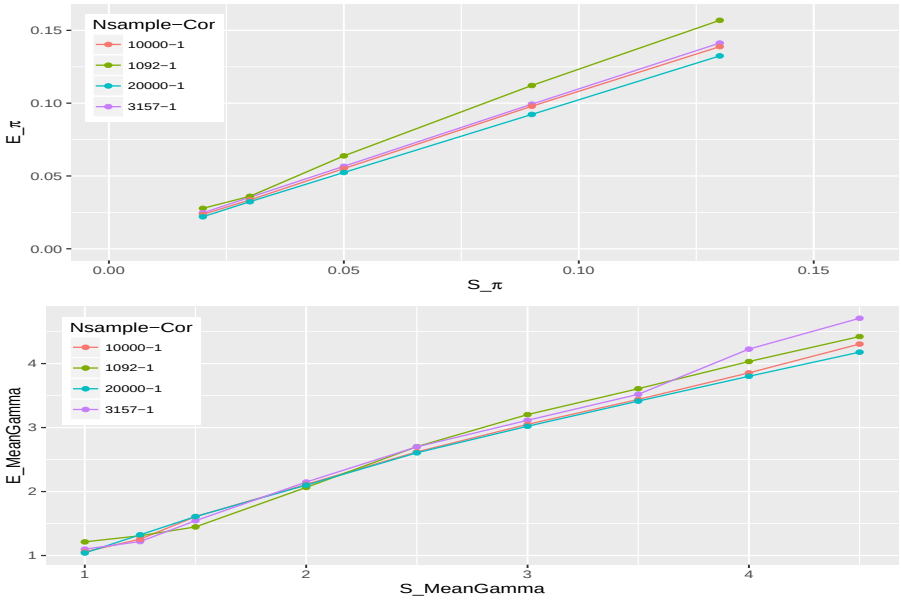

Figure 1: Correlations between estimated and simulated values for one CC class with different sample sizes. X and Y axes describe simulated (S) and estimated (E) values respectively. The top picture is for mean relative risks (MeanRRs) while the bottom picture is for the proportion of risk genes (π). Legends show sample sizes and correlations.

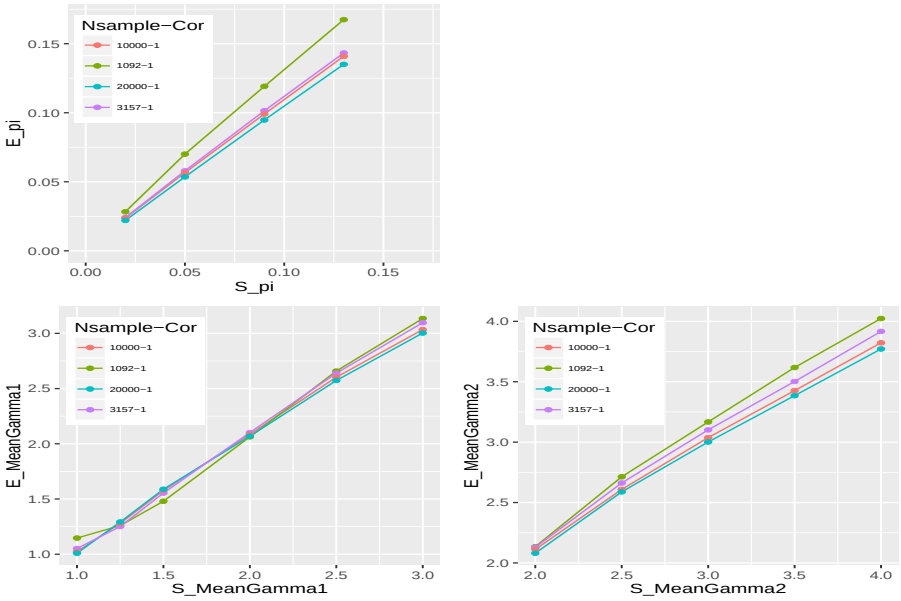

Figure 2: Correlations between estimated and simulated values for two CC class with different sample sizes. X and Y axes describe simulated (S) and estimated (E) values respectively. A range of mean relative risks for two classes (MeanRR1 and MeanRRs) and risk-gene proportions (π) were used in the simulation process. Legends show sample sizes and correlations.

2.2.2 Evaluate the performance on integrated DN and CC data

We next evaluated the main model used in this study as described in Equation 2 by combining different values of DN and CC data. We calculated the ratio of simulated risk genes inside the gene sets determined with different FDR thresholds. These values were called observed FDR (oFDR, see the Methods). Overall, high correlations were observed between FDRs and oFDRs (Figure 3). For high π values, oFDRs and FDRs were nearly equal. For small π values (e.g., $\pi = 0.02$) oFDRs were higher than FDRs when mean RRs of de novo data were small. In addition, oFDRs were equal to zero for many cases when FDRs were small (Figure 3). The reason for this was because there were only a few significant genes with small π values and small mean RRs.

Further investigation into estimated and simulated values was also carried out. We saw high correlations between estimated and simulated values (Table S1). Slightly over and under estimation were observed for mean RRs and proportions of risk genes. This was expected and usually did not have a large effect on the final results as discussed in the previous work (He et al., 2013) as well as shown in TPRs and FPRs results above (Figure 3).

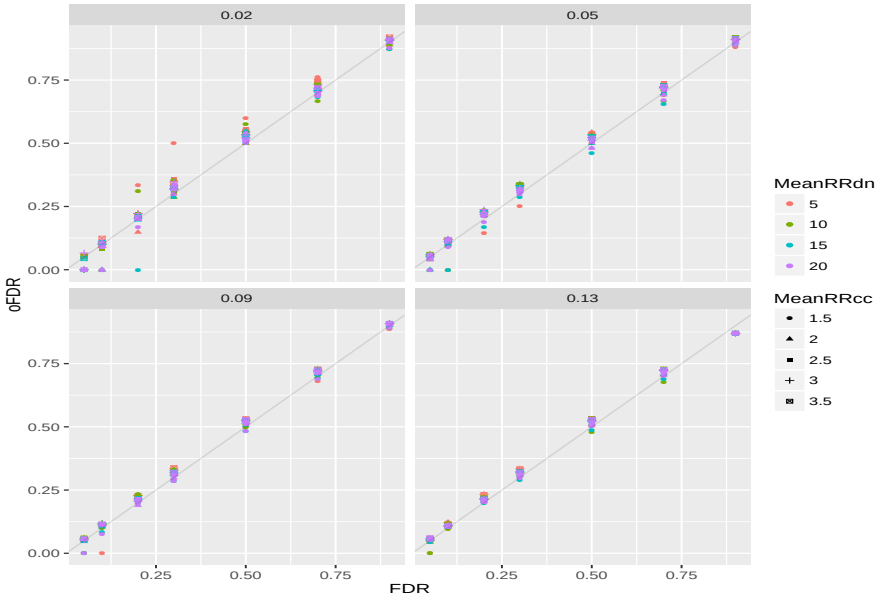

Figure 3: Observed false discovery rates (oFDR) with different FDR thresholds for different π values (0.02, 0.05, 0.09 and 0.13).

2.3 Schizophrenia data sets

The SCZ data sets were refined into non heterogeneous populations and then `extTADA` was used to integratively analyze. The CC data were also separated into variants present/absent in the Exome Aggregation Consortium (ExAC) (Lek et al., 2015), termed InExAC and NotExAC respectively. In total, there were 6,699 cases, 13,028 controls, 1077 trio/quad families used in this analysis. In addition, we used 731 families including uninfected siblings as controls for de novo data (Table S2, Figure S2).

2.3.1 Extract data sets to analyze integratively

De novo mutations and case-control variants were tested to select classes and samples for the meta analysis of the `extTADA` pipeline. For case-control data, there were multiple populations and the data were also sequenced from different centers. Therefore, we classified the data as follows. Firstly for the 4,929 cases and 6,232 controls of the Sweden population, we clustered all cases and controls into different groups and then calculated p values (using the *lm, glm* function in R) between cases and controls by adjusting and not adjusting for covariates. We aimed to obtain populations for which results would not be much different when using or excluding covariates. We chose a clustering process that resulted in highly similar results when either incorporating or excluding covariates. The

clustering process divided the data set into three groups as in Figure 4: Group 1 = 3,157 cases + 4,672 controls; Group 2 = 681 cases + 367 controls; and Group 3 = 1,091 cases + 1,193 controls. Only Groups 1 and 3 were used in the next stage because Group 2 showed a large difference between the two sets of adjusted and unadjusted results. In addition, similar to Genovese et al. (2016), InExAC variants did not show significant differences between cases and controls but NoExAC variants showed some significant differences (Figure 4). Secondly, from the data of the UK10K project (Singh et al., 2016), only case-control data from the UK population was used. Because high significance was observed for NoExAC variants, we therefore used only NoExAC variants in all main analyses. However, we also used both InExAC and NoExAC variants to form a better picture of the genetic architecture of the disease in a comparison section.

For de novo mutations, we calculated the sample-adjusted ratios of mutation counts between 1,077 cases and 731 controls. Similar to Takata et al. (2016), the highest ratio was observed for silentCFPK (2.57), followed by MiD (2.3), LoF (1.83) and missense, silent ($\sim$ 1.3) mutations (Figure S5). Three classes (LoF, silentCFPK, and MiD) were used in next steps.

2.3.2 Integrated analysis of de novo mutations and case-control variants

Three categories of de novo mutations and two categories of case/control variants were used in the integration analysis process. They included LoF, MiD and silentCFPK denovo mutations; and LoF and MiD case-control variants. We only used MiD and LoF variants from case-control data. In addition, these variants also showed similar enrichment in our current analysis as shown in Figure 4; therefore MiD and LoF case-control variants were pooled in order to maximize the case-control information. There were four populations in total. The four populations comprise one de novo population and three case-control populations: two from the clustering process of the Sweden data and the data from the UK10K project.

`extTADA` automatically estimated all genetic parameters of SCZ based on the current data set. All chains showed convergences (Figure S6). The proportion of risk genes was approximately 8.01% (CI = (4.59%, 12.9%)). Regarding de novo classes, LoF had the highest mean RR, which was 12.25 with CI = (4.78, 22.22). Two other de novo classes had approximate mean RRs: 1.22 (CI = (1, 2.16)) and 1.44 (CI = (1, 3.16)) for silentCFPK and MiD respectively. For MiD+LoF case-control variants, two Sweden populations had nearly equal values of mean RRs: 2.09 (CI = 1.04, 3.54)) and 2.44 (CI = 1.04, 5.73)) respectively; however the signal was weak for the UK population with mean RR only approximately 1.04 (CI = 1, 1.19), (Table 2, Figure 5).

FDRs of genes were also reported by `extTADA`. There were only four genes (SETD1A, TAF13, PRRC2A, RB1CC1) having FDR < 0.1 (Table S4). Both SETD1A (FDR = 0.0033) and TAF13 (FDR = 0.026) showed individually significant genes. SETD1A had been confirmed as the highest statistically significant

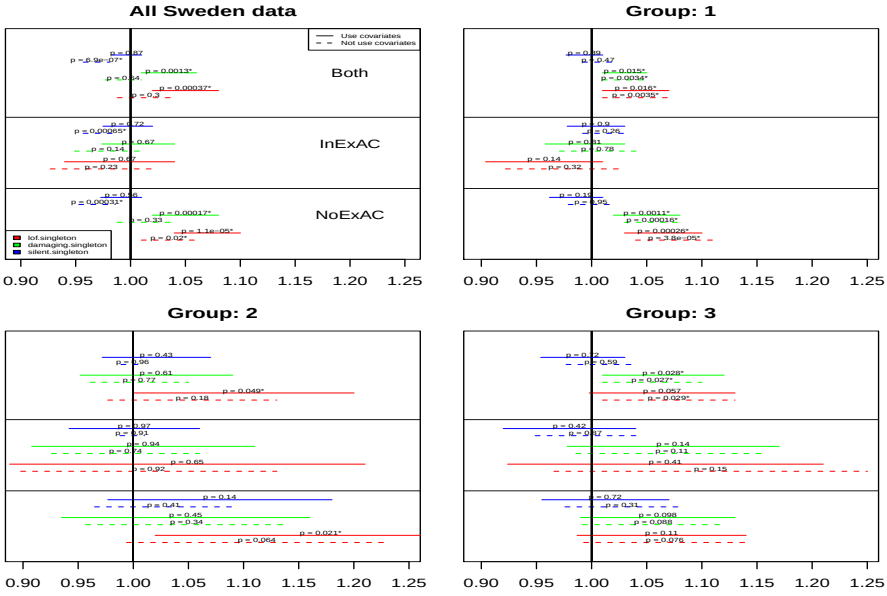

Figure 4: Odds ratios for the analysis of all case-control samples. Top left picture shows odds ratios for all Sweden samples while the three other pictures show odds ratios for three groups after the clustering process. Only group 1 and 3 are used in the current analysis because there are strong differences between results using covariates and not using covariates in group 2. P values were calculated for variants in (InExAC), not in (NoExAC) the ExAC database, and all variants (Both).

gene of SCZ in previous studies (Singh et al., 2016; Takata et al., 2016) while TAF13 was only reported as a potential risk gene in the study of Fromer et al. (2014). Interestingly for the RB1CC1 gene, rare duplications were reported to be associated with SCZ with very high odds ratio (8.58) in the study of Degenhardt et al. (2013), but has not been reported in other studies since. In addition, as discussed by the authors, duplications at this gene were also observed by Cooper et al. (2011) with an odds ratio = 5.29 in a study of 15,767 children with ID and/or DD.

If we increased the FDR threshold to 0.3 as in the previous ASD study of De Rubeis et al. (2014), there were 13 genes (SETD1A, TAF13, RB1CC1, PRRC2A, VPS13C, MKI67, RARG, ITSN1, KIAA1109, DARC, URB2, HSPA8, KLHL17, ST3GAL6, SHANK1, EPHA5, LPHN2, NIPBL, KDM5B, TNRC18, ARFGEF1, MIF, HIST1H1E, BLNK) which were significant with this threshold. Of these genes, EPHA5, KDM5B and ARFGEF1 did not have any de novo mutations (Table S4).

Parameter	Estimated mode	lCI	uCI
SCZ_pi0 (%)	8.01	4.59	12.9
SCZ_meanRR_silentCFPKdenovo	1.22	1.00	2.16
SCZ_meanRR_MiDdenovo	1.44	1.00	3.16
SCZ_meanRR_LoFdenovo	12.25	4.79	22.22
SCZ_meanRR_MiD+LoFccPop1	2.09	1.04	3.54
SCZ_meanRR_MiD+LoFccPop2	2.44	1.05	5.73
SCZ_meanRR_MiD+LoFccPop3	1.04	1	1.19
ASD_pi0 (%)	4.59	3.19	6.01
ASD_meanRR_MiDdenovo	3.67	1.98	8.68
ASD_meanRR_LoFdenovo	23.4	13.63	36.94
ASD_meanRR_LoFcc	4.18	2.04	9.96
ID_pi0 (%)	2.76	2.07	3.7
ID_meanRR_MiDdenovo	28.61	16.18	41.86
ID_meanRR_LoFdenovo	96.04	67.57	130.73
EPI_pi0 (%)	1.65	0.8	3.21
EPI_meanRR_MiDdenovo	47.5	19.77	87.32
EPI_meanRR_LoFdenovo	77	37.19	138.24
DD_pi0 (%)	2.87	2.34	3.49
DD_meanRR_MiDdenovo	22.55	13.19	32.53
DD_meanRR_LoFdenovo	86.53	65.79	111.61

Table 2: Estimated parameters for de novo and case-control SCZ data and four other diseases: ID, EPI, ASD and DD. These results are obtained by sampling 20,000 times of three MCMC chains. The two last columns show the lower (lCI) and upper (uCI) values of CIs.

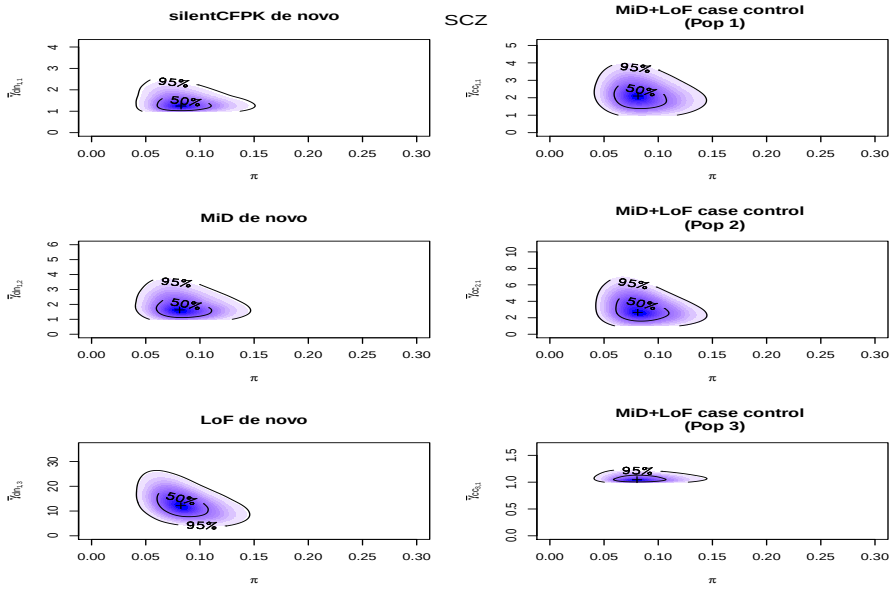

Figure 5: The densities of the proportion of risk genes and mean relative risks for SCZ data. These are obtained after 20,000 iterations of three MCMC chains. The first two case-control populations are derived from the Sweden data set while the third case-control population is the UK population.

2.3.3 Test enrichment of SCZ-risk genes in gene sets

From `extTADA`, we extracted the FDR of each gene and used the information of (1 - FDR) to test the enrichment of gene sets.

2.3.3.1 Top SCZ significant genes from `extTADA` are enriched in known gene sets

We first tested 161 known gene sets (Table S5). Significant results were observed for 60 gene sets which included gene sets reported by Genovese et al. (2016). The most significant gene sets were genes flanking SNPs and Indels of DD and ASD, missense constrained genes, targets of fragile X mental retardation protein (FMRP) genes, loss-of-function intolerant genes (pLI09), RBFOX13, CELF4, CHD8 promoter targets, RBFOX2, abnormal behavior, abnormal sensory capabilities|reflexes|nociception, abnormal motor capabilities|coordination|movement, abnormal emotion|affect behavior, abnormal nervous system morphology, ARC (all $p < 8.0e\text{-}04$), Table 3). The very significant result for the abnormal behavior gene set replicated the latest result of Pardinas et al. (2017). Furthermore, the gene set obtained from common variants (Pardinas et al., 2017) was also enriched in our meta-analysis information ($p < 5.3e\text{-}03$). This showed the con-

vergence (even not very strong in our current study) between rare-variant and common-variant signal. The significant results of gene sets of other neurodevelopmental disorders (NDs: DD, ASD, ID, EPI) showed overlapping signal between SCZ and these NDs. In addition, novel results were observed for essential genes, known epilepsy genes (p < 1.7e-03, Table 3). The essential gene set was just reported recently by Ji et al. (2016) as ASD risk genes.

2.3.3.2 Top SCZ genes are enriched in other gene sets from a data-driven approach

To identify more gene sets enriched in the `extTADA` results for SCZ, we tested 1745 gene sets from different data bases and the 161 gene sets above. Significant results were observed in 161 gene sets including 36 gene sets in the above 161 gene sets (Table S6) . The top significant gene sets in the known gene sets still had the lowest p values in these results (Table S6).

The most interesting result for GO gene sets was GO:0051179/localization (p = 5.2e-03). This gene set was reported by Murphy and Bentez-Burraco (2016) in a study relating to language evolution and SCZ (Table S6).

Gene set	P value	Adjusted p value	Gene set	P value	Adjusted p value
DD.allDenovoMiDandLoF	1e-07	2.0e-06	PSD-95_(core)	0.0016	8.3e-03
celf4	1e-07	2.0e-06	abnormal_excitatory_postsynaptic_currents	0.0017	8.3e-03
constrained	1e-07	2.0e-06	abnormal_learning|memory|conditioning	0.0017	8.3e-03
pLI09	1e-07	2.0e-06	abnormal_associative_learning	0.0024	1.1e-02
rbfox13	1e-07	2.0e-06	abnormal_social_investigation	0.0027	1.2e-02
rbfox2	1e-07	2.0e-06	abnormal_synapse_morphology	0.0027	1.2e-02
FMRP_targets	1e-07	2.0e-06	abnormal_neuron_morphology	0.0028	1.2e-02
abnormal_behavior	1e-07	2.0e-06	abnormal_neuron_physiology	0.0032	1.4e-02
abnormal_sensory_capabilities|reflexes|nociception	2.2e-06	3.9e-05	abnormal_brain_morphology	0.0045	1.9e-02
AST.allDenovoMiDandLoF	2.6e-06	4.2e-05	abnormal_CNS_synaptic_transmission	0.0055	2.2e-02
abnormal_motor_capabilities|coordination|movement	3.2e-06	4.7e-05	PSD_(human_core)	0.0064	2.5e-02
chd8.human_brain	9e-06	1.2e-04	abnormal_aggression-related_behavior	0.0073	2.8e-02
abnormal_emotion|affect_behavior	1.2e-05	1.5e-04	abnormal_brain_size	0.0083	2.9e-02
abnormal_nervous_system_morphology	2.7e-05	3.1e-04	abnormal_consumption_behavior	0.0084	2.9e-02
ARC	7.5e-05	8.0e-04	abnormal_parental_behavior	0.0079	2.9e-02
synaptome	0.00012	1.2e-03	abnormal_spatial_learning	0.0081	2.9e-02
abnormal_social|conspecific_interaction	0.00013	1.2e-03	abnormal_forebrain_morphology	0.0093	3.2e-02
essentialGenes	0.00019	1.7e-03	abnormal_innervation	0.0099	3.3e-02
list.EPI.43genes.2017.Epi4K.2017	2e-04	1.7e-03	abnormal_response_to_new_environment	0.013	4.2e-02
mir137	0.00023	1.8e-03	abnormal_telencephalon_morphology	0.013	4.2e-02
NMDAR_network	0.00023	1.8e-03	abnormal_corpus_callosum_morphology	0.014	4.3e-02
mGluR5	0.00035	2.6e-03	abnormal_discrimination_learning	0.014	4.3e-02
abnormal_fear|anxiety-related_behavior	0.00061	4.2e-03	abnormal_temporal_lobe_morphology	0.014	4.3e-02
abnormal_cued_conditioning_behavior	0.00062	4.2e-03	abnormal_contextual_conditioning_behavior	0.016	4.6e-02
abnormal_synaptic_transmission	0.00069	4.4e-03	abnormal_inhibitory_postsynaptic_currents	0.016	4.6e-02
seizures	0.00072	4.5e-03	abnormal_response_to_novelty	0.016	4.6e-02
abnormal_behavioral_response_to_xenobiotic	0.00077	4.5e-03	abnormal_brain_vasculature_morphology	0.017	4.7e-02
ID.allDenovoMiDandLoF	0.00079	4.5e-03	abnormal_excitatory_postsynaptic_potential	0.017	4.7e-02
Padinas2017_extTable9.genes	0.00095	5.3e-03	Cav2_channels	0.018	4.8e-02
ID.allKnownGenes	0.00099	5.3e-03	abnormal_cerebrum_morphology	0.018	4.8e-02

Table 3: Enrichment of 161 known gene sets from `extTADA` results. These p values were obtained by 10,000,000 simulations, and then adjusted by using the method of Benjamini and Hochberg (1995). The information for these gene sets is summarised in Table S5.

2.3.4 Identify number of risk genes for SCZ studies with different sample sizes

We estimated the number of risk genes using the genetic architecture of SCZ inferred from the current data sets. Different samples sizes (500-20000 and 1000-50000 for families and cases/controls respectively) were simulated. The number of risk genes with $FDR \leq 0.05$ ranged from 0 to 238. Based on this calculation, we would expect > 50 risk genes if the total sample sizes of families and cases (controls = cases) were larger than 14,000 (Figure 6).

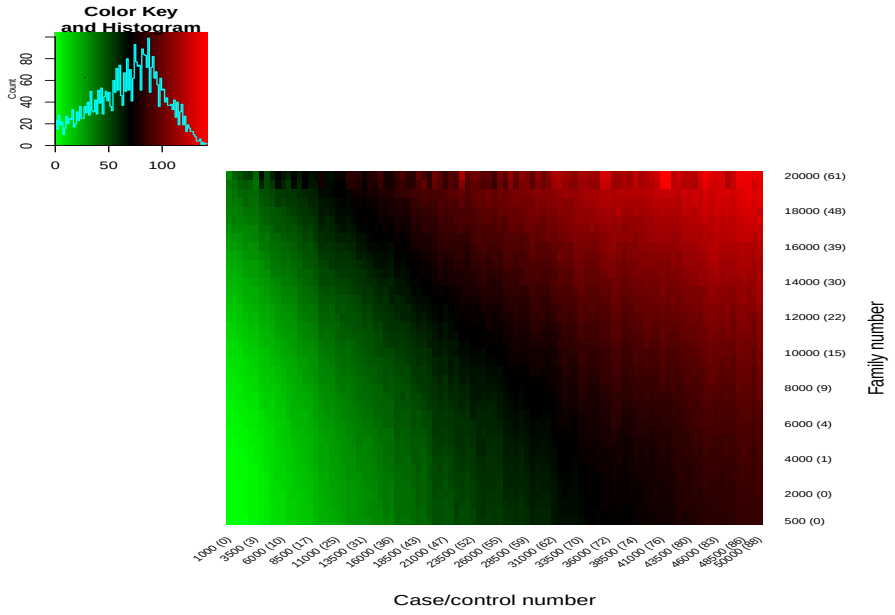

Figure 6: Number of risk genes with different sample sizes based on genetic architecture predicted by extTADA. Case/control number is only for cases (or controls); therefore if Case/control number = 10,000 this means total cases+controls = 20,000. The numbers in brackets show risk-gene numbers if we use only case-control data or only de novo mutation data. These results were obtained by resampling 50 times for each combination of sample sizes.

2.3.5 Test for single classes of SCZ data

To see the genetic architecture of single classes, and also to test the performance of the pipeline for smaller number of classes, `extTADA` was used to estimate parameters separately for four single classes: only single class of silentCFPK, MiD and LoF de novo mutations, and only MiD+LoF case-control variants. Overall, the modes of the proportions of risk genes were less than 8% and CIs were between 0 to 23.68% (Table S7). Regarding the π values, both LoF de

novo mutations and MiD+LoF case-control variants showed strong overlapping CIs and their estimated modes were also not much different: 5.48% (CI = (1.24%, 20.62%)) and 6.9% (CI = (2.96%, 13.59%)) respectively. The π values of silentCFPK and MiD de novo mutations were scattered and we also did not see clear convergent results for these two classes. Mean RRs for these classes were similar to those estimated using the integrative model above, but their ranges were large (Table S7). This situation was similar to what observed in simulated data: using a single category might result in unreliable results, especially for small mean RRs (Figure S3).

2.3.6 Test genetic architecture of SCZ using both InExAC and NoExAC variants

To see any difference in the genetic architecture of SCZ if all variants are used, we pooled all InExAC and NoExAC case-control variants and re-ran `extTADA` the same way we did as for only NoExAC variants above. The genetic parameters were similar to NoExAC based results (Table S8). However, there were only three genes with FDR < 0.1, even though SETD1A and TAF13 were still the top significant genes (Table S9).

2.3.7 Test the influence of mutation rates to the analyzing results of SCZ

The observed counts of silent mutations were lower than expected; therefore, we multiplied all mutation rates by 0.81 to balance between the observed and expected counts of synonymous de novo mutations, and then used `extTADA` to re-analyze the SCZ data. As expected, genetic parameters for de novo mutations were slightly higher than in the original analysis; as a result, the proportion of risk genes also increased to 9.37% (CI = (5.47%, 15.12%)). Parameters of case-control variants were not much different (Table S10). The most interesting finding was that the number of significant genes increased to 3 and 6 genes for FDR < 0.05 and < 0.1 respectively. All top significant genes in the original analysis were among the top genes here (Table S11).

2.4 Estimate genetic parameters of other neurodevelopmental diseases using extTADA

We also used the current pipeline to infer genetic architectures of ASD, EPI, DD and ID. Sample sizes of these diseases are presented in Table S2, Figure S2. Overall, except for EPI, the parameter results of the three other diseases showed strong convergence (Figure 7, Table 2). This might be because EPI had small family numbers, 392 families compared with > 1000 families for other diseases. The numbers of risk genes (π) in these diseases were lower than that of SCZ (Figure 7, Table 2). For ASD, the 95% CI was between 3.19% and 6.01% (mode = 4.59%) which overlapped with the result 550-1000 genes estimated in the original TADA model (He et al., 2013) using only LoF de novo information.

For ID, π was smaller than that of ASD; estimated value was 2.76% (2.1% to 3.7% for the 95% CI). The proportion of risk genes of DD was approximately 2.87% (CI = c(2.34%, 3.49%) which was similar to that of ID. The lowest π value, 1.65% (95% CI = (0.8%, 3.21%)), was observed for EPI. Mean RRs of de novo mutations in these diseases were much higher than those of SCZ. This was expected because of the strong signal of de novo mutations (Table S3). ID, DD had the highest LoF-mutation mean RRs which were 96.04 and 86.53 (CI = (67.57,130.73) and CI = (65.79, 111.6)) respectively. Even though the mean RR of LoF mutations of EPI, which was 77 (CI = (37.19, 138.24)), was lower than that of ID; this value for MiD mutations (47.5, CI = (19.77, 87.32)) was much higher than those of other diseases. The mean RR of EPI had previously been estimated by Epi4K Consortium and Epilepsy Phenome/Genome Project (2013). Their result was = 81 which was in the CI of our current results. For ASD, mean RRs for de novo mutations were much lower than for these other diseases (Figure 7, Table 2).

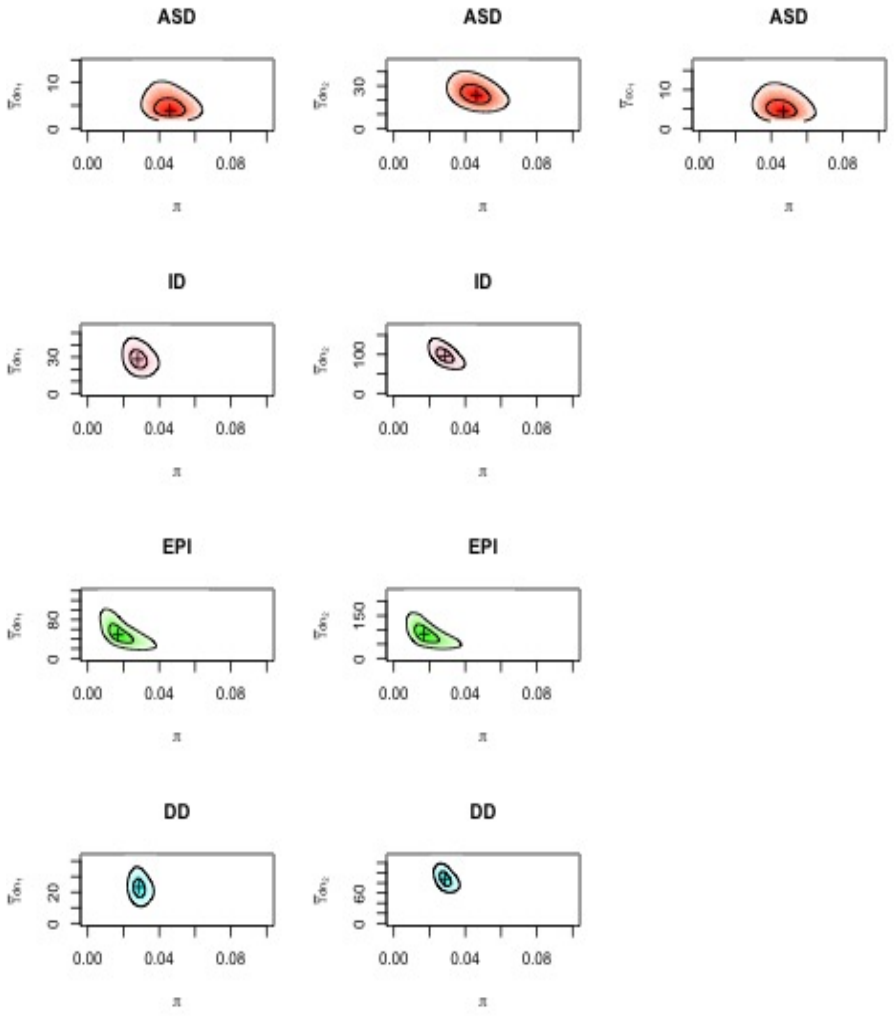

Figure 7: The densities of the proportion of risk genes and π mean relative risks (γ) for ASD, EPI, ID and DD data. For ASD, there are two de novo (dn) classes and one case-control (cc) class. For other diseases, only two de novo classes are publicly available for our current study.

2.5 Novel risk genes in ID and DD diseases.

The results of de novo mutations for ID and DD have been recently reported by using other statistical methods (McRae et al., 2016; Lelieveld et al., 2016); therefore, we aimed to use `extTADA` to identify novel statistically significant

genes from these two latest data sets. The results of `extTADA` for these diseases are shown in Table S12, S13. Genes with FDR ≤ 0.05 were extracted for the two diseases. There were 58 and 164 genes for ID and DD respectively. For the ID disease, 15/58 genes (TCF7L2, USP7, ATP8A1, FBXO11, KDM2B, MED12L, MAST1, MFN1, TNPO2, CLTC, CEP85L, AGO1, AGO2, SLC6A1-AS1, POU3F3) were not on the list of known ID genes as well as 10 significant genes reported by Lelieveld et al. (2016). Of the 15 genes, 6 genes (FBXO11, MFN1, TNPO2, CLTC, CEP85L, AGO2) were very significant (FDR < 0.01). The total MiD and LoF de novo counts of these 15 genes were only between 1 and 3. This might be the reason that they were not reported as significant genes in the recent study of Lelieveld et al. (2016). Regarding the DD, from the 164 discovered genes, only 59 genes were in the list reported by McRae et al. (2016) while 101 genes were novel. Similar to ID, the total MiD+LoF de novo counts of these 101 genes were not high (between 2 and 6). Surprisingly, there were 58 of the 101 genes with FDRs < 0.01.

3 Discussion

In this work, we have built an integrative pipeline (extTADA) between case-control variants and de novo mutations to infer genetic architecture for schizophrenia and four other neurodevelopmental disorders. The pipeline is based on our previous work for autism studies (TADA). Even though `extTADA` is based on TADA, it is a more flexible pipeline. It is using another strategy to obtain genetic parameters. `extTADA` uses the information of all classes of variants to obtain genetic information which is different from He et al. (2013) and De Rubeis et al. (2014) in which LoF de novo mutations play an important role to obtain genetic information. Using a MCMC method, `extTADA` estimates all mean relative risks and the proportion of risk genes simultaneously without using any previous risk gene sets or prior information (Figure S2). We are assuming that different variant classes have similar proportions of risk genes in a large population, based on some convergent results between de novo mutations and case-control rare variants in SCZ (Fromer et al., 2014; Purcell et al., 2014; Singh et al., 2016). One important point is that `extTADA` is able to use for multiple populations. It divides the data into separate small populations and combines the information across these populations. This is an advantage of the new pipeline compared with TADA because sequencing data are usually from different countries/centers. In our current work, we are using `extTADA` to apply for the multi-pop issue of case-control data; however it can be used for multi-pop issues for both de novo data and case-control data. All integrative information is used to test the enrichment of gene sets by using the FDR results of genes. Finally, `extTADA` can be used as TADA if users have prior information relating to the architecture of the tested disease.

Current study's results replicate previous studies and supply more information about SCZ. Firstly, SETD1A is the most significant gene (FDR $\sim 1.5 \times 10^{-3}$) which was reported by Singh et al. (2016); Takata et al. (2016).

In addition, TAF13 is always a significant gene if we adjust mutation rates or add InExAC variants. Of the two genes RB1CC1 and PRRC2A with FDR < 0.1, RB1CC1 was reported as the most significant SCZ gene in a study of copy-number variation in SCZ (Degenhardt et al., 2013). Secondly, significantly overlapping results between the top genes in this study and known gene sets such as constrained, de novo SNPs and INDEL of ASD/DD, FMRP, pLI09 genes show similar trends as the study of Genovese et al. (2016). Apart from those results, many GO gene sets also showed significantly results (Table 3). Thirdly, from this study, rare-variant genetic architecture of SCZ is described in details. It is complex, and may be more complex than those of ASD, ID, DD and EPI (we have both case-control and de novo data for SCZ while only de novo for ID, DD and EPI; therefore the comparison might not totally exact). It can be seen that the number of risk genes for the disease ($\sim 8\%$) is higher than those of the four other diseases (Figure 5 and 7, Table 2). Except for LoF de novo mutations, mean RRs of case-control variants were slightly larger than those of de novo mutations. One important point in our current results is that we see the convergence in the proportion of risk genes between the LoF de novo mutations and case-control variants (Table S7). Based on this information and SCZ case/control rare variants have been shown enrichment in genes containing SCZ de novo mutations (Genovese et al., 2016; Purcell et al., 2014), this shows that the SCZ-gene information is highly overlapping between disruptive de novo mutations and case-control variants. We may probably see a similar trend for the two other classes of de novo mutations if family numbers are large. Finally, we also see that adding InExAC variants for case/control data do not improve the prediction results (Table S8, S9).

Integrating multiple classes to infer genetic parameters can obtain more reliable information. As can be seen in our current work, it was easier to obtain convergent results if multiple classes are integrated into the estimation process, but the results for single classes do not show very strong converges in SCZ (Figure 5, S6, Table S7), and also in simulation data (Figure S3, S4). This situation is very important because `extTADA` (and also TADA) is developed for rare variants. Counts are not high in these classes of variants; therefore using information from multiple sources may result in more reliable results. As can be seen in Figure 1 and 2, , and S4 the information from multiple classes can help obtain more accurate results especially in cases of small effect sizes. We do not compare the results of the current pipeline and TADA on SCZ because `extTADA` uses all available information (all classes) while TADA uses specific class/classes (e.g., LoF variants) to obtain genetic parameters for the disease. In addition, TADA requires prior information from known gene sets; however, there is not specific gene sets for SCZ.

The pipeline can be applied to other diseases to obtain genetic parameters as well as to identify novel significant genes. As be seen in the current study, we are able to use `extTADA` to infer genetic parameters for four other neurodevelopmental diseases ASD, EPI, DD and ID (Table 2, Figure 7). The ASD results of `extTADA` are comparable with those of He et al. (2013); De Rubeis et al. (2014), even though no risk-gene set was used as prior information in the

current estimation process. Regarding the mean RR of LoF mutations of EPI, the result of EuroEPINOMICS-RES Consortium et al. (2014) ($\sim$ 86) is in the current CI and approximate to the estimated value of our study (Table 2). In addition, many novel significant genes which were missed in recent studies are discovered by `extTADA` (101 for DD and 15 for DD).

Apart from that, genetic architecture can be distinct between classes; for example, recently, Sifrim et al. (2016) have shown some evidence for this hypothesis for de novo protein-truncating variants (PTVs) and inherited PTVs in congenital heart defects (CHDs). Therefore, `extTADA` can be flexibly used to infer genetic architecture of any specific classes or combining multiple classes together (e.g., only case-control/inherited variants, only de novo mutations, or LoF or MiD variants), and can be applied to other diseases as we did for intellectual disorder, epilepsy, developmental disorder and autism spectrum disorder in this study. However, count information should be high in order to obtain reliable results if only some classes are used.

There are limits in the current SCZ study. Firstly, the model is developed to use for the combination of non heterogeneous populations. This causes the case-control sample size smaller after the process of population adjustment. Even though we are assuming that there are not differences between populations for de novo data and do not adjust (similar to previous studies of De Rubeis et al. (2014); He et al. (2013); Singh et al. (2016)), the differentiation between populations might happen in this type of data. Secondly, compared with four other diseases, SCZ de novo counts are small (Table S3) while SCZ family size is not large in this study (1,077 families). Therefore, the de novo signal is probably not comparable with that of case-control signal (Table S7). Finally, we are assuming that de novo mutations and case-control variants are convergent to the same proportion of risk genes. We can definitely see different proportions of risk genes if single class is used in the estimation process because of sample collections, noise of data. If the assumption is violated then the current results may probably not totally reflect genetic architecture of SCZ. However, with overlapping results between de novo mutations and case-control rare variants reported recently in SCZ (Purcell et al., 2014; Fromer et al., 2014; Genovese et al., 2016; Singh et al., 2016) and also in the current study between LoF de novo mutations and MiD+LoF case/control variants (Table S7), this assumption can be reliable.

4 Data and methods

A workflow of all data used in this study is described in Figure S2.

4.1 Data

4.1.1 Simulation data

The simulation method described in the TADA paper (He et al., 2013) was used to simulate only one case-control (CC) class, two CC classes and a combination between one CC and one de novo (DN) class. Different mean RRs and risk-gene proportions were used in this process.

4.1.2 Variant data of SCZ, ID, DD, EPI and ASD

High-quality variants were obtained from original analyses as described in Table S2. Variants were annotated using Plink/Seq (using RefSeq gene transcripts, UCSC Genome Browser, http://genome.ucsc.edu) as described in Fromer et al. (2014). After that, SnpSift version 4.2 (Cingolani et al., 2012) was used to further annotate these variants using dbnsfp31a (Liu et al., 2015). Variants were groups into different categories. Loss of function (LoF) class comprised of nonsense, essential splice, and frameshift variants. Missense damaging (MiD) were defined as missense by Plink/Seq and damaging by results of all of 7 methods Genovese et al. (2016) from dbnsfp31a: SIFT, $Polyphen2\_HDIV$, $Polyphen2\_HVAR$, LRT, PROVEAN, MutationTaster and MutationAssessor. Recently, Takata et al. (2016) reported significant results for synonymous mutations in regulatory regions; therefore, this category was also analyzed. To annotate synonymous variants within DNase I hypersensitive sites (DHSs) as Takata et al. (2016), the file *wgEncodeOpenChromDnaseCerebrumfrontalocPk.narrowPeak.gz* was downloaded from `http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeOpenChromDnase/` on April 20, 2016. After that, BEDTools (Quinlan and Hall, 2010) was used to intersect silent variants/mutations with the DHSs. Based on analyzing results of Genovese et al. (2016) in which significant signal was seen for singleton variants, only case-control singleton variants were used in this study.

The data from Exome Aggregation Consortium (ExAC) (Lek et al., 2015) were used to annotate variants inside ExAC (InExAC or not private) and not inside ExAC (NoExAC or private). On April 20, 2016, the file *ExAC.r0.3.nonpsych.sites.vcf.gz* was downloaded from `ftp://ftp.broadinstitute.org/pub/ExAC_release/release0.3/subsets/`. After that, BEDTools was used to obtain variants inside (InExAC) or outside this file (NonExAC).

4.1.3 Gene sets

Multiple resources were used to obtain gene sets in our study. Firstly, we used known gene sets with prior evidence for involvement in schizophrenia and autism from a variety of sources. Secondly, to identify possible novel significant gene sets, we collected genes sets from available data bases.

4.1.3.1 Known gene sets

These gene sets and their abbreviations are presented in Table S5.

- Gene sets were enriched for SCZ ultra rare variants which were described detailedly in the study of Genovese et al. (2016): missense constrained genes (constrained) from Samocha et al. (2014), loss-of-function tolerance genes (pLI90) from Lek et al. (2015), RBFOX2 and RBFOX1/3 genes (rbfox2, rbfox13) from Weyn-Vanhentenryck et al. (2014), Fragile mental retardation protein targets genes (fmrp) from Darnell et al. (2011), CELF4 genes (celf4) from Wagnon et al. (2012), synaptic genes (synaptome) from Pirooznia et al. (2012), microRNA-137 (mir137) from Robinson et al. (2015), PSD-95 complex genes (psd95) from Bayés et al. (2011), ARC and NMDA receptors (NMDARs) genes from Kirov et al. (2012), de novo copy number variants in SCZ, ASD, bipolar as presented in Supplementary Table 5 of Genovese et al. (2016).

- Promoter targets of CHD8 from Cotney et al. (2015).

- Known ID gene set was from the Sup Table 4 of Lelieveld et al. (2016) and 10 novel genes reported by Lelieveld et al. (2016).

- Gene sets from MiD and LoF de novo mutations of ASD, EPI, DD, ID.

- The essential gene set from the supplementary data set 2 of Ji et al. (2016).

- Lists of human accelerated regions (HARs) and primate accelerated regions (PARs) (Lindblad-Toh et al., 2011) were downloaded from `http://www.broadinstitute.org/scientific-community/science/projects/mammals-models/29-mammals-project-supplementary-info` on May 11, 2016. The coordinates of these regions were converted to hg19 using Liftover tool (Kent et al., 2002). We used a similar approach as Xu et al. (2015) to obtain genes nearby HARs. Genes in regions flanking 100 kb of the HARs/PARs were extracted to use in this study.

- List of known epilepsy genes was obtained from Supplementary Table 3 of Phenome et al. (2017).

- 134 central nervous system (CNS) related gene sets were obtained from Pardinas et al. (2017). Steps which were used to obtain the gene sets were described in Pocklington et al. (2015).

- List of common-variant genes was obtained from Extended Table 9 of Pardinas et al. (2017).

  We finally obtained 161 gene sets from this step after removing overlapping gene sets between previous studies and the 134 gene sets.

4.1.3.2 Other gene sets

We also used multiple data sets to identify novel gene sets overlapping with the current gene sets. All gene sets from GO data basse (Consortium et al., 2015), and other gene sets collected by The Molecular Signatures Database (MSigDB) (Subramanian et al., 2005): KEGG, REACTOME and C3: motif gene sets. To increase the power of this process, we only used gene sets whole lengths were between 100 to 4995 genes. Totally, there were 1717 gene sets. These gene sets and the above gene sets above were used in this data-drive approach.

4.2 Methods

4.2.1 `extTADA` pipeline: analyze de novo, transmission and case-control data

4.2.1.1 `extTADA` for one de novo population and one case/control population

`extTADA` was summarised in Table 1 and Figure S1 in which $x_d \sim Pois(2N_d\mu, \gamma_{dn}), x_{ca} \sim Pois(qN_1\gamma_{cc})$, $x_{cn} \sim Pois(qN_0)$ and $\gamma_{dn} \sim Gamma(\bar{\gamma}_{dn}\beta_{dn}, \beta_{dn})$, $\gamma_{cc} \sim Gamma(\bar{\gamma}_{cc}\beta_{cc}, \beta_{cc})$, $q \sim Gamma(\rho, \nu)$.

Let K be the number of categories (e.g., missense, MiD, LoF), $x_i = (x_{i1}, .., x_{iK})$ be the vector of counts at the i^{th} given gene. The Bayes Factor for each j^{th} category to test two hypotheses: $H_0 : \gamma = 1$ versus $H_1 : \gamma \neq 1$ was:

$$\begin{aligned} B_{ij} &= \frac{P(x_{ij}|H_1)}{P(x_{ij}|H_0)} \\ &= \frac{\int P(x_{ij}|\gamma,q)P(q|H_1)P(\gamma|H_1)dqd\gamma}{\int P(x_{ij}|\gamma,q)P(q|H_0)P(\gamma|H_0)dqd\gamma} \\ &\underset{Because\ \gamma=1 for\ H_0}{=} \frac{\int P(x_{ij}|\gamma,q)P(q|H_1)P(\gamma|H_1)dqd\gamma}{\int P(x_{ij}|q)P(q|H_0)dq} \end{aligned} \tag{3}$$

In Equation 3, $x_{ij} = x_d$ and $x_{ij} = (x_{ca}, x_{cn})$ for de novo, and case-control data respectively. In addition, the integral across q was not used for de novo data because there was not this parameter in de novo data.

As in He et al. (2013), the BF for the i^{th} gene for combining all categories was:

$$B_i = \prod_{j=1}^{K} B_{ij} \tag{4}$$

To calculate BFs, hyper parameters in Table 1 were needed to know in advance. Let ϕ_{1j} and ϕ_{0j} be hyperparameters for H_1 and H_0 respectively. A mixture model of the two hypotheses was used to infer parameters using information across the number of tested genes (m) as in Equation 5.

$$P(x|\phi_1, \phi_0) = \prod_{i=1}^{m} \left[\pi \prod_{j=1}^{K} P(x_{ij}|\phi_{1j}) + (1-\pi) \prod_{j=1}^{K} P(x_{ij}|\phi_{0j}) \right] \tag{5}$$

The Equation 5 was calculated across categories as described in calculating BFs in Equation 4. The hyperparameters $\phi_{1j} = (\gamma_{j(dn)}, \gamma_{j(cc)}, \beta_{j(dn)}, \beta_{j(cc)}, \rho_j, \nu_j)$ were estimated using a Markov chain Monte Carlo (MCMC) method named Hamiltonian Monte Carlo (HMC) implemented in the `rstan` package (Carpenter et al., 2015; R Core Team, 2015). However, Equation 5 was simplified by removing $q \sim Gamma(\rho, \nu)$ in the estimation process of parameters.

4.2.1.2 Simplified approximate case/control model

For case-control (inheritance) data, $\frac{\rho}{\nu}$ represented for mean of q, and ν controlled the dispersion of q; therefore as in the previous study of De Rubeis et al. (2014), ν was heuristically chosen (in all current study, 200 was used) and $\frac{\rho}{\nu}$ = the mean frequency across genes by using both case and control data.

The case-control model was deployed as follows:

$$P(x_{ca}, x_{cn}|H_j) \quad = P(x_{ca}|x_{ca} + x_{cn}, H_j)P(x_{ca} + x_{cn}|H_j) \tag{6}$$

Because of $x_{ca} \sim Pois(N_1 q \gamma_{cc})$ and $x_{cn} \sim Pois(N_0 q)$, assuming that x_{ca} and x_{cn} were **independent**, the case data could be modelled as:

$x_{ca}|x_{ca} + x_{cn}, H_j \sim Binomial(x_{ca} + x_{cn}, \theta|H_j)$

with $\theta|H_1 = \frac{N_1\gamma_{cc}}{N_1\gamma_{cc}+N_0}$ and $\theta|H_0 = \frac{N_1}{N_1+N_0}$

The marginal likelihood was

$P(x_{ca}|x_{ca} + x_{cn}, H_j) = \int P(x_{ca}|x_{ca} + x_{cn}, \gamma_{cc}, H_j)P(\gamma_{cc}|H_j)d\gamma_{cc}$

Based on simulation results, the first part $P(x_{ca}|x_{ca} + x_{cn}, H_j)$ can be used to approximately infer mean RRs ($\bar{\gamma}_{cc}$); therefore only this part was used in the estimation process in Equation 5.

4.2.1.3 Control of the proportion of protective vs risk variants using the mean and dispersion parameters of relative risks

If $\bar{\gamma}$ and β were small then we would see a high proportion of protective variants. To control for the proportion of protective variants, we tested the relationship between β and $\bar{\gamma}$. We set this proportion very low (0.5%) and built a nonlinear relationship for β and $\bar{\gamma}$ values as in Equation 7 (Figure S7). The *nls* in the *R* version of 3.3.0 (R Core Team, 2016) was used to estimate a, b and c. These estimated values were 6.83, −1.29 and −0.58 respectively.

$$\beta = e^{a*\bar{\gamma}^b + c} \tag{7}$$

4.2.1.4 `extTADA` for multiple populations

To extend the work for multiple populations, we used the same approach as the integration of information across categories in one population. Let Ndn_{pop}, Cdn and Ncc_{pop}, Ccc be the number of populations, categories for de novo and case-control data respectively. The total Bayes Factor of a given gene was the product of Bayes Factors of all populations as in Equation 1, and all hyper parameters were estimated using Equation 2.

4.2.1.5 Predict the number of risk genes

BFs of genes were calculated using Equation 1. The original case-control model was used in this calculation; however, we changed the order of the integral of parameters to not rely on q because the range of this parameter was not frequently known in advance (Sup Information 6.3). After that, the BFs were converted to false discovery rates (FDRs) using the method of Newton et al. (2004) as described in De Rubeis et al. (2014). The number of risk genes could be predicted based on a threshold(s) defined by users.

4.2.2 Use simulation data to test model

To calculate the ability of the model in predicting significant genes, we simulated multiple combinations between one CC category, two CC categories, one CC category and DN one category data. For CC data, the original case-control model in TADA (He et al., 2013) was used to simulate case-control data and then case-control parameters were estimated using the approximate model. The frequency of SCZ case-control LoF variants was used to calculate prior information of $q \sim Gamma(\rho, \nu)$ as described in Table 1. For DN data, we used exactly the original model of TADA in both the simulation and estimation process.

Different sample sizes were used. For CC data, to see the performance of the approximate model, we used four sample sizes: 1092 cases plus 1193 controls, 3157 cases plus 4672 controls, 10000 cases plus 10000 controls, 20000 cases plus 20000 controls. The first two sample sizes were exactly the same as the two sample sizes from Sweden data in current study. The last two sample sizes were used to see whether the model would be better if sample sizes increased. For DN and CC data, we used exactly the sample sizes of the largest groups in our current data sets: family numbers = 1077, case numbers = 3157 and control numbers = 4672.

To see correlations between simulated and estimated parameters, the Spearman correlation method (Spearman, 1904) was used. To see the performance of the estimation process of parameters inside the model, we compared between expected FDRs and observed FDRs (oFDRs).

We defined oFDR for a FDR threshold as follows. Let G be the set of significant genes under the FDR threshold, and n_1 be the length of G. Let n_2 be the number of real risk genes (information from simulated data) inside G. oFDR for the FDR threshold was the ratio of n_2 and n_1 (oFDR = n_2/n_1). Estimated paramters from `extTADA` were used in this calculation.

For each combination of simulated parameters, we re-ran 100 times and obtained the medians of estimated values to use for inferences.

We also used different priors of hyper parameters (e.g., $\bar{\bar{\gamma}}, \bar{\beta}$)) in the simulation process and chose the most reliable priors corresponding with ranges of $\bar{\gamma}$. Because $\bar{\beta}$ mainly controled the dispersion of hyper parameters, $\bar{\bar{\gamma}}$ was set equal to 1, and only $\bar{\beta}$ was tested.

4.2.3 Calculate mutation rates

We used the methodology which was based on trinucleotide context, depth of coverage as described in Fromer et al. (2014) to obtain mutation rates (MTs) for different classes. There were genes whole mutation rates were equal to 0 (0-MT genes). To adjust for this situation for each mutation class, we calculated the minimum MT of genes having this value > 0, then this minimum value divided by 10 was used as MTs of 0-MT genes.

4.2.4 Analyze SCZ data

4.2.4.1 Obtain non-heterogeneous populations for case-control data of SCZ

The case-control data sets were divided into three big populations: Finland, United Kingdom and Sweden. For the Sweden population, this was a large data set and was also sequenced at different centers (Genovese et al., 2016), therefore we divided this population as follows.

A simple combination between a clustering process using a multivariate normal mixture model and a data analyzing strategy using linear and generalized linear models was used to divide the Sweden data into non-heterogeneous populations. Genovese et al. (2016) recently analyzed all case-control data sets by adjusting for multiple covariates: genotype gender of individuals (SEX), 20 principal components (PCs), year of birth of individuals (BIRTH), Aligent kit used in wet-labs (KIT) by using linear regression and generalized linear regression models as in Equation 8. They reported significant results for NonExAC LoF and MiD variants; therefore, this information was used in this step. We defined homogeneous populations as populations which were not much affected by the covariates. Thus, for the populations, analyzing results using Equation 8 (adjusting covariates) would not be much different from those results using Equation 9 (not adjusting covariates). The `mclust` package Version 5.2 (Fraley and Raftery, 1999) which uses a multivariate normal mixture model was used to divide 11,161 samples (4,929 cases and 6,232 controls) into different groups. To see all situations of the grouping process, we used `mclust` with three strategies on 11,161 samples: grouping all 20 PCs, grouping all 20 PCs and total counts, and grouping only the first three PCs. The number of groups were set between 2 and 6. For each clustering time, Equation 8 and 9 were used to calculate p values for each variant category of each group from the clustering results (p1 and p2 respectively); then, Spearman correlation (Spearman, 1904) between p-value results from the two Equations (cPvalue) was calculated. Next, to filter reliable results from the clustering process, we set criteria:

- cPvalue ≥ 0.85 and p-values for NonExAC ≤ 0.005.
- Ratio p1/p2 from Equation 8 and 9 had to between 0.1 and 1.

From results satisfied the above criteria, we manually chose groups which

had similar results between Equation 9 and 8.

$$\begin{aligned} logit(P(SCZ = 1)) &\sim count + countAll + sex + birth + kit + \sum_{i=1}^{20} PC_i \\ count &\sim SCZ + countAll + sex + birth + kit + \sum_{i=1}^{20} PC_i \end{aligned} \tag{8}$$

$$\begin{aligned} SCZ &\sim count \\ count &\sim SCZ \end{aligned} \tag{9}$$

For the data from the UK10K project (Singh et al., 2016), we divided the data into two separate populations England and Finland, and tested NoExAC variants in these populations by calculating population-size-adjusted ratios between cases and controls. The ratios were 0.91 and 0.95 for the UK data. Regarding the Finland data, the ratio for MiD variants was only 0.41 which were extremely low. This could be a special case for the population or might be because of other technical reasons. We did not use this population in the next stage because it showed a different trend with other populations.

4.2.4.2 Estimate genetic parameters for SCZ

De novo mutations and case-control variants from the non-heterogeneous populations were integratively analyzed. Three de novo classes (MiD, LoF and silentCFPK mutations) and two case-control classes (MiD and LoF variants) were used in Equation 5 to obtain genetic parameters for SCZ. Case-control MiD and LoF variants were pooled into one class in the estimation process.

4.2.4.3 Estimate number of risk genes for SCZ

Based on estimated genetic parameters from the data sets available, the number of risk genes were predicted as described in the `extTADA` pipeline above. Different thresholds of FDRs were used to report their corresponding risk-gene numbers.

4.2.4.4 Test enrichment in known gene sets

Based on the `extTADA` results, we tested the enrichment of gene sets by using gene FDRs as follows. At each gene, we calculated pFDR = 1 - FDR. For each tested gene set, we calculated the mean of pFDRs (m_0). After that, we randomly choose gene sets n times (n = 10 millions in this study) from the whole genes and recalculated the means of pFDRs of the chosen gene sets (vector m). The p value for the gene set was calculated as: $p = \frac{length(m[m>m0])+1}{length(m)+1}$. To correct for multiple tests, the p values were adjusted using the method of Benjamini and Hochberg (1995) for all the number of tests.

4.2.4.5 Predict number of risk genes for different sample sizes

Based on the genetic architecture of SCZ, we predicted the number of risk genes for the disease. To simplify the calculation, we assumed that sample sizes of cases and controls were the same and only one de novo and case-control population. In addition, a threshold FDR = 0.05 was used in this process to predict a number of individually significant genes. Therefore, a grid of different simulated counts of family numbers between 500 and 20000 and case/control numbers between 1000 and 50000 were generated. From these simulated counts, we inferred how many risk genes with FDR $\leq$ 0.05.

4.2.4.6 Test for single classes

To have a general picture of all classes, `extTADA` was used to test for single classes (LoF/MiD/silentCFPK de novo mutations, LoF/MiD case-control variants only). All parameters were set as the integration analysis.

4.2.4.7 Test genetic architecture of SCZ using both InExAC and NoExAC variants

To test whether InExAC variants could increase (or decrease) the strength of identifying significant genes, we pooled all InExAC and NoExAC case-control variants and then used `extTADA` to analyze this pooled data set.

4.2.4.8 Test the influence of mutation rates to the analyzing results of SCZ

The de novo data in current study were from different sources; therefore, de novo counts could be affected by differences in coverage, technologies. We therefore tested the analyzing results by adjusting for mutation rates by using synonymous mutations. We divided the observed counts by expected counts (= 2 x family numbers * total mutation rates), and then used this ratio to adjust for all mutation rates. The new mutation rates and the original data (NoExAC) were re-analyzed using `extTADA`.

4.2.5 Use `extTADA` to predict genetic parameters of other neurodevelopmental diseases

Use exTADA, we analyzed the integration architecture of genetics for four other neurodevelopmental diseases: EPI, ID, DD and ASD. For ASD, genetic parameters were estimated simultaneously for both de novo and case-control data. For the three other diseases, the estimation process was only carried out for de novo data because there were not rare case-control data publicly available.

4.2.6 Infer parameters using MCMC results

The `rstan` package (Carpenter et al., 2015) was used to run MCMC processes. For simulation data, 5,000 times and a single chain were used. For real data,

20,000 times and three independent chains were used. In addition, for SCZ data we used two steps to obtain final results. Firstly, 10,000 times were run to obtain parameters. After that, we used Equation 7 to calculate β values from estimated mean RRs. Finally, `extTADA` was re-run 20,000 times on the SCZ data with calculated β values set as constants to re-estimate mean RRs and the proportions of risk genes. For each MCMC process, a burning period = a half of total running times was used to assure that chains did not rely on their initial values. For example, we ran and removed 2,500 burning times before the 5,000 running times for simulation data.

We just chose 1,000 samples of each chain from MCMC results to do further analyses. For example, with a chain with 20,000 run times, the step to obtain a sample was 20 run times. For all estimated parameters from MCMC chains, the convergence of each parameter was diagnosed using the estimated potential scale reduction statistic ($\hat{R}$) introduced in `Stan` (Carpenter et al., 2015). To produce heatmap plots, modes as well as the credible intervals (CIs) of estimated parameters, the Locfit (Loader, 2007) was used. The mode values were used as our estimated values for other calculations.

5 Acknowledgements

Research reported in this paper was supported by the Office of Research Infrastructure of the National Institutes of Health under award number S10OD018522. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

6 Supplementary information

6.1 Sup Table

Parameter		Q50	Q5	Q95
π	0.02	0.0224	0.0125	0.0253
	0.05	0.0535	0.0351	0.0611
	0.09	0.0965	0.0752	0.1063
	0.13	0.1381	0.11	0.149
$\bar{\gamma}_{DN}$	5	4.265	3.5608	4.947
	10	8.575	5.7255	10.4417
	15	13.23	9.9955	15.925
	20	17.07	14.2005	20.3087
$\bar{\gamma}_{CC}$	1.5	1.64	1.5938	1.7888
	2	2.21	2.1638	2.2662
	2.5	2.76	2.7138	2.8575
	3	3.225	3.14	3.31
	3.5	3.675	3.5812	3.7663

Table S1: Simulated and estimated values of de novo (DN) and case-control (CC). Q50, Q5 and Q95 are for quantile values of 0.5, 0.05 and 0.95 respectively.

Source	Disease	DN	DN control	Case	Control
Fromer et al. (2014)	SCZ	617			
Girard et al. (2011)	SCZ	14			
Gulsuner et al. (2013)	SCZ	105	84		
McCarthy et al. (2014)	SCZ	57			
Xu et al. (2012)	SCZ	231	34		
Guipponi et al. (2014)	SCZ	53			
Genovese et al. (2016)	SCZ			4954/4248	6239/5865
Singh et al. (2016)	SCZ			1745/1353	6789/4769
McRae et al. (2016)	DD	4293			
EuroEPINOMICS-RES Consortium et al. (2014)	EPI	365			
De Ligt et al. (2012)	ID	100			
Hamdan et al. (2014)	ID	41			
Rauch et al. (2012)	ID	51	20		
Lelieveld et al. (2016)	ID	820			
Turner et al. (2016)	ASD	5122			
De Rubeis et al. (2014)	ASD			404	3654
Iossifov et al. (2012)	ASD		343		
ORoak et al. (2012)	ASD		50		
Sanders et al. (2012)	ASD		200		

Table S2: De novo and case/control data. For ASD studies, Turner et al. (2016) integrated previous results in their study; therefore only de novo meta data in this study are shown in the table. In addition, for ASD case-control data, only one homogeneous Sweden population from De Rubeis et al. (2014) was used. For case-control data of SCZ, after correcting for the population stratification, only 4,248 cases (3,157 + 1,091) + 5,865 (4,672 + 1,193) controls from Genovese et al. (2016) and 1,353 cases + 4,769 controls from Singh et al. (2016) are used in this study.

Disease	Mutation	Count	Sample size	Mutation count per sample size
SCZ	silentCFPK	50	1077	0.05
	MiD	105	1077	0.1
	LoF	116	1077	0.11
ASD	MiD	620	5122	0.12
	LoF	638	5122	0.12
ID	MiD	222	1022	0.22
	LoF	230	1022	0.23
EPI	MiD	67	356	0.19
	LoF	58	356	0.16
DD	MiD	1056	4293	0.25
	LoF	1078	4293	0.25

Table S3: De novo mutation counts of categories and their mutation counts per sample size for schizophrenia (SCZ), autism spectrum disorder (ASD), epilepsy (EPI), intellectual disorder (ID) and developmental disorder (DD).

Table S4: extTADA results of SCZ data.

Gene set name	Abbreviation	Author
Missense constrained genes	constrained	Samocha et al. (2014)
Loss-of-function tolerance genes	pLI90	Lek et al. (2015)
RBFOX2 and RBFOX1/3 genes	rbfox2, rbfox13	Weyn-Vanhentenryck et al. (2014)
FMRP genes	fmrp	Darnell et al. (2011)
CELF4 genes	celf4	Wagnon et al. (2012)
synaptic genes	synaptome	Pirooznia et al. (2012)
microRNA-137	mir137	Robinson et al. (2015)
PSD-95 complex genes	psd95	Bayés et al. (2011)
ARC and NMDA receptors genes	nmdarc	Kirov et al. (2012)
Essential genes	essential	Ji et al. (2016)
Human accelerated regions and primate accelerated regions	HARs, PARS	Lindblad-Toh et al. (2011)
Known ID gene sets	IDallKnownGenes	Lelieveld et al. (2016)
Voltage-gated Calcium Channel Genes	vacc	
CHD8 promoter targets	chd8 hNSC, chd8 hNSC specific, chd8 human brain, chd8 hNSC human brain, chd8 hNSC human mouse	Cotney et al. (2015)
De novo copy number variants		Genovese et al. (2016)
ASD	CNV.denovo.gain/loss.asd	
Bipolar	CNV.denovo.gain/loss.bd	
SCZ	CNV.denovo.gain/loss.scz	
MiD and LoF de novo mutations		
DD	DD.allDenovoMiDandLoF	
ASD	ASD.allDenovoMiDandLoF	
EPI	EPI.allDenovoMiDandLoF	
ID	ID.allDenovoMiDandLoF	

Table S5: Known gene sets used in this study.

Gene set	P value	Adjusted p value
DD.allDenovoMiDandLoF	1e-07	2.3e-05
celf4	1e-07	2.3e-05
constrained	1e-07	2.3e-05
pLI09	1e-07	2.3e-05
rbfox13	1e-07	2.3e-05
rbfox2	1e-07	2.3e-05
FMRP_targets	1e-07	2.3e-05
abnormal_behavior	1e-07	2.3e-05
GGGAGGRR_V$MAZ_Q6	3e-07	6.3e-05
abnormal_sensory_capabilities|reflexes|nociception	2.2e-06	4.1e-04
AST.allDenovoMiDandLoF	2.6e-06	4.4e-04

Gene set	P value	Adjusted p value
abnormal_motor_capabilities|coordination|movement	3.2e-06	5.0e-04
chd8.human_brain	9e-06	1.3e-03
ACAGGGT,MIR-10A,MIR-10B	9.9e-06	1.3e-03
abnormal_emotion|affect_behavior	1.2e-05	1.5e-03
GO:0016043	1.4e-05	1.6e-03
GO:0045202	2.1e-05	2.3e-03
GO:0071840	2.4e-05	2.5e-03
abnormal_nervous_system_morphology	2.7e-05	2.7e-03
CAGGTG_V$E12_Q6	3.3e-05	3.1e-03
GO:0008104	5.3e-05	4.7e-03
GO:0051179	6.1e-05	5.2e-03
GO:0006996	6.7e-05	5.5e-03
ARC	7.5e-05	5.9e-03
GO:0043234	7.8e-05	5.9e-03
CTTTGT_V$LEF1_Q2	9.1e-05	6.5e-03
AACTTT_UNKNOWN	9.4e-05	6.5e-03
GO:0048519	0.00011	7.4e-03
synaptome	0.00012	7.8e-03
abnormal_social|conspecific_interaction	0.00013	8.1e-03
GGATTA_V$PITX2_Q2	0.00014	8.5e-03
KEGG_AXON_GUIDANCE	0.00017	1.0e-02
GO:0043005	0.00018	1.0e-02
essentialGenes	0.00019	1.0e-02
list.EPI.43genes.2017.Epi4K.2017	2e-04	1.0e-02
GO:0045211	2e-04	1.0e-02
mir137	0.00023	1.1e-02
NMDAR_network	0.00023	1.1e-02
GO:0044456	0.00023	1.1e-02
GO:0022836	0.00027	1.2e-02
GO:0022839	0.00027	1.2e-02
GO:0033036	0.00029	1.3e-02
GO:0034702	0.00029	1.3e-02
AATGTGA,MIR-23A,MIR-23B	0.00031	1.3e-02
GO:0097060	0.00032	1.3e-02
GO:0044765	0.00034	1.4e-02
mGluR5	0.00035	1.4e-02
GO:0015276	0.00038	1.5e-02
GO:0022834	0.00038	1.5e-02
GO:0048193	0.00041	1.5e-02
CTTTGA_V$LEF1_Q2	0.00044	1.6e-02
GO:0097458	0.00049	1.8e-02
GO:0019226	0.00053	1.9e-02
GO:0022892	0.00054	1.9e-02
GO:0005261	0.00058	1.9e-02
GO:0008022	0.00057	1.9e-02

Gene set	P value	Adjusted p value
abnormal_fear|anxiety-related_behavior	0.00061	2.0e-02
abnormal_cued_conditioning_behavior	0.00062	2.0e-02
GO:0005215	0.00064	2.0e-02
GO:0048592	0.00063	2.0e-02
REACTOME_TRANSMISSION_ACROSS_CHEMICAL_SYNAPSES	0.00065	2.0e-02
abnormal_synaptic_transmission	0.00069	2.1e-02
GO:0048523	0.00069	2.1e-02
seizures	0.00072	2.1e-02
GO:0035637	0.00073	2.1e-02
abnormal_behavioral_response_to_xenobiotic	0.00077	2.2e-02
ID.allDenovoMiDandLoF	0.00079	2.2e-02
GO:0007268	0.00079	2.2e-02
Padinas2017_extTable9.genes	0.00095	2.5e-02
GO:0005886	0.00094	2.5e-02
ID.allKnownGenes	0.00099	2.6e-02
GO:0000904	0.00099	2.6e-02
GO:0007399	0.001	2.6e-02
GO:0006810	0.0012	3.0e-02
GO:0022612	0.0012	3.0e-02
GO:0015031	0.0013	3.2e-02
GO:0016568	0.0013	3.2e-02
GO:0048589	0.0014	3.3e-02
GO:0051234	0.0014	3.3e-02
REACTOME_NEURONAL_SYSTEM	0.0014	3.3e-02
GO:0071944	0.0015	3.5e-02
PSD-95_(core)	0.0016	3.6e-02
GO:0042995	0.0016	3.6e-02
abnormal_excitatory_postsynaptic_currents	0.0017	3.7e-02
abnormal_learning|memory|conditioning	0.0017	3.7e-02
GO:0005216	0.0017	3.7e-02
GO:0007154	0.0017	3.7e-02
GO:0007267	0.0018	3.8e-02
GO:0010646	0.0018	3.8e-02
GO:0023052	0.0019	3.9e-02
GO:0030662	0.0019	3.9e-02
GO:0044700	0.0019	3.9e-02
GO:0000139	0.0021	4.2e-02
GO:0007519	0.0021	4.2e-02
GO:0045184	0.0021	4.2e-02
abnormal_associative_learning	0.0024	4.7e-02
GO:0022838	0.0025	4.7e-02
GO:0023051	0.0025	4.7e-02
GO:0048731	0.0025	4.7e-02
GO:0032991	0.0026	4.9e-02
abnormal_social_investigation	0.0027	4.9e-02

Gene set	P value	Adjusted p value
abnormal_synapse_morphology	0.0027	4.9e-02
GO:0010629	0.0027	4.9e-02

Table S6: Enrichment of gene sets from different data bases with SCZ genes from `extTADA` results. These p values were obtained by 10,000,000 simulations, and then adjusted by using the method of Benjamini and Hochberg (1995).

Parameters	Estimated mode	lCI	uCI
SCZ_pi_silentCFPKdn	0.0056	0	0.1977
SCZ_hyperGammaMean_silentCFPKdn	1.5802	1.001	21.5139
SCZ_pi_MiDdn	0.012	0	0.2368
SCZ_hyperGammaMean_MiDdn	1.7486	1	17.8548
SCZ_pi_LoFdn	0.0548	0.0124	0.2062
SCZ_hyperGammaMean_LoFdn	11.1857	3.3973	31.3602
SCZ_pi_MiD+LoFcc	0.069	0.0296	0.1359
SCZ_hyperGammaMean_MiD+LoFcc	2.0176	1.2133	5.3694
SCZ_hyperGammaMean_MiD+LoFcc	3.2288	1.2372	17.1478
SCZ_hyperGammaMean_MiD+LoFcc	1.0691	1.0002	2.9574

Table S7: Genetic parameters for SCZ data if single class is used in the analysis.

Parameters	Estimated mode	lCI	uCI
SCZ_pi0	0.0732	0.0306	0.1506
SCZ_meanRR_silentCFPKdenovo	1.2353	1.0021	3.6086
SCZ_meanRR_MiDdenovo	1.4459	1.0008	4.7004
SCZ_meanRR_LoFdenovo	12.0403	4.6136	25.8786
SCZ_meanRR_MiD+LoFccPop1	1.5856	1.1255	4.0881
SCZ_meanRR_MiD+LoFccPop2	1.7361	1.0438	4.8856
SCZ_meanRR_MiD+LoFccPop3	1.0698	1.0001	2.9991

Table S8: SCZ genetic parameters using all variants in and not in ExAC database (InExAC + NoExAC).

Table S9: extTADA results of SCZ data uisng all variants in and not in ExAC database (InExAC + NoExAC).

Parameters	Estimated mode	lCI	uCI
SCZ_pi0	0.0937	0.0547	0.1512
SCZ_meanRR_silentCFPKdenovo	1.3068	1.0005	2.7489
SCZ_meanRR_MiDdenovo	2.2246	1.0006	5.3491
SCZ_meanRR_LoFdenovo	15.1491	5.8606	27.3941
SCZ_meanRR_MiD+LoFccPop1	1.8677	1.0374	3.0736
SCZ_meanRR_MiD+LoFccPop2	2.2632	1.003	4.9168
SCZ_meanRR_MiD+LoFccPop3	1.0372	1.0002	1.1807

Table S10: SCZ genetic parameters after adjusting mutation rates (NoExAC).

Table S11: extTADA results of SCZ data after adjusting mutation rates.

Table S12: extTADA results of ID data.

Table S13: extTADA results of DD data.

Table S14: extTADA results of ASD data.

6.2 Sup Figure

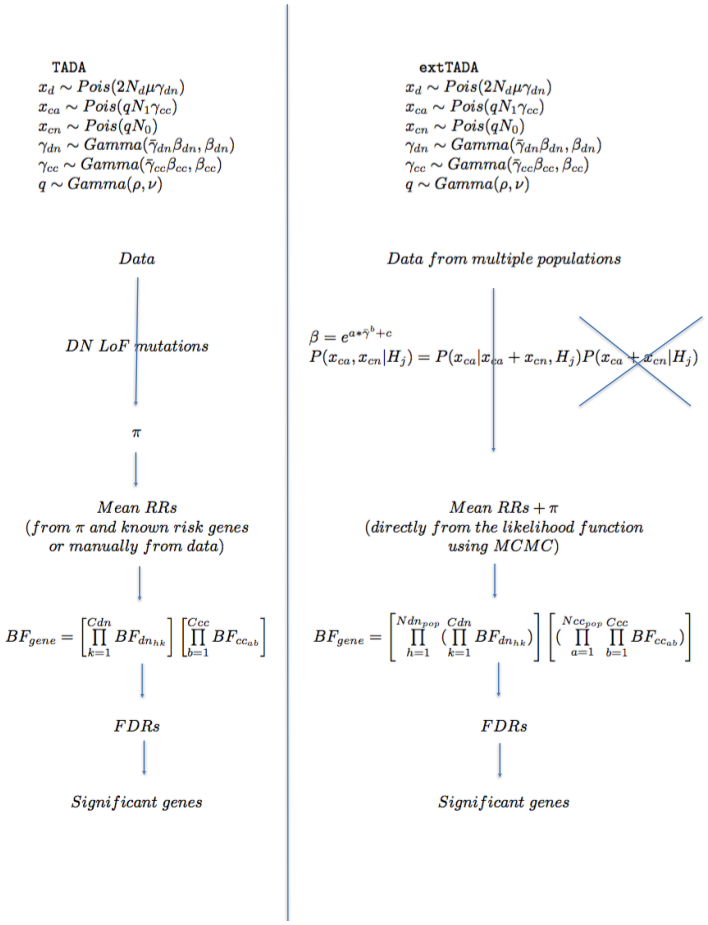

Figure S1: Comparison between `TADA` and `extTADA`. They both use the same model for de novo data (x_{dn} and case/control (x_{ca}, x_{cn}) data. `extTADA` combines all categories to obtain parameters while `TADA` is based on LoF mutations. `extTADA` uses an approximate model for case-control data, and constrains β and $\bar{\gamma}$ in the estimation process. `extTADA` is designed to work for multiple populations. `TADA` can be used inside `extTADA`.

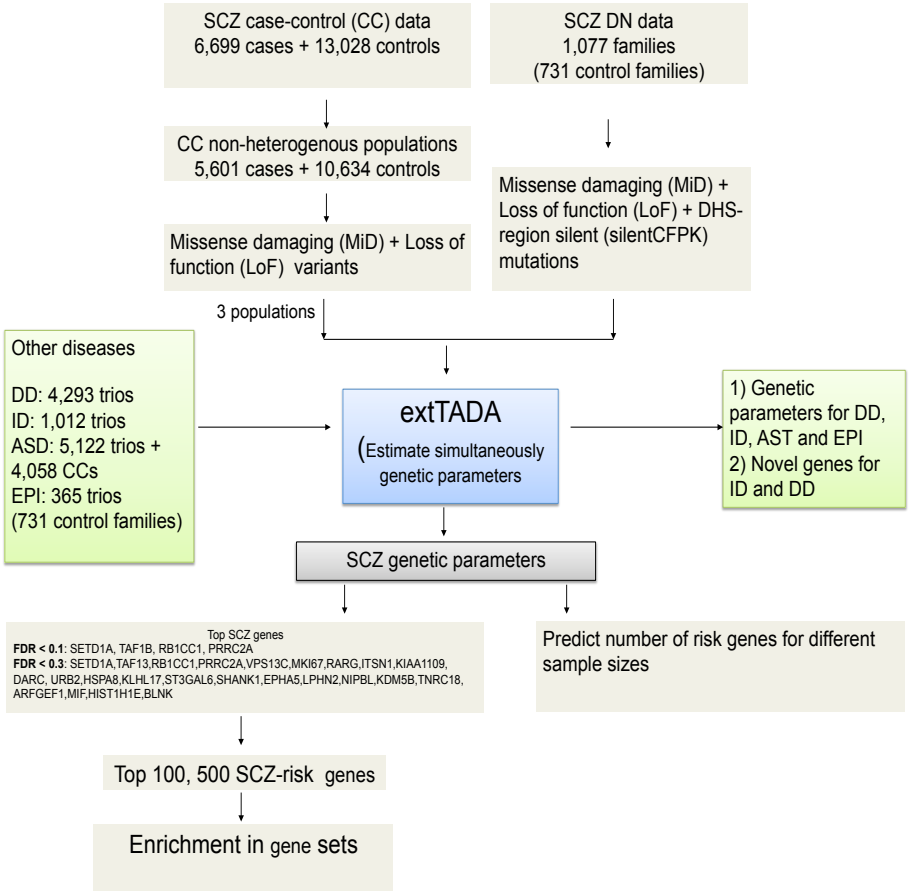

Figure S2: Workflow of data analysis.

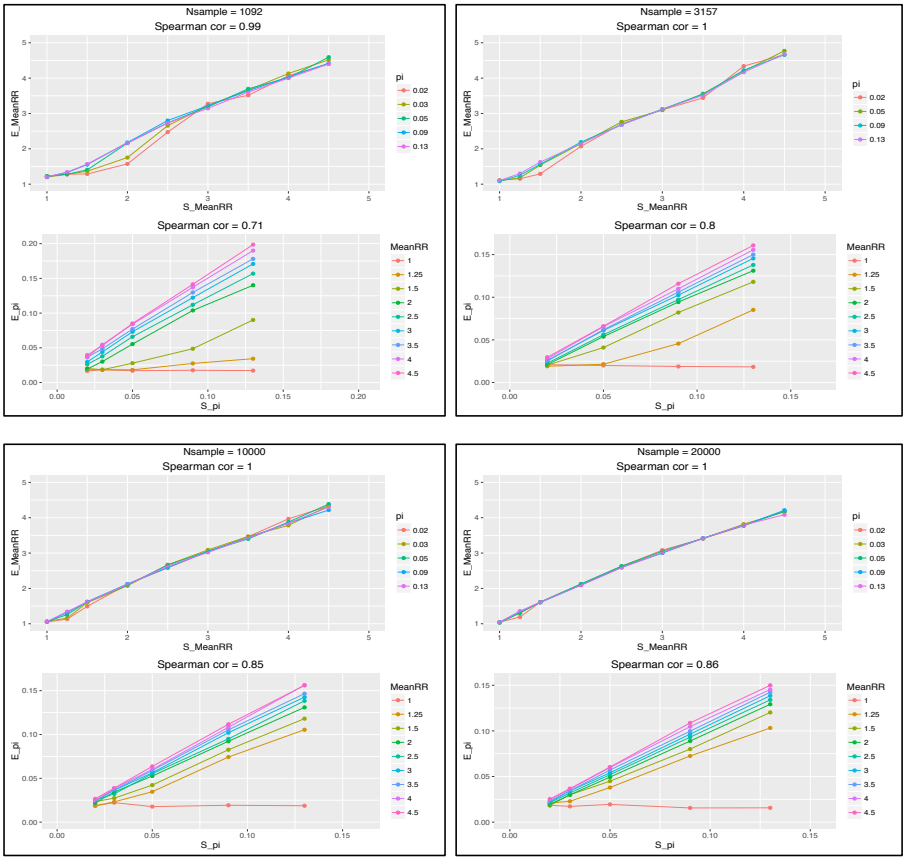

Figure S3: Correlation between simulated and estimated values for one-category case/control data.

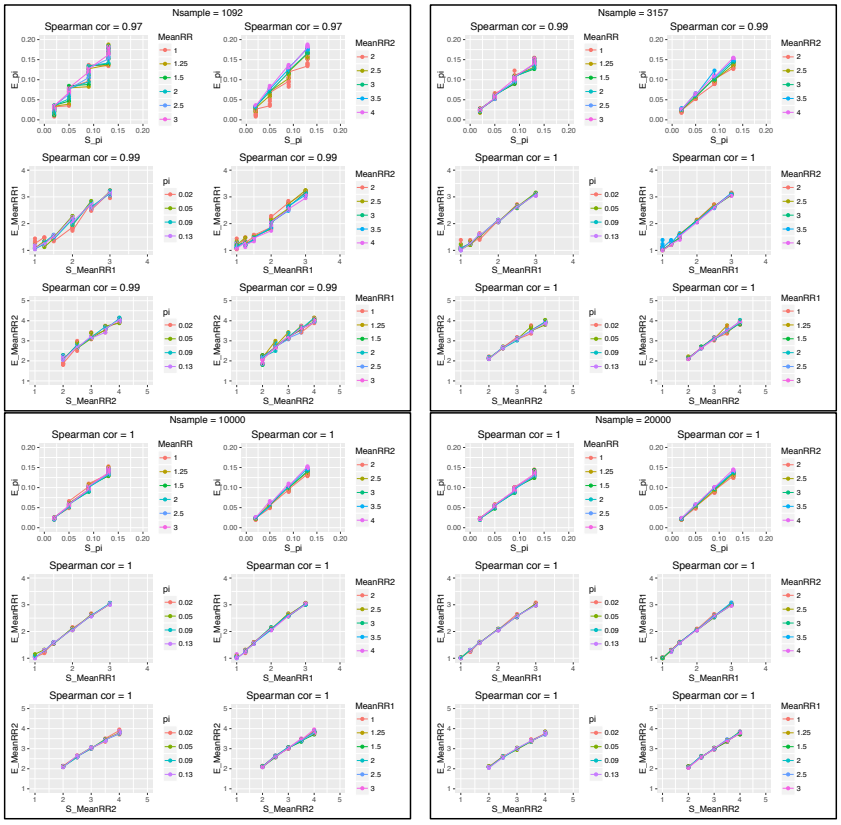

Figure S4: Correlation between simulated and estimated values for two-category case/control data.

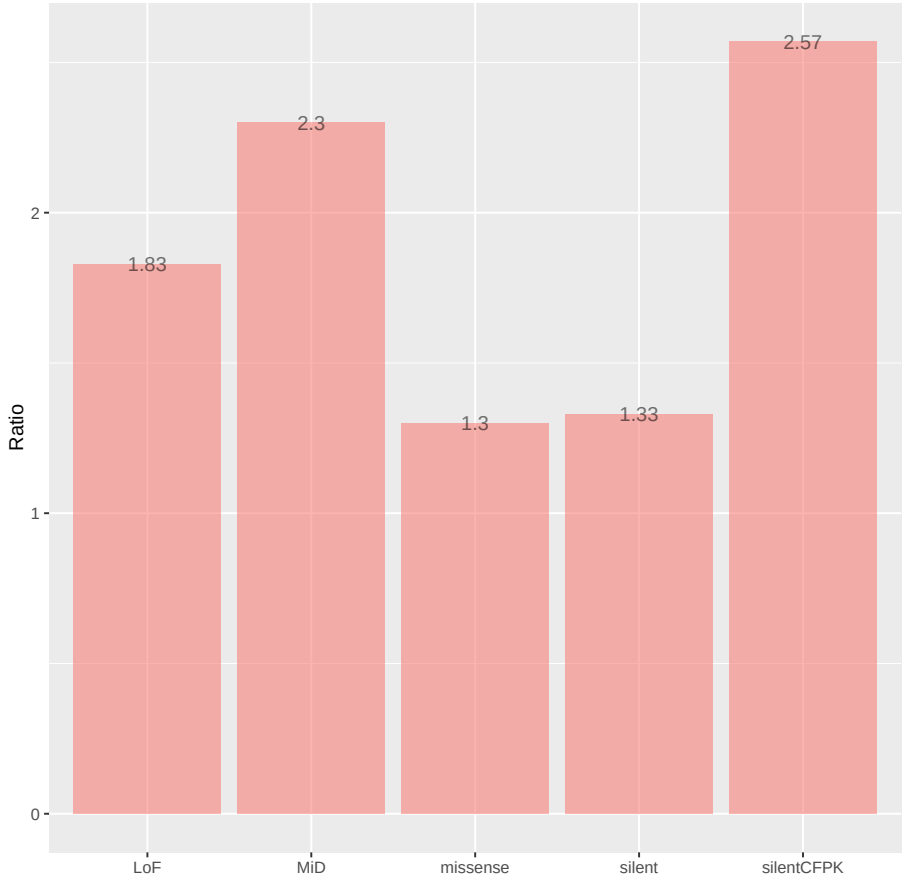

Figure S5: Ratios of de novo mutations between SCZ probands and controls (unaffected siblings). "silentCFPK" describes for silent mutations within frontal cortex-derived DHS (silentCerebrumfrontalocPk.narrowPeak). MiD mutations are missense mutations derived from 7 methods.

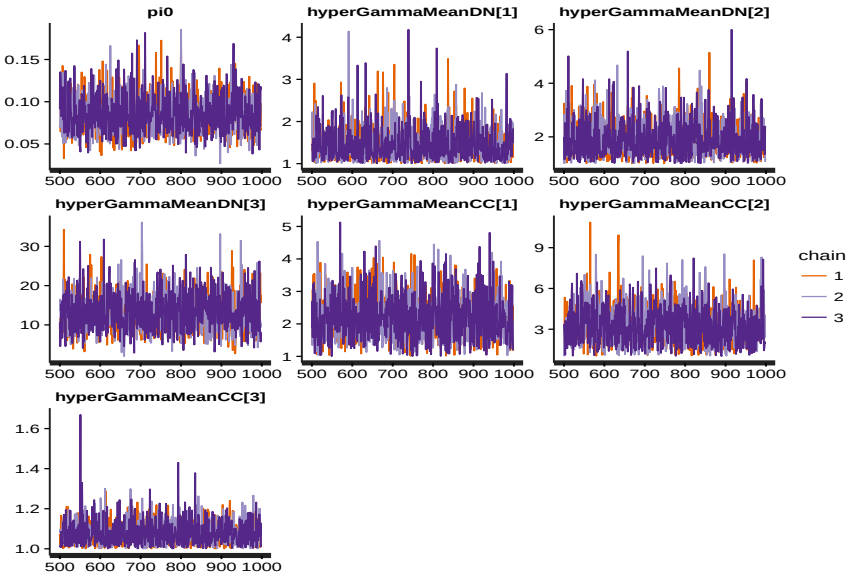

Figure S6: MCMC results for SCZ data.

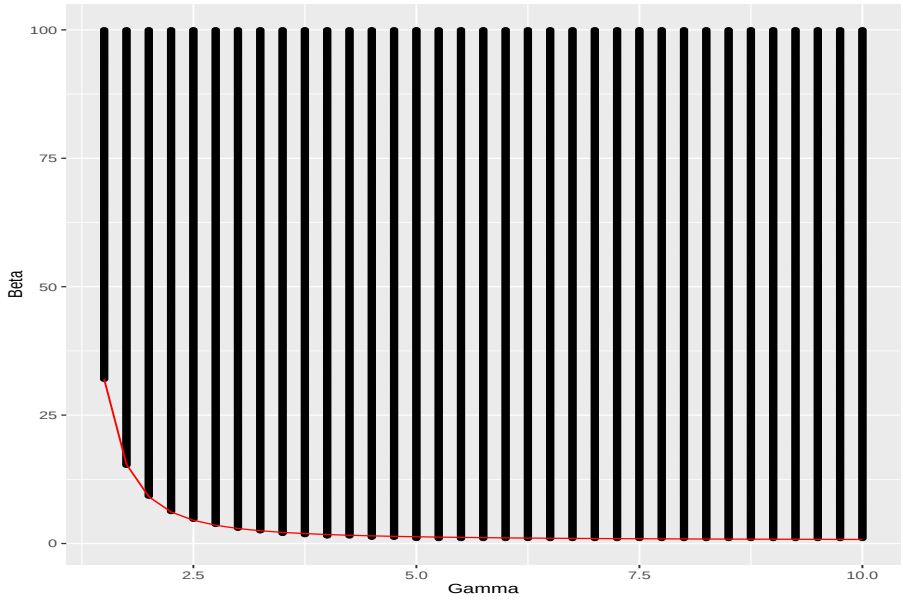

Figure S7: A grid of β and γ values. Points on the red line are corresponding with the proportion of protective variants less than 0.0%.

6.3 Sup Information

6.3.0.1 Calculate Bayes Factor for case/control data

At a given gene, Bayes Factor for each class was calculated as $BF = \frac{P(x_1,x_0|H_1)}{P(x_1,x_0|H_0)}$. The probability for each model $(H_j, j = 0, 1)$ was calculated order to rely only γ parameters as follows.

$$P(x_{ca}, x_{cn}|H_j) = P(x_{cn}|H_j)P(x_{ca}|x_{cn}, H_j) \tag{10}$$

- The first part $P(x_{cn}|H_j)$ was the same as De Rubeis et al. (2014):

$$P(x_{cn}|H_j) = \int P(x_{cn}|q, H_j)P(q|\rho, \nu, H_j)dq = NegBin(x_{cn}|\rho, \frac{N_0}{\nu + N_0}), j = 0, 1 \tag{11}$$

- The second part:

$$\begin{aligned} P(x_{ca}|H_j, x_{cn}) &= \int P(x_{ca}|q, \gamma_{cc})P(q|H_j, x_{cn})P(\gamma_{cc}|H_j)dqd\gamma_{cc} \\ &= \int [P(x_{ca}|q, \gamma_{cc})P(q|H_j, x_{cn})dq]\, P(\gamma_{cc}|H_j)d\gamma_{cc} \\ &= \int NegBin(x_{ca}|\rho + x_{cn}, \tfrac{N_0+\nu}{N_1\gamma_{cc}+N_0+\nu})P(\gamma_{cc}|H_j)d\gamma_{cc} \end{aligned} \tag{12}$$

To identify the lower and upper limits of γ_{CC} for the integral, we randomly sampled 10,000 times values from the $Gamma(\bar{\gamma}_{cc} * \beta_{cc}, \beta_{cc})$ and used the minimum and maximum values for the lower and upper limits respectively.

References

À. Bayés, L. N. van de Lagemaat, M. O. Collins, M. D. Croning, I. R. Whittle, J. S. Choudhary, and S. G. Grant. Characterization of the proteome, diseases and evolution of the human postsynaptic density. *Nature neuroscience*, 14(1): 19–21, 2011.

Y. Benjamini 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)*, pages 289–300, 1995.

B. Carpenter, A. Gelman, M. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. A. Brubaker, J. Guo, P. Li, and A. Riddell. Stan: a probabilistic programming language. *Journal of Statistical Software*, 2015.

P. Cingolani, A. Platts, L. L. Wang, M. Coon, T. Nguyen, L. Wang, S. J. Land, X. Lu, and D. M. Ruden. 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*, 6(2):80–92, 2012.

G. O. Consortium et al. Gene ontology consortium: going forward. *Nucleic acids research*, 43(D1):D1049–D1056, 2015.

G. M. Cooper, B. P. Coe, S. Girirajan, J. A. Rosenfeld, T. H. Vu, C. Baker, C. Williams, H. Stalker, R. Hamid, V. Hannig, et al. A copy number variation morbidity map of developmental delay. *Nature genetics*, 43(9):838–846, 2011.

J. Cotney, R. A. Muhle, S. J. Sanders, L. Liu, A. J. Willsey, W. Niu, W. Liu, L. Klei, J. Lei, J. Yin, et al. The autism-associated chromatin modifier chd8 regulates other autism risk genes during human neurodevelopment. *Nature communications*, 6, 2015.

J. C. Darnell, S. J. Van Driesche, C. Zhang, K. Y. S. Hung, A. Mele, C. E. Fraser, E. F. Stone, C. Chen, J. J. Fak, S. W. Chi, et al. Fmrp stalls ribosomal translocation on mrnas linked to synaptic function and autism. *Cell*, 146(2): 247–261, 2011.

J. De Ligt, M. H. Willemsen, B. W. van Bon, T. Kleefstra, H. G. Yntema, T. Kroes, A. T. Vulto-van Silfhout, D. A. Koolen, P. de Vries, C. Gilissen, et al. Diagnostic exome sequencing in persons with severe intellectual disability. *New England Journal of Medicine*, 367(20):1921–1929, 2012.

S. De Rubeis, X. He, A. P. Goldberg, C. S. Poultney, K. Samocha, A. E. Cicek, Y. Kou, L. Liu, M. Fromer, S. Walker, et al. Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature*, 515(7526):209–215, 2014.

F. Degenhardt, L. Priebe, S. Meier, L. Lennertz, F. Streit, S. Witt, A. Hofmann, T. Becker, R. Mössner, W. Maier, et al. Duplications in rb1cc1 are associated with schizophrenia; identification in large european sample sets. *Translational psychiatry*, 3(11):e326, 2013.

Epi4K Consortium and Epilepsy Phenome/Genome Project. De novo mutations in epileptic encephalopathies. *Nature*, 501(7466):217–221, 2013.

EuroEPINOMICS-RES Consortium, Epilepsy Phenome/Genome Project, and Epi4K Consortium. De novo mutations in synaptic transmission genes including dnm1 cause epileptic encephalopathies. *The American Journal of Human Genetics*, 95(4):360–370, 2014.

C. Fraley and A. E. Raftery. Mclust: Software for model-based cluster analysis. *Journal of Classification*, 16(2):297–306, 1999.

M. Fromer, A. J. Pocklington, D. H. Kavanagh, H. J. Williams, S. Dwyer, P. Gormley, L. Georgieva, E. Rees, P. Palta, D. M. Ruderfer, et al. De novo mutations in schizophrenia implicate synaptic networks. *Nature*, 506(7487): 179–184, 2014.

G. Genovese, M. Fromer, E. A. Stahl, D. M. Ruderfer, K. Chambert, M. Landen, J. L. Moran, S. M. Purcell, P. Sklar, P. F. Sullivan, C. M. Hultman, and S. A. McCarroll. Increased burden of ultra-rare protein-altering variants among 4,877 individuals with schizophrenia. *Nat Neurosci*, advance online publication:–, 10 2016. URL http://dx.doi.org/10.1038/nn.4402.

S. L. Girard, J. Gauthier, A. Noreau, L. Xiong, S. Zhou, L. Jouan, A. Dionne-Laporte, D. Spiegelman, E. Henrion, O. Diallo, et al. Increased exonic de novo mutation rate in individuals with schizophrenia. *Nature genetics*, 43(9): 860–863, 2011.

M. Guipponi, F. A. Santoni, V. Setola, C. Gehrig, M. Rotharmel, M. Cuenca, O. Guillin, D. Dikeos, G. Georgantopoulos, G. Papadimitriou, et al. Exome sequencing in 53 sporadic cases of schizophrenia identifies 18 putative candidate genes. *PloS one*, 9(11):e112745, 2014.

S. Gulsuner, T. Walsh, A. C. Watts, M. K. Lee, A. M. Thornton, S. Casadei, C. Rippey, H. Shahin, V. L. Nimgaonkar, R. C. Go, et al. Spatial and temporal mapping of de novo mutations in schizophrenia to a fetal prefrontal cortical network. *Cell*, 154(3):518–529, 2013.

F. F. Hamdan, M. Srour, J.-M. Capo-Chichi, H. Daoud, C. Nassif, L. Patry, C. Massicotte, A. Ambalavanan, D. Spiegelman, O. Diallo, et al. De novo mutations in moderate or severe intellectual disability. *PLoS Genet*, 10(10): e1004772, 2014.

X. He, S. J. Sanders, L. Liu, S. De Rubeis, E. T. Lim, J. S. Sutcliffe, G. D. Schellenberg, R. A. Gibbs, M. J. Daly, J. D. Buxbaum, et al. Integrated model

of de novo and inherited genetic variants yields greater power to identify risk genes. *PLoS Genet*, 9(8):e1003671, 2013.

I. Iossifov, M. Ronemus, D. Levy, Z. Wang, I. Hakker, J. Rosenbaum, B. Yamrom, Y.-h. Lee, G. Narzisi, A. Leotta, et al. De novo gene disruptions in children on the autistic spectrum. *Neuron*, 74(2):285–299, 2012.

X. Ji, R. L. Kember, C. D. Brown, and M. Buan. Increased burden of deleterious variants in essential genes in autism spectrum disorder. *Proceedings of the National Academy of Sciences*, 2016. doi: 10.1073/pnas.1613195113.

W. J. Kent, C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, and D. Haussler. The human genome browser at ucsc. *Genome research*, 12(6):996–1006, 2002.

G. Kirov, A. Pocklington, P. Holmans, D. Ivanov, M. Ikeda, D. Ruderfer, J. Moran, K. Chambert, D. Toncheva, L. Georgieva, et al. De novo cnv analysis implicates specific abnormalities of postsynaptic signalling complexes in the pathogenesis of schizophrenia. *Molecular psychiatry*, 17(2):142–153, 2012.

M. Lek, K. Karczewski, E. Minikel, K. Samocha, E. Banks, T. Fennell, A. O'Donnell-Luria, J. Ware, A. Hill, B. Cummings, et al. Analysis of protein-coding genetic variation in 60,706 humans. *bioRxiv*, page 030338, 2015.

S. H. Lelieveld, M. R. Reijnders, R. Pfundt, H. G. Yntema, E.-J. Kamsteeg, P. de Vries, B. B. de Vries, M. H. Willemsen, T. Kleefstra, K. Löhner, et al. Meta-analysis of 2,104 trios provides support for 10 new genes for intellectual disability. *Nature Neuroscience*, 19(9):1194–1196, 2016.

P. Lichtenstein, B. H. Yip, C. Björk, Y. Pawitan, T. D. Cannon, P. F. Sullivan, and C. M. Hultman. Common genetic determinants of schizophrenia and bipolar disorder in swedish families: a population-based study. *The Lancet*, 373(9659):234–239, 2009.

K. Lindblad-Toh, M. Garber, O. Zuk, M. F. Lin, B. J. Parker, S. Washietl, P. Kheradpour, J. Ernst, G. Jordan, E. Mauceli, et al. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature*, 478(7370):476–482, 2011.

X. Liu, C. Wu, C. Li, and E. Boerwinkle. dbnsfp v3. 0: A one-stop database of functional predictions and annotations for human nonsynonymous and splice-site snvs. *Human mutation*, 2015.

C. Loader. Locfit: Local regression, likelihood and density estimation. *R package version*, 1, 2007.

S. E. McCarthy, J. Gillis, M. Kramer, J. Lihm, S. Yoon, Y. Berstein, M. Mistry, P. Pavlidis, R. Solomon, E. Ghiban, et al. De novo mutations in schizophrenia implicate chromatin remodeling and support a genetic overlap with autism and intellectual disability. *Molecular psychiatry*, 19(6):652, 2014.

J. F. McRae, S. Clayton, T. W. Fitzgerald, J. Kaplanis, E. Prigmore, D. Rajan, A. Sifrim, S. Aitken, N. Akawi, M. Alvi, et al. Prevalence, phenotype and architecture of developmental disorders caused by de novo mutation. *bioRxiv*, page 049056, 2016.

E. Murphy and A. Bentez-Burraco. Bridging the gap between genes and language deficits in schizophrenia: An oscillopathic approach. *Frontiers in Human Neuroscience*, 10:422, 2016. ISSN 1662-5161. doi: 10.3389/fnhum.2016.00422. URL http://journal.frontiersin.org/article/10.3389/fnhum.2016.00422.

M. A. Newton, A. Noueiry, D. Sarkar, and P. Ahlquist. Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics*, 5(2):155–176, 2004.

B. J. ORoak, L. Vives, S. Girirajan, E. Karakoc, N. Krumm, B. P. Coe, R. Levy, A. Ko, C. Lee, J. D. Smith, et al. Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature*, 485(7397): 246–250, 2012.

A. Pardinas, P. Holmans, A. Pocklington, V. Escott-Price, R. Stephan, N. Carrera, B. Sophie, C. Darren, M. Hamshere, H. Jun, et al. Common schizophrenia alleles are enriched in mutation-intolerant genes and maintained by background selection. *bioRxiv*, 2017.

E. Phenome et al. Ultra-rare genetic variation in common epilepsies: a case-control sequencing study. *The Lancet Neurology*, 16(2):135–143, 2017.

M. Pirooznia, T. Wang, D. Avramopoulos, D. Valle, G. Thomas, R. L. Huganir, F. S. Goes, J. B. Potash, and P. P. Zandi. Synaptomedb: an ontology-based knowledgebase for synaptic genes. *Bioinformatics*, 28(6):897–899, 2012.

A. J. Pocklington, E. Rees, J. T. Walters, J. Han, D. H. Kavanagh, K. D. Chambert, P. Holmans, J. L. Moran, S. A. McCarroll, G. Kirov, et al. Novel findings from cnvs implicate inhibitory and excitatory signaling complexes in schizophrenia. *Neuron*, 86(5):1203–1214, 2015.

S. M. Purcell, N. R. Wray, J. L. Stone, P. M. Visscher, M. C. O'Donovan, P. F. Sullivan, P. Sklar, D. M. Ruderfer, A. McQuillin, D. W. Morris, et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, 460(7256):748–752, 2009.

S. M. Purcell, J. L. Moran, M. Fromer, D. Ruderfer, N. Solovieff, P. Roussos, C. ODushlaine, K. Chambert, S. E. Bergen, A. Kähler, et al. A polygenic burden of rare disruptive mutations in schizophrenia. *Nature*, 506(7487):185–190, 2014.

A. R. Quinlan and I. M. Hall. Bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, 2010.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015. URL https://www.R-project.org/.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016. URL https://www.R-project.org/.

A. Rauch, D. Wieczorek, E. Graf, T. Wieland, S. Endele, T. Schwarzmayr, B. Albrecht, D. Bartholdi, J. Beygo, N. Di Donato, et al. Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. *The Lancet*, 380(9854):1674–1682, 2012.

E. B. Robinson, B. M. Neale, and S. E. Hyman. Genetic research in autism spectrum disorders. *Current opinion in pediatrics*, 27(6):685, 2015.

K. E. Samocha, E. B. Robinson, S. J. Sanders, C. Stevens, A. Sabo, L. M. McGrath, J. A. Kosmicki, K. Rehnström, S. Mallick, A. Kirby, et al. A framework for the interpretation of de novo mutation in human disease. *Nature genetics*, 46(9):944–950, 2014.

S. J. Sanders, M. T. Murtha, A. R. Gupta, J. D. Murdoch, M. J. Raubeson, A. J. Willsey, A. G. Ercan-Sencicek, N. M. DiLullo, N. N. Parikshak, J. L. Stein, et al. De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature*, 485(7397):237–241, 2012.

A. Sifrim, M.-P. Hitz, A. Wilsdon, J. Breckpot, S. H. Al Turki, B. Thienpont, J. McRae, T. W. Fitzgerald, T. Singh, G. J. Swaminathan, et al. Distinct genetic architectures for syndromic and nonsyndromic congenital heart defects identified by exome sequencing. *Nature Genetics*, 2016.

T. Singh, M. I. Kurki, D. Curtis, S. M. Purcell, L. Crooks, J. McRae, J. Suvisaari, H. Chheda, D. Blackwood, G. Breen, et al. Rare loss-of-function variants in setd1a are associated with schizophrenia and developmental disorders. *Nature neuroscience*, 2016.

C. Spearman. The proof and measurement of association between two things. *The American journal of psychology*, 15(1):72–101, 1904.

H. Stefansson, R. A. Ophoff, S. Steinberg, O. A. Andreassen, S. Cichon, D. Rujescu, T. Werge, O. P. Pietiläinen, O. Mors, P. B. Mortensen, et al. Common variants conferring risk of schizophrenia. *Nature*, 460(7256):744–747, 2009.

A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102 (43):15545–15550, 2005.

P. F. Sullivan, K. S. Kendler, and M. C. Neale. Schizophrenia as a complex trait: evidence from a meta-analysis of twin studies. *Archives of general psychiatry*, 60(12):1187–1192, 2003.

A. Takata, I. Ionita-Laza, J. A. Gogos, B. Xu, and M. Karayiorgou. De novo synonymous mutations in regulatory elements contribute to the genetic etiology of autism and schizophrenia. *Neuron*, 89(5):940–947, 2016.

T. N. Turner, Q. Yi, N. Krumm, J. Huddleston, K. Hoekzema, H. A. Stessman, A.-L. Doebley, R. A. Bernier, D. A. Nickerson, and E. E. Eichler. denovo-db: a compendium of human de novo variants. *Nucleic Acids Research*, page gkw865, 2016.

J. L. Wagnon, M. Briese, W. Sun, C. L. Mahaffey, T. Curk, G. Rot, J. Ule, and W. N. Frankel. Celf4 regulates translation and local abundance of a vast set of mrnas, including genes associated with regulation of synaptic function. *PLoS Genet*, 8(11):e1003067, 2012.

S. M. Weyn-Vanhentenryck, A. Mele, Q. Yan, S. Sun, N. Farny, Z. Zhang, C. Xue, M. Herre, P. A. Silver, M. Q. Zhang, et al. Hits-clip and integrative modeling define the rbfox splicing-regulatory network linked to brain development and autism. *Cell reports*, 6(6):1139–1152, 2014.

B. Xu, I. Ionita-Laza, J. L. Roos, B. Boone, S. Woodrick, Y. Sun, S. Levy, J. A. Gogos, and M. Karayiorgou. De novo gene mutations highlight patterns of genetic and neural complexity in schizophrenia. *Nature genetics*, 44(12): 1365–1369, 2012.

K. Xu, E. E. Schadt, K. S. Pollard, P. Roussos, and J. T. Dudley. Genomic and network patterns of schizophrenia genetic variation in human evolutionary accelerated regions. *Molecular biology and evolution*, 32(5):1148–1160, 2015.