Bayesian Integrated analysis of multiple types of ..

October 11, 2016

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, however, while sets of genes have been implicated through study of rare variation, very few individual risk genes have been identified. Here, we apply a 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 modified the model used previously in ASD studies. To simplify the complexity of the model, an approximation for the case-control model in which case variants are conditional on total counts is used. In addition, instead of using only one class of de novo mutation as in the previous studies, all classes of de novo mutations and case-control variants are used to infer genetic parameters. These parameters are estimated using a Markov Chain Monte Carlo method. We applied this method to 1,024 trios and 4,954 cases/6,239 controls. We defined four variant annotation categories: disruptive (nonsense, frameshift, essential splice site mutations) and MiD de novos (predicting damaging by seven algorithms), disruptive and missense damaging case/control singletons. We estimated that 8.4% of approximate 20,000 estimated genes are risk genes (95% credible interval 3.5-16%), with mean effect sizes (95% CIs) of 14.21 (5.04- 25.65) for disruptive de novos, 1.99 (1-3.99) for missense damaging de novos, 1.79 (1-2.94) for disruptive case/control singletons, and 1.56 (1-2.46) for missense damaging case/control singletons. Our analysis identified only three gene with FDR¡0.1, SETD1A, TAF13 (FDR¡0.05) and RB1CC1. We further analyzed the top 100 genes, with FDR¡=0.496, for enrichment in several candidate gene sets. Significant results are observed in gene sets previously implicated in schizophrenia (including in a subset of these data): FMRP, Rbfox1/2/3, constrained, de novo mutations in ASD (all p values less than 7.8x10-4), and synaptic (p = 1.3x10-3). Overall, our results replicate previous studies for known gene sets as well as the single gene SETD1A indicating the robustness of the approach. We also used this approach to infer genetic architecture for four other psychiatric diseases (intellectual disorder, epilepsy, developmental disorder and autism spectrum disorder) using 7,072 trios and 4058 cases/controls available. We anticipate this approach will improve

our power to detect schizophrenia risk genes (and also for other diseases) as more data is included.

NOTE

- All these results are in Figure **??**, and Table 5. P-values in the abstract are adjusted using the method Benjamini & Hochberg (1995), NOT the method "Bonferroni".
- Author list please? Should I add any people?
- The file below describes the method to obtain p-values for gene sets (Inside Model on GitHub, it would be slightly different because of choosing random gene sets):

  `intersect_with_differentGeneSet_2classes.ipynb`

Data

CaseControl for clustering: /hpc/users/nguyet26/psychen/methods/extTADA/Re$_{a}nnotate/Sweden2_{C}aseControl$

De novo: /hpc/users/nguyet26/psychen/methods/extTADA/Re$_{a}nnotate/DenovoData$

Contents

1 Introduction **3**

2 Data and methods **4**

2.1 Data . . . 4

2.1.1 Simulation data . . . 4

2.1.2 Variant data of SCZ, ID, DD, EPI and AST . . . 4

2.1.3 Gene sets . . . 5

2.2 Methods . . . 6

2.2.1 extTADA pipeline: analyse de novo, transmission and case-control data . . . 6

2.2.2 Use simulation data to test model . . . 10

2.2.3 Calculate mutation rates . . . 11

2.2.4 Analyse SCZ data . . . 11

2.2.5 Use extTADA to predict genetic parameters of other psychiatric diseases . . . 13

2.2.6 Infer parameters using MCMC results . . . 14

3 Results **14**

3.1 Simulated data . . . 14

3.2 Schizophrenia data sets . . . 16

3.2.1 Extract data sets to analyse integratively . . . 16

3.2.2 Integrated analysis of de novo mutations and case-control variants . . . 17

3.2.3 Test enrichment of SCZ-risk genes in known gene sets . . . 20

3.2.4 Identify number of risk genes for SCZ studies with different sample sizes . . . 21

3.2.5 Test for single classes or combination only two classes of SCZ data . . . 21
3.3 Estimate genetic parameters of other psychiatric diseases using extTADA . . . 22

4 Discussion **24**

5 Supplementary information **27**
5.1 Sup Table . . . 27
5.2 Sup Figure . . . 29

1 Introduction

Schizophrenia (SCZ) is a complex psychiatric disorder including positive symtoms (hallucinations, delusions, thought and movement disorders), negative symtoms (not feel happy in daily life, reduced speaking) and cognitive symtoms. In spite of the high reduction of reproductive fecundity and a life time 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 arichitecture 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 producing of high-quality next-generation sequencing data, genetic parameters of the disease will be estimated more accurate, especially for rarer variants.

Using rare variants, de novo mutations as well as combining these two classes to identify risk genes for SCZ has been successful in identifying specific genes (Singh et al., 2016; Takata et al., 2016) or implicating gene sets for this disease (Purcell et al., 2014; Fromer et al., 2014). However, the genetic architecture of SCZ for rare variants, de novo mutations as well as the integration between these two classes has not been performed. Such analyses would help understand further insight into the characteristics of this disease. In addition, it would facilitate the obtaining a highly confident set of risk genes for SCZ. As a result, a better picture of biological pathways specific for the disease could be determined.

In this study, we modify a Bayesian meta-analysis framework (TADA, Transmission And De novo Association) which was developed for autism spectrum disorder (AST) (He et al., 2013) to use it in identifying the genetic architecture of rare variants and de novo mutations of SCZ. The new framework (extTADA, extended Transmission And De novo Association) assummes that all variant/-mutation classes play important roles in the genetic architecture of the disease; therefore they are used to simultaneously estimate genetic parameters. In ext-TADA, an approximate model for case-control data was used instead of the original case-control model in TADA in order to facilitate the estimation process of parameters which is carried out by using a Markov Chain Monte Carlo (MCMC) method. In this study, extTADA was used to integratively analysis 4,929 cases and 6,232 controls and 1,024 trios for the SCZ disease. It predicted

mean relative risks (RRs) of different variant categories as well the proportion of risk genes for the disease. Based on these results, SCZ-risk gene sets were determined with different thresholds, and the gene sets showed enrichments in known gene sets. Finally, extTADA was used to estimate genetic parameters for other neuropsychiatric diseases: intellectual disability (ID), autism spectrum disorder (AST), epilepsy (EPI) and developmental disorder (DD) by using a total of available 7,072 trios and 4058 cases/controls.

2 Data and methods

A workflow of all data used in this study is described in Figure S1.

2.1 Data

2.1.1 Simulation data

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

2.1.2 Variant data of SCZ, ID, DD, EPI and AST

High-quality variants were obtained from original analyses as described in Table 1. Some modifications were used to obtain final sets of variants in our current study. For only AST data, De Rubeis et al. (2014) annotated and used them in their TADA pipeline; therefore the same their annotations as their work were used. For other data set, some steps were used to obtain a final data set. Variants were annotated using Plink/Seq (using RefSeq gene stranscripts, 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 7 methods from dbnsfp31a: SIFT, $Polyphen2\_HDIV$, $Polyphen2\_HVAR$, LRT, PROVEAN, MutationTaster and MutationAssessor. To annotate synonymous variants within regular regions (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/

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		
Genovese et al. (2016)	SCZ			4954	6239
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		
De Rubeis et al. (2014)	AST	2231		404	3654
Iossifov et al. (2012)	AST		343		
ORoak et al. (2012)	AST		50		
Sanders et al. (2012)	AST		200		

Table 1: De novo, transmitted/non-transmitted and case/control data. For AST studies, De Rubeis et al. (2014) integrated previous results in their study; therefore only de novo meta data in this study are shown in the table. In addition, for AST 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 3,157 cases and 4,672 controls are used in this study.

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), only case-control singleton variants were used in this study.

To annotate private variants, the data from Exome Aggregation Consortium (ExAC) (Lek et al., 2015) were used. 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).

2.1.3 Gene sets

Multiple resources were used to obtain gene sets in our study. We used 3,733 gene sets whose lengths ranged between 5 and 4,994 genes.

- The first 19 gene sets in the list of gene sets used by Genovese

et al. (2016).

- 3,689 gene sets with sizes between 25 to 4,995. were extracted from Gene Ontology database (Consortium et al., 2015).
- Human accelerated regions (HARs)

  Lists of 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.
- SCZ and chromatin gene set from Table 1 and 2 of the transcriptome-wide association study from the Psychiatric Genomics Consortium (Gusev et al., 2016).
- SCZ expression gene set from Supplementary Data File 3 from The CommonMind Consortium (Fromer et al., 2016)
- Gene sets from Table 6 and 7 from the SCZ GWAS and eQTL study of Zhu et al. (2016)
- Get set which expresses differently in SCZ-associated vs. control hiPSC-derived neurons from Sup Table 2 of Roussos et al. (2016)

2.2 Methods

2.2.1 extTADA pipeline: analyse de novo, transmission and case-control data

We used an integrated approach in which de novo and case control information was used to infer risk genes. The current study is a framework which is extended from the The Transmission and Disequilibrium Association (TADA) model proposed by He et al. (2013); De Rubeis et al. (2014). For a given gene, all variants of a class (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 was assummed to be risk genes which were represented by the H_1 model. Under this model, relative risks (γ) were assummed to follow a probability distribution. The model H_0 described for non-risk genes of the genes; and relative risks (γ) of genes were set to equal to 1. As in He et al. (2013), we modeled de novo (x_d) and case (x_{ca}) control (x_{cn}) data as Equation 1:

$$\begin{aligned} x_d &\sim Pois(2N_d\mu\gamma_{dn}) \\ x_1 &\sim Pois(qN_1\gamma_{cc}) \\ x_0 &\sim Pois(qN_0) \end{aligned} \tag{1}$$

in which N_d, N_1, N_0 are sample sizes of trios, cases and controls respectively; γ_{dn} and γ_{cc} are relative risks for de novo mutations and case-control variants, μ are mutation rates.

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

$$\begin{aligned} B_{ij} &= \frac{\int P(x_{ij}|\gamma,q)P(q|H_1)P(\gamma|H_1)dqd\gamma}{\int P(x_{ij}|\gamma,q)P(q|H_0)P(\gamma|H_0)dqd\gamma} \\ &\overset{\gamma_{H_0}=1}{=} \frac{\int P(x_{ij}|\gamma,q)P(q|H_1)P(\gamma|H_1)dqd\gamma}{\int P(x_{ij}|q)P(q|H_0)dq} \end{aligned} \tag{2}$$

In Equation 2, $x_{ij} = x_d$ and $x_{ij} = (x_{ca}, x_{cn})$ for de novo, and case-control data respectively. In addion, the integral across q was not used for de novo data.

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

$$B_i = \prod_{j=1}^{K} B_{ij} \tag{3}$$

Gamma distributions were assummed as prior distributions for γ_{dn} and γ_{cc} as in Equation 4.

$$\begin{aligned} \gamma_{dn} &\sim Gamma(\bar{\gamma_{dn}}\beta_{dn}, \beta_{dn}) \\ \gamma_{cc} &\sim Gamma(\bar{\gamma_{cc}}\beta_{cc}, \beta_{cc}) \\ q &\sim Gamma(\rho, \nu) \end{aligned} \tag{4}$$

Regarding priors of the parameter q, He et al. (2013) used different values for H_1 and H_0 ((ρ_1, ν_1) and (ρ_0, ν_0) respectively) ; however, it was challenging to estimate these parameters independently as discussed in De Rubeis et al. (2014). Therefore, simplified parameters as a current TADA version (De Rubeis et al., 2014) were used in this study: $\rho_1 = \rho_0 = \rho$ and $\nu_1 = \nu_0 = \nu$.

To calculate BFs, hyper parameters in Equation 4 were need 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 gene (m) as in Equation 5.

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

Use approximate model for case-control data

The Equation 5 was different from the original TADA model because we integrated all categories into the mixture model as described in our method for calculating BFs in Equation 3. To obtain hyperparameters $\phi_{1j} = (\gamma_{j(dn)}, \gamma_j, \beta_{j(dn)}, \beta_j, \rho_j, \nu_j)$, we used 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 complex with multiple parameters; therefore, the Equation was simplified to avoid sampling directly $q \sim Gamma(\rho, \nu)$:

- For de novo data, the same as Equation 1.
- 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 heuristicly chosen (in all current study, 200 was used) and $\frac{\rho}{\nu}$ = the mean frequency across genes by using both case and control data.
  - **Approximate (simplify) case/control model**

The case-control model was deployed as follows:

$$\begin{aligned} P(x_{ca}, x_{cn}|H_j) &= P(x_{ca}, x_{ca} + x_{cn}|H_j) \\ &= P(x_{ca}|x_{ca} + x_{cn}, H_j)P(x_{ca} + x_{cn}|H_j) \end{aligned} \tag{6}$$

The first part: $P(x_{ca}|x_{ca} + x_{cn}, H_j)$
Because of $x_{ca} \sim Pois(N_1 q \gamma_{cc})$ and $x_{cn} \sim Pois(N_0 q)$, we assumed that x_{ca} and x_{cn} were **independent**, we had:
$x_{ca}|x_{ca} + x_{cn}, H_j \sim Binomial(x_{ca} + x_{cn}, \theta|H_j)$
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
$P(x_{ca}|x_{ca} + x_{cn}, H_j) = \int P(x_{ca}|x_{ca} + x_{cn}, \gamma_{cc}, H_j)P(\gamma|x_{ca} + x_{cn}, H_j)d\gamma_{cc}$
Based on simulation results, the first part can be approximately used to infer mean RRs ($\bar{\gamma_{cc}}$); therefore only this part was used in the estimation process in Equation 5.

Control the proportion of risk genes using the mean and dispersion parameters of relative risks

If $\bar{\gamma}$ and β were small then we would see high 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 S2). 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 are $6.82722, -1.2918269$ and -0.5783759 respectively.

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

The prior information used in this study is presented in Table 2.

<table>
<tr><td rowspan="2">$x_{dn} \sim P(2N_{dn}\mu\gamma_{dn})$</td><td rowspan="2">$\gamma_{dn} \sim Gamma(\bar{\gamma_{dn}} * \beta_{dn}, \beta_{dn})$</td><td>$\bar{\gamma_{dn}} \sim Gamma(1, 0.05)$
$\beta_{dn} = e^{a*\bar{\gamma}^b_{dn}+c}$</td></tr>
<tr><td></td></tr>
<tr><td rowspan="3">$x_{ca} \sim P(N_1 q\gamma_{cc})$</td><td>$\gamma_{cc} \sim Gamma(\bar{\gamma_{cc}} * \beta_{cc}, \beta_{cc})$</td><td>$\bar{\gamma_{cc}} \sim Gamma(1, 0.05)$
$\beta_{cc} = e^{a*\bar{\gamma}^b_{cc}+c}$</td></tr>
<tr><td>$q \sim Gamma(\rho, \nu)$</td><td>$\frac{\rho}{\nu} = mean(x_{cn} + x_{ca})$
$\nu = 200$</td></tr>
<tr><td></td></tr>
<tr><td>$x_{cn} \sim P(N_0 q)$</td><td>$q \sim Gamma(\rho, \nu)$</td><td>$\frac{\rho}{\nu} = mean(x_{cn} + x_{ca})$
$\rho = 200$</td></tr>
<tr><td>$\pi \sim \ Beta(1, 5)$</td><td colspan="2"></td></tr>
</table>

Table 2: Prior information used in all analyses.

2.2.1.1 Predict the number of risk genes

BFs of genes were calculated using Equation 3. The original case-control model was used in this calculation. After that, the BFs were converted to false discovery rates (FDRs) as described in De Rubeis et al. (2014). The number of risk genes could be predicted based on a threshold(s) defined by users.

2.2.2 Use simulation data to test model

We simulated multiple combinations between CC and DN 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 was 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 2. For DN data, we used exactly the original model of TADA in both the simulation and estimation process.

To see the performance of the estimation process of parameters inside the model, we calculated true positive rates (TPRs) and false positive rates (FPRs) as follows. For each set of simulated parameters, we recorded genes (sGene) whose false discovery rates (FDRs) were $\leq$ a threshold. Then we used extTADA to re-estimate the simulated parameters and recorded genes (rGene) with the same FDR threshold. Because low proportions of risk genes could result in zero or very low the number of risk genes if we set a very small FDR

threshold, therefore, a threshold FDR = 0.1 was used to obtain a reasonable number of genes for any situations.

- TPR = $\frac{the\ number\ of\ overlapping\ genes\ of\ sGene\ and\ rGene}{the\ number\ of\ sGene}$.
- FPR = $\frac{the\ number\ of\ rGene\ not\ in\ sGene}{the\ total\ gene\ (1{,}941) - the\ number\ of\ sGene}$.

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

2.2.3 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.

2.2.4 Analyse SCZ data

2.2.4.1 Obtain a homogeneous population for case-control data of SCZ

A simple combination between a clustering process using a multivariate normal mixture model and a data analysing strategy using linear and generalized linear models was used to obtain a homegeneous population used in this study. Genovese et al. (2016) recently analysed 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 regresion and generalized linear regression models as in Equation 8. They reported significant results for NonExAC LoF and MiD variants. We defined a homogeneous population as a population which was not much affected by the covariates. Thus, for the population, analysing 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 a group which had similar results between Equation 9 and 8.

$$\begin{aligned} &logit(P(SCZ = 1)) \sim count + countAll + sex + birth + kit + PC1 + ... + PC20 \\ &count \sim SCZ + countAll + sex + birth + kit + PC1 + ... + PC20 \end{aligned} \tag{8}$$

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

2.2.4.2 Estimate genetic parameters for SCZ

De novo mutations and case-control variants from the non-heteregenous population were integratively analysed. 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. Firstly, genetic parameters of the five classes (3 DN + 2 CC classes) were estimated. Then we pooled the data of two CC classes together because they had nearly equal mean RRs, and estimate parameters for four classes (3 DN classes + 1 pooled CC class).

2.2.4.3 Estimate number of risk genes for SCZ

Based on estimated genetic parameters from the four classes, 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.

2.2.4.4 Test enrichment in known gene sets

Based on FDRs results, we chose the top 100, 500 genes having the smallest FDRs to see their overlaps with known gene sets. The hypergeometric test in R (Kachitvichyanukul and Schmeiser, 1985) was used to obtain p values of the overlapping significance. To correct for multiple tests, the p values were divided by the total number of tests.

2.2.4.5 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 assummed that sample sizes of cases and controls were the same. 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 ≤ 0.05.

2.2.4.6 Test for single classes or combination of two classes

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

2.2.5 Use extTADA to predict genetic parameters of other psychiatric diseases

Use exTADA, we analysed the integration architecture of genetics for four other psychiatric diseases: EPI, ID, DD and AST. For AST, genetic parameters were estimated simulatenously 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.

2.2.6 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, 50,000 times and three independent chains were used. In addition, for SCZ data we used two steps to obtain final results. Firstly, 50,000 times were run to obtain parameters. After that, we used Equation S2 to calculate β values from estimated mean RRs. Finally, extTADA was re-run 50,000 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 burnning 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 for each chain from MCMC results to do further analyses. For example, with a chain with 50,000 run times, the step to obtain a sample was 50 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 confidence intervals (CIs) of estimated paramters, the Locfit (Loader, 2007) was used. The mode values were considerred as our estimated values for other calculations.

3 Results

3.1 Simulated data

We simulated data of one CC class and one DN class. For CC data, the original model of TADA was used and then the CC approximate model was used in the estimation process.

The majority of TPRs was $> 0.9\%$ with the median TPR for each specific value of each parameter $> 95\%$ (Table 3 and Figure 1). All FPRs were smaller 10^{-4}. This showed that the case-control

approximate model was able to represent the orginal case-control model in the estimation process. Low mean RRs ($\sim$ 2) showed some large outliers for FPRs and small outliers for TPRs.

High correlations (> 0.9 were seen for simulated values and estimated values (Figure S3). Interestingly, case-control data showed a higher correlation than de novo data. This could be because CC sample sizes (3,157 cases and 4,672 controls) were much larger than the DN sample size (1,024). Slightly over and under estimation were observed for mean RRs and proportions of risk genes. This was expected and usually did not much affect to final results as disscused in the previous work (He et al., 2013).

Parameter		5%	50%	95%	5%	50%	95%
π	0.01	1	1	1	2.94e-05	3.08e-05	3.42e-05
	0.05	0.96	1	1	5.85e-05	8.52e-05	1.28e-04
	0.09	0.96	1	1	8.11e-05	1.33e-04	1.96e-04
	0.13	0.94	0.99	1	1.02e-04	1.81e-04	2.69e-04
$\bar{\gamma}_{DN}$	2	0.96	1	1	3.04e-05	9.32e-05	1.98e-04
	7	0.94	0.99	1	3.25e-05	1.06e-04	2.04e-04
	12	0.96	1	1	3.26e-05	1.31e-04	2.26e-04
	20	0.97	1	1	9.20e-05	1.57e-04	2.67e-04
$\bar{\gamma}_{CC}$	1.5	0.96	1	1	5.67e-05	8.61e-05	2.47e-04
	4	0.95	0.97	1	2.99e-05	1.03e-04	1.70e-04
	6	0.98	0.99	1	3.25e-05	1.27e-04	1.97e-04
	8	1	1	1	1.07e-04	1.66e-04	2.70e-04

Table 3: Different quantiles of TPRs and FPRs for simulation data. The first three quantiles are for TPRs and the second three quantiles are for FPRs.

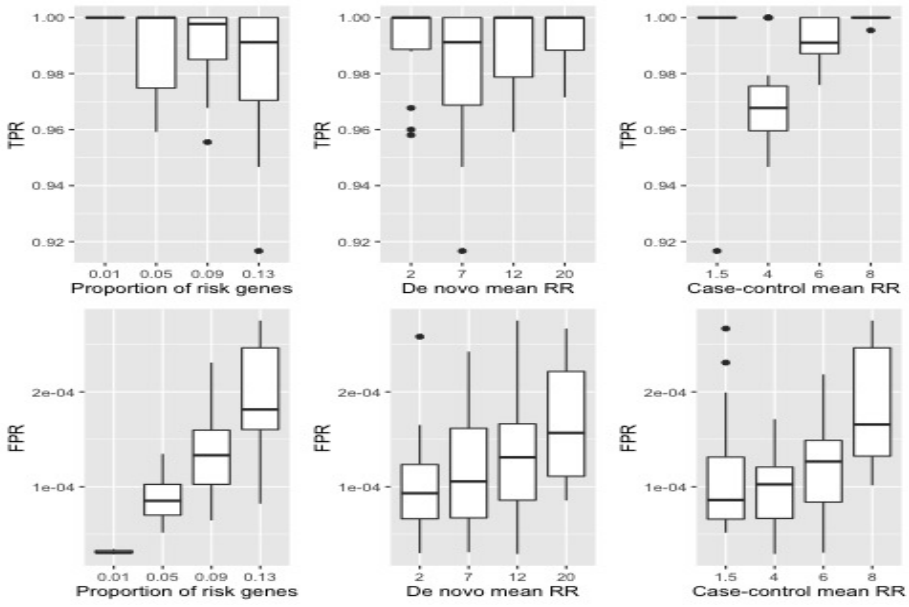

Figure 1: True postive rates (TPRs) and false postive rate (FPRs) for simulated data.

3.2 Schizophrenia data sets

3.2.1 Extract data sets to analyse integratively

De novo mutations and case-control variants were tested to choose classes and samples for the meta analysis of the extTADA pipeline. For case-control data, we clusterred 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 no adjusting covariates. One group which consisted of 3157 cases and 4672 controls showed highest similar results between adjusting and non adjusting results was considered as non heteregeneous group; therefore it was used in our next analyses (Figure 2). Similar to Genovese et al. (2016), non-private variants did not show significant differences between cases and controls. In addition, all silent variants (all and within the regulatory region) also had a non significant trend in all private and non private classes. Therefore, only LoF and MiD case-control variants of the 3,157 cases and 4,672 controls were used (Figure 2). For de novo mutations, Fisher's exact test was

used to compare between 1,024 cases (from trios) and 731 controls. Significant results were seen in four classes: LoF, missense, MiD mutations, and silent mutations within frontal cortex-derived DHS (silentCFPK). The highest odd ratio was observed for silentCFPK (2.63), followed by MiD (2.44), LoF (1.94) and missense (1.77) mutations (Table S1) . There classes: LoF, silentCFPK, MiD (we used this class, instead of missense DNs, to be comparable with case-control data) were used in next steps.

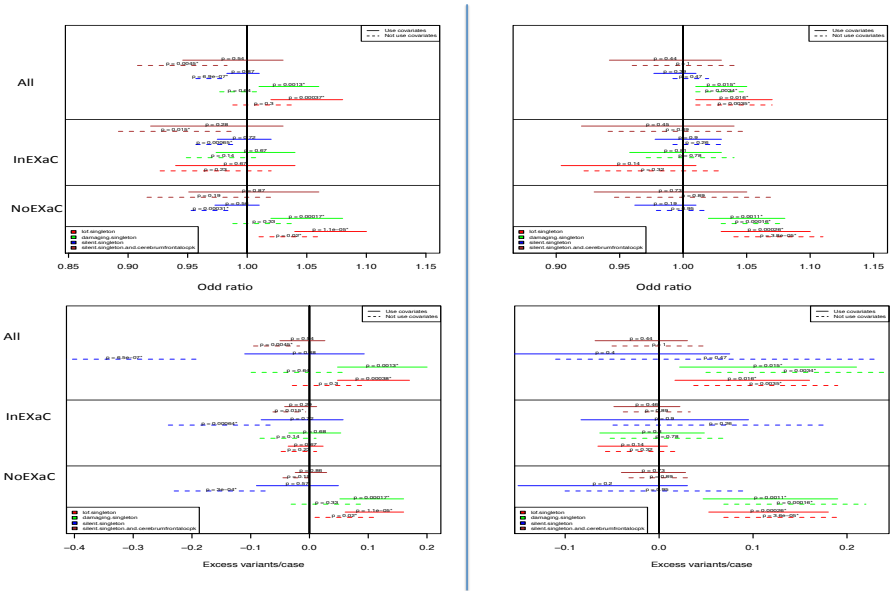

Figure 2: Odd ratios and excess variants in cases for the analysing of all case-control samples. Left panels show results for all samples before adjustment for population stratification while right panels describe results for only one homogeneous population. Top pictures are results of modeling SCZ status (yes/no) as a function of variant counts (and covariates) using a generalized linear regression model. Bottom pictures are results of modelling variant counts as a function of SCZ status (and covariates) using a linear regression model.

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

Five categories of de novo mutations and case-control variants were used in the integration analysis process. They included LoF, MiD

and silentCFPK denovo mutations as well as LoF and MiD case-control variants. SilentCFPK case-control variants were not used in this process because this class did not show significantly excess counts in cases against controls as showed in Figure 2. extTADA automatically estimated all genetic parameters of SCZ based on the current data set. Trace plots of parameters are in Sup Figure S4. The proportion of risk genes was aprroximately 8.08% (CI = (3.71%, 14.94%)). Regarding de novo classes, LoF had the highest mean RR, which was 10.6242 with CI = (4.0365, 23.8542). Two other de novo classes had approximate mean RRs: 1.3859 (CI = (1.002, 2.7961) and 1.5594 (CI = (1.0015, 3.5689)) for silentCFPK and MiD respectively. In constrast with the large difference between LoF and other classes in mean RRs of de novo mutations, these parameters were not much different between LoF and MiD case-control variants. They were 2.0875, CI = (1.05, 3.8705) and 1.7496, CI = (1.0191, 3.2144) for the two classes in that order (Table 4, Figure 3).

Based on the similar results of two case-control classes, to increase the power of the analysis, we combined two LoF and MiD case-control classes into one case-control class. After that, extTADA was used to estimate parameters for a new integration of four classes (three de novos and one case-control). The proportion of risk genes obtained from this integration was not different (8.67%) but the CI was much larger (CI = (4.22%, 18.84%)). This might be because there were some non-overlapping signals between two case-control classes. Mean RRs of this new estimation were similar to the five-category estimation as presented in Table 4.

We then calculated FDRs for each genes using estimated genetic parameters. For five-category model, there was only one gene having FDR < 0.05 or < 0.1. This gene, SETD1A (FDR $= 0.038$), had been confirmed as a SCZ-risk gene in previous studies Singh et al. (2016); Takata et al. (2016). For four-category model, SETD1A was still the only gene having FDR < 0.05; however, there were two other genes TAF1B (FDR $= 0.053$) and MKI67 (FDR $= 0.085$) having FDR < 0.1. If we increased the FDR threshold to 0.3 as the previous AST study De Rubeis et al. (2014), there were 13 genes (SETD1A, TAF13, MKI67, ST3GAL6, VPS13C, RB1CC1, PRRC2A, LPHN2, DARC, DUOXA2, KLHL17, HSPA8, ZDHHC5, XPR1) which were significant with this threshold.

AST_pi0	0.0409	0.0255	0.0637
AST_hyperGammaMeanDN[1]	5.7237	3.1296	8.5704
AST_hyperGammaMeanDN[2]	21.0151	11.6042	35.5736
AST_hyperGammaMeanCC[1]	6.7062	2.6702	13.804
ID_pi0	0.0278	0.0173	0.0474
ID_hyperGammaMeanDN[1]	16.1312	9.4684	25.594
ID_hyperGammaMeanDN[2]	137.5875	78.9605	221.3481
EPI_pi0	0.0155	0.0075	0.0326
EPI_hyperGammaMeanDN[1]	46.4968	22.8005	87.3737
EPI_hyperGammaMeanDN[2]	76.3407	36.8854	141.7984
DD_pi0	0.033	0.0275	0.0401
DD_hyperGammaMeanDN[1]	7.2099	5.5068	9.1343
DD_hyperGammaMeanDN[2]	78.0916	59.0825	98.6762
SCZ_pi0	0.0808	0.0371	0.1494
SCZ_hyperGammaMeanDN[1]	1.3859	1.002	2.7961
SCZ_hyperGammaMeanDN[2]	1.5594	1.0015	3.5689
SCZ_hyperGammaMeanDN[3]	10.6242	4.0365	23.8542
SCZ_hyperGammaMeanCC[1]	1.7496	1.0191	3.2144
SCZ_hyperGammaMeanCC[2]	2.0875	1.05	3.8705
SCZ_combine_CC_pi0	0.0867	0.0422	0.1884
SCZ_combine_CC_hyperGammaMeanDN[1]	1.3139	1.001	2.6771
SCZ_combine_CC_hyperGammaMeanDN[2]	1.5115	1.0045	3.3582
SCZ_combine_CC_hyperGammaMeanDN[3]	10.4081	3.1495	21.1959
SCZ_combine_CC_hyperGammaMeanCC[1]	1.714	1.0308	3.1804
SCZ_combine_CC_combine_DN_pi0	0.1062	0.0525	0.217
SCZ_combine_CC_combine_DN_hyperGammaMeanDN[1]	1.2109	1.0004	3.6365
SCZ_combine_CC_combine_DN_hyperGammaMeanDN[2]	8.6068	3.1476	19.4025
SCZ_combine_CC_combine_DN_hyperGammaMeanCC[1]	1.6526	1.0122	4.6376

Table 4: Estimated parameters for de novo and case-control SCZ data and four other diseases: ID, EPI, AST and DD. These results are obtained by running sampling 50000 MCMC times.

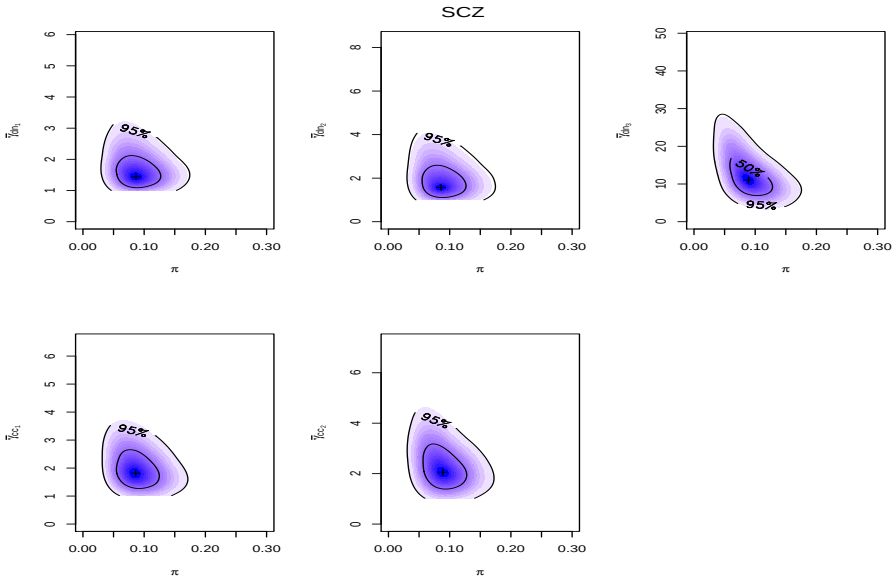

Figure 3: Posterior probability results of the proportion of risk genes and relative risks for SCZ data.

3.2.3 Test enrichment of SCZ-risk genes in known gene sets

We tested the enrichment of known gene sets with the gene set generated from the current analysis. The top 100 and 500 smallest FDR genes from extTADA results were used. Regarding the top 100 genes, significant overlapps were observed for five gene sets: constrained genes, genes flanking SNPs and Indels of DD and AST, FMRP and pLI09 (all $p < 0.006$, Table 5). For the top 500 genes, significant results were also observed in more other 18 gene sets including 14 GO gene sets and RBFOX2, RBFOX13 and CELF4 gene sets (Table 5). The results of not GO gene sets replicated previous results of Genovese et al. (2016). The most significant result for GO gene sets was GO:0051179 (p = 3.3e-05). This gene set was reported by Murphy and Bentez-Burraco (2016) in a study genes relating to language evolution and SCZ.

Gene set	gene number	Top 100	Top 500
constrained	936	2.1e-06	6.3e-10
SNPsINdel.denovo.dd	3937	4.3e-05	0
fmrp	1194	4.3e-05	0
SNPsINdel.denovo.aut	2789	0.0014	0
pLI09	3231	0.0054	0
rbfox2	2892	1	2.6e-10
rbfox13	3226	1	9.8e-07
celf4	2461	1	5.2e-06
GO:0051179	3789	1	3.3e-05
GO:0043228	3642	1	0.0045
GO:0043232	3642	1	0.0045
GO:0016043	3555	1	0.0095
GO:0045202	465	1	0.011
GO:0006810	3067	1	0.019
GO:0071840	3668	1	0.019
GO:0000166	2265	1	0.022
GO:1901265	2266	1	0.022
GO:0051234	3113	1	0.024
GO:0036094	2402	1	0.026
GO:0005215	1129	1	0.03
synaptome	1812	1	0.03
GO:0055085	1047	1	0.032
GO:1901565	865	1	0.046

Table 5: Test overlapping gene sets with extTADA results for two classes.

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

We estimated 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 168. Based on this calculation, we would expect > 50 risk genes if total sample sizes of families and cases (controls = cases) were larger than 16,500 (Figure 4).

3.2.5 Test for single classes or combination only two classes of SCZ data

To see genetic architecture of single classes or combining two classes, and also to test the performance of the pipeline for smaller number of classes, extTADA was used to estimate parameters separately for 5 single classes, and 4 combinations of MiD and LoF mutations/-variations. Overall, the modes of the proportions of risk genes were

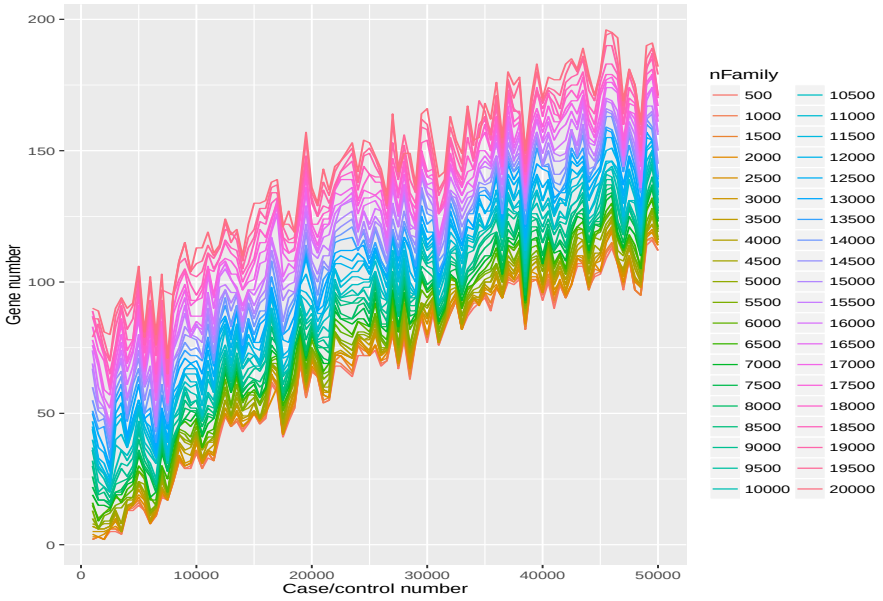

Figure 4: Number of risk genes with different sample sizes based on genetic architecture predicted by extTADA. Case/control number is only for cases (or controls); therefore if Case/control number = 10,000 means total cases+controls = 20,000.

less than 0.1 and CIs were between 0.001 to 0.273 (Table S3). For single class, only LoF de novo mutations show a weak convergence, other classes did not converge (Figure S5). This could be because strong signal was observed for this class in current data set (Table S2). For combination two classes, all convergences, even not very strong, ware observed (Figure S5) with the peaks of estimation proportions of risk genes close to the value ~ 0.81 which was predicted for combination of all classes inside one model (Table S3).

3.3 Estimate genetic parameters of other psychiatric diseases using extTADA

We also used the current pipeline to infer genetic architectures of AST, EPI, DD and ID. The number of risk genes (π) in these diseases was lower than that in SCZ (Figure 5, Table 4). For AST, the 95% confidence interval was between 2.55% and 6.37% which overlapped with the result 550-1000 genes estimated in the original

TADA model (He et al., 2013) using only LoF de novo information. However, the estimated mode value was 4.09% which was smaller the value estimated by TADA ($> 5\%$) using only LoF de novo mutations. For ID, π was smaller than that of AST; estimtated value was 2.78% (1.73% to 4.74% for the 95% CI). The lowest of π values, 1.55% (95% CI = (0.75%, 3.26%)), was observed for EPI. Mean RRs of de novo mutations in these disease were much higher than those of SCZ. This was expected because of lower π values. ID had the highest LoF-mutation mean RR which was 137.59 (CI = (78.96, 221.35)). Even though the mean RR of LoF mutations of EPI, which was 76.34 (CI = (36.89, 141.80)), was lower than that of ID; this value for MiD (46.5, CI = (22.8, 87.37)) was much higher than those of other diseases. The mean RR = 81 for EPI estimated by Epi4K Consortium and Epilepsy Phenome/Genome Project (2013) was in the CI of our current results. For AST, mean RRs for de novo mutations were much slower than these other diseases (Figure 5, Table 4). DD had π smaller than that of AST, but its mean RRs were very high (7.03 for MiD and 78 for LoF muations (Figure 5, Table 4). Because of its high sample size and high mutation rates per samples (Table S2), estimated results of DD showed reliable convergences (Figure 5).

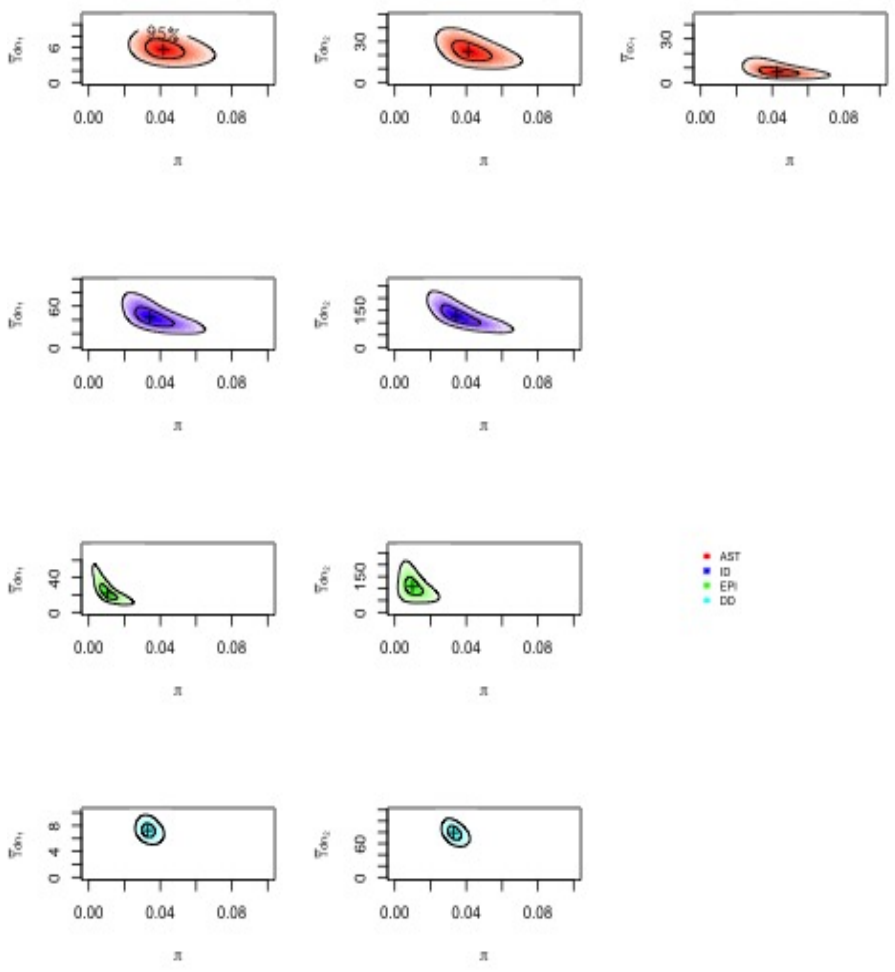

Figure 5: Posterior probability results of the proportion of risk genes and relative risks for AST, EPI, ID and DD data. For AST, there are two de novo classes and one case-control class. For other diseases, only two de novo classes are publicly available for our current study.

4 Discussion

In this work, we have modified an integrative model between case-control variants and de novo mutations used in autism studies (TADA)

to apply to schizophrenia data (extTADA). Even though extTADA is based on TADA, 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. We are assumming that different variant classes have similar the proportion of risk genes in the 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).

Current study's results replicate previous studies and supply more information about SCZ. Firstly, SETD1A is the most significant gene (FDR < 0.05) which was reported by Singh et al. (2016); Takata et al. (2016). In addion, two other genes: TAF13 and MIK67 also show as SCZ-risk candidates with FDR < 0.1. Secondly, significantly overlapping results between the top genes in this study and known gene sets such as constrained, DN SNPs and INDEL AST, FMRP, pLI09 genes show similar trends as the study of Genovese et al. (2016). Moreover, in this study, we tested for all the de novo SNPs and INDEL genes of DD, and have seen that this gene set is very strong overlap with the top genes from extTADA. Apart from those results, 14 GO gene sets significantly overlap with the top 500 extTADA genes (but not for the top 100 genes, mean that they strongly present between 100 and 500 top SCZ risk genes) (Table 5). Finally, from this study, rare-variant genetic architecture of SCZ is described detailedly. It is complex, and more complex than those of AST, ID, DD and EPI. 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 3 and 5, Table 4). Except for LoF de novo mutations, mean RRs of case-control variants were slightly larger than those of de novo mutations. In addition, the nearly equal values of mean RRs of LoF and MiD (~ 2 case-control variants replicated results reported by Genovese et al. (2016) using other statistics.

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 covergent results if multiple classes are integrated

into the estimation process (Figure 3, S4, S5). This situation is very important because extTADA (and also TADA) is developed for rare variants. Count information is usually not strong in these classes of variants. In addition, if we use one class to obtain the proportion of risk genes then it might not represent for other classes (Figure S5, Table S3). 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.

The pipeline can be used for other diseases. As be seen in the current study, we are able to use extTADA to infer genetic parameters for four other psychiatric diseases AST, EPI, DD and ID (Table 4, Figure 5). The results of AST are comparable with results of the He et al. (2013); De Rubeis et al. (2014), even though no risk-gene set was used as prior information in the estimation process. Regarding the mean RR of LoF mutations of EPI, the result of EuroEPINOMICS-RES Consortium et al. (2014) (~ 86) is in the current CI and approximate to the estimated value of our study (Table 4). Apart from that, genetic architecture can be distinct between classes; for example, recently Sifrim et al. (2016) have shown that for de novo protein-truncating variants (PTVs) and inherited PTVs in congenital heart defects (CHDs). Therefore, extTADA can be flexibily 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 strong in order to obtain reliable results if just some classes are used.

There are limits in the current SCZ study. Firstly, the model is developed to use for a non heterozyous population. This causes the case-control sample size smaller after the process of populuation adjustment. Even though we are assumming 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 S2) while SCZ trio size is not large in this study (1,024 trios). Therefore, the de

novo signal is probably not comparable with that of case-control signal. Finally, we are assumming 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 reflect exact 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), this assumption can be reliable.

5 Supplementary information

5.1 Sup Table

Mutation	dnControl	dnCase	lCI	uCI	odd ratio	p value
lof	43	111	1.34	2.87	1.94	0.000302
missense	334	612	1.45	2.15	1.77	7.04e-09
silent	134	227	0.994	1.62	1.27	0.0552
MiD	31	100	1.6	3.83	2.44	1.21e-05
silentCFPK	14	50	1.42	5.19	2.63	0.00109

Table S1: De novo mutations in trios and unaffected siblings. "silentCFPK" describes for silent mutations within frontal cortex-derived DHS (silentCerebrumfrontalocPk.narrowPeak). MiD mutations are missense mutations derived from 7 methods. Lower/upper confidence intervals (lCI/uCI) and odd ratios are calculated using Fisher's exact test.

Disease	Mutation	Count	Sample size	Mutation rate per sample
SCZ	silentCFPK	50	1024	0.05
	MiD	100	1024	0.1
	LoF	111	1024	0.11
AST	Missense	828	2231	0.37
	LoF	308	2231	0.14
ID	MiD	54	192	0.28
	LoF	63	192	0.33
EPI	MiD	273	156	1.75
	LoF	58	156	0.37
DD	MiD	4442	4293	1.03
	LoF	1080	4293	0.25

Table S2: De novo mutation counts of categories and their mutation rates per sample for schizophrenia (SCZ), autism spectrum disorder (AST), epilepsy (EPI), intellectual disorder (ID) and developmental disorder (DD).

LoF DN	0.073	0.012	0.273
MiD DN	0.015	0.001	0.225
silentCFPK DN	0.009	0.001	0.178
LoF CC	0.072	0.009	0.294
MiD CC	0.059	0.004	0.278
LoF+MiD DN	0.05	0.009	0.192
LoF+MiD CC	0.083	0.028	0.219
LoF DN+CC	0.09	0.038	0.239
MiD DN+CC	0.093	0.039	0.242

Table S3: Posterior mode and confidence intervals (95%) of proportions of risk genes for single classes and for combining two any classes.

5.2 Sup Figure

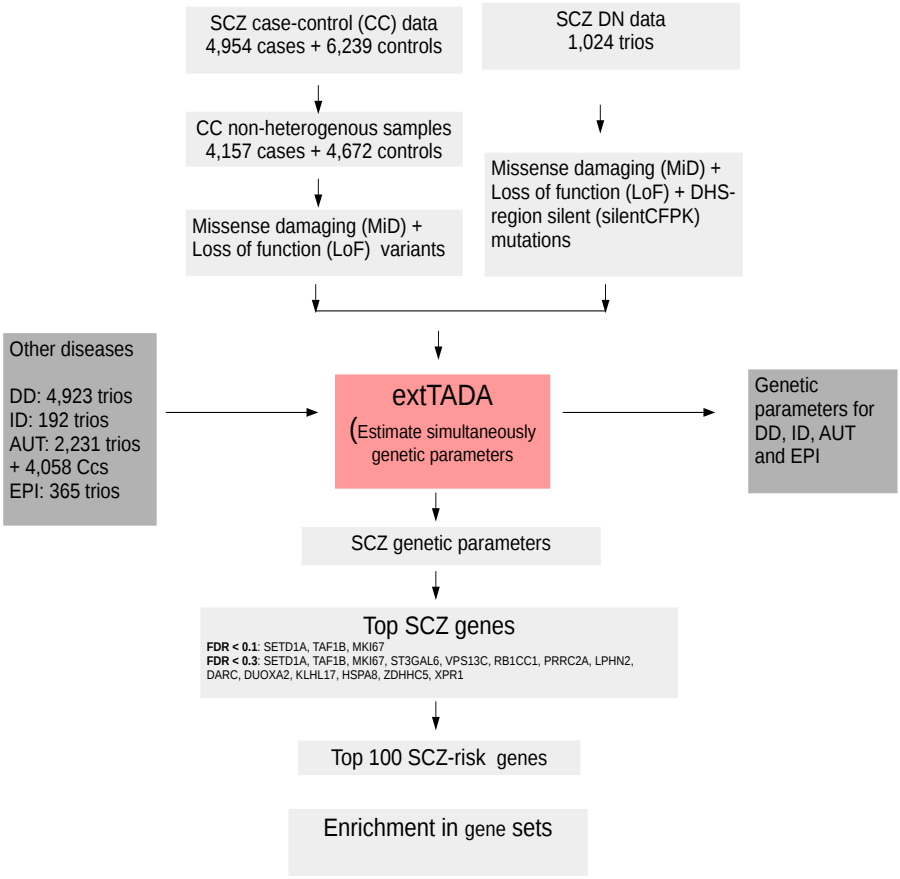

Figure S1: Workflow of data analysis.

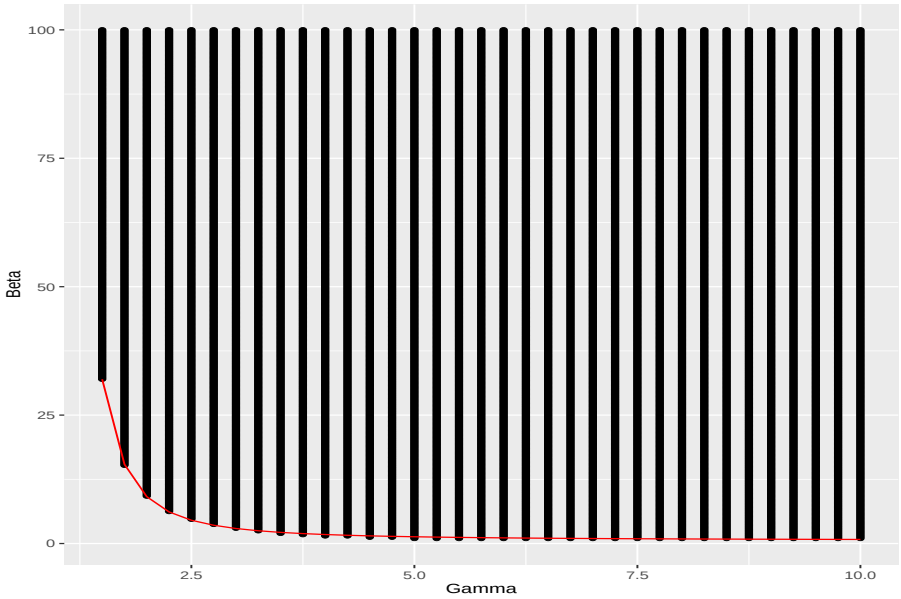

Figure S2: A grid of β and γ values. Points on the red line are corresponding with the proportion of protective variants less than 0.0%.

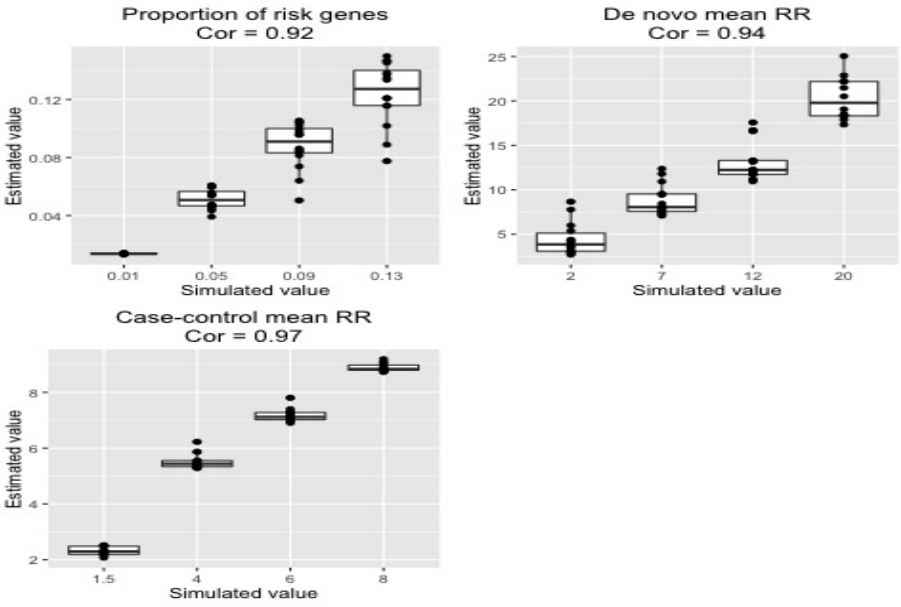

Figure S3: Estimated values for simulated data.

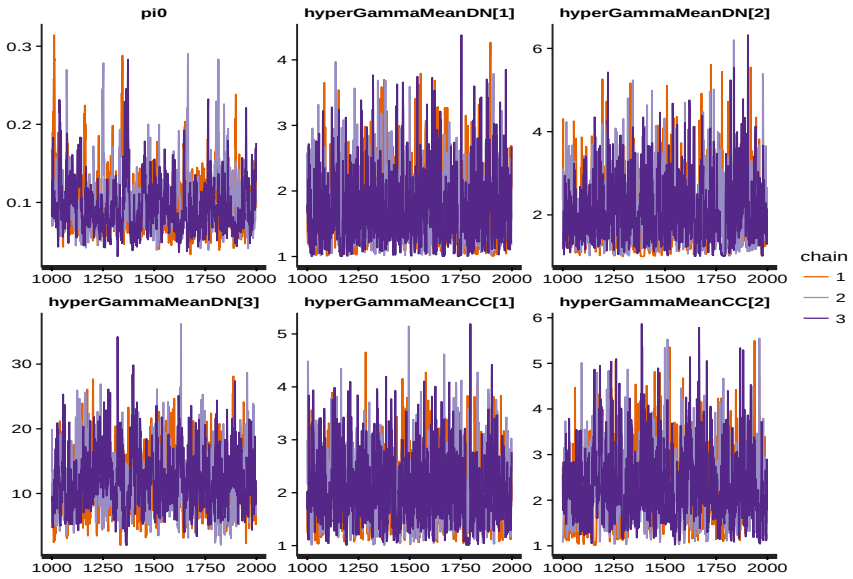

Figure S4: MCMC results for SCZ data

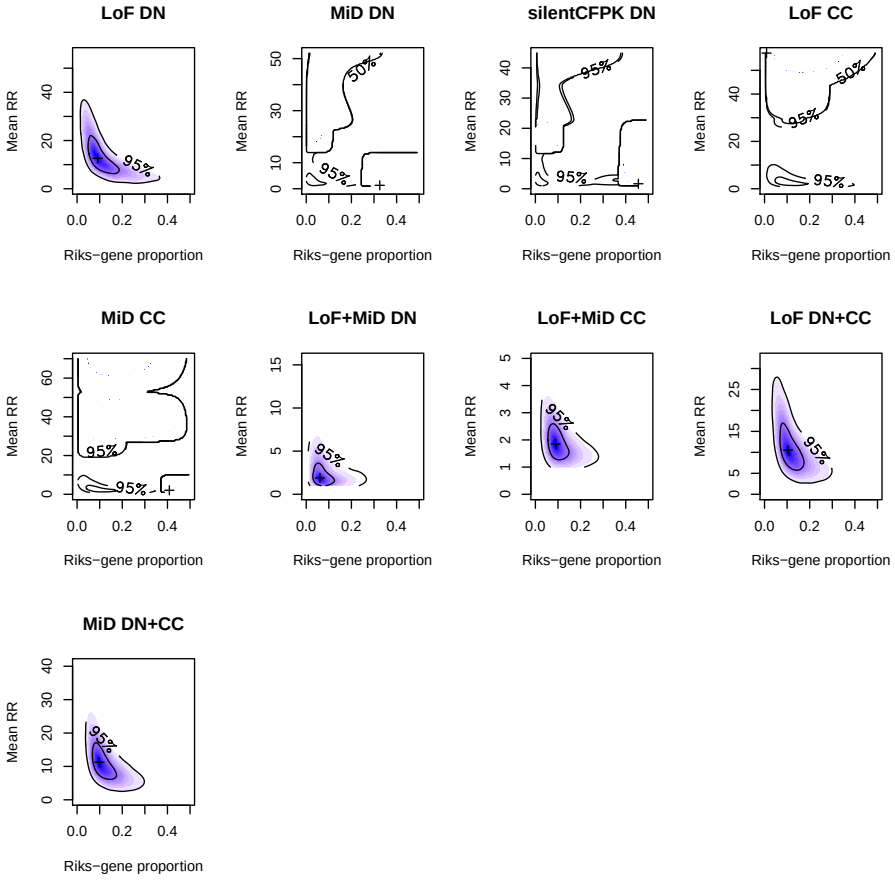

Figure S5: Estimation results for single classes or combination between two single classes of SCZ data using extTADA. Graphs only show the proportion of risk genes and one representative class for each estimation process.

References

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.

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.

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.

M. Fromer, P. Roussos, S. K. Sieberts, J. S. Johnson, D. H. Kavanagh, T. M. Perumal, D. M. Ruderfer, E. C. Oh, A. Topol, H. R. Shah, et al. Gene expression elucidates functional impact of polygenic risk for schizophrenia. *bioRxiv*, page 052209, 2016.

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.

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.

A. Gusev, N. Mancuso, H. K. Finucane, Y. Reshef, L. Song, A. Safi, E. Oh, S. McCaroll, B. Neale, R. Ophoff, et al. Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights. *bioRxiv*, page 067355, 2016.

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.

V. Kachitvichyanukul and B. Schmeiser. Computer generation of hypergeometric random variates. *Journal of Statistical Computation and Simulation*, 22(2):127–145, 1985.

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.

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.

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.

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.

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.

P. Roussos, B. Guennewig, D. Kaczorowski, G. Barry, and K. J. Brennand. Schizophrenia hipsc neurons display expression changes that are enriched for disease risk variants and a blunted activity-dependent response. *bioRxiv*, page 062885, 2016.

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.

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.

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.

Z. Zhu, F. Zhang, H. Hu, A. Bakshi, M. R. Robinson, J. E. Powell, G. W. Montgomery, M. E. Goddard, N. R. Wray, P. M. Visscher, et al. Integration of summary data from gwas and eqtl studies predicts complex trait gene targets. *Nature genetics*, 2016.