MAIN PAPER

August 3, 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 mployed the same model used previously in ASD studies. However, 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 missense damaging 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 anticipate this approach will improve our power to detect schizophrenia risk genes as more data is included.

NOTE

- All these results are in Figure 5, and Table 6. 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`

Contents

1 Introduction

2 Data and methods

2.1 Data

2.1.1 Simulated data

2.1.2 Schizophrenia data

Source	De novo	De novo control	Non/Transmitted	Case	Control
Fromer et al. (2014)	617		617		
Girard et al. (2011)	14				
Gulsuner et al. (2013)	105				
McCarthy et al. (2014)	57				
Xu et al. (2012)	231	34			
Rauch et al. (2012)		20 (ID)			
Giulio et al. (2016)				4954	6239
Total	1024	54	617	4954	6239

Table 1: De novo, transmitted/non-transmitted and case/control data. De novo trios are from schizophrenia (SCZ), intellectual disability (ID) studies.

These variants were annotated using Plink/Seq 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, splice, and frameshift variants. Missense damaging 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.

The file *wgEncodeOpenChromDnaseCerebellumocPk.narrowPeak.gz* was downloaded from http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeOpenChromDnase/ on April 20, 2016. Then, BEDTools (Quinlan and Hall, 2010) was used to intersect silent variants/mutations with the DHS regions.

2.1.3 Gene sets

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 used in this study.

Other gene sets

We also testest 18 gene sets described in Giulio et al (2016):

- Missense constrained genes from Table 2 of Samocha et al. (2014).

- Loss-of-function intolerant genes (Lek et al., 2015) from `ftp://ftp.broadinstitute.org/pub/ExAC_release/release0.3/functional_gene_constraint/fordist_cleaned_nonpsych_z_pli_rec_null_data.txt` on May 12, 2016.

2.2 Methods

2.2.1 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 for different classes.

For synonymous mutation rates within the frontal cortex-derived DHS, Takata et al. (2016) showed that there were 23 these mutations in a total of 154 silent mutations observed in controls. Therefore, we multiplied this proportion 23/154 with mutation rates of synonymous mutations to obtain mutation rates for this class.

2.2.2 Simulated data

We simulated ata sets of case/control variants alone, de novo mutations alone, and case/control variants and de novo mutations together. To see the influence of β on simulation results, we tested the simulation on two situations. Firstly, $\beta = 4$, which is the same as the case/control β in previous studies, was used. After that, we limitted the percentage of protective variants by using a simple function to constrain β and γ.

2.2.3 Obtain a homogeneous population for case-control data

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. Giulio 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 1. They reported significant results for private lof and damaging missense 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 1 (adjusting covariates) would not vary those results using Equation 2 (not adjusting covariates). The mclust package Version 5.2 (Fraley and Raftery, 1999) was used to cluster 11,161 samples (4,929 cases and 6,232 controls) into different groups. We tested the clustering results on three data

sets of the 11,161 samples. They were all 20 PCs, 20 PCs and total counts, and only the first three PCs. The number of groups were set between 2 and 6. For each clustering time, Equation 1 and 2 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 (rPvalue) was calculated. Next, to choose reliable results from the clustering process, we set criteria:

- rPvalue ≥ 0.85 and p-values for NonEXAC 0.005.
- Ratio p1/p2 from Equation 1 and 2 had to between 0.1 and 1.

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

$$\begin{aligned} SCZ &\sim count \\ count &\sim SCZ \end{aligned} \tag{2}$$

2.2.4 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 framework which is extended from the The Transmission and Disequilibrium Association (TADA) model proposed by He et al. (2013). For a given gene, all variants of a class (e.g., LoF, missense damaging) 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_1) control (x_0) data as Equation 3:

$$\begin{aligned} x_d &\sim Pois(2N\mu\gamma_d) \\ x_1 &\sim Pois(qN_1\gamma) \\ x_0 &\sim Pois(qN_0) \end{aligned} \tag{3}$$

in which N_d, N_1, N_0 are sample sizes of trios, cases and controls respectively; γ_d and γ are relative risks for de novo mutations and case-control variants.

At i^{th} gene, two hypotheses: $H_0 : \gamma = 1$ versus $H_1 : \gamma > 1$ are compared using The Bayes Factor (BF):

$$\begin{aligned} B_i &= \frac{P(x|H_1)}{P(x|H_0)} \\ &= \frac{\prod_{j=1}^{K} P(x_{ij}|H_1)}{\prod_{j=1}^{K} P(x_{ij}|H_0)} \ (K: \ number \ of \ categories) \\ &= \prod_{j=1}^{K} \frac{P(x_{ij}|H_1)}{P(x_{ij}|H_0)} \ (Independence \ between \ categories) \\ &= \prod_{j=1}^{K} B_{ij} \end{aligned} \tag{4}$$

Where B_{ij} is the BF of the gene for j^{th} category:

$$\begin{aligned} B_{ij} &= \frac{\int P(x_{ij}|\gamma,q)P(q|H_1)P(\gamma|H_1)dqd\gamma}{\int P(x_{ij}|\gamma,q)P(q|H_0)P(\gamma|H_0)dqd\gamma} \\ &\underset{=}{\gamma_{H_0}=1} \frac{\int P(x_{ij}|\gamma,q)P(q|H_1)P(\gamma|H_1)dqd\gamma}{\int P(x_{ij}|q)P(q|H_0)dq} \end{aligned} \tag{5}$$

Or $BF_{ij} = BF_{ij(dn)}BF_{ij(CC)}$

The same as He et al. (2013), gamma distributions are assummed as prior distributions for γ_d and γ as in 6.

$$\begin{aligned} &\gamma_d \sim Gamma(\bar{\gamma_d}\beta_d, \beta_d) \\ &\gamma \sim Gamma(\bar{\gamma}\beta, \beta) \\ &q \sim Gamma(\rho, \nu) \end{aligned} \tag{6}$$

Regarding priors of thhe 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) ; howerver, 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, we need to know hyper parameters in Equation 6. Let ϕ_{1j} and ϕ_{0j} be hyperparameters for H_1 and H_0 respectively. A mixture model of the two hypotheses were used to infer parameters using information across the number of tested gene (m) as in Equation 7.

$$P(x|\phi_1, \phi_0) = \prod_{i=1}^{m} \left[\pi \prod_{j=1}^{K} P(x_{ij}|\phi_{1j}) + (1-\pi) \prod_{j=1}^{K} P(x_{ij}|\phi_{0j}) \right] \tag{7}$$

In the Equation 7, differently from the original TADA model, we integrated all categories into the mixture model as described in our method for calculating BFs in Equation 4. 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 7 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 3.
- For case-control (inheritance) data:
  - $\frac{\rho}{\nu}$ represented for prior mean of q and ν controlled the dispersion of the prior of q; therefore as in the previous study of De Rubeis et al. (2014) we chose $\nu = 200$ and $\frac{\rho}{\nu}$ = the mean frequency across genes by using both case and control data.
  - **Approximate (simplify) case/control model**

$$\begin{aligned} P(x_1, x_0|H_j) &= P(x_1, x_1 + x_0|H_j) \\ &= P(x_1|x_1 + x_0, H_j)P(x_1 + x_0|H_j) \end{aligned} \tag{8}$$

    * The first part: $P(x_1|x_1 + x_0, H_j)$
    Because of $x_1 \sim Pois(N_1 q\gamma)$ and $x_0 \sim Pois(N_0 q)$, we assumed that x_1 and x_0 were **independent**, we had:
    $x_1|x_1 + x_0, H_j \sim Binomial(x_1 + x_0, \theta|H_j)$
    with $\theta|H_1 = \frac{N_1\gamma}{N_1\gamma+N_0}$ and $\theta|H_0 = \frac{N_1}{N_1+N_0}$
    The marginal likelihood was
    $P(x_1|x_1 + x_0, H_j) = \int P(x_1|x_1 + x_0, \gamma, H_j)P(\gamma|x_1 + x_0, H_j)d\gamma$
    * The second part $P(x_1 + x_0|H_j)$ was not used in the estimation process in Equation 7

Change the order of integrals to rely only on relative risks

$$P(x_1, x_0|H_j) = P(x_0|H_j)P(x_1|x_0, H_j) \tag{9}$$

- The first part $P(x_0|H_j)$ was the same as De Rubeis et al. (2014):

$$P(x_0|H_j) = \int P(x_0|q, H_j)P(q|\rho, \nu, H_j)dq = NegBin(x_0|\rho, \frac{N_0}{\nu + N_0}), j = 0, 1 \tag{10}$$

- The second part:

$$\begin{aligned} P(x_1|H_j, x_0) &= \int P(x_1|q, \gamma)P(q|H_j, x_0)P(\gamma|H_j)dqd\gamma \\ &= \int [P(x_1|q, \gamma)P(q|H_j, x_0)dq]\, P(\gamma|H_j)d\gamma \\ &= \int NegBin(x_1|\rho + x_0, \tfrac{N_0+\nu}{N_1\gamma+N_0+\nu})P(\gamma|H_j)d\gamma \end{aligned} \tag{11}$$

The second line in Equation 11 is because $P(q|H_j, x_0)$ is the posterior probability of q after seeing the data x_0 with $q|H_j, x_0 \sim Gamma(\rho + x_0, \nu + N_0)$ (De Rubeis et al., 2014).

In Equation 11

$$\begin{aligned} x_d &\sim Pois(2N_d\gamma_d) \\ \gamma_d &\sim Gamma(\bar{\gamma_d}\beta_d, \beta_d) \\ \bar{\gamma_d} &\sim Normal(15, 10) \\ \beta_d &\sim Normal(\beta_{d_s}, 0.01) \end{aligned} \tag{12}$$

$x \sim Pois(2N_{dn}\mu\gamma_{dn})$	$\gamma_{dn} \sim Gamma(\bar{\gamma_{dn}}\beta, \beta)$	$\bar{\gamma_{dn}} \sim Normal(15, 15)$ $\beta \sim Normal(1, 0.1)$
$x_1 \sim Pois(N_1 q\gamma)$	$\gamma \sim Gamma(\bar{\gamma}\beta, \beta)$ $q \sim Gamma(\rho, \nu)$	$\bar{\gamma} \sim Gamma(1, 0.1)$ $\beta \sim Normal(\beta_0, 0.1)$ $\rho = mean(x_0), \nu = 200$
$x_0 \sim Pois(N_0 q)$	$q \sim Gamma(\rho, \nu)$	$\rho = mean(x_0), \nu = 200$

(13)

$x_{dn} \sim P(2N_{dn}\mu\gamma_{dn})$	$\gamma_{dn} \sim Gamma(\bar{\gamma_{dn}} * \beta_{dn}, \beta_{dn})$	$\bar{\gamma_{dn}} \sim Gamma(1, 0.1)$ $\beta = 4$
$x_1 \sim P(N_1 q\gamma_{cc})$	$\gamma_{cc} \sim Gamma(\bar{\gamma_{cc}} * \beta_{cc}, \beta_{cc})$	$\bar{\gamma_{cc}} \sim Gamma(1, 0.1)$ $\beta_{cc} = 4$
	$q \sim Gamma(\rho, \nu)$	$\frac{\rho}{\nu} = mean$ $\nu = 200$
$x_0 \sim P(N_0 q)$	$q \sim Gamma(\rho, \nu)$	$\frac{\rho}{\nu} = mean$ $\rho = 200$

(14)

2.2.5 Set parameters for CC

To control for the proportion of protective variants, we tested the relationship between β and γ. We set this proportion very low (0.5%) and built a nonlinear relationship for β and γ values as in Equation 15 (Figure 1). 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.

$$\beta = e^{a*\gamma^b + c} \tag{15}$$

3 Results

3.1 Simulated data

3.1.1 Only case-control data

To test the approximate model for case-control data, we simulated different combinations of $\bar{\gamma}_{cc}$ and π. Table 2 shows correlations between simulated values and estimated values.

The Equation 15 was used to

Check priors for the approximate case-control model

3.1.2 For de novo and case/control data

Using prior results for CC above, we used the simulation method in TADA package to simulate for different combinations of RRs of CC and DN data. We then

Prior	Pcor	RRcor	BetaCor	AdjustBeta(Yes/No)
GammaPriorLowerGamma.0.	0.18	0.54	-0.03	1
GammaPriorLowerGamma.1	**0.97**	**0.99**	**1**	**1**
GammaPriorLowerGamma.0.5	0.18	0.53	-0.03	1
NormalPriorLowerGamma.0.	0.18	0.39	-0.01	1
NormalPriorLowerGamma.1.	0.96	0.66	0.93	1
NormalPriorLowerGamma.0.5	0.18	0.39	-0.01	1
GammaPriorLowerGamma.0.	0.46	0.59	NA	0
GammaPriorLowerGamma.1	0.83	0.99	NA	0
GammaPriorLowerGamma.0.5	0.65	0.83	NA	0
NormalPriorLowerGamma.0.	0.46	0.66	NA	0
NormalPriorLowerGamma.1.	0.81	0.99	NA	0
NormalPriorLowerGamma.0.5	0.56	0.79	NA	0

Table 2: Simulation results for the approximate case-control model. The table shows correlations between π, γ, β (Pcor, RRcor, BetaCor respectively) for simulated and estimated values. The last column shows an option to describe whether β values are adjusted by using γ values as in Equation 15.

re-estimated parameters using the approximate CC model and denovo model. Correlations between simulated and median estimated values were calculated. They were high for π and CC relative risks (0.93, 0.95 and 0.93 for π and two CC relative risks respectively), but relative high for γ of de novo data (0.74 and 0.65 respectively (Figure 3). Table 3 shows different percentiles for estimated values and their corresponding simulated values.

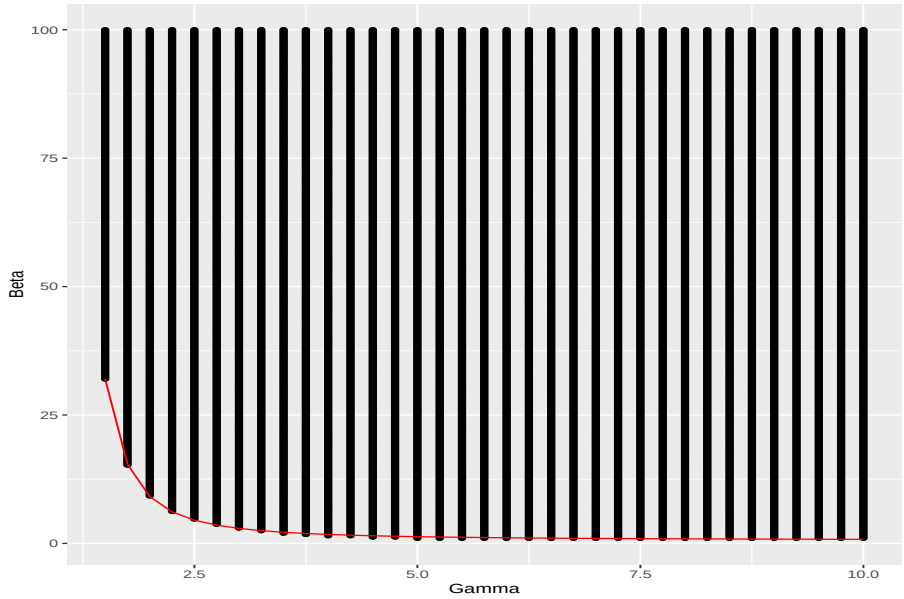

Figure 1: A grid of β and γ values. Points on the red line are corresponding with the proportion of protective variants less than 0.0%.

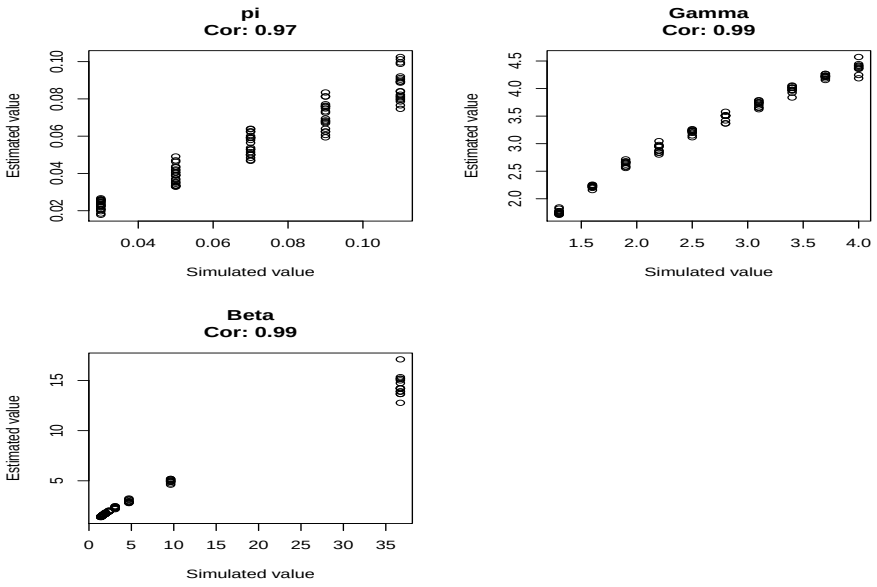

Figure 2: Correlations between simulated and median estimated values.

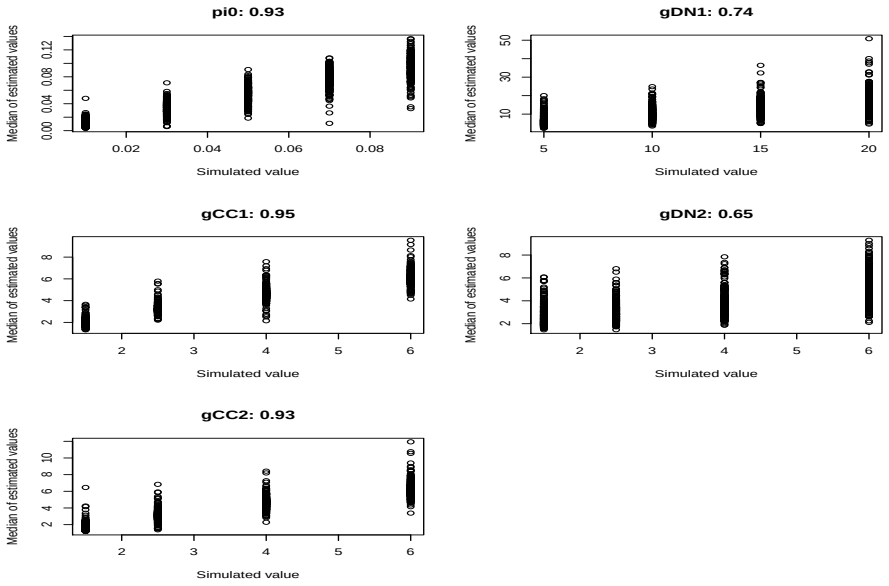

Figure 3: Correlations between simulated data from original TADA model and the median estimated values of the approximate model.

Parameters	Simulated value	5%	50%	95%
pi0	0.01	0.006	0.013	0.022
pi0	0.03	0.017	0.034	0.048
pi0	0.05	0.033	0.056	0.074
pi0	0.07	0.053	0.078	0.099
pi0	0.09	0.066	0.096	0.119
gDN1	5	3.245	6.092	12.664
gDN1	10	5.669	9.88	16.535
gDN1	15	7.809	14.025	20.643
gDN1	20	8.759	17.734	26.393
gCC1	1.5	1.591	1.956	2.547
gCC1	2.5	2.674	3.22	4.025
gCC1	4	3.928	4.645	5.568
gCC1	6	5.098	6.117	7.184
gDN2	1.5	1.722	2.508	4.49
gDN2	2.5	2.01	2.921	4.617
gDN2	4	2.368	3.65	6.217
gDN2	6	2.968	5.034	7.594
gCC2	1.5	1.292	1.709	2.471
gCC2	2.5	2.176	3.128	4.187
gCC2	4	3.564	4.64	5.831
gCC2	6	4.975	6.133	7.821

Table 3: Different percentiles of estimated values for simulation data of two classes.

3.2 Schizophrenia data sets

3.2.1 Enrichment results for different classes of de novo mutations and case-control variants

Results for three classes of de novo mutation are showed in Table 4. Significant results were observed for LoF, missense damaging mutations (p = $3.02x10^{-4}$ and $1.21x10^{-5}$ respectively). As reported by Takata et al. (2016), silent mutations hitting the DHS $Cerebrum_frontal_OC$ also had similar results (p = $1.09x10^{-3}$); however this trend was not observed for all silent mutations (p = 0.0552).

De novo counts for different classes are in Table 4.

rownames(outDNData)	dnControl	dnCase	V1	V2	odds.ratio	V4
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
damaging missense	31	100	1.6	3.83	2.44	1.21e-05
silentCerebellumocPk.narrowPeak	14	30	0.788	3.18	1.55	0.216
silentCerebrumfrontalocPk.narrowPeak	14	50	1.42	5.19	2.63	0.00109
silentFrontalcortexocPk.narrowPeak	21	55	1.13	3.37	1.92	0.0122

Table 4: De novo mutations in trios and unaffected siblings. "Silent FCdDHS" describes for silent mutations within frontal cortex-derived DHS. Missense damaging mutations are missense mutations derived from 7 methods.

Case-control data were clustered into different groups and Equation 1 and 2 were used to calculate p values for the largest population after each clustering process. Similar results before and after adjusting for covariates were similar for the homogeneous population (Figure 4).

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

All six categories (LoF, damaging missense and silentCFPK) of de novo mutations and case-control variants were used in the integration analysis process. However, as showed in Figure 4, silentCFPK case-control variants did not significantly show excess counts in cases against controls. Therefore, RR of this category was set 1 and β was set 1000. The five other categories which showed enrichment in their own category were used in the integration analysis process. They included LoF, damaging missense and silentCFPK denovo mutations as well as LoF and damaging missense case-control variants. Therefore, the real

3.2.3 TADA results

pi = 0.06172522 gDN1 = 17.58647

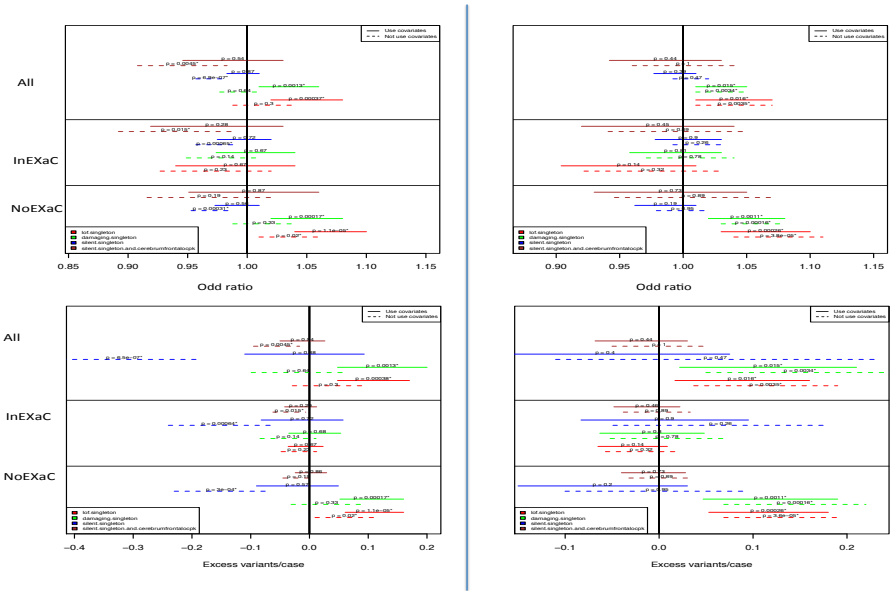

Figure 4: Odd ratios and excess variants in cases for the analysing of all case-control samples. Left panels show results for all samples before adjustment for population stratification while right panels describe results for only one homogeneous population. Top pictures are results of modeling SCZ status (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.

3.3 Enrichment analyses

We tested the enrichment of the schizophrenia gene set with FDR < 0.3 in xx other gene sets. Highest enrichment was observed in the FMRP gene set (3.99992e-05) followed by RBFOX2, constrained, RBFOX13 and synaptome (5.99988e-05, 7.99984e-05, 0.0002399952, 0.0009399812 respectively). We aslo saw significant results in SNPs and Indel de novo gene set of autism (0.005959881), as well as PSD (0.01101978), and CELF4 (0.01705966).

The results were not significant in CNV de novo gene sets of SCZ, ASD, BD, CHD, EPI, and the SCZ GWAS gene set.

pi0	0.041	0.168	0.093
hyperGammaMeanDN[1]	1	2.718	1.629
hyperGammaMeanDN[2]	1	3.568	1.937
hyperGammaMeanDN[3]	3.274	20.441	11.296
hyperGammaMeanCC[1]	1.006	3.217	1.92
hyperGammaMeanCC[2]	1.01	3.737	2.193

Table 5: Estimated parameters for de novo and case-control SCZ data. These results are obtained by running sampling 50000 MCMC times.

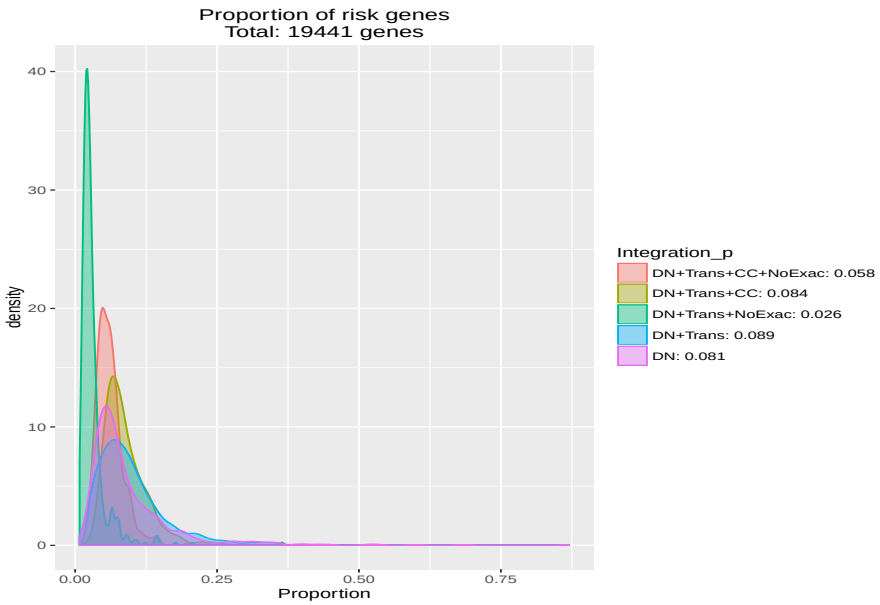

Figure 5: MCMC results of proportion of risk genes for the combination of 2 classes

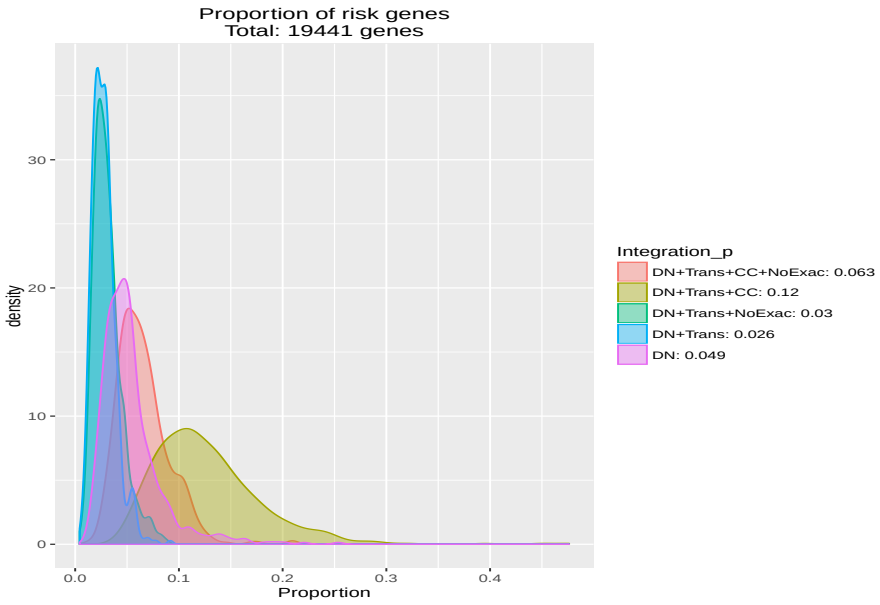

Figure 6: MCMC results of proportion of risk genes for the combination of 3 classes

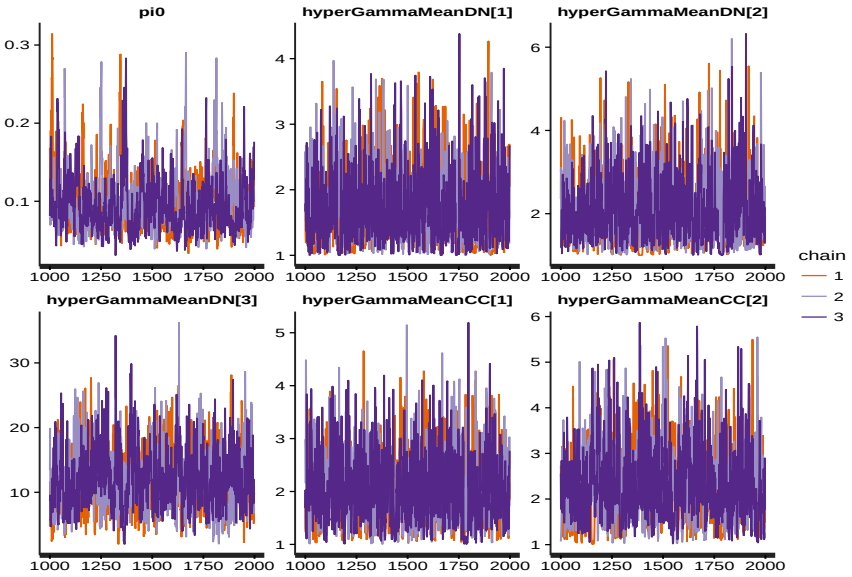

Figure 7: MCMC results for SCZ data

Gene set	TransTop100	TransnoexacTop100	TransCCnoexacTop100	TransCCTop100	Trans(FDR0.3)	Transnoexac(FDR0.3)	TransCCnoexac(FDR0.3)	TransCC(FDR0.3)
FromSZGR.160byLewis.gene	1	0.53	1	1	1	1	1	1
FromSZGR.1974GO$_neurodevelopment.gene$	0.0016	6e-04	0.0016	4e-04	0.24	0.016	0.15	0.25
FromSZGR.coreGeneSet.gene	1	1	1	1	1	1	1	1
FromSZGR.173byNg.gene	0.54	0.042	1	0.55	1	1	1	1
FromSZGR.75genesByCOR.gene	1	0.29	1	1	1	1	1	1
antipsychotics-combined.set.gene	0.22	0.081	0.0052	0.22	1	1	1	0.38
geneInPARs.txt	0.12	0.016	0.016	0.045	0.25	0.17	0.06	0.0062
geneInHARs.txt	0.32	0.15	0.16	0.78	0.27	0.19	0.36	0.2
listMcRae2016.txt	0.0022	0.34	0.0012	0.007	1	1	0.0018	0.12
celf4.txt	0.0014	0.24	2e-04	0.017	0.76	1	0.12	0.0078
constrained.txt	2e-04	0.017	2e-04	2e-04	0.4	0.035	2e-04	0.012
CNV.denovo.gain.asd.txt	0.49	0.026	0.11	0.11	0.46	1	0.23	0.5
CNV.denovo.gain.bd.txt	1	1	1	1	1	1	1	1
CNV.denovo.gain.scz.txt	1	1	1	0.56	1	1	1	1
CNV.denovo.loss.asd.txt	0.38	0.38	0.13	0.55	0.095	0.29	0.57	0.78
CNV.denovo.loss.bd.txt	0.45	0.43	0.43	0.43	1	1	1	1
CNV.denovo.loss.scz.txt	0.2	0.57	0.2	0.19	1	1	0.12	0.21
SNPsINdel.denovo.aut.txt	2e-04	2e-04	2e-04	2e-04	0.18	0.059	0.0024	0.043
SNPsINdel.denovo.chd.txt	0.026	0.11	0.001	0.0066	0.006	0.072	0.014	1
SNPsINdel.denovo.epi.txt	0.059	0.018	0.002	0.19	0.16	1	0.2	1
SNPsINdel.denovo.id.txt	0.034	0.034	0.11	0.34	1	1	1	1
SNPsINdel.denovo.scz.txt	2e-04	2e-04	2e-04	2e-04	2e-04	2e-04	2e-04	2e-04
fmrp.txt	0.0024	0.027	2e-04	2e-04	0.49	0.34	2e-04	0.002
gwas.txt	0.34	0.33	0.06	0.33	0.042	0.026	0.066	0.12
rbfox13.txt	2e-04	0.021	2e-04	2e-04	1	0.69	0.011	0.0042
rbfox2.txt	2e-04	0.015	2e-04	2e-04	0.82	0.65	0.004	0.0068
synaptome.txt	0.027	0.015	4e-04	2e-04	0.63	1	0.41	0.11
psd.txt	0.068	0.0018	6e-04	0.0028	0.53	1	0.29	0.32
psd95.txt	0.42	0.41	0.41	0.41	1	1	1	1
pLI09.txt	2e-04	0.0022	2e-04	4e-04	0.86	0.31	0.0026	0.002

Table 6: Test overlapping gene sets with extTADA results for two classes.

Gene set	TransCCnoexacTop100	TransCCTop100	TransTop100	TransnoexacTop100	TransCCnoexac(FDR0.3)	TransCC(FDR0.3)	Trans(FDR0.3)	Transnoexac(FDR0.3)
FromSZGR.160byLewis.gene	1	1	1	1	1	1	1	1
FromSZGR.1974GO$_{n}$*eurodevelopment.gene*	0.002	0.0012	2e-04	2e-04	0.56	0.22	0.47	1
FromSZGR.coreGeneSet.gene	1	1	1	1	1	1	1	1
FromSZGR.173byNg.gene	1	1	0.55	0.55	1	1	1	1
FromSZGR.75genesByCOR.gene	1	1	1	1	1	1	1	1
antipsychotics-combined.set.gene	0.22	0.024	0.024	0.023	0.14	0.13	1	1
geneInPARs.txt	0.12	0.12	0.27	0.25	0.21	0.24	0.17	0.048
geneInHARs.txt	0.31	0.54	0.54	0.78	1	1	0.19	1
listMcRae2016.txt	0.07	0.068	0.065	0.066	0.036	0.15	1	1
celf4.txt	0.039	0.059	0.34	0.25	0.69	0.29	0.59	1
constrained.txt	2e-04	0.0062	2e-04	4e-04	2e-04	0.002	0.28	0.097
CNV.denovo.gain.asd.txt	0.053	0.19	0.1	0.11	0.39	0.35	1	1
CNV.denovo.gain.bd.txt	1	0.57	0.54	1	1	1	1	1
CNV.denovo.gain.scz.txt	0.2	0.19	0.55	1	1	0.26	1	1
CNV.denovo.loss.asd.txt	0.37	0.23	0.026	0.061	1	0.56	0.048	0.095
CNV.denovo.loss.bd.txt	1	1	0.43	0.44	1	1	1	1
CNV.denovo.loss.scz.txt	0.2	1	0.56	0.2	1	1	1	1
SNPsINdel.denovo.aut.txt	8e-04	0.025	6e-04	2e-04	0.028	0.052	0.063	1
SNPsINdel.denovo.chd.txt	0.32	0.69	0.32	0.11	0.1	0.35	0.08	1
SNPsINdel.denovo.epi.txt	0.064	0.18	0.19	0.065	0.13	0.43	1	1
SNPsINdel.denovo.id.txt	0.032	0.33	0.1	0.0066	0.11	0.068	1	1
SNPsINdel.denovo.scz.txt	2e-04	2e-04	2e-04	2e-04	2e-04	2e-04	2e-04	0.0014
fmrp.txt	2e-04	2e-04	2e-04	2e-04	2e-04	2e-04	0.054	0.12
gwas.txt	1	1	0.33	0.33	1	1	0.026	0.0084
rbfox13.txt	8e-04	0.0058	0.002	0.0018	0.0064	0.02	0.69	1
rbfox2.txt	0.0022	0.011	0.0022	0.004	0.37	0.066	0.65	0.27
synaptome.txt	6e-04	0.0026	0.028	0.0032	0.18	0.014	1	1
psd.txt	2e-04	0.071	0.072	8e-04	0.12	0.24	1	1
psd95.txt	1	0.41	0.42	0.1	1	1	1	1
pLI09.txt	6e-04	2e-04	2e-04	0.0044	0.0014	0.0012	0.31	0.31

Table 7: Test overlapping gene sets with extTADA results for three classes

4 Discussion

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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

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.