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

Hoang T Nguyen[1], Amanda Dobbyn[1], Laura M Huckins[1], Douglas M Ruderfer[1,2], Gulio Genovese[3,4], Menachem Fromer[1,5], Xinyi Xu[6], Joseph Buxbaum[6], Christina Hultman[7], Pamela Sklar[1,3], Shaun M Purcell[1,3,8], Xin He[9], Patrick F Sullivan[7,10], and Eli A Stahl* [1,3]

[1]Division of Psychiatric Genomics, Department of Genetics and Genomic Sciences, Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, New York 10029, USA.
[2]Vanderbilt University Medical Center, Vanderbilt University, Nashville, TN 37235, USA.
[3]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA.
[4]Department of Genetics, Harvard Medical School, Massachusetts, USA.
[5]Verily Life Sciences, USA.
[6]Seaver Autism Center, Icahn School of Medicine at Mount Sinai, New York, New York 10029, USA.
[7]Department of Medical Epidemiology and Biostatistics, Karolinska Instituet, Stockholm, Sweden.
[8]Sleep Center, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, USA.
[9]Department of Human Genetics, University of Chicago, Chicago, IL 60637, USA.
[10]Departments of Genetics and Psychiatry, University of North Carolina, Chapel Hill, North Carolina 27599-7264, USA.

May 8, 2017

*eli.stahl@mssm.edu

Abstract

Integrating rare variation from family and case/control studies has successfully implicated specific genes contributing to risk of autism spectrum disorder (ASD). In schizophrenia (SCZ), however, while sets of genes have been implicated through study of rare variation, very few individual risk genes have been identified. Here, we apply hierarchical Bayesian modeling of rare variation in schizophrenia and describe the proportion of risk genes and distribution of risk variant effect sizes across multiple variant annotation categories. Briefly, we developed a pipeline based on the previous work used in ASD studies to jointly estimate genetic parameters for one or multiple combined populations of any disease. We applied this method to the largest available collection for rare variants in schizophrenia (1,077 families, 6,699 cases and 13,028 controls). We defined five variant annotation categories: disruptive (nonsense, frameshift, essential splice site mutations), damaging (predicting damaging by seven algorithms), silentFCPk (silent mutations within frontal cortex-derived DHS peaks) de novo mutations, and disruptive and damaging missense case/control singletons. We estimated that 8.01% of genes are risk genes (95% credible interval, CI, 4.59-12.9%), with mean effect sizes (95% CIs) of 12.25 (4.8- 22.22) for disruptive de novos, 1.44 (1-3.16) for missense damaging de novos, and 1.22 (1-2.16) for silentFCPk de novos. The mean effect sizes of damaging and disruptive singleton variants for three case-control populations were 2.09 (1.04-3.54), 2.44 (1.04, 5.73) and 1.04 (1-1.19) respectively. Our analysis identified only two known SCZ risk genes with FDR<0.05: SETD1A and TAF13; and two other genes with FDR < 0.1: RB1CC1 and PRRC2A. We further used FDRs to directly analyze candidate gene sets for the enrichment of Bayesian support. Significant enrichments were observed for essential genes, which were found enriched among autism genes in a recent study, and central nervous system (CNS) related genes, in addition to gene sets previously found to be enriched (including in these data). We conduct power analyses under our inferred model for SCZ, estimating the number of risk gene discoveries as more data become available, and quantifying the greater value of case/control over trio samples for novel rare variant risk gene discovery. We also applied the method to four other neurodevelopmental disorders: autism spectrum disorder (ASD), intellectual disorder (ID), developmental disorder (DD) and epilepsy (EPI), in total 10,792 families, and 4,058 cases and controls. The predicted proportions of risk genes in these diseases were smaller than that in SCZ, 4.6% in ASD, and $< 3\%$ for the other disorders. We report 164 and 58 genes with FDR < 0.05 for DD and ID, respectively, 101 and 15 of which are novel. Overall, replication of previous results confirms the robustness of our approach, and our method is able to identify novel risk genes for SCZ as well as for other diseases.

Contents

1 Introduction **5**

2 Results **6**
2.1 The extTADA pipeline . . . 6
2.2 Evaluating extTADA on simulated data . . . 8
2.3 extTADA Analyses of Schizophrenia . . . 9
2.3.1 SCZ data . . . 10
2.3.2 Rare variant genetic architecture of SCZ . . . 10
2.3.3 Enrichment of gene sets in extTADA SCZ risk genes candidates . . . 14
2.3.4 Power analysis for SCZ exome sequencing studies across sample sizes . . . 17
2.4 extTADA Analyses of Other Neurodevelopmental Disorders . . . 18
2.4.1 Rare variant genetic architectures of ASD, ID, DD, EPI . 18
2.4.2 Novel risk genes in ID and DD . . . 20
2.4.3 Multiple gene sets are enriched in top significant genes across neurodevelopmental diseases . . . 21

3 Discussion **21**

4 Data and methods **24**
4.1 Data . . . 24
4.1.1 Variant data of SCZ, ID, DD, EPI and ASD . . . 24
4.1.2 Gene sets . . . 25
4.2 Methods . . . 26
4.2.1 extTADA pipeline: extended transmission (case-control) and de novo analysis . . . 26
4.2.2 Testing the model on simulated data . . . 28
4.2.3 Calculate mutation rates . . . 29
4.2.4 Analyze SCZ data . . . 30
4.2.5 Use extTADA to predict genetic parameters of other neurodevelopmental diseases . . . 32
4.2.6 Infer parameters using MCMC results . . . 32

5 Acknowledgements **33**

6 Supplementary information **34**
6.1 Supplementary Tables . . . 34
6.2 Sup Figure . . . 45
6.3 Sup Information . . . 56
6.3.1 Sup Results . . . 56
6.3.2 Sup methods . . . 57

1 Introduction

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

Rare variants in case/control samples and de novo mutations have been successfully leveraged to implicate biologically relevant gene sets for this disease (Purcell et al., 2014; Fromer et al., 2014; Genovese et al., 2016), and to identify a handful specific SCZ risk genes (Singh et al., 2016; Takata et al., 2016). However, the genetic architecture of SCZ for rare variants and de novo mutations remains unknown. Rare variant genetic architecture analyses could help gain further insights into this disease, for example by using the estimated number of risk genes to calibrate gene discovery false discovery rates, or by using the distribution of effect sizes to estimate power for rare variant association studies. A better understanding of our certainty in sets of risk genes for SCZ will provide a better picture of biological pathways specific for the disease.

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

In this study, we used `extTADA` to analyze the largest available exome-sequence data, including 19,727 (6,699+13,028) case+control samples and 1,077 trio/quad families for SCZ. We estimated mean relative risks (RRs) of different variant annotation categories as well as the proportion of risk genes for disease. Based on this analysis, SCZ risk gene sets determined with different false dis-

covery rate (FDR) thresholds were tested for enrichment in known and novel gene sets. Analysis of separate classes of variants/mutations in terms of annotation and rarity helps provide a detailed picture of the disease's rare variant genetic architecture, allowing for example power analyses for risk gene discovery as more data become available. Finally, we used available data for four other neurodevelopmental diseases: intellectual disability (ID), autism spectrum disorder (ASD), epilepsy (EPI) and developmental disorder (DD), totaling 10,792 trios and 4058 cases/control samples. We are able to identify additional new significant genes for ID and DD based on `extTADA` results.

2 Results

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

2.1 The `extTADA` pipeline

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

In summary, for a given gene, all variants of a given annotation category (e.g. loss-of-function) were collapsed and considered as a single count. Let q, γ and μ be the population frequency of rare heterozygous genotypes for case/control (equivalently, transmitted/nontransmitted) data, the mean relative risk (RR) of the variants, and sum of mutation rates of de novo variants, respectively. At each gene, two hypotheses $H_0 : \gamma = 1$ and $H_1 : \gamma \neq 1$ were compared. A fraction of the genes π, assumed to be risk genes, were represented by the H_1

model. Under this model, mean relative risks (γ) were assumed to follow a probability distribution across genes. The model H_0 described non-risk genes, for which relative risks (γ) equal 1. As in He et al. (2013), we modeled de novo (x_d) and case (x_{ca}) control (x_{cn}) data as Poisson distributions and their hyper parameters as following Gamma distributions priors. In addition, in `extTADA`, we used a Beta distribution prior for π and constrain π to be less than 0.5, and a nonlinear function for the variance parameter of γ to constrain mean RRs above 1 (i.e. so that variants are not implied by the model to be protective). Model parameters for `TADA` are shown in in Table 1.

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

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

At each gene, a Bayes Factor (BF_{gene}) can be calculated for each category to compare models H_1 and H_0 ($BF = P(data|H_1)/P(data|H_0)$). BF_{gene} can be calculated as the product of BFs across multiple variant categories, either DN and CC data or multiple annotation categories. Data could be from heterogeneous population samples; therefore, we extended `TADA`'s BF_{gene} as the product of BFs of all variant categories including population samples as in Equation 1,

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

in which Ndn_{pop}, Ncc_{pop} are the numbers of DN and CC population samples, and C_{dn}, C_{cc} are the number of annotation categories in DN and CC data. To infer significant genes, BFs were converted to false discovery rates (FDRs) using the approach of Newton et al. (2004).

To calculate BFs in Equation 1, hyper parameters for different categories in Table 1 are needed in advance. These were jointly estimated based on a mixture model of the two hypotheses as in Equation 2,

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

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

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

(j = 0, 1)

To simplify the estimation process in Equation 2, we approximated the original `TADA` model for CC data $P(x_{ca}, x_{cn}|H_j)$ using a new model in which case counts were conditioned on total counts: $P(x_{ca}|x_{ca}+x_{cn}, H_j)$ (see Methods and Figure S1).

`extTADA` used Markov Chain Monte Carlo (MCMC) for Bayesian analysis. We extracted posterior density samples from at least two MCMC chains. Posterior modes were reported as parameter estimates for all analyses, with 95% credible intervals (CIs).

2.2 Evaluating `extTADA` on simulated data

In order to assess `extTADA` in a realistic use case, we analyzed the main model used in this study as described in Equation 2 on simulated DN and CC data with one variant category each. We also analyzed simulated CC data with one or two variant categories, to examine inference on a single variant class as well as to assess the conditional probability approximation for CC data (Figures S3, S4, S5 and S6, Supplementary Results 6.3). Trinucleotide context dependent mutation rate estimates (Samocha et al., 2014; Fromer et al., 2014; De Rubeis et al., 2014) were used for denovo data for both simulation and estimation. We tested sample sizes ranging from that of the available data, 1,077 trios and 3,157 cases (equal controls) (see below), and larger sample sizes of up to 20,000 cases (see Supplementary Results 6.3).

We saw little bias in parameter estimation (Table S1 and S2). Slight under and over estimation were observed for risk gene proportions and CC mean RRs, respectively, specifically for large simulated CC mean RRs ; we note that these conditions appear outside the range of our SCZ analyses. Some bias can be expected in Bayesian analysis and not expected have a large effect on the risk gene identification results (He et al., 2013). We assessed this directly by calculating observed FDR (oFDR, i.e. the proportion of genes meeting a given FDR significance threshold that are true simulated risk genes). We observed

high correlations between oFDR and the FDR significance thresholds over wide parameter ranges (Figure 1). Only for small π (e.g., $\pi = 0.02$) oFDRs were higher than FDRs when de novo mean RRs were small (~ 5). We also saw oFDR were equal to zero for some cases with small FDR, when very small numbers of FDR-significant genes were all true risk genes. We also ran `extTADA` on null data, $\pi = 0$ and $\bar{\gamma} = 1$ for both DN and CC data (Table S3). MCMC chains tended not to converge, π estimates trended to very small values, and Bayes factors and FDRs identified almost no FDR-significant genes as expected (Table S3).

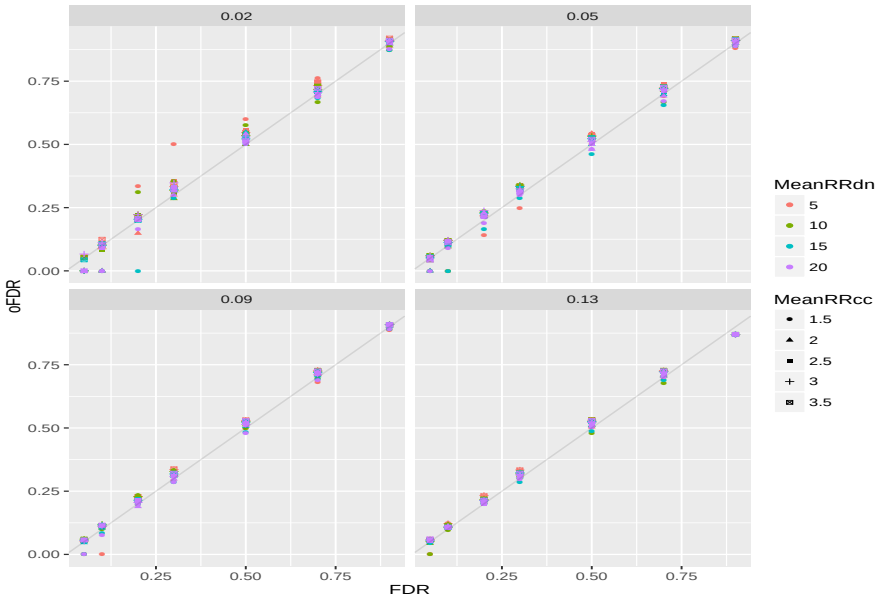

Figure 1: Observed false discovery rates (oFDR) with different FDR thresholds for different π values (0.02, 0.05, 0.09 and 0.13).

2.3 `extTADA` Analyses of Schizophrenia

We next applied `extTADA` to available DN and CC SCZ data (Figure S2), for inference of rare variant genetic architecture parameters, and for genic association. In total, there were 6,699 cases, 13,028 controls, 1077 trio/quad families used in this analysis (Table S12). Primary analyses included three variant categories for DN data, LoF, MiD and silentFCPk, and a single category of CC singletons (Purcell et al., 2014; Genovese et al., 2016) not present in the Exome Aggregation Consortium (ExAC) (Lek et al., 2015) (termed NoExAC), LoF+MiD. An array of secondary `extTADA` analyses were conducted to help validate and dissect our results.

2.3.1 SCZ data

De novo mutations and case-control variants were tested to select classes and samples for the `extTADA` pipeline. Since currently `extTADA` requires integer counts data, adjustment for ancestry and technical covariates is not possible. For case-control data, there were multiple population samples and sequencing centers; therefore, the data were restricted to non-heterogeneous population samples. First, for the 4,929 cases and 6,232 controls of the Sweden population sample, we clustered all cases and controls into different groups and then tested for case-control differences with and without adjustment for covariates. We aimed to generate clusters yielding very similar results with and without adjustment for covariates. The clustering process divided the data set into three groups as in Figure S7: Group 1, 3,157 cases + 4,672 controls; Group 2, 681 cases + 367 controls; and Group 3, 1,091 cases + 1,193 controls. Only Groups 1 and 3 were used in the next stage because Group 2 showed some difference between adjusted and unadjusted results and was relatively small. As in Genovese et al. (2016), NoExAC variants showed case-control significant differences and InExAC variants did not (Figure S7). Second, only UK and Finnish sample case/control summary counts were available from the UK10K project data (Singh et al., 2016), and we used only the larger UK population sample. Again significance of case-control differences was observed only for NoExAC singleton variants; therefore, we used only NoExAC singletons in primary `extTADA` analyses, however we also used all singletons in secondary analyses for comparison.

For de novo mutations, we calculated the sample-adjusted ratios of mutation counts between 1,077 cases and 731 controls (Table S12). Similar to Takata et al. (2016), the highest ratio was observed for silentFCPk (2.57), followed by MiD (2.3), LoF (1.83) and missense, silent ($\sim$ 1.3) mutations (Figure S8). Three classes (LoF, MiD and silentFCPk) were used in `extTADA` analyses.

2.3.2 Rare variant genetic architecture of SCZ

Three categories of de novo mutations and one category of case/control variants were used in integrative analysis using `extTADA`. They included LoF, MiD and silentFCPk denovo mutations; and LoF+MiD case-control variants. LoF and MiD variants showed similar enrichment in our case-control data analysis (Figure S7); we pooled them in order to maximize the case-control information. There were four population samples in total: one de novo population, and three case-control populations including two Sweden clusters and the UK data from the UK10K project.

`extTADA` generated samples from the joint posterior density of all genetic parameters for SCZ. All MCMC chains showed convergences (Figure S9). The estimated proportion of risk genes was 8.01% (95% CI = (4.59%, 12.9%)). LoF de novo variants had the highest estimated mean RR, 12.25 (4.78, 22.22). Two other de novo classes had estimated mean RRs 1.22 (1, 2.16) for silentFCPk and 1.44 (1, 3.16) for MiD. For MiD+LoF case-control variants, two Sweden populations had nearly equal values of mean RRs: 2.09 (1.04, 3.54) and 2.44

(1.04, 5.73); however the signal was weak for the UK population with mean RR 1.04 (1, 1.19), (Table 2, Figure 2).

Parameter	Estimated mode	lCI	uCI
SCZ_pi (%)	8.01	4.59	12.9
SCZ_meanRR_silentFCPk_denovo	1.22	1.00	2.16
SCZ_meanRR_MiD_denovo	1.44	1.00	3.16
SCZ_meanRR_LoF_denovo	12.25	4.79	22.22
SCZ_meanRR_MiD+LoF_CCpop1	2.09	1.04	3.54
SCZ_meanRR_MiD+LoF_CCpop2	2.44	1.05	5.73
SCZ_meanRR_MiD+LoF_CCpop3	1.04	1	1.19

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

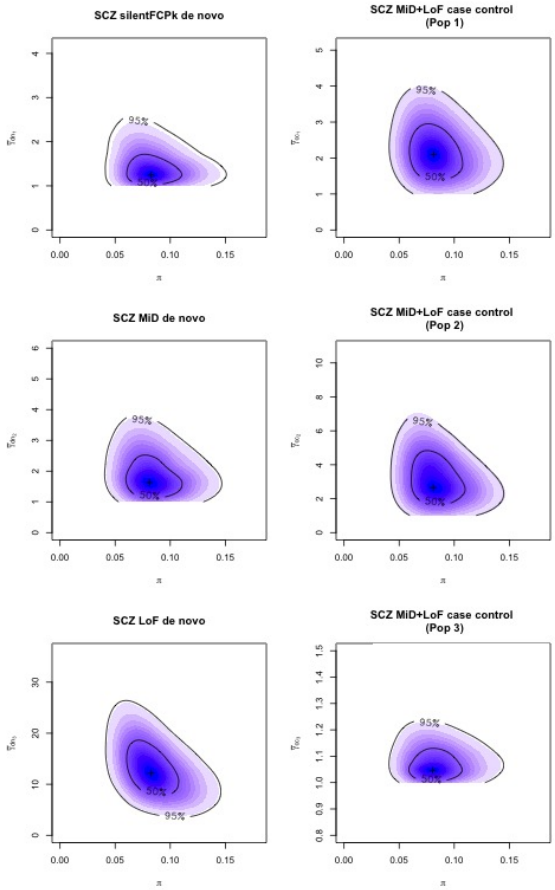

Figure 2: The densities of the proportion of risk genes and mean relative risks for SCZ data. These are obtained after 20,000 iterations of three MCMC chains. The first two case-control populations are derived from the Sweden data set while the third case-control population is the UK population.

To test the performance of the pipeline on individual data types and to assess their contribution to the overall results, we ran `extTADA` separately on each of four single variant classes: silentFCPk, MiD and LoF de novo mutations, and MiD+LoF case-control variants (Table S4). All parameter estimates were consistent with the integrative analysis, with broad credible intervals. The much larger $\bar{\gamma}$ CIs than in integrative analyses demonstrates `extTADA`'s borrowing of information across data types (also observed in simulation, Figure S4).

We also assessed the sensitivity of genetic parameter inference in several secondary analyses. We observed that synonymous de novo mutation counts were lower than expected, suggesting that mutation rates may be systemati-

cally underestimated. Adjusting mutation rates by a factor 0.81, DNM mean RR estimates slightly increased as expected, and the estimated proportion of risk genes increased slightly to 9.37% (5.47-15.12%), while case-control parameters were highly similar (Table S5). Above we assummed that different case-control population samples may have different mean RRs, which could be due to clinical ascertainment, stratification or population specific genetic architectures. Analysis using a single mean RR parameter for all three case-control samples yielded similar π and DNM mean RRs and an intermediate CC MiD+LoF mean RR with relatively narrower credible interval, $\bar{\gamma}_{CC} = 1.93$ (1.08-3.21) (Table S6, Figure S11). Considering all CC singleton variants (not just those absent from ExAC) in `extTADA` also generated similar genetic parameter estimates, with predictably slightly lower case-control mean RRs (Table S7). We note that these alternative analyses also slightly impact support for individual genes as described below.

2.3.2.1 Identifying SCZ risk genes using `extTADA`

`extTADA` also generates Bayes factors for all genes, from which we calculated posterior probabilities of association (PPAs) (Stephens and Balding, 2009) and false discovery rates (FDRs) (Benjamini and Hochberg, 1995) (Table S8, which includes supporting data as well as association results). Four genes achieved PPA > 0.8 and FDR < 0.1 (SETD1A, TAF13, PRRC2A, RB1CC1). Two genes SETD1A (FDR $= 0.0033$) and TAF13 (FDR $= 0.026$) were individually significant at FDR < 0.05. SETD1A has been confirmed as the highest statistically significant gene of SCZ in previous studies (Singh et al., 2016; Takata et al., 2016), while TAF13 was only reported as a potential risk gene in the study of Fromer et al. (2014). Interestingly for the RB1CC1 gene, rare duplications were reported to be associated with SCZ with very high odds ratio (8.58) in the study of Degenhardt et al. (2013), but has not been reported in other studies since. In addition, as discussed by the authors, duplications at this gene were also observed by Cooper et al. (2011) with an odds ratio $= 5.29$ in a study of 15,767 children with ID and/or DD. If we increase the FDR threshold to 0.3 as in the previous ASD study of De Rubeis et al. (2014), we identify 24 candidate SCZ risk genes (SETD1A, TAF13, RB1CC1, PRRC2A, VPS13C, MKI67, RARG, ITSN1, KIAA1109, DARC, URB2, HSPA8, KLHL17, ST3GAL6, SHANK1, EPHA5, LPHN2, NIPBL, KDM5B, TNRC18, ARFGEF1, MIF, HIST1H1E, BLNK). Of these, EPHA5, KDM5B and ARFGEF1 did not have any de novo mutations (Table S8). We note that still more genes showed substantial support for the alternative hypothesis over the null under the model (Jeffreys, 1998) (58 genes with PPA > 0.5, corresponding to BF > 11.49, FDR < 0.391; Table S8).

Secondary `extTADA` analyses had predictable effects on risk gene identification. Considering all CC singleton variants (not just those absent from ExAC) decreased the impact of CC data and yielded slightly fewer significant genes (three and seventeen genes with FDR < 0.1, 0.3, respectively). Using a single CC $\bar{\gamma}$ parameter for the model also resulted in 4 and 22 significant genes for FDR < 0.1 and 0.3 respectively. Mutation rate adjustment increased support

for individual genes with DNMs, increasing the findings to three and six genes at FDR < 0.05, < 0.1, respectively, including (Table S9). Generally the top genes were consistent across analyses, specifically SETD1A and TAF13 were always the top significant genes (FDR < 0.05 in all analyses).

2.3.3 Enrichment of gene sets in `extTADA` SCZ risk genes candidates

From `extTADA`, we extracted the FDR of each gene to test the enrichment of gene sets. We used gene set mean FDR to test for significant enrichment in comparison to random gene sets, and empirical P-values were FDR corrected (Benjamini and Hochberg, 1995).

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

We first tested 161 gene sets previously implicated in SCZ genetics or with strong genetic evidence relevant to SCZ rare variation (Table S10) (Purcell et al., 2014; Genovese et al., 2016; Pardinas et al., 2017; Ji et al., 2016; Epi4K Consortium and Epilepsy Phenome/Genome Project, 2013; Lin et al., 2012). FDR-significant results were observed for 61 gene sets including those reported using these data (Purcell et al., 2014; Fromer et al., 2014; Genovese et al., 2016) (Table 3). The most significant gene sets were genes harboring de novo SNPs and Indels in DD and ASD, missense constrained and loss-of-function intolerant (pLI09) genes, targets of the fragile X mental retardation protein (FMRP) and CELF4 genes, targets of RBFOX1/3 and RBFOX2 splicing factors, CHD8 promoter targets, and post-synaptic density activity-regulated cytoskeleton-associated (ARC), NMDA-receptor (NMDAR) and mGluR5 complexes (all P $<$ 8.0e-04, FDR $<$ 4.5e-03), Table 3). Genes exhibiting allelic bias in neuronal RNA-seq data Lin et al. (2012) were also strongly enriched in SCZ `extTADA` results (P = 1.1e-05, FDR = 1.4e-04). Significant enrichments were also obtained for several gene sets enriched in the recent SCZ GWAS of Pardinas et al. (2017), including the mouse mutant gene sets with psychiatric-relevant phenotypes including abnormal behavior, and abnormal nervous system morphology and physiology, as well as genome-wide significant genes from the SCZ gene-level GWAS itself (Pardinas et al., 2017) (P = 9.4e-03, FDR = 5.0e-03), showing convergence with common-variant genetic signal in genes hit by rare variation in SCZ. In addition, novel results were observed for essential genes, and known epilepsy genes ($p \leq$ 2.0e-04, FDR $\leq$ 1.6e-03; Table 3). The essential gene set was just reported recently by Ji et al. (2016) as ASD risk genes. De novo genes for other neurodevelopmental diseases (see below) were also strongly enriched in SCZ (DD, P = 1.0e-07, FDR = 2.3e-06; ASD, P = 2.1e-06, FDR = 3.4e-05; ID, P = 7.9e-04, FDR = 4.4e-03).

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

To test more novel gene sets for enrichment in the SCZ `extTADA` results, we tested 1,878 gene sets from several data bases, and FDR-adjusted for the full set of 1,717 + 161 = 1,878 gene sets tested (Tables S11). We used GO, KEGG, REACTOME and C3 sets from MSigDB (http://software.broadinstitute.org/gsea/msigdb), filtered for sets including greater than 100 genes (see Methods for details).

Significant results were observed in 103 gene sets including 36 gene sets in the above 161 gene sets. The top known gene sets still had the lowest p values in these results. We observed significant enrichment of several C3 conserved non-coding motif genesets showing brain specific expression (Xie et al., 2005): GGGAGGRR_V$MAZ_Q6, genes containing the conserved M24 GGGAGGRR motif, a MAZ transcription factor binding site; ACAGGGT,MIR-10A,MIR-10B, including microRNA MIR10A/B targets; M12 CAGGTG_V$E12_Q6, E12/TCF3 targets; M17 AACTTT_UNKNOWN, IRF1 targets; and M13 CTTTGT_V$LEF1_Q2, LEF1 targets ($P \leq$ 1.5e-04, FDR $<$ 0.01; Table S11). Relatively specific significant GO gene sets included GO:0045202/synapse and GO:0043005/neuron projection ($P \leq$ 2e-04, FDR $\leq$ 0.01). GO:0051179/localization ($P =$ 6.4e-05, FDR = 5.2e-03) was reported by Murphy and Benítez-Burraco (2016) in a study relating to language evolution and SCZ.

Gene set	P value	FDR	Gene set	P value	FDR
FMRP_targets	1.0e-07	2.3e-06	PSD-95_(core)	1.6e-03	7.9e-03
rbfox13	1.0e-07	2.3e-06	abnormal_learning|memory|conditioning	1.7e-03	8.2e-03
constrained	1.0e-07	2.3e-06	abnormal_excitatory_postsynaptic_currents	1.7e-03	8.2e-03
celf4	1.0e-07	2.3e-06	abnormal_associative_learning	2.4e-03	1.1e-02
pLI09	1.0e-07	2.3e-06	abnormal_synapse_morphology	2.7e-03	1.2e-02
rbfox2	1.0e-07	2.3e-06	abnormal_social_investigation	2.8e-03	1.2e-02
DD.allDenovoMiDandLoF	1.0e-07	2.3e-06	abnormal_neuron_morphology	2.8e-03	1.2e-02
abnormal_behavior	3.0e-07	6.0e-06	abnormal_neuron_physiology	3.2e-03	1.3e-02
abnormal_sensory_capabilities|reflexes|nociception	1.9e-06	3.4e-05	abnormal_brain_morphology	4.5e-03	1.8e-02
AST.allDenovoMiDandLoF	2.1e-06	3.4e-05	abnormal_CNS_synaptic_transmission	5.5e-03	2.1e-02
abnormal_motor_capabilities|coordination|movement	2.8e-06	4.1e-05	PSD_(human_core)	6.4e-03	2.4e-02
chd8.human_brain	9.2e-06	1.2e-04	abnormal_aggression-related_behavior	7.2e-03	2.7e-02
AlleleBiasedExpression.Neuron	1.1e-05	1.4e-04	abnormal_parental_behavior	8.0e-03	2.9e-02
abnormal_emotion|affect_behavior	1.3e-05	1.5e-04	abnormal_spatial_learning	8.2e-03	2.9e-02
abnormal_nervous_system_morphology	2.7e-05	2.8e-04	abnormal_brain_size	8.3e-03	2.9e-02
ARC	7.8e-05	7.8e-04	abnormal_consumption_behavior	8.4e-03	2.9e-02
synaptome	1.2e-04	1.1e-03	abnormal_forebrain_morphology	9.3e-03	3.1e-02
abnormal_social|conspecific_interaction	1.3e-04	1.1e-03	abnormal_innervation	9.9e-03	3.2e-02
essentialGenes	1.8e-04	1.5e-03	abnormal_telencephalon_morphology	1.3e-02	4.1e-02
Known_EPI_genes	2.0e-04	1.6e-03	abnormal_response_to_new_environment	1.3e-02	4.1e-02
mir137	2.4e-04	1.8e-03	abnormal_corpus_callosum_morphology	1.4e-02	4.1e-02
NMDAR_network	2.5e-04	1.8e-03	abnormal_temporal_lobe_morphology	1.4e-02	4.1e-02
mGluR5	3.7e-04	2.5e-03	abnormal_discrimination_learning	1.4e-02	4.3e-02
abnormal_fear|anxiety-related_behavior	6.0e-04	3.9e-03	abnormal_contextual_conditioning_behavior	1.6e-02	4.5e-02
abnormal_cued_conditioning_behavior	6.1e-04	3.9e-03	abnormal_inhibitory_postsynaptic_currents	1.6e-02	4.5e-02
abnormal_synaptic_transmission	7.0e-04	4.3e-03	abnormal_response_to_novelty	1.6e-02	4.6e-02
seizures	7.3e-04	4.3e-03	abnormal_brain_vasculature_morphology	1.7e-02	4.6e-02
abnormal_behavioral_response_to_xenobiotic	7.7e-04	4.4e-03	abnormal_excitatory_postsynaptic_potential	1.7e-02	4.7e-02
ID.allDenovoMiDandLoF	7.9e-04	4.4e-03	abnormal_cerebrum_morphology	1.8e-02	4.8e-02
GWAS_(Pardinas_et_al_2017)	9.4e-04	5.0e-03	Cav2_channels	1.8e-02	4.8e-02
ID.allKnownGenes	9.9e-04	5.1e-03			

Table 3: Enrichment of 161 known gene sets from `extTADA` results. These P values were obtained by 10,000,000 simulations, and then adjusted by using the method of Benjamini and Hochberg (1995). The information for these gene sets is summarised in Table S10.

2.3.4 Power analysis for SCZ exome sequencing studies across sample sizes

We simulated risk gene discovery using `extTADA` using the genetic architecture of SCZ inferred from the current data. Different samples sizes from 500-20,000 trio families and 1,000-50,000 cases (controls = cases) were simulated as in our validation analyses, using parameters from the posterior distribution samples given the SCZ data. The number of risk genes with $FDR \leq 0.05$ ranged from 0 to 238. Based on this analysis, we expect > 50 risk genes with total sample sizes of trio families plus case-control pairs $\sim$ 24,000 (Figure 3). The results imply that, assuming sequencing costs are proportional to the number of individuals, generating case-control data is more efficient than trio data despite the larger relative risks of de novo mutations.

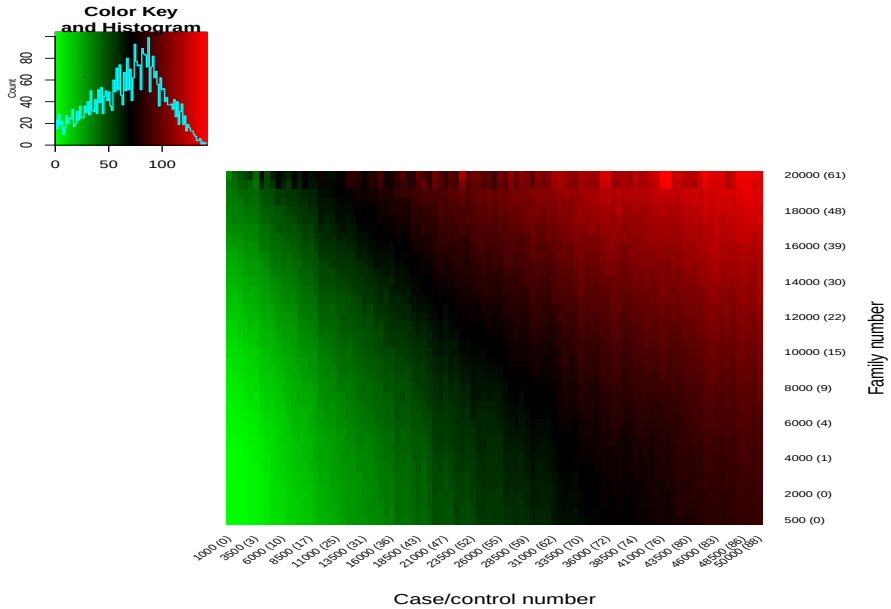

Figure 3: Number of risk genes with different sample sizes based on genetic architecture predicted by extTADA. Case/control number is only for cases (or controls); therefore if Case/control number = 10,000 this means total cases+controls = 20,000. The numbers in brackets show risk-gene numbers if we use only case-control data or only de novo mutation data.

2.4 extTADA Analyses of Other Neurodevelopmental Disorders

We also used the current pipeline to infer rare variant genetic architecture parameters from available data for autism spectrum disorder (ASD), intellectual disability (ID), developmental disorders (DD), and epilepsy (EPI). Sample sizes of these diseases are presented in Table S12, Figure S2. Numbers of trios ranged from 365 for EPI, 1,112 for ID, 4,293 for DD, 5,122 trios for ASD. As previously reported (see references in Table S12, these data have strong signals for de novo mutations contributing to disease (Table S13). Only ASD data included case-control samples (404 cases, 3,654 controls) from the Swedish PAGES study of the Autism Sequencing Consortium (De Rubeis et al., 2014) (see Methods for details).

2.4.1 Rare variant genetic architectures of ASD, ID, DD, EPI

extTADA genetic parameter estimates are presented in Figure 4 and Table 4. MCMC analyses showed good convergence, except for the small sample size EPI (392 families compared with > 1000 families for other diseases). The numbers of risk genes (π) in these diseases were lower than that of SCZ (Figure 4, Tables 2 & 4). For ASD, the estimated proportion of risk genes π was 4.59% (95% CI 3.19% - 6.01%), consistent with the result of 550-1000 genes estimated in the original TADA model (He et al., 2013) using only LoF de novo data. For ID, π was smaller than that of ASD; estimated value was 2.76% (2.1% - 3.7%). For DD $\pi = 2.87\%$ (2.34% - 3.49%) was similar to that of ID. The estimated π value for EPI, 1.65% (0.8% - 3.21%) was the lowest but with a broad credible interval owing to its much smaller sample size. Mean RRs of de novo mutations in all four neurodevelopmental diseases were much higher than those of SCZ. This was expected because of the strong signal of de novo mutations in these data for other diseases. For ASD, estimated mean RRs for de novo mutations were consistent with previous results and much lower than for the other diseases. ID and DD had the highest estimated de novo LoF mean RRs, 96.0 (68 - 131) and 86.5 (66 - 112), respectively. Even though the EPI estimated de novo LoF mean RR, 77.0 (37 - 138), was slightly lower than those of ID and DD, the estimate for EPI de novo MiD mean RR, 48 (20 - 87) was somewhat higher than those of other diseases. The previously estimated (Epi4K Consortium and Epilepsy Phenome/Genome Project, 2013) EPI MiD mean RR of 81 is consistent with the current results, and it will be of interest to see if this result remains consistent in additional data in the future.

Parameter	Estimated mode	lCI	uCI
ASD_pi (%)	4.59	3.19	6.01
ASD_meanRR_MiDdenovo	3.67	1.98	8.68
ASD_meanRR_LoFdenovo	23.4	13.63	36.94
ASD_meanRR_LoFcc	4.18	2.04	9.96
ID_pi (%)	2.76	2.07	3.7
ID_meanRR_MiDdenovo	28.61	16.18	41.86
ID_meanRR_LoFdenovo	96.04	67.57	130.73
EPI_pi (%)	1.65	0.8	3.21
EPI_meanRR_MiDdenovo	47.5	19.77	87.32
EPI_meanRR_LoFdenovo	77	37.19	138.24
DD_pi (%)	2.87	2.34	3.49
DD_meanRR_MiDdenovo	22.55	13.19	32.53
DD_meanRR_LoFdenovo	86.53	65.79	111.61

Table 4: Estimated parameters for de novo and case-control SCZ data and four other diseases: ID, EPI, ASD and DD. These results are obtained by sampling 20,000 times of three MCMC chains. The two last columns show the lower (lCI) and upper (uCI) values of CIs.

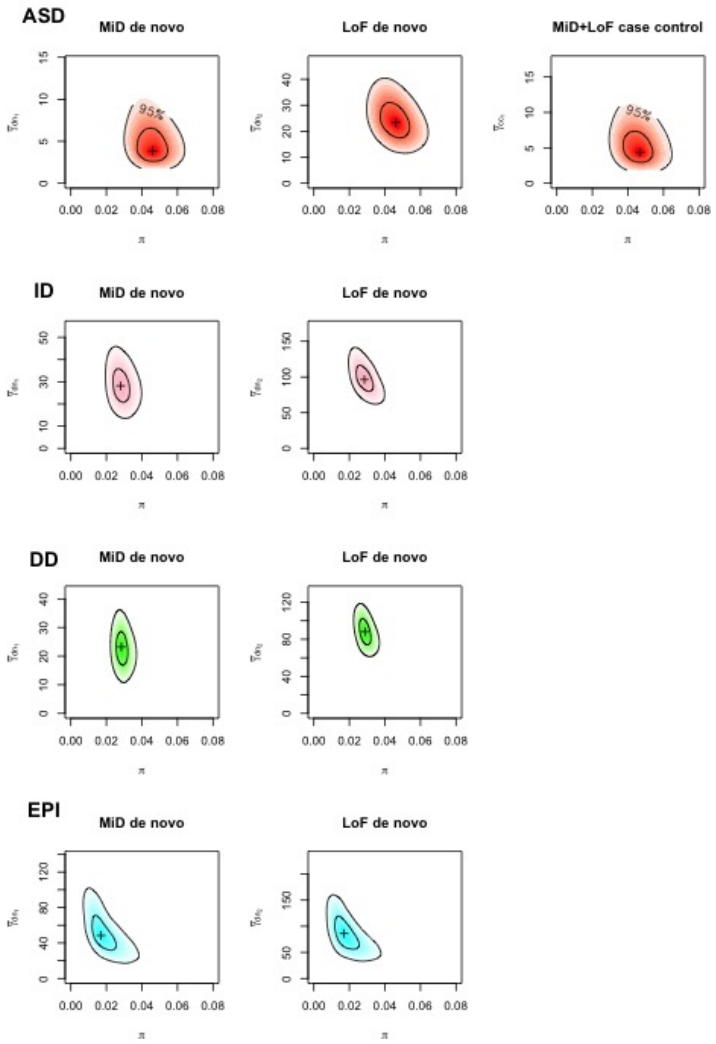

Figure 4: The densities of the proportion of risk genes and π mean relative risks (γ) for ASD, EPI, ID and DD data. For ASD, there are two de novo (dn) classes and one case-control (cc) class. For other diseases, only two de novo classes are publicly available for our current study.

2.4.2 Novel risk genes in ID and DD

The `extTADA` risk gene results of the four disorders ID, DD, ASD and EPI are presented in Tables S14, S15, S16 and S17. Results of other de novo mutation methods using these same data have been recently reported (Lelieveld et al., 2016; Deciphering Developmental Disorders Study, 2017); nevertheless, `extTADA` identified novel genes with strong statistical support from these recent

data. There were 58 and 73 genes for ID with FDR ≤ 0.05 and 0.1, respectively, and 164 and 201 genes for DD. In ID 15 of 58 FDR ≤ 0.05 genes (TCF7L2, USP7, ATP8A1, FBXO11, KDM2B, MED12L, MAST1, MFN1, TNPO2, CLTC, CEP85L, AGO1, AGO2, SLC6A1-AS1, POU3F3) were not on the list of previously reported known and novel ID genes (Lelieveld et al., 2016). Of the 15 genes, six (TNPO2, AGO2, CLTC, CEP85L, FBXO11, MFN1) were strongly significant (FDR < 0.01); these are genes hit by two or three MiD or LoF de novos but were not identified by the simulation based analyses of Lelieveld et al. (2016). In DD, only 59 of 164 FDR ≤ 0.05 genes were reported by Deciphering Developmental Disorders Study (2017); 101 genes are novel. Similar to ID, the total MiD+LoF de novo counts of these 101 genes were not high (between two and six). Surprisingly, there were 58 of the 101 genes with FDRs < 0.01.

2.4.3 Multiple gene sets are enriched in top significant genes across neurodevelopmental diseases

We also tested for gene set enrichment in the four NDs and combined this information with the SCZ gene-set information above (Tables S18 and S19, Figures 5 and S12). First, we tested 161 known or strong-candidate gene sets tested in SCZ (see Methods for details). The numbers of significant gene sets (FDR < 0.05) were 51, 74, 29 and 17 for ID, DD, ASD and EPI respectively. There were five gene sets significant across five diseases; these included Cav2 channels, FMRP targets, NMDAR network, PSD95, abnormal excitatory postsynaptic currents (all FDR ≤ 0.0097). Second, we tested our 1,877 data-driven gene sets; only one gene set which was significant in all five diseases after FDR adjustment: NMDAR network genes (all FDR ≤ 0.024). FMRP target genes were also very high significant across ASD, ID, DD, SCZ (all FDR $\leq$ 3.1e-05) but not significant for EPI (FDR $= 0.058$, Figure S12, Table S19).

The number of significant gene sets was not as high in EPI as in the other diseases, likely due to its smaller sample size and power; therefore, we removed this disorder and repeated our assessment of significant gene sets overlap in the four disorders SCZ, DD, ID and ASD. Twelve gene sets were significant in all four disorders. These consisted of the five gene sets above and seven other gene sets: constrained genes (constrained and pLI09), rbfox1/3 and rbfox2 targets, CHD8 targets (chd8 human brain), and the mouse mutant gene sets abnormal social investigation and abnormal brain size. In an analysis of all 1,877 data-driven gene sets, FMRP targets, constrained and pLI09 genes, and NMDAR-network genes remained significant across the four disorders. In addition, one other gene set, GO:0016568/chromatin organization, was also enriched for each of SCZ, ASD, DD and ID (Table S19, Table S18).

3 Discussion

In this work, we have built an integrative pipeline `extTADA` for Bayesian analysis of de novo mutations and rare case-control variants, to infer genetic architecture

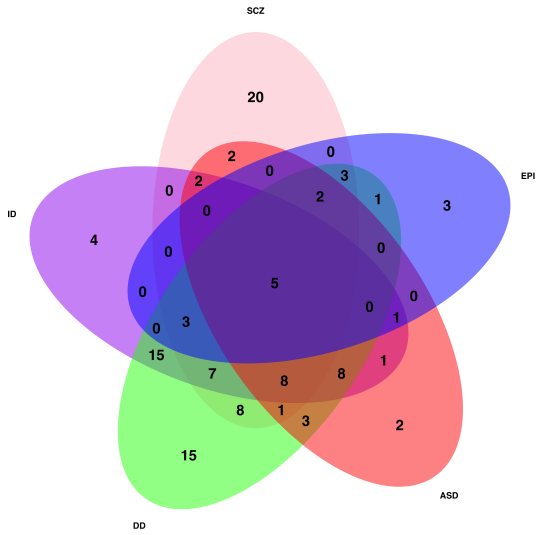

Figure 5: The overlaps of significant gene sets in SCZ, ASD, EPI, DD and ID from the analysis of the 161 genes.

parameters and identify risk genes. We applied `extTADA` to available data in schizophrenia and four other neurodevelopmental disorders (Figure S2). The pipeline is based on our previous work in autism sequencing studies, `TADA` (He et al., 2013; De Rubeis et al., 2014), and conducts fully Bayesian analysis of a simple rare variant genetic architecture model. Unlike `TADA`, which was developed for studies where LoF de novo mutations have strong discernible effects, we developed `extTADA` for schizophrenia, where de novo and case-control variants have more subtle effects discernible only at the level of gene set analysis. `extTADA` borrows information across all annotation categories and between de novo and case-control samples in genetic parameter inference, critical for sparse rare variant sequence data, and we hope that it will be generally useful for rare variant analyses across complex traits.

Using Markov Chain Monte Carlo, `extTADA` samples from the joint posterior density of risk gene proportion and mean relative risk parameters. Inference of rare variant genetic architexture is of great interest in its own right (Zuk et al., 2014), but of course risk gene discovery is one of the most important objectives of genetics. We provide Bayesian statistical support for risk gene status in the form of Bayes factors for each gene, and we further calculate posterior probabilities (Stephens and Balding, 2009) and false discoery rates (Benjamini and Hochberg, 1995). Although we use `TADA` for inference of genetic parameters, and joint analysis certainly impacts genetic parameter estimation (see the primary analysis vs single class analyses in Tables S8 and S4), we found that the empirical Bayesian approach of calculating genic BFs from model parameter point estimates (He et al., 2013) is highly similar to joint posterior

mean genic BFs (see Methods). Therefore, the approach of He et al. (2013) is a good one if model parameters are known approximately, and we maintain this functionality in `extTADA` if users have prior information on the rare variant genetic architecture of the tested disease.

As in all Bayesian and Likelihood analyses, we must specify a statistical model; the true model underlying the data is unknown and could in principle yield different results. This is addressed by analyzing a simple model that can allow illustrative, interpretable results, and by assessing the sensitivity of results to a range of alternative model specifications. `extTADA` uses relatively agnostic hyper-parameter prior distributions (Figure S2), without assuming known parameters and without any previously known risk gene seeds. Still, `extTADA` makes important assumptions, both in common with `TADA` and uniquely. First, both models assume Poisson distributed counts data and Gamma distributed mean relative risks across genes for analytical convenience, making alternative model specification inconvenient. Poisson counts are likely to be a good approximation for genetic counts data (He et al., 2013), assuming linkage disequilibrium can be ignored, and that stratification has been adequately addressed. Alternatives should be explored for Gamma distributed mean relative risk distributions. Poisson de novo muation counts further assume known mutation rates, uncertainty in which may introduce bias for multiple reasons; in our data, mutation rate adjustment for silent de novo count rates was actually anti-conservative S9. Differences between de novo studies is not unlikely even though previous studies of De Rubeis et al. (2014); Singh et al. (2016) did not adjust mutation rates to account for it. The ability to incorporate covariates, perhaps with Gaussian sample frequency data and Gaussian effect sizes, would be an important further extention of TADA-like models.

Second, `extTADA` assumes that different variant classes share risk genes such that the mixture model parameter π applies to all data types, facilitating borrowing of information across classes. This is supported by convergent de novo and case-control rare variant results in SCZ (Fromer et al., 2014; Purcell et al., 2014; Singh et al., 2016; Genovese et al., 2016) (Table S4); however, some evidence exists for disjoint risk genes for de novo vs case-control protein-truncating variants e.g. in congenital heart disease (CHD) Sifrim et al. (2016). We emphasize that we do consider multiple population samples as different categories in `extTADA`, since sequence data are very often from different countries and/or centers. (Here we used multiple categories of case-control data but multiple de novo categories could be important as well.)

The current study replicated previous studies, and supplies new information about SCZ. First, SETD1A (Singh et al., 2016; Takata et al., 2016) is the most significant gene across analyses (FDR $\sim 1.5\times10^{-3}$), TAF13 (Fromer et al., 2014) is also significant across analyses. Of two genes with FDR < 0.1, RB1CC1 was reported in a study of copy-number variation in SCZ (Degenhardt et al., 2013). Second, we found substantial overlap of top genes in this study and gene sets known from previous reports on these same SCZ data Genovese et al. (2016). Several conserved non-coding motif gene sets (Xie et al., 2005) and a few GO gene sets were also significant (Table 3). Third, in this study, we describe in

detail the rare variant genetic architecture of SCZ. It appears more complex than those of ASD, ID, DD and EPI; the estimated risk gene proportion for SCZ ($\sim$ 8%) is higher than those of the four other diseases (Figure 2 and 4, Tables 2 and 4). We also see that disease risk information is concentrated in ultra-rare variants not present in the ExAC database (Kosmicki et al., 2016; Genovese et al., 2016) (Table S7). Finally, we see substantial overlap between de novo and case-control, and common variant (Pardinas et al., 2017) genes in SCZ.

We used `extTADA` to infer genetic parameters for four other neurodevelopmental diseases ASD, EPI, DD and ID (Table 4, Figure 4). The ASD results of `extTADA` are comparable to previous results (He et al., 2013; De Rubeis et al., 2014). We note the exceptionally high de novo missense damaging mean RR estimated for EPI, also consistent with previous analyses (EuroEPINOMICS-RES Consortium et al., 2014). We also highlight the sharing of gene sets enriched across multiple neurodevelopmental diseases (Figure 5), including diverse synaptic gene sets, and possible distinguishing EPI as less similar to the other disorders. Multi-phenotype analyses leveraging shared this could have higher power to to detect novel risk genes. Finally, importantly, many novel significant genes which were missed in recent studies are discovered by `extTADA` (101 for DD and 15 for DD).

4 Data and methods

4.1 Data

Figure S2 shows the workflow of all data used in this study.

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

High-quality variants were obtained from published analyses (Table S12). Variants were annotated using Plink/Seq (using RefSeq gene transcripts, UCSC Genome Browser, `http://genome.ucsc.edu`) as described in Fromer et al. (2014). SnpSift version 4.2 (Cingolani et al., 2012) was used to further annotate these variants using dbnsfp31a (Liu et al., 2015). Variants were grouped into different categories as follows. Loss of function (LoF): nonsense, essential splice, and frameshift variants. Missense damaging (MiD): defined as missense by Plink/Seq and damaging by all of 7 methods (Genovese et al., 2016)- SIFT, $Polyphen2_HDIV$, $Polyphen2_HVAR$, LRT, PROVEAN, MutationTaster and MutationAssessor. Recently, Takata et al. (2016) reported significant results for synonymous mutations in regulatory regions; therefore, this category was also analyzed. To annotate synonymous variants within DNase I hypersensitive sites (DHS) as Takata et al. (2016), the file *wgEncodeOpenChromDnaseCerebrumfrontalocPk.narrowPeak.gz* was downloaded from `http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeOpenChromDnase/` on April 20, 2016. Based on previous results with SCZ exomes Purcell et al. (2014); Genovese et al.

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

4.1.2 Gene sets

Multiple resources were used to obtain gene sets for our study. First, we used known gene sets with prior evidence for involvement in schizophrenia and autism from several sources. Second, to identify possible novel significant gene sets, we collected genes sets from available data bases (see below).

4.1.2.1 Known gene sets

These gene sets and their abbreviations are presented in Table S10.

- Gene sets enriched for ultra rare variants in SCZ which were described in detailed in Genovese et al. (2016): missense constrained genes (constrained) from Samocha et al. (2014), loss-of-function tolerance genes (pLI90) from Lek et al. (2015), RBFOX2 and RBFOX1/3 target genes (rbfox2, rbfox13) from Weyn-Vanhentenryck et al. (2014), Fragile X mental retardation protein target genes (fmrp) from Darnell et al. (2011), CELF4 target genes (celf4) from Wagnon et al. (2012), synaptic genes (synaptome) from Pirooznia et al. (2012), microRNA-137 (mir137) from Robinson et al. (2015), PSD-95 complex genes (psd95) from Bayés et al. (2011), ARC and NMDA receptor complexes (arc, nmdar) genes from Kirov et al. (2012), de novo copy number variants in SCZ, ASD, bipolar as presented in Supplementary Table 5 of Genovese et al. (2016).
- Allelic-biased expression genes in neurons from Table S3 of Lin et al. (2012).
- Promoter targets of CHD8 from Cotney et al. (2015).
- Known ID gene set was from the Sup Table 4 of Lelieveld et al. (2016) and the 10 novel genes reported by Lelieveld et al. (2016).
- Gene sets from MiD and LoF de novo mutations of ASD, EPI, DD, ID.
- The essential gene set from the supplementary data set 2 of Ji et al. (2016).
- Lists of human accelerated regions (HARs) and primate accelerated regions (PARs) (Lindblad-Toh et al., 2011) were downloaded from `http://www.broadinstitute.org/scientific-community/science/projects/mammals-models/29-mammals-project-supplementary-info` on May 11, 2016. The coordinates of these regions were converted to hg19 using

Liftover tool (Kent et al., 2002). We used a similar approach as Xu et al. (2015) to obtain genes nearby HARs. Genes in regions flanking 100 kb of the HARs/PARs were extracted to use in this study (geneInHARs, geneInPARs).

- List of known epilepsy genes was obtained from Supplementary Table 3 of Phenome et al. (2017).
- List of common-variant genes was obtained from Extended Table 9 of Pardinas et al. (2017).
- 134 gene sets from mouse mutants with central nervous system (CNS) phenotypes were obtained from Pardinas et al. (2017). Steps which were used to obtain the gene sets were described in Pocklington et al. (2015). We finally obtained 134 gene sets from this step after removing overlapping gene sets between previous studies and the 161 gene sets.

In the gene-set tests for a given disease, we removed the list of known genes and the list of de novo mutation genes for that disease. As a result, we tested 161 known gene sets for ASD, DD and SCZ; and 159 gene sets for EPI and ID.

4.1.2.2 Other gene sets

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

4.2 Methods

4.2.1 `extTADA` pipeline: extended transmission (case-control) and de novo analysis

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

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

Let K be the number of categories (e.g., LoF, MiD), and $x_i = (x_{i1}, .., x_{iK})$ be the vector of counts at the ith given gene. The Bayes Factor for each jth

category to test two hypotheses: $H_0 : \gamma = 1$ versus $H_1 : \gamma \neq 1$ was:

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

In Equation 3, $x_{ij} = x_d$ for de novo data and $x_{ij} = (x_{ca}, x_{cn})$ for case-control data. In addition, the integral over q was not applicable for de novo data because there is no q parameter for de novo data.

As in He et al. (2013), the BF for the ith gene combining all categories is:

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

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

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

Equation 5 was calculated across categories as in Equation 4.

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

The hyperparameters $\phi_{1j} = (\gamma_{j(dn)}, \gamma_{j(cc)}, \beta_{j(dn)}, \beta_{j(cc)}, \rho_j, \nu_j)$ were estimated using a Hamiltonian Monte Carlo (HMC) Markov chain Monte Carlo (MCMC) method implemented in the `rstan` package (Carpenter et al., 2015; R Core Team, 2016). However, the model was first simplified by removing q (see below).

4.2.1.2 Simplified approximate case-control model

For case-control (transmitted) data, $q \sim Gamma(\rho, \nu)$, and hyper-parameters ρ and ν controlled the mean and dispersion of q; therefore, as in the previous studies (He et al., 2013; De Rubeis et al., 2014), ν was heuristically chosen (200 was used in all analyses) and $\frac{\rho}{\nu}$ = the mean frequency across genes in both cases and controls.

We simplified the case-control model by expressing it as

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

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

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

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

The marginal likelihood was

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

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

4.2.1.3 Control of an implied proportion of protective variants using the relative risk dispersion hyper-parameter

If $\bar{\gamma}$ and β were small then we could see a high proportion of protective variants when $\bar{\gamma}$ is not large. Although this might be of biological interest, it is not currently accounted for in the model. To control the proportion of protective variants, we tested the relationship between β and $\bar{\gamma}$ in determining $\int_0^1 Gamma(\bar{\gamma}_{dn}\beta_{dn}, \beta_{dn})$. We set this proportion very low (0.5%) (Figure S10) and built a nonlinear relationship $\beta = e^{a*\bar{\gamma}^b+c}$. The R package *nls* was used to estimate a, b and c, as $6.83, -1.29$ and -0.58 respectively.

4.2.1.4 Power analyses for `extTADA` risk gene identification

We simulated DN and CC data for ranges of sample sizes, using random samples from the posterior density of our primary genetic architecture inference analysis. The original case-control model was used in this calculation; however, we changed the order of the integral of parameters to not rely on q because the range of this parameter was not frequently known in advance (Sup Information 6.3). BFs of genes were calculated according to Equation 1, and Newton et al. (2004) false discovery rates (FDRs) were calculated following De Rubeis et al. (2014). Posterior probability (PP) for each gene was calculated as $PP = \pi * BF/(1-\pi+\pi*BF)$ (Stephens and Balding, 2009). The number of risk genes could be predicted based on the FDR threshold, for which we chose 0.05.

4.2.2 Testing the model on simulated data

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

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

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

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

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

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

4.2.2.1 Test NULL model ($\pi = 0, \bar{\gamma} = 1$)

We also tested the situation in which no signal of both de novo mutations and rare case-control variants was present. We simulated one DN category and one CC category with $\pi = 0$, $\bar{\gamma} = 1$. To see the influence of prior information of $\bar{\gamma}$ ($\bar{\gamma} \sim Gamma(1, \bar{\beta})$) on these results, we used different values of $\bar{\beta}$.

4.2.3 Calculate mutation rates

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

4.2.4 Analyze SCZ data

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

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

A simple combination between a clustering process using a multivariate normal mixture model and a data analyzing strategy using linear and generalized linear models was used to divide the Sweden data into non-heterogeneous populations. Genovese et al. (2016) recently analyzed all case-control data sets by adjusting for multiple covariates: genotype gender of individuals (SEX), 20 principal components (PCs), year of birth of individuals (BIRTH), Aligent kit used in wet-labs (KIT) by using linear regression and generalized linear regression models as in Equation 7. They reported significant results for NonExAC LoF and MiD variants; therefore, this information was used in this step. We defined homogeneous populations as populations which were not much affected by the covariates. Thus, for the populations, analyzing results using Equation 7 (adjusting covariates) would not be much different from those results using Equation 8 (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 7 and 8 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 7 and 8 had to between 0.1 and 1.

From results satisfied the above criteria, we manually chose groups which had similar results between Equation 8 and 7.

$$\begin{aligned} &logit(P(SCZ = 1)) \sim count + countAll + sex + birth + kit + \sum_{i=1}^{20} PC_i \\ &count \sim SCZ + countAll + sex + birth + kit + \sum_{i=1}^{20} PC_i \end{aligned} \tag{7}$$

$$\begin{aligned} SCZ \quad &\sim count \\ count \quad &\sim SCZ \end{aligned} \tag{8}$$

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

4.2.4.2 Estimate genetic parameters for SCZ

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

4.2.4.3 Estimate number of risk genes for SCZ

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

4.2.4.4 Test enrichment in known gene sets

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

4.2.4.5 Predict number of risk genes for different sample sizes

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

4.2.4.6 Test for single classes

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

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

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

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

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

4.2.4.9 Test `extTADA` with the same mean relative risks for case-control data

To test the performance of the model when mean RRs ($\bar{\gamma}_{CC}$) were equal, we re-ran the analysis for SCZ data with an adjustment inside the model: $\gamma_{CC}^{ij} \sim Gamma(\bar{\gamma}_{CC} * \beta_{CC}, \beta_{CC})$ (γ_{CC}^{ij} was the relative risk at the ith gene in the jth population).

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

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

4.2.6 Infer parameters using MCMC results

The `rstan` package (Carpenter et al., 2015) was used to run MCMC processes. For simulation data, 5,000 times and a single chain were used. For real data, 20,000 times and three independent chains were used. In addition, for SCZ data we used two steps to obtain final results. Firstly, 10,000 times were run to obtain

parameters. After that, we calculated β values from estimated mean RRs as the Equation described in Table 1. Finally, `extTADA` was re-run 20,000 times on the SCZ data with calculated β values set as constants to re-estimate mean RRs and the proportions of risk genes. For each MCMC process, a burning period = a half of total running times was used to assure that chains did not rely on their initial values. For example, we ran and removed 2,500 burning times before the 5,000 running times for simulation data.

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

5 Acknowledgements

This work was supported in part through the computational resources and staff expertise provided by Scientific Computing at the Icahn School of Medicine at Mount Sinai, and by NIH grant R01MH105554 to E.A.S. The Sweden exome sequencing data generation and analysis are supported by the Stanley Center for Psychiatric Research and NIH grant R01 MH077139 to C.H., P.S. and P.F.S. We are deeply grateful for the participation of all subjects contributing to this research.

6 Supplementary information

6.1 Supplementary Tables

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

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

pi	dn_RR	cc_RR	e.pi Q50	Q5	Q95	e.dn_RR Q50	Q5	Q95	e.cc_RR Q50	Q5	Q95
0.02	5	1.5	0.0126	0.0051	0.0224	4.72	2.65	9.22	1.83	1.70	2.53
0.02	5	2	0.0158	0.0054	0.0394	4.08	2.33	11.88	2.27	1.88	3.01
0.02	5	2.5	0.0230	0.0093	0.0419	3.49	2.05	10.25	2.83	2.22	3.43
0.02	5	3	0.0219	0.0123	0.0355	3.52	2.06	8.49	3.31	2.74	4.25
0.02	5	3.5	0.0269	0.0171	0.0373	3.76	2.10	9.59	3.80	3.15	4.73
0.02	10	1.5	0.0185	0.0085	0.0280	5.72	2.82	10.08	1.80	1.54	2.57
0.02	10	2	0.0176	0.0076	0.0373	5.22	2.55	13.43	2.24	1.85	3.05
0.02	10	2.5	0.0218	0.0118	0.0342	5.13	2.39	17.29	2.80	2.27	3.57
0.02	10	3	0.0227	0.0137	0.0344	7.29	2.35	15.88	3.28	2.64	4.23
0.02	10	3.5	0.0255	0.0163	0.0328	7.87	2.79	14.28	3.73	3.07	4.69
0.02	15	1.5	0.0152	0.0046	0.0315	11.63	4.65	28.84	1.77	1.48	2.67
0.02	15	2	0.0213	0.0091	0.0348	10.19	3.32	23.63	2.25	1.86	2.84
0.02	15	2.5	0.0230	0.0118	0.0377	9.73	4.00	20.84	2.69	2.06	3.50
0.02	15	3	0.0226	0.0128	0.0370	10.51	3.53	21.52	3.18	2.55	4.11
0.02	15	3.5	0.0240	0.0140	0.0364	10.56	3.58	20.62	3.62	3.04	4.70
0.02	20	1.5	0.0138	0.0062	0.0398	14.96	5.47	47.26	1.68	1.35	2.29
0.02	20	2	0.0188	0.0079	0.0363	14.10	4.79	36.13	2.28	1.81	3.20
0.02	20	2.5	0.0233	0.0110	0.0343	13.61	5.67	26.36	2.76	2.12	3.54
0.02	20	3	0.0243	0.0140	0.0371	13.89	6.44	23.95	3.41	2.63	4.30
0.02	20	3.5	0.0240	0.0146	0.0352	14.87	6.90	24.81	3.77	2.96	4.65
0.05	5	1.5	0.0343	0.0120	0.0688	4.55	2.26	14.06	1.71	1.51	2.17
0.05	5	2	0.0479	0.0279	0.0699	4.86	2.32	10.44	2.21	1.88	2.60
0.05	5	2.5	0.0556	0.0351	0.0743	4.59	2.06	8.03	2.81	2.31	3.17
0.05	5	3	0.0558	0.0427	0.0722	4.36	2.08	8.18	3.35	2.91	3.74
0.05	5	3.5	0.0621	0.0435	0.0727	3.65	1.89	7.36	3.73	3.22	4.48
0.05	10	1.5	0.0381	0.0161	0.0723	9.20	3.92	15.41	1.74	1.45	2.18
0.05	10	2	0.0531	0.0293	0.0801	8.71	3.70	12.99	2.26	1.91	2.71
0.05	10	2.5	0.0528	0.0386	0.0727	8.76	4.15	14.48	2.74	2.47	3.11
0.05	10	3	0.0569	0.0416	0.0737	8.22	4.47	13.57	3.25	2.83	3.72
0.05	10	3.5	0.0615	0.0491	0.0733	8.06	3.97	13.17	3.66	3.30	4.29
0.05	15	1.5	0.0406	0.0182	0.0877	13.51	6.94	24.71	1.67	1.43	1.98
0.05	15	2	0.0489	0.0311	0.0723	14.04	8.16	22.70	2.19	1.90	2.61
0.05	15	2.5	0.0522	0.0327	0.0734	13.13	8.28	20.66	2.72	2.38	3.11
0.05	15	3	0.0577	0.0449	0.0732	12.37	7.27	18.59	3.19	2.83	3.75
0.05	15	3.5	0.0607	0.0465	0.0756	11.97	8.23	18.55	3.61	3.11	4.29
0.05	20	1.5	0.0418	0.0205	0.0814	18.37	9.74	32.56	1.63	1.37	1.97
0.05	20	2	0.0482	0.0325	0.0697	17.08	10.14	29.26	2.27	1.91	2.60
0.05	20	2.5	0.0537	0.0406	0.0733	16.59	10.57	23.23	2.77	2.29	3.06
0.05	20	3	0.0569	0.0424	0.0770	16.15	10.37	24.32	3.23	2.84	3.75
0.05	20	3.5	0.0596	0.0449	0.0765	15.50	10.23	21.45	3.75	3.19	4.61
0.09	5	1.5	0.0767	0.0404	0.1207	4.46	2.17	9.59	1.66	1.51	1.97
0.09	5	2	0.0904	0.0666	0.1115	4.52	2.04	7.33	2.23	2.03	2.54
0.09	5	2.5	0.0963	0.0753	0.1256	4.70	2.52	7.54	2.79	2.49	3.11
0.09	5	3	0.1040	0.0879	0.1217	3.90	2.08	6.71	3.19	2.85	3.68
0.09	5	3.5	0.1039	0.0876	0.1211	4.22	2.34	7.93	3.70	3.35	4.13
0.09	10	1.5	0.0778	0.0423	0.1208	10.01	5.56	17.73	1.64	1.46	1.93
0.09	10	2	0.0925	0.0660	0.1196	9.26	5.85	13.45	2.16	1.96	2.49
0.09	10	2.5	0.0963	0.0729	0.1170	9.30	7.16	12.50	2.82	2.40	3.18
0.09	10	3	0.0992	0.0831	0.1189	9.25	6.11	12.76	3.22	2.95	3.61
0.09	10	3.5	0.1070	0.0885	0.1222	8.29	5.81	10.94	3.67	3.36	4.20
0.09	15	1.5	0.0822	0.0507	0.1257	14.59	9.22	22.62	1.61	1.43	1.89
0.09	15	2	0.0911	0.0668	0.1217	14.35	9.39	20.13	2.16	1.94	2.45
0.09	15	2.5	0.0978	0.0754	0.1202	13.77	10.40	17.99	2.72	2.40	3.00
0.09	15	3	0.0997	0.0844	0.1206	13.50	10.60	16.88	3.13	2.82	3.49
0.09	15	3.5	0.1036	0.0861	0.1229	12.95	9.89	16.86	3.60	3.20	4.15
0.09	20	1.5	0.0804	0.0495	0.1236	19.92	13.06	31.58	1.60	1.38	1.82
0.09	20	2	0.0920	0.0694	0.1205	18.18	12.71	24.69	2.21	1.95	2.51
0.09	20	2.5	0.0958	0.0742	0.1166	18.28	13.76	22.90	2.75	2.49	3.05
0.09	20	3	0.0974	0.0816	0.1202	17.55	13.32	22.38	3.28	2.95	3.59
0.09	20	3.5	0.1067	0.0925	0.1171	16.49	13.66	20.83	3.68	3.32	4.21
0.13	5	1.5	0.1163	0.0720	0.1671	4.87	2.51	8.11	1.65	1.49	1.83
0.13	5	2	0.1250	0.0991	0.1603	5.15	2.82	7.72	2.22	2.04	2.53
0.13	5	2.5	0.1387	0.1173	0.1654	4.65	2.51	7.04	2.77	2.52	3.08
0.13	5	3	0.1469	0.1220	0.1649	4.40	2.83	6.17	3.20	2.93	3.47
0.13	5	3.5	0.1467	0.1293	0.1747	4.45	2.46	6.13	3.69	3.36	4.25
0.13	10	1.5	0.1094	0.0707	0.1660	10.69	7.35	17.96	1.68	1.53	1.84
0.13	10	2	0.1306	0.1113	0.1529	9.40	6.89	13.22	2.17	2.01	2.36
0.13	10	2.5	0.1432	0.1197	0.1595	9.15	7.21	11.97	2.73	2.52	3.01
0.13	10	3	0.1457	0.1308	0.1682	8.89	6.64	11.08	3.23	3.00	3.54
0.13	10	3.5	0.1497	0.1320	0.1728	8.54	6.62	10.61	3.60	3.26	3.97
0.13	15	1.5	0.1180	0.0778	0.1677	15.08	10.41	22.63	1.60	1.46	1.80
0.13	15	2	0.1277	0.1044	0.1593	14.45	11.56	18.45	2.15	1.96	2.40
0.13	15	2.5	0.1380	0.1124	0.1625	14.34	11.23	17.93	2.72	2.45	3.04
0.13	15	3	0.1432	0.1254	0.1667	13.41	11.03	16.81	3.13	2.87	3.63
0.13	15	3.5	0.1488	0.1281	0.1674	13.00	10.35	16.30	3.56	3.20	3.98
0.13	20	1.5	0.1203	0.0862	0.1765	19.72	13.93	26.77	1.61	1.48	1.83
0.13	20	2	0.1325	0.1093	0.1546	18.54	15.11	23.43	2.21	1.99	2.39
0.13	20	2.5	0.1351	0.1130	0.1601	18.38	14.63	22.97	2.79	2.50	3.00
0.13	20	3	0.1434	0.1256	0.1645	18.43	14.94	22.27	3.24	2.94	3.60
0.13	20	3.5	0.1488	0.1320	0.1637	16.81	13.99	20.18	3.64	3.34	4.11

Table S2: Estimated values for the cases in Table S1, for each uniqe set of parameter values. The first three columns are simulated values. The following columns show estimated π, de novo mean relative risk (dn RR) and case-contorl (cc) RR; for each parameter, shown are median (Q50) and 5th and 95th %-iles (Q5 and Q95) estimates over 100 simulation replicates.

$\bar{\beta}_{DN}$	$\bar{\beta}_{CC}$	e.π	e.$\bar{\beta}_{DN}$	e.$\bar{\beta}_{CC}$	e.β_{DN}	e.β_{CC}	FDR0.01	FDR0.05	FDR0.1	FDR0.25	FDR0.5
0.01	0.11	0.0008985	12.16	2.37	0.82	1.38	0	0	0	0	0
0.01	0.14	0.0013925	7.76	2.02	0.84	2.08	0	0	0	0	0
0.01	0.2	0.0011444	7.38	1.66	0.86	3.2	0	0	0	0	0
0.01	0.33	0.0014319	10.32	1.46	0.83	5.06	0	0	0	0	0
0.01	1	0.0010192	6.12	1.26	0.87	24.04	0	0	0	0	0
0.02	0.11	0.0012389	5.7	1.72	0.88	2.07	0	0	0	0	0
0.02	0.14	0.00339	6.25	1.6	0.88	5.1	0	0	0	0	0
0.02	0.2	0.0036757	12.62	1.53	0.83	4.77	0	0	0	0	0
0.02	0.33	0.0040126	3.34	1.32	1.14	15.47	0	0	0	0	0
0.02	1	0.0057346	5.27	1.15	0.92	51.7	0	0	0	0	0
0.03	0.11	0.0012311	7.23	1.63	0.87	2.43	0	0	0	0	0
0.03	0.14	0.0009967	6.37	1.61	0.87	3.88	0	0	0	0	0
0.03	0.2	0.0022818	5.16	1.55	0.92	5.4	0	0	0	0	0
0.03	0.33	0.0110319	4.16	1.35	1.02	16.06	0	0	0	0	2
0.03	1	0.004111	3.75	1.19	1.03	42.34	0	0	0	0	0
0.05	0.11	0.0018204	5.78	1.38	0.9	5.92	0	0	0	0	0
0.05	0.14	0.0015779	7.84	2.04	0.86	2.14	0	0	0	0	0
0.05	0.2	0.0034645	4.75	1.34	0.94	9.15	0	0	0	0	0
0.05	0.33	0.0123621	1.75	1.24	2.27	24.09	0	0	0	0	0
0.05	1	0.0035687	3.63	1.18	1.03	47.33	0	0	0	0	0

Table S3: Estimated values in the case $\pi = 0$ and $\bar{\gamma} = 1$. The first two columns are $\bar{\beta}$ values (prior information of $\bar{\gamma}$: $\bar{\gamma} \sim Gamma(1, \bar{\beta})$). The third to the seventh columns are genetic parameters estimated from `extTADA`. Next columns are the number of risk genes estimated with the corresponding FDR values in the header.

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

Table S4: Genetic parameters for SCZ data if single class is used in the analysis.

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

Table S5: SCZ genetic parameters after adjusting mutation rates (NoExAC).

Parameter	Mode	lCI	uCI
pi0	0.0821	0.0487	0.1398
hyperGammaMeanDN[1]	1.2199	1.0001	2.2
hyperGammaMeanDN[2]	1.4407	1.0043	2.9893
hyperGammaMeanDN[3]	11.9591	4.1894	23.9414
hyperGammaMeanCC	1.9498	1.0845	3.2072

Table S6: Estimated genetic parameters for SCZ data with the same mean RRs for case-control data.

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

Table S7: SCZ genetic parameters using all variants in and not in ExAC database (InExAC + NoExAC).

Table S8: extTADA results of SCZ risk gene identification (See LongSupTables.xlsx Download).

Table S9: extTADA results of SCZ risk gene identification after adjusting mutation rates (See LongSupTables.xlsx Download).

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

Table S10: Abbreviations of known gene sets used in this study.

Gene set	P value	FDR
FMRP_targets	1.0e-07	2.7e-05
rbfox13	1.0e-07	2.7e-05
constrained	1.0e-07	2.7e-05
celf4	1.0e-07	2.7e-05
pLI09	1.0e-07	2.7e-05
rbfox2	1.0e-07	2.7e-05
DD.allDenovoMiDandLoF	1.0e-07	2.7e-05
abnormal_behavior	3.0e-07	7.0e-05
GGGAGGRR_V$MAZ_Q6	7.0e-07	1.5e-04
abnormal_sensory_capabilities|reflexes|nociception	1.9e-06	3.6e-04
AST.allDenovoMiDandLoF	2.1e-06	3.6e-04
abnormal_motor_capabilities|coordination|movement	2.8e-06	4.4e-04
chd8.human_brain	9.2e-06	1.3e-03
ACAGGGT,MIR-10A,MIR-10B	1.1e-05	1.4e-03
AlleleBiasedExpression.Neuron	1.1e-05	1.4e-03
GO:0016043	1.2e-05	1.4e-03
abnormal_emotion|affect_behavior	1.3e-05	1.5e-03
GO:0045202	2.2e-05	2.3e-03
GO:0071840	2.6e-05	2.5e-03
abnormal_nervous_system_morphology	2.7e-05	2.5e-03
CAGGTG_V$E12_Q6	2.9e-05	2.6e-03
GO:0008104	5.7e-05	4.9e-03
GO:0051179	6.4e-05	5.2e-03
GO:0006996	7.1e-05	5.6e-03
GO:0043234	7.4e-05	5.6e-03
ARC	7.8e-05	5.7e-03
AACTTT_UNKNOWN	8.8e-05	6.1e-03
CTTTGT_V$LEF1_Q2	9.3e-05	6.3e-03
GO:0048519	1.0e-04	6.8e-03
synaptome	1.2e-04	7.6e-03
abnormal_social|conspecific_interaction	1.3e-04	7.7e-03
GGATTA_V$PITX2_Q2	1.5e-04	8.6e-03
KEGG_AXON_GUIDANCE	1.7e-04	9.5e-03
GO:0043005	1.8e-04	9.8e-03
essentialGenes	1.8e-04	9.8e-03
Known_EPI_genes	2.0e-04	1.0e-02
GO:0045211	2.1e-04	1.0e-02
GO:0044456	2.3e-04	1.1e-02
mir137	2.4e-04	1.2e-02
NMDAR_network	2.5e-04	1.2e-02
GO:0022839	2.7e-04	1.2e-02
GO:0022836	2.7e-04	1.2e-02
GO:0034702	2.8e-04	1.2e-02
GO:0033036	2.9e-04	1.2e-02
AATGTGA,MIR-23A,MIR-23B	3.1e-04	1.3e-02

Gene set	P value	FDR
GO:0097060	3.2e-04	1.3e-02
GO:0044765	3.4e-04	1.4e-02
mGluR5	3.7e-04	1.4e-02
GO:0022834	3.7e-04	1.4e-02
GO:0015276	3.8e-04	1.4e-02
GO:0048193	4.2e-04	1.5e-02
CTTTGA_V$LEF1_Q2	4.5e-04	1.6e-02
GO:0097458	5.0e-04	1.8e-02
GO:0019226	5.1e-04	1.8e-02
GO:0022892	5.4e-04	1.8e-02
GO:0005261	5.8e-04	1.9e-02
GO:0008022	5.9e-04	1.9e-02
abnormal_fear|anxiety-related_behavior	6.0e-04	1.9e-02
abnormal_cued_conditioning_behavior	6.1e-04	1.9e-02
GO:0005215	6.2e-04	1.9e-02
GO:0048592	6.3e-04	1.9e-02
REACTOME_TRANSMISSION_ACROSS_CHEMICAL_SYNAPSES	6.6e-04	2.0e-02
GO:0048523	6.8e-04	2.0e-02
abnormal_synaptic_transmission	7.0e-04	2.1e-02
GO:0035637	7.2e-04	2.1e-02
seizures	7.3e-04	2.1e-02
abnormal_behavioral_response_to_xenobiotic	7.7e-04	2.2e-02
ID.allDenovoMiDandLoF	7.9e-04	2.2e-02
GO:0007268	7.9e-04	2.2e-02
GO:0005886	9.3e-04	2.5e-02
GWAS_(Pardinas_et_al_2017)	9.4e-04	2.5e-02
GO:0000904	9.8e-04	2.5e-02
ID.allKnownGenes	9.9e-04	2.5e-02
GO:0007399	1.0e-03	2.5e-02
GO:0022612	1.2e-03	3.0e-02
GO:0006810	1.2e-03	3.1e-02
GO:0015031	1.3e-03	3.1e-02
GO:0016568	1.3e-03	3.2e-02
GO:0048589	1.4e-03	3.3e-02
REACTOME_NEURONAL_SYSTEM	1.4e-03	3.3e-02
GO:0051234	1.4e-03	3.3e-02
GO:0071944	1.5e-03	3.5e-02
PSD-95_(core)	1.6e-03	3.6e-02
GO:0042995	1.6e-03	3.6e-02
abnormal_learning|memory|conditioning	1.7e-03	3.7e-02
abnormal_excitatory_postsynaptic_currents	1.7e-03	3.7e-02
GO:0007154	1.7e-03	3.7e-02
GO:0005216	1.7e-03	3.7e-02
GO:0010646	1.8e-03	3.7e-02
GO:0007267	1.9e-03	3.8e-02

Gene set	P value	FDR
GO:0030662	1.9e-03	3.8e-02
GO:0044700	1.9e-03	3.8e-02
GO:0023052	1.9e-03	3.8e-02
GO:0045184	2.0e-03	4.1e-02
GO:0007519	2.1e-03	4.1e-02
GO:0000139	2.1e-03	4.1e-02
abnormal_associative_learning	2.4e-03	4.6e-02
GO:0023051	2.4e-03	4.7e-02
GO:0022838	2.5e-03	4.7e-02
GO:0048731	2.6e-03	4.8e-02
GO:0032991	2.6e-03	4.9e-02
abnormal_synapse_morphology	2.7e-03	4.9e-02
GO:0010629	2.7e-03	4.9e-02

Table S11: Enrichment of gene sets from different databases with SCZ genes from `extTADA` results. These p values were obtained by 10,000,000 simulations, and then adjusted by using the method of Benjamini and Hochberg (1995).

Source	Disease	DN	DN control	Case	Control
Fromer et al. (2014)	SCZ	617			
Girard et al. (2011)	SCZ	14			
Gulsuner et al. (2013)	SCZ	105	84		
McCarthy et al. (2014)	SCZ	57			
Xu et al. (2012)	SCZ	231	34		
Guipponi et al. (2014)	SCZ	53			
Genovese et al. (2016)	SCZ			4954/4248	6239/5865
Singh et al. (2016)	SCZ			1745/1353	6789/4769
Deciphering Developmental Disorders Study (2017)	DD	4293			
EuroEPINOMICS-RES Consortium et al. (2014)	EPI	365			
De Ligt et al. (2012)	ID	100			
Hamdan et al. (2014)	ID	41			
Rauch et al. (2012)	ID	51	20		
Lelieveld et al. (2016)	ID	820			
Turner et al. (2016)	ASD	5122			
De Rubeis et al. (2014)	ASD			404	3654
Iossifov et al. (2012)	ASD		343		
ORoak et al. (2012)	ASD		50		
Sanders et al. (2012)	ASD		200		

Table S12: De novo and case/control data. For ASD studies, Turner et al. (2016) integrated previous results in their study; therefore only de novo meta data in this study are shown in the table. In addition, for ASD case-control data, only one homogeneous Sweden population from De Rubeis et al. (2014) was used. For case-control data of SCZ, after correcting for the population stratification, only 4,248 cases (3,157 + 1,091) + 5,865 (4,672 + 1,193) controls from Genovese et al. (2016) and 1,353 cases + 4,769 controls from Singh et al. (2016) are used in this study.

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

Table S13: De novo mutation counts of categories and their mutation counts per sample size for schizophrenia (SCZ), autism spectrum disorder (ASD), epilepsy (EPI), intellectual disorder (ID) and developmental disorder (DD).

Table S14: extTADA risk gene identification results of ID data (See LongSupTables.xlsx Download).

Table S15: extTADA risk gene identification results of DD data (See LongSupTables.xlsx Download).

Table S16: extTADA risk gene identification results of ASD data (See LongSupTables.xlsx Download).

Table S17: extTADA risk gene identification results of EPI data (See LongSupTables.xlsx Download).

Table S18: The p values of enrichment tests for 161 known gene sets in SCZ, DD, ID, ASD and EPI (See LongSupTables.xlsx Download).

Table S19: The p values of enrichment tests for whole gene sets in SCZ, DD, ID, ASD and EPI (See LongSupTables.xlsx Download).

6.2 Sup Figure

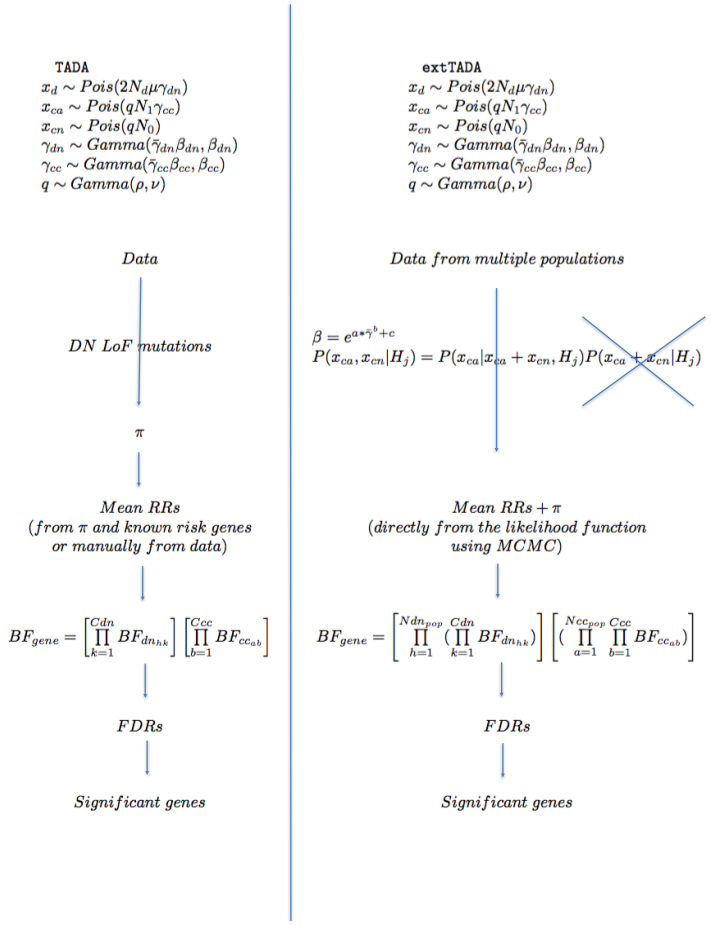

Figure S1: Comparison between `TADA` and `extTADA`. They both use the same model for de novo data (x_{dn} and case/control (x_{ca}, x_{cn}) data. `extTADA` combines all categories to obtain parameters while `TADA` is based on LoF mutations. `extTADA` uses an approximate model for case-control data, and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

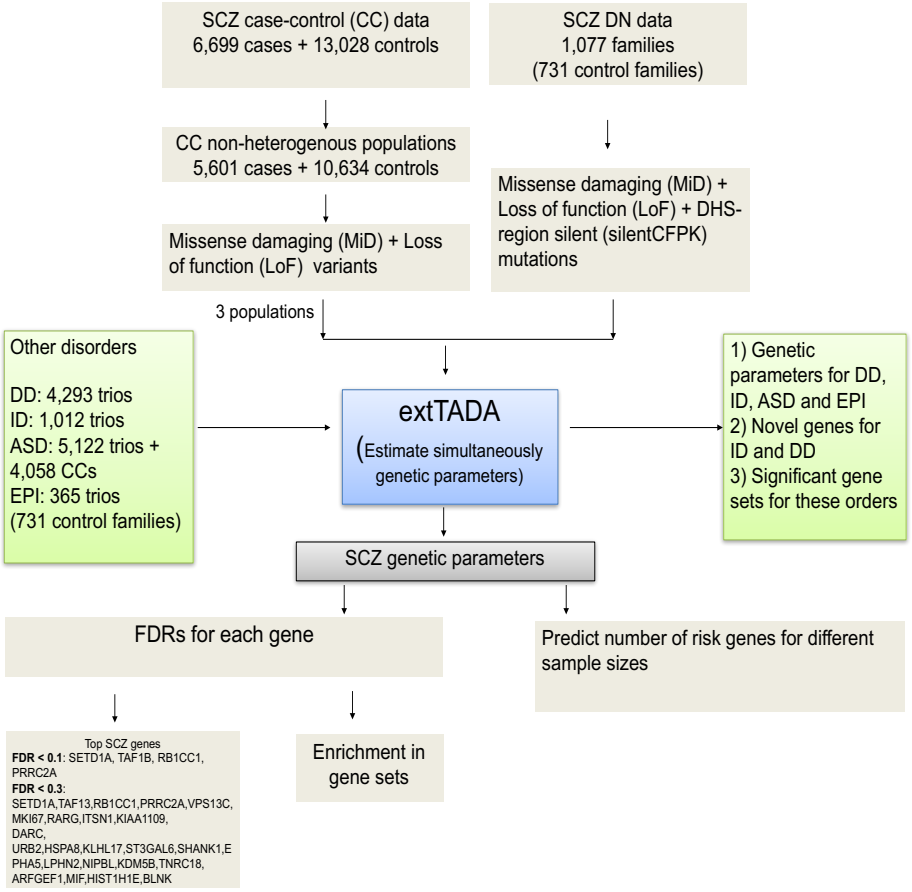

Figure S2: Workflow of data analysis.

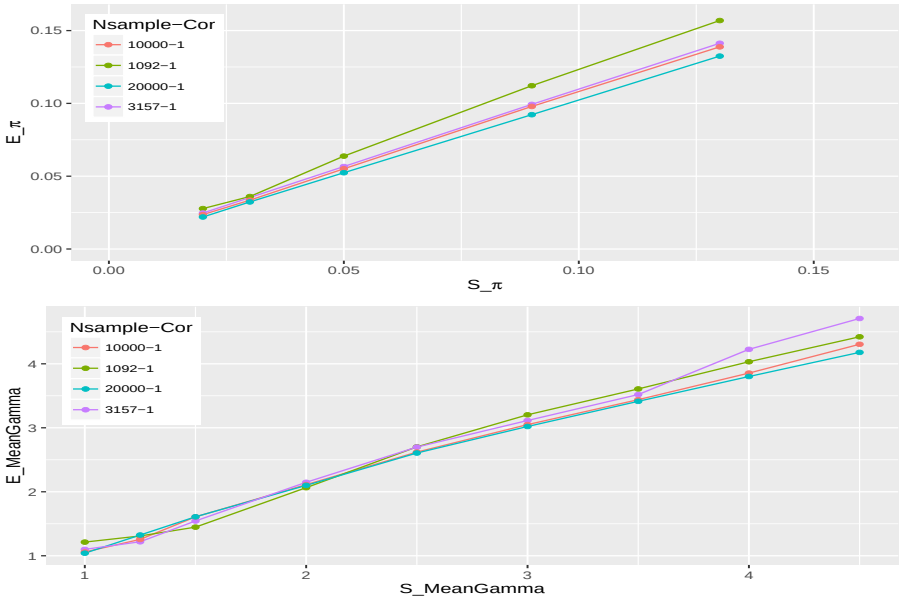

Figure S3: Correlations between estimated and simulated values for one CC class with different sample sizes. X and Y axes describe simulated (S) and estimated (E) values respectively. The top picture is for mean relative risks (MeanRRs) while the bottom picture is for the proportion of risk genes (π). Legends show sample sizes and correlations.

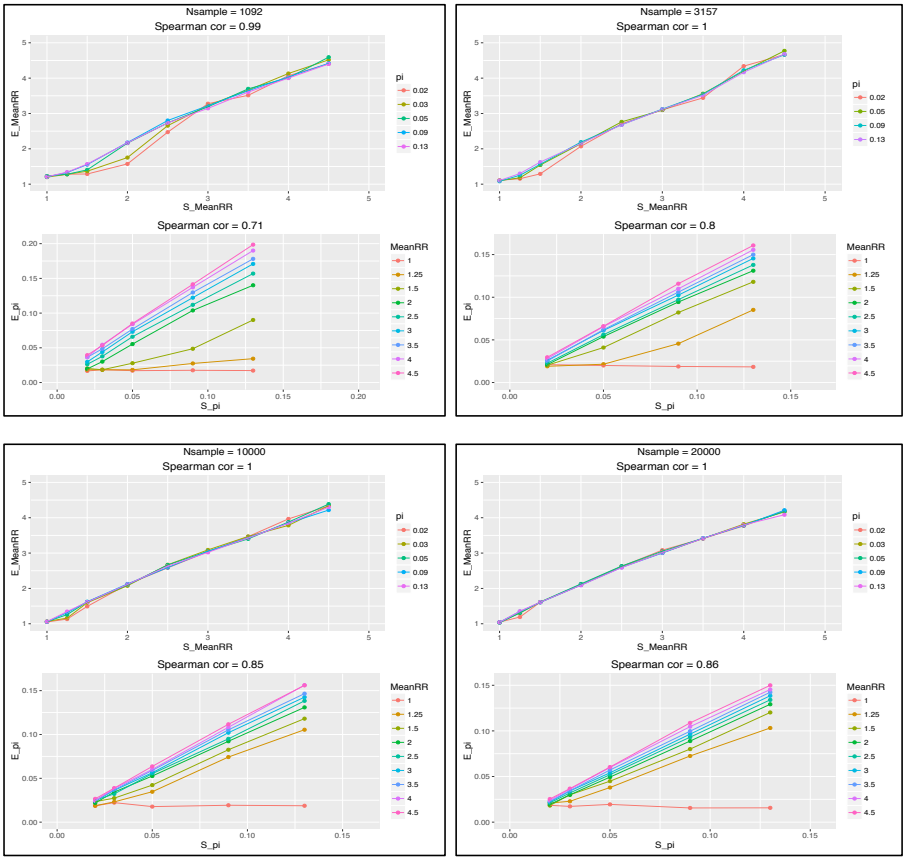

Figure S4: Correlation between simulated and estimated values for one-category case/control data.

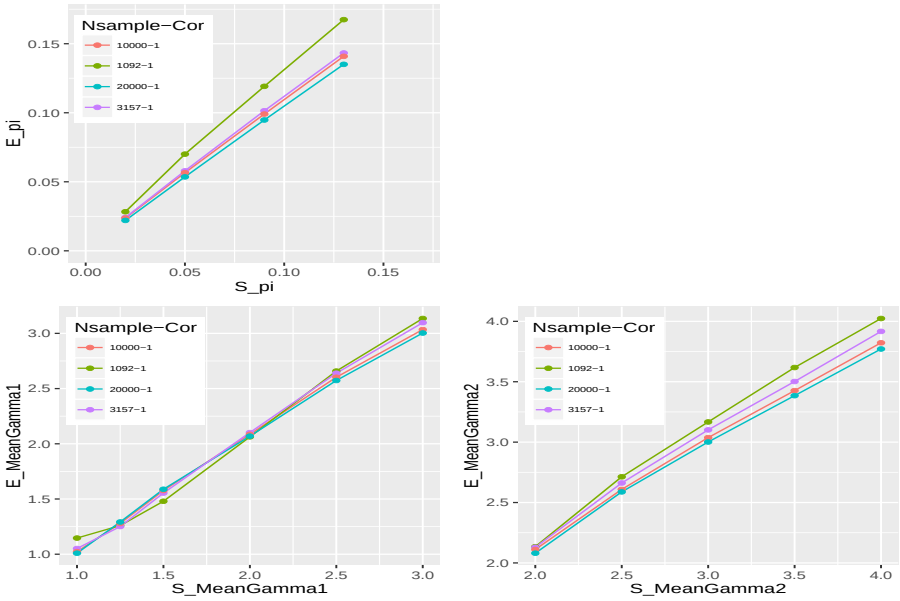

Figure S5: Correlations between estimated and simulated values for two CC class with different sample sizes. X and Y axes describe simulated (S) and estimated (E) values respectively. A range of mean relative risks for two classes (MeanGamma1 and MeanGamma2) and risk-gene proportions (π) were used in the simulation process. Legends show sample sizes and correlations.

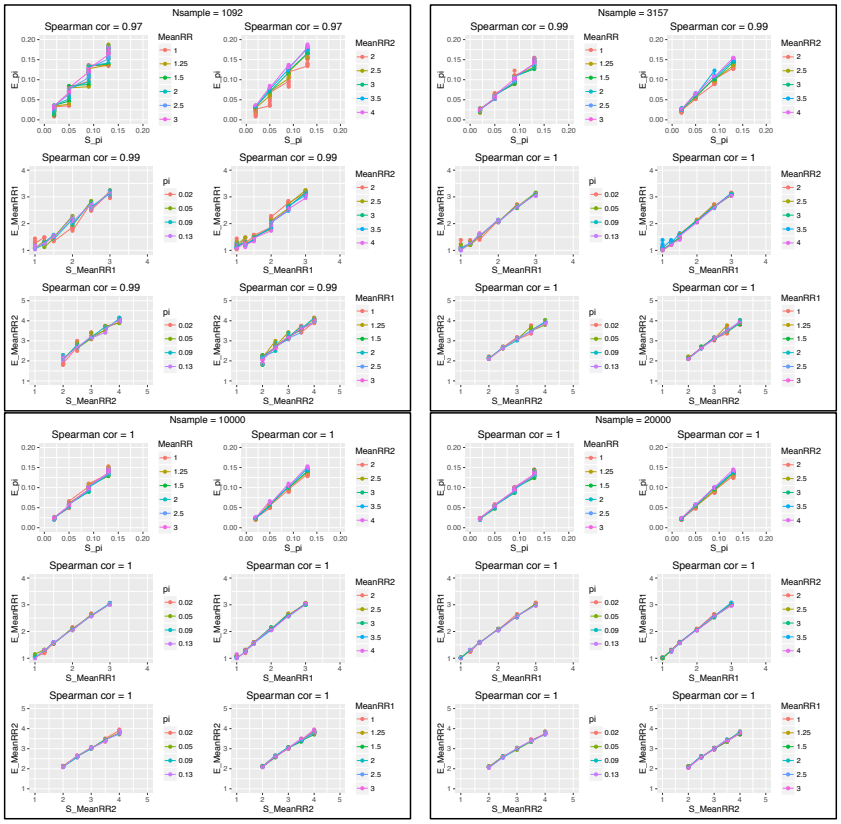

Figure S6: Correlation between simulated and estimated values for two-category case/control data.

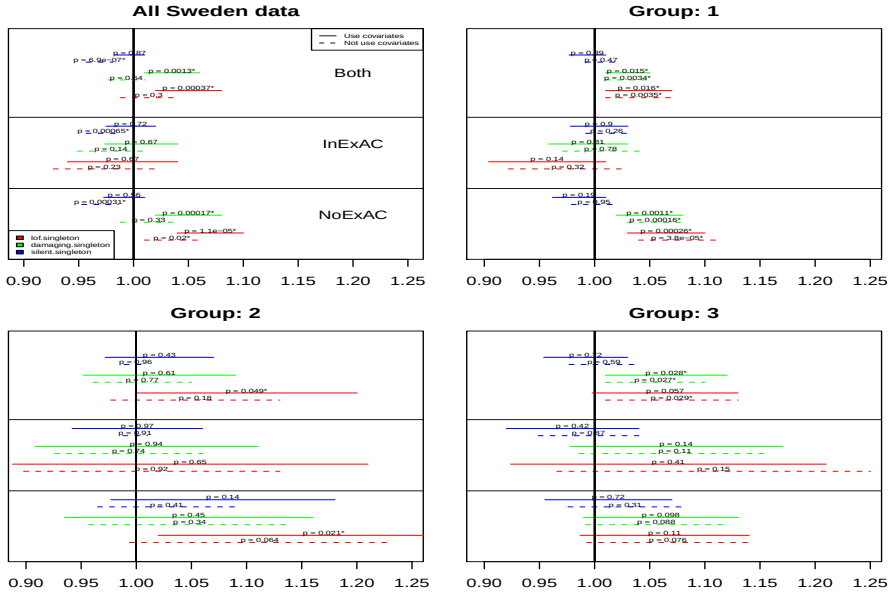

Figure S7: Odds ratios for the analysis of all case-control samples. Top left picture shows odds ratios for all Sweden samples while the three other pictures show odds ratios for three groups after the clustering process. Only group 1 and 3 are used in the current analysis because there are strong differences between results using covariates and not using covariates in group 2. P values were calculated for variants in (InExAC), not in (NoExAC) the ExAC database, and all variants (Both).

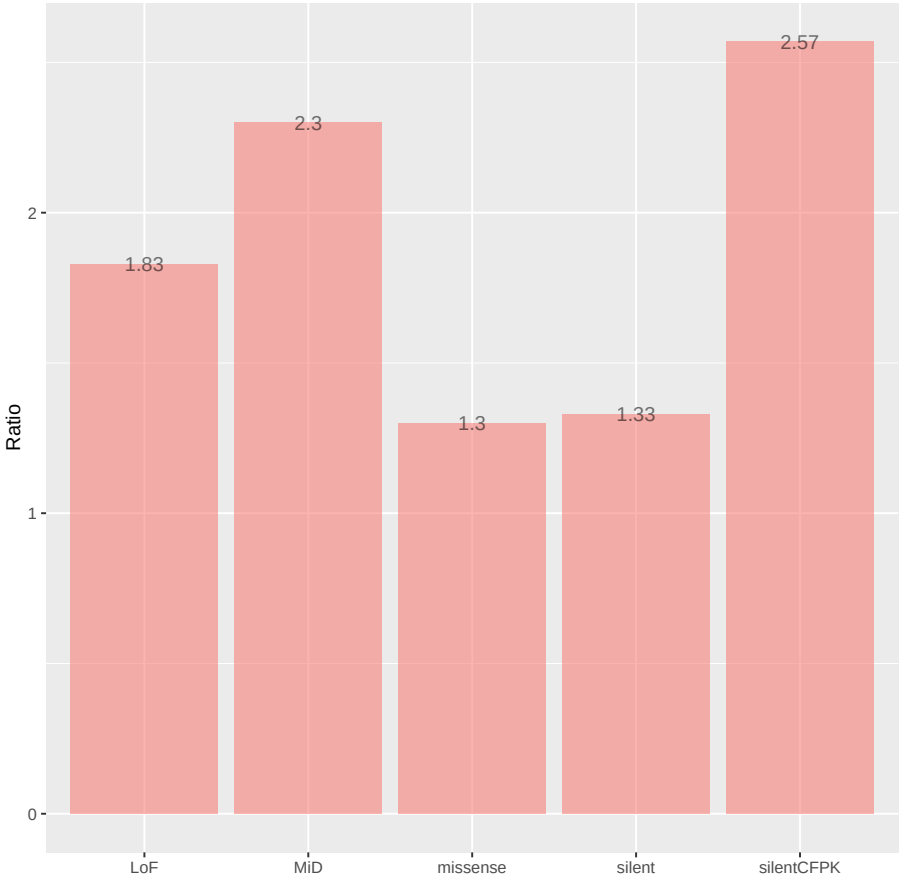

Figure S8: Ratios of de novo mutations between SCZ probands and controls (unaffected siblings). "silentFCPk" describes for silent mutations within frontal cortex-derived DHS (silentCerebrumfrontalocPk.narrowPeak). MiD mutations are missense mutations derived from 7 methods.

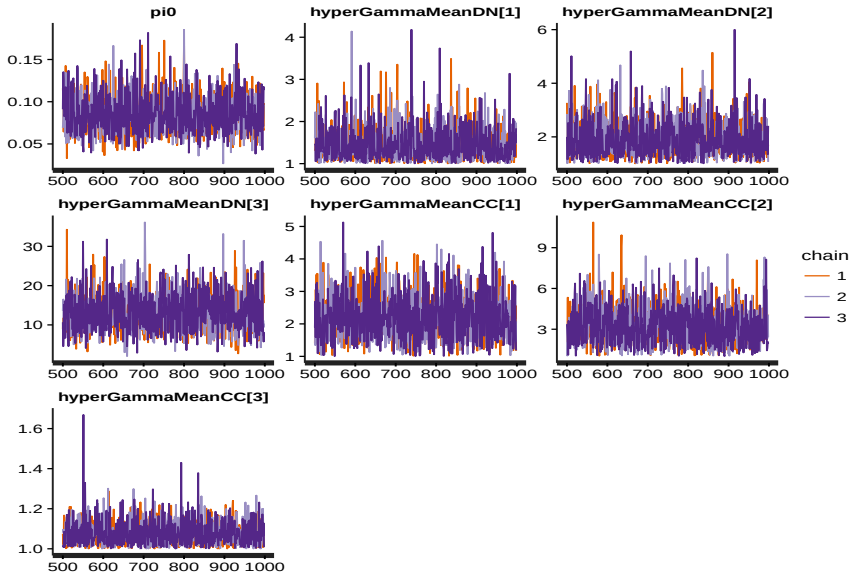

Figure S9: MCMC results for SCZ data.

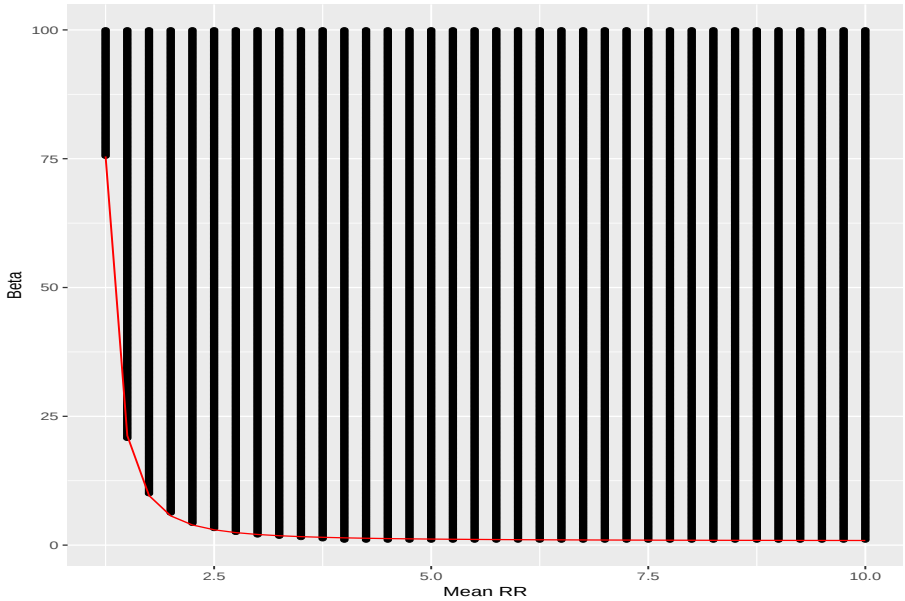

Figure S10: A grid of β and $\bar{\gamma}$ values. Points on the red line are corresponding with the proportion of protective variants less than 0.05%.

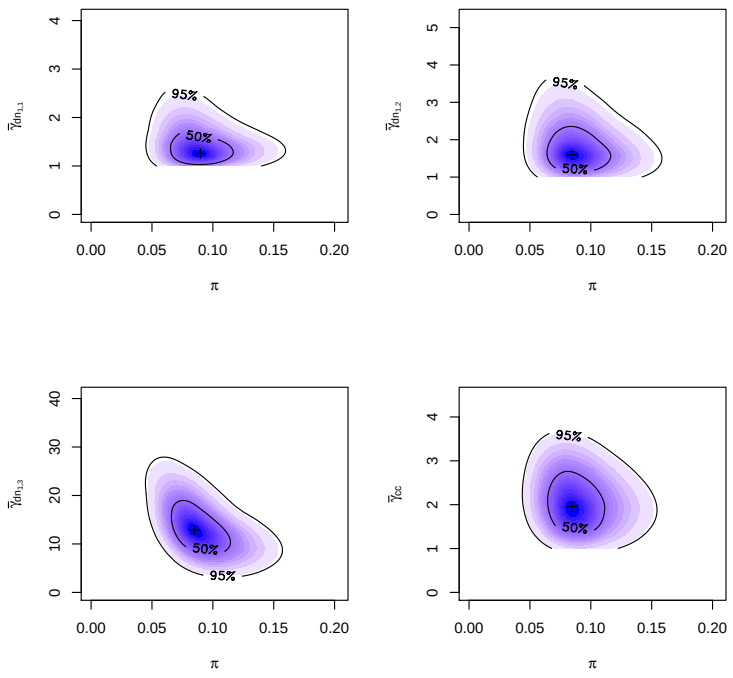

Figure S11: SCZ genetic parameters when mean RRs of case-control data are equal.

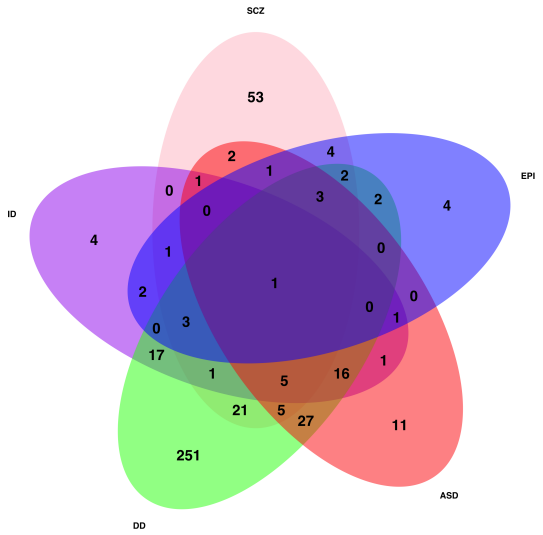

Figure S12: The overlaps of significant gene sets in SCZ, ASD, EPI, DD and ID in the genes collected from multiple databases.

6.3 Sup Information

6.3.1 Sup Results

6.3.1.1 Simulation case-control data only

To evaluate the performance of the approximate CC model for different parameter values, we simulated a single CC sample with either one or two variant/annotation classes. We tested sample sizes ranging from that of the available data, 1,092 each cases and controls (ASD), and 3,157 cases and controls (SCZ), to larger sample sizes of 10,000 cases and controls, and 20,000 cases and controls.

Overall, high correlations (~ 1) between estimated and simulated parameter values indicate little bias in inference based on CC data (Figure S3 and S5). Slight over estimation was observed for the sample size of 1092, especially for risk-gene proportions.

An additional analysis was carried out to assess the performance of specific simulated values. Correlations were calculated for each mean RR and π value. For one CC class, mean RRs were estimated well by the model with correlations ~ 1 (Figure S4). However, the proportion of risk genes was affected by mean RRs. They were estimated well when mean RRs were between 1.5 and 3.5, but underestimated with smaller mean RRs and slightly overestimated with larger mean RRs (Figure S4). For two CC classes, high correlations (≥ 0.97) between simulated and estimated values were seen for all parameters. In addition, small mean RRs of a given class did not directly affect the estimated values of proportions of risk genes (Figure S6).

The issue of poor estimation for one class, but good estimation for $>$ one

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

6.3.2 Sup methods

6.3.2.1 Calculate Bayes Factor for case/control data

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

$$P(x_{ca}, x_{cn}|H_j) = P(x_{cn}|H_j)P(x_{ca}|x_{cn}, H_j) \tag{9}$$

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

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

- The second part:

$$\begin{aligned} P(x_{ca}|H_j, x_{cn}) &= \int P(x_{ca}|q, \gamma_{cc})P(q|H_j, x_{cn})P(\gamma_{cc}|H_j)dqd\gamma_{cc} \\ &= \int [P(x_{ca}|q, \gamma_{cc})P(q|H_j, x_{cn})dq]\, P(\gamma_{cc}|H_j)d\gamma_{cc} \\ &= \int NegBin(x_{ca}|\rho + x_{cn}, \tfrac{N_0+\nu}{N_1\gamma_{cc}+N_0+\nu})P(\gamma_{cc}|H_j)d\gamma_{cc} \end{aligned} \tag{11}$$

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

References

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

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300, 1995.

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

P. Cingolani, A. Platts, L. L. Wang, M. Coon, T. Nguyen, L. Wang, S. J. Land, X. Lu, and D. M. Ruden. A program for annotating and predicting the effects of single nucleotide polymorphisms, snpeff: Snps in the genome of drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*, 6(2):80–92, 2012.

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

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

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

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

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

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

Deciphering Developmental Disorders Study. Prevalence and architecture of de novo mutations in developmental disorders. *Nature*, 542(7642):433–438, 2017.

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

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

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

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

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

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

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

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

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

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

X. He, S. J. Sanders, L. Liu, S. De Rubeis, E. T. Lim, J. S. Sutcliffe, G. D. Schellenberg, R. A. Gibbs, M. J. Daly, J. D. Buxbaum, et al. Integrated model of de novo and inherited genetic variants yields greater power to identify risk genes. *PLoS Genet*, 9(8):e1003671, 2013.

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

H. Jeffreys. *The theory of probability*. OUP Oxford, 1998.

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

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

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

J. Kosmicki, K. Samocha, D. Howrigan, S. Sanders, K. Slowikowski, M. Lek, K. Karczewski, D. Cutler, B. Devlin, K. Roeder, et al. Refining the role of de novo protein truncating variants in neurodevelopmental disorders using population reference samples. *bioRxiv*, page 052886, 2016.

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

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

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

M. Lin, A. Hrabovsky, E. Pedrosa, T. Wang, D. Zheng, and H. M. Lachman. Allele-biased expression in differentiating human neurons: implications for neuropsychiatric disorders. *PLoS One*, 7(8):e44017, 2012.

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.

E. Murphy and A. Benítez-Burraco. Language deficits in schizophrenia and autism as related oscillatory connectomopathies: an evolutionary account. *Neuroscience & Biobehavioral Reviews*, 2016.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

M. Stephens and D. J. Balding. Bayesian statistical methods for genetic association studies. *Nature Reviews Genetics*, 10(10):681–690, 2009.

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

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

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

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

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

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

X. Xie, J. Lu, E. Kulbokas, T. R. Golub, V. Mootha, K. Lindblad-Toh, E. S. Lander, and M. Kellis. Systematic discovery of regulatory motifs in human promoters and 3 utrs by comparison of several mammals. *Nature*, 434(7031): 338–345, 2005.

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.

O. Zuk, S. F. Schaffner, K. Samocha, R. Do, E. Hechter, S. Kathiresan, M. J. Daly, B. M. Neale, S. R. Sunyaev, and E. S. Lander. Searching for missing heritability: designing rare variant association studies. *Proceedings of the National Academy of Sciences*, 111(4):E455–E464, 2014.