MAIN PAPER

May 17, 2016

Abstract

Integrating de novo mutations, inherited and case-control variants has been successfull in identifying genetic architecture of autism spectrum disorder. Here, this approach is modified and applied to schizophrenia cohorts including 1024 trios, 4954 cases and 6239 controls. We identify 12 autosomal genes at a false discovery rate (FDR) < 0.1. Increasing FDR to 0.3, 48 risk genes are determined. The set of these 48 genes shows enrichment in

Contents

1 Introduction

2 Data and methods

2.1 Data

2.1.1 Simulated data

2.1.2 Schizophrenia data

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

These variants were annotated using Plink/Seq as described in Fromer et al. (2014). After that, SnpSift version 4.2 (Cingolani et al., 2012) was used to further annotate these variants using dbnsfp31a (Liu et al., 2015). Variants were groups into different categories. Loss of function (LoF) class comprised of nonsense, splice, and frameshift variants. Missense damaging were defined as missense by Plink/Seq and damaging by results of 7 methods from dbnsfp31a: SIFT, $Polyphen2_HDIV$, $Polyphen2_HVAR$, LRT, PROVEAN, MutationTaster and MutationAssessor.

2.1.3 Gene sets

Human accelerated regions (HARs)

Lists of HARs and primate accelerated regions (PARs) (Lindblad-Toh et al., 2011) were downloaded from

http://www.broadinstitute.org/scientific-community/science/projects/mammals-models/29-mammals-project-supplementary-info

on May 11, 2016. The coordinates of these regions were converted to hg19 using Liftover tool (Kent et al., 2002). We used a similar approach as Xu et al. (2015) to obtain genes nearby HARs. Genes in regions flanking 100 kb of the HARs/PARs were used in this study.

Other gene sets

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

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

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

2.2 Methods

We use the model developed in The Transmission and Disequilibrium Association (TADA) test (He et al., 2013) to describe de novo (x_d) and case (x_1) control (x_0) data as Equation 1:

$$\begin{aligned} &x_d \sim Pois(2N\mu\gamma_d) \\ &x_1 \sim Pois(qN_1\gamma) \\ &x_0 \sim Pois(qN_0) \end{aligned} \tag{1}$$

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

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

$$\begin{aligned} B_i &= \frac{P(x|H_1)}{P(x|H_0)} \\ &= \frac{\prod_{j=1}^{K} P(x_{ij}|H_1)}{\prod_{j=1}^{K} P(x_{ij}|H_0)} \\ &= \prod_{j=1}^{K} \frac{P(x_{ij}|H_1)}{P(x_{ij}|H_0)} \; (Independence\ between\ categories) \\ &= \prod_{j=1}^{K} B_{ij} \end{aligned} \tag{2}$$

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

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

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

The same as He et al. (2013), gamma distributions are assummed as prior distributions for γ_d and γ as in 4.

$$\begin{aligned} &\gamma_d \sim Gamma(\bar{\gamma}_d\beta_d, \beta_d) \\ &\gamma \sim Gamma(\bar{\gamma}\beta, \beta) \\ &q \sim Gamma(\rho, \nu) \end{aligned} \tag{4}$$

However, instead of using different ρ_1, ν_1 and ρ_0, ν_0 parameters for H_1 and H_0 (He et al., 2013); we use simplified parameters as a current TADA version (De Rubeis et al., 2014), in which $\rho_1 = \rho_0 = \rho$ and $\nu_1 = \nu_0 = \nu$.

To calculate BFs, we need to know hyper parameters in Equation 4. Let ϕ_{1j} and ϕ_{0j} be hyperparameters for H_1 and H_0 respectively. Similar to He et al. (2013), we assume a mixture model for all genes, with a probability π for a gene being a risk gene. However, we integrate all categories into the mixture model, and the marginal likelihood as in Eq 5.

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

To obtain hyperparameters $\phi_{1j} = (\gamma_{j(dn)}, \gamma_j, \beta_{j(dn)}, \beta_j, \rho_j, \nu_j)$, we use a Markov chain Monte Carlo (MCMC) method named Hamiltonian Monte Carlo (HMC) implemented in the `rstan` package (Carpenter et al., 2015; R Core Team, 2015). However, Equation 5 is complex with multiple parameters; therefore, we simplify the Equation to avoid sampling directly $q \sim Gamma(\rho, \nu)$:

- For de novo data, the same as Equation 1.
- For case-control (inheritance) data, we infer ρ, ν from control data.

  Approximation method

  $$\begin{aligned} P(x_1, x_0|H_j) \quad &= P(x_1, x_1 + x_0|H_j) \\ &= P(x_1|x_1 + x_0, H_j)P(x_1 + x_0|H_j) \end{aligned} \tag{6}$$

  - The first part: $P(x_1|x_1 + x_0, H_j)$
    Because of $x_1 \sim Pois(N_1 q\gamma)$ and $x_0 \sim Pois(N_0 q)$, we assume that x_1 and x_0 are **independent**, we have:
    $x_1|x_1 + x_0, H_j \sim Binomial(x_1 + x_0, \theta|H_j)$
    with $\theta|H_1 = \frac{N_1\gamma}{N_1\gamma + N_0}$ and $\theta|H_0 = \frac{N_1}{N_1 + N_0}$
    The marginal likelihood is:
    $P(x_1|x_1 + x_0, H_j) = \int P(x_1|x_1 + x_0, \gamma, H_j)P(\gamma|x_1 + x_0, H_j)d\gamma$
  - The second part $P(x_1 + x_0|H_j)$ is not used in the estimation process in Equation 5

Change the order of integrals to rely only on relative risks

$$P(x_1, x_0|H_j) = P(x_0|H_j)P(x_1|x_0, H_j) \tag{7}$$

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

$$P(x_0|H_j) = \int P(x_0|q, H_j)P(q|\rho, \nu, H_j)dq = NegBin(x_0|\rho, \frac{N_0}{\nu + N_0}), j = 0, 1 \tag{8}$$

- The second part:

$$\begin{aligned} P(x_1|H_j, x_0) \quad &= \int P(x_1|q, \gamma)P(q|H_j, x_0)P(\gamma|H_j)dqd\gamma \\ &= \int \left[P(x_1|q, \gamma)P(q|H_j, x_0)dq\right] P(\gamma|H_j)d\gamma \\ &= \int NegBin(x_1|\rho + x_0, \tfrac{N_0+\nu}{N_1\gamma + N_0 + \nu})P(\gamma|H_j)d\gamma \end{aligned} \tag{9}$$

The second line in Equation 9 is because $P(q|H_j, x_0)$ is the posterior probability of q after seeing the data x_0 with $q|H_j, x_0 \sim Gamma(\rho + x_0, \nu + N_0)$ (De Rubeis et al., 2014).

In Equation 9

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

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

(11)

3 Results

3.1 Simulated data

3.2 Schizophrenia data sets

3.3 Enrichment analyses

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

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

4 Discussion

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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.

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.