Integration of single cell gene expression data in Bayesian association analysis of rare variants

Guojie Zhong[1,2], Yoolim A. Choi[3], Yufeng Shen[1,3,4]

1. Department of Systems Biology
2. Integrated Program in Cellular, Molecular, and Biomedical Studies
3. Department of Biomedical Informatics
4. JP Sulzberger Columbia Genome Center

Columbia University Irving Medical Center, New York, NY

Contact: Y.S., ys2411@cumc.columbia.edu

Abstract

We present VBASS, a Bayesian method that integrates single-cell expression and de novo variant (DNV) data to improve power of disease risk gene discovery. VBASS models disease risk prior as a function of expression profiles, approximated by deep neural networks. It learns the weights of neural networks and parameters of Poisson likelihood models of DNV counts jointly from expression and genetics data. On simulated data, VBASS shows proper error rate control and better power than state-of-the-art methods. We applied VBASS to published datasets and identified more candidate risk genes with supports from literature or data from independent cohorts.

Keywords

Bayesian association analysis; Machine Learning; Rare variants; Single cell expression

Background

About 3% of children are born with congenital anomalies or will develop neurodevelopmental disorders (NDD)[1]. Given the severe consequence of these conditions on reproductive fitness, risk variants with large effect are under strong negative selection and therefore have low frequency in the population. Recent genetics studies identified hundreds of risk genes of these conditions, largely by rare *de novo* variants[2-11], however, the majority of risk genes remain unidentified[10,12-15], due to challenges in statistical power in analysis of rare variants[16].

Cell-type specific gene expression has long been used qualitatively for interpretation of biological mechanisms in developmental biology and genetics. Previously we have shown that high expression in developing heart and diaphragm is associated with increased burden of *de novo* coding variants in congenital heart disease (CHD)[7] and congenital diaphragmatic hernia[9], respectively. We have also shown that cell-type specific expression in brain is associated with plausibility of autism spectrum disorders (ASD) risk genes[17,18]. It is clear that gene expression profile can inform association analysis of rare variants for risk gene discovery. However, the ability to improve power in gene discovery using expression data has been hindered by the lack of rigorous statistical methods and cell type specific expression data from relevant tissues during development. Recent efforts in cell atlas of human and model organisms have been generating large amount of single cell expression data of adult tissues[19,20] in addition to an increase in various developmental stages[21-24]. Here we describe a novel computational method that leverages expression data with probabilistic models to improve statistical power of risk gene discovery.

VBASS (Variational inference Bayesian ASSociation), takes a vector of expression profile, such as cell-type specific expression from single cell RNA-seq and models the priors of risk genes as a function of expression profile of multiple cell types. VBASS uses deep neural networks to approximate the function and uses semi-supervised variational inference to estimate the parameters. Although optimized for scRNA-seq data, VBASS could also be applied to bulk RNA seq data with a simplified framework. We compared the performances of VBASS with extTADA under two conditions (bulk and scRNA-seq data) by both simulated and published *de novo* variants datasets to assess error control and statistical power and showed their better performances.

Methods

The probabilistic model of VBASS

VBASS is a Bayesian mixture model with learnable priors (**Figure 1**). VBASS assumes the number of genetic variants of interest (LGD or Dmis *de novo* variants) in the gene d_{gv} are drawn independently through this generative process, given M_{gv} being the aggregated background mutation rate for variant type v in gene g and x_g being the cell type specific gene expression profiles in gene g:

$$\pi_g = f_E(x_g)$$

$$y_g \sim Bernoulli(\pi_g)$$

$$k_{gv} = \begin{cases} \overline{k_v}\ if\ y_g = 1 \\ 1\ otherwise \end{cases}$$

$$\theta_{gv} = \begin{cases} \overline{\theta_v}\ if\ y_g = 1 \\ 1\ otherwise \end{cases}$$

$$\gamma_{gv} \sim \begin{cases} Gamma\left(k_{gv}, \theta_{gv}\right)\ if\ y_g = 1 \\ 1\ otherwise \end{cases}$$

$$d_{gv} \sim \begin{cases} Poisson\left(\gamma_{gv} * M_{gv}\right)\ if\ y_g = 1 \\ Poisson\left(M_{gv}\right)\ otherwise \end{cases}$$

π_g is a gene specific prior probability of being disease risk. y_g is a binary random variable that indicates whether the gene is a risk gene or not. It is also used to generate a mixture of posterior probabilities on effect size γ_{gv}.

We use neural network f_E to infer π_g from gene expression data x_g with KL penalty of a fixed Bernoulli prior (Figure 1). By default, we used a 32-dim encoding module, followed by a 2-dim sampler module for π, respectively. Each module consists of a linear layer followed by ELU activation and layer normalization layers. We apply the same reparameterization trick as conventional variational autoencoders in f_E with Bernoulli sampler[25].

γ_{gv} is a random variable that denotes the enrichment rate of damage variant v in the patient cohort, which is also known as the relative risk of this gene. γ_{gv} is drawn independently through Gamma distribution $p(\gamma_{gv}|k_{gv}, \theta_{gv})$. k_{gv}, θ_{gv} are conditioned on y_g, under null they are equal to 1 while under alternative, they are equal to $\overline{k_v}$ and $\overline{\theta_v}$, respectively. We assume $\overline{k_v}$ and $\overline{\theta_v}$ are shared across all disease risk genes to reduce the number of parameters.

The loss function is given by the evidence lower bound (ELBO),

$$ELBO = -KL[q(y|x)||p(y)] - \mathbb{E}_{q(y|x)} \log\left(p(d|y)\right)$$

The KL penalty term regularized the gene-specific prior π by the hyperparameter $\bar{\pi}$ (**Figure 1**), which reflects the average proportion of risk genes:

$$KL[q(y|x)||p(y)] = KL[Bernoulli(y|\pi; f_E(x))||Bernoulli(y|\bar{\pi})]$$

The expectation term quantified the log likelihood of d conditioned on y integrated on the distributions parameterized by π:

$$\mathbb{E}_{q(y|x)} \log\left(p(d|y)\right) = \int Poisson\left(d\middle|\gamma * M; \overline{k_v}, \overline{\theta_v}\right) * Bernoulli(y|\pi; f_E(x))dy$$

$f_E, \overline{k_v}, \overline{\theta_v}$ are the parameters to learn, we use stochastic gradient decent to estimate them. The estimated parameters were used to calculate the posterior probability of association (PPA) for each gene being risk or not:

$$PPA = \frac{\pi_g * GammaPoisson\left(d_{gv}\middle|\overline{k_v}, \overline{\theta_v}, M_{gv}\right)}{\pi_g * GammaPoisson\left(d_{gv}\middle|\overline{k_v}, \overline{\theta_v}, M_{gv}\right) + \left(1 - \pi_g\right) * Poisson\left(d_{gv}\middle|M_{gv}\right)}$$

For conditions where gene expression data x_g is a scalar, i.e., bulk RNA-seq data or average expression data of a certain cell type in scRNA-seq data, we could rewrite f_E as a function with sigmoid shape, corresponding to a linear transformation with sigmoid activation:

$$f_E(x_g) = \bar{\pi} * Sigmoid(x_g|A, B, C)$$

$$Sigmoid(x_g|A,B,C) = C + \frac{L}{1 + A * \exp{(-x_g + B)}}$$
$$L = (1 - C) * \frac{A}{\log(\exp(A) + \exp(A * B)) - \log{(\exp(A * B) + 1)}}$$
while the other parts of the model remain the same.

Given PPA of all genes, we calculate Bayesian false discovery rate (FDR) by estimated false discovery proportion following the method described in He et al., 2013[12]:

$$\mathrm{FDR}_k = \frac{\sum_{i=1}^{k}(1 - \mathrm{PPA}_i)}{k}$$

Where i is the rank index of genes (start with highest PPA), and FDR_k is the estimated FDR of the gene ranked at k.

Parameter inference for VBASS

The parameters of VBASS could be inferenced with either unsupervised or semi-supervised training. For the scalar version, there are only six parameters to be estimated, $\overline{\pi}, \overline{\theta_v}, \overline{k_v}, A, B, C$, which is possible for complete unsupervised training via MCMC. In practice, we used rstan package with 4 chains and 2000 iterations. For the neural network version, it is better to train in a semi-supervised manner to avoid converging issues. In practice, we trained the model with two training steps. First, we pre-trained our model using known risk genes labeled as positives and genes that harbor LGD variants in control cohort as negatives, replacing the Bernoulli KL penalty with cross-entropy loss[26]. The known risk genes (59 in total) were randomly picked from SFARI[27] (release 2021 Q4) scored 1 genes, while negative controls (86 in total) were picked from genes with LGD variants in a control cohort[14] (Table S1). During pre-training we set large learning rate to make the model converge faster. The parameters estimated from pre-training were then used as initial values in the second step, unsupervised training, which uses all genes without labels with reduced learning rate after each epoch. In practice, we used 50 epochs of semi-supervised pretraining and 60 epochs of unsupervised training. After training, we calculated PPA for all genes using the estimated parameters. For the simulation dataset, we estimated FDR on all genes to measure the statistical power. For the real dataset, we removed the known risk genes selected as positives in training when we estimate FDR to identify candidate risk genes.

***De novo* variants (DNV) and gene expression data**

We obtained DNV data sets from a publication on congenital heart disease (CHD)[13] of 2,645 parent-offspring trios (Table S2) and a preprint on autism spectrum disorder (ASD)[11] of 16,616 trios (Table S3). The latter is a combined data set from exome or whole genome sequencing data of the SPARK consortium[28], Simons Simplex Collection[29], Autism Sequencing Consortium[30], and MSSNG[31]. The gene expression rank was based on bulk RNA-seq data of mouse developing heart at E14.5, inspired from previous publications[6,7]. We obtained single cell RNA-seq data of human fetal midbrain and prefrontal cortex from two publications[21,22]. We used the combination of developmental time and cell ontology annotations as described in the two publications to define cell types. Small clusters of cell types with less than 10 single cells were removed. For each gene and cell type, we calculate the proportion of cells that express the gene as input to VBASS.

Annotation of *de novo* variants and background mutation rate calculation

We used ANNOVAR[32] and VEP[33] to annotate variants, protein-coding consequences, and predicted damaging scores for missense variants. We classified variants as LGD (likely gene disrupting, including frameshift, stop gained/lost, start lost, splice acceptor/donor), Dmis (Damage missense variants, defined by REVEL[34] score ≥ 0.5), missense, or synonymous. For each variant type, we calculated the expected number of variants based on a background mutation rate model[7,35] given the sample size. In-frame deletions/insertions (multiple of 3 nucleotides) and other splice region variants were excluded in the following analysis. Variants in olfactory receptor genes, HLA genes or MUC gene family were excluded in further analysis.

Generation of simulation datasets

We simulated two datasets to test VBASS's performance with bulk and scRNA-seq datasets, respectively. For the first scenario, we first estimate the parameters based on real dataset and then used the estimated hyperparameters to generate the simulated dataset based on the Bayesian mixture model. Specifically, we randomly assigned 3.7% of genes as risk gene, then we drew the covariates (gene expression rank) of risk genes from the sigmoid distribution function. The *de novo* damage variants were drawn from Gamma-Poisson distribution with relative risk of 20 and 12 for LoF and Dmis, respectively. For non-risk genes, we drew covariates from a uniform distribution and *de novo* variants from Poisson distribution. We did the simulation under different sample sizes ranging from 2,645 to 20,000. For each sample size setting, we simulated 100 datasets and fit both models on each simulated dataset independently to estimate the hyperparameters, which were used to calculate the posterior probability of association (PPA) and then a Bayesian false discovery rate (FDR) by false discovery proportion implied by it. We performed single-tail Poisson tests independently on each simulated dataset to show the baseline statistical power, where the FDR were calculated by the Benjamini-Hochberg (BH) method.

For the second scenario, the simulation is based on real single cell dataset, where we created a non-linear function that maps cell-type specific expression to prior of being risk with following steps. First, we did a singular value decomposition (SVD) on the expression data of 59 known ASD risk genes (picked randomly from SFARI[27] scored 1 genes) and 86 negative control genes (picked randomly from genes with LGD variants in control cohort[14]) (Table S3). Next, we fit a logistic regression model with elastic net penalty on the eigen vectors that explain 95% of the variance. The regression model was applied to all other genes and the output probabilities were squared and scaled to have an average of 3.2%, which matches the average proportion of risk genes estimated from extTADA model. This value served as a simulated prior of being risk, from which disease risk genes were randomly sampled. The *de novo* damage variants were drawn from Gamma-Poisson distribution for disease risk genes while Poisson distribution for non-risk genes with same sample size and relative risk as in real ASD dataset. We performed the simulation 50 times with same simulated prior and disease risk genes, then estimated the hyperparameters and calculate the PPA and Bayesian false discovery rate (FDR) independently on each simulated dataset for both models.

Results

VBASS model disease risk association with both genetics and expression data

VBASS is a Bayesian mixture model with learnable priors (**Figure 1**). We model the number of genetic variants of interest (e.g., LGD or Dmis *de novo* variants) in the gene as a sample drawn independently through mixture of Poisson and Gamma-Poisson (Negative binomial)

Distributions. Such Gamma-Poisson distribution has been proved useful in modeling the sparse de novo variant data[12,36]. Instead of using a naïve prior that all genes share the same probability of being disease risk, we assume that this prior should be gene specific and could be inferred from the spatiotemporal expression data of fetal development of corresponding organ. In VBASS, we model this prior π_g as a function of expression profiles x_g that could be approximated with a neural network f_E (**Figure 1** and **Methods**). With such approximation it is possible to take the advantage of the state-of-art stochastic gradient descent method[37]. y_g is a binary random variable that indicates the risk status of a gene, which follows a Bernoulli distribution of π_g and constrained with Kullback–Leibler divergence over average proportion of disease risk gene, $\bar{\pi}$. This penalty term could be replaced by a cross-entropy loss term if the label of gene is known, making it possible for semi-supervised training[26]. γ_{gv} is a random variable that denotes the relative risk of damage variant v in the patient cohort. It is drawn independently through a Gamma distribution $p(\gamma_{gv}|k_{gv}, \theta_{gv})$. We assume this distribution is shared across all disease risk genes (**Methods**).

VBASS could also take bulk RNA seq data of certain organ or cell type as input when prior knowledge of its disease risk association is available. For example, the increased burden of damage variants of high heart expression genes in CHD[7]. In that case x_g is a scalar and f_E could be parameterized by three parameters (A, B, C) that corresponds for a linear transformation followed by sigmoid activation (**Methods**). This sigmoid-shape function could quantify the fact that genes with higher expression in the corresponding organ or cell type are more likely to harbor disease risk variants.

We trained VBASS in a semi-supervised manner with stochastic gradient descent method to estimate the parameters (**Methods**). While for the simplified version with scalar input, VBASS can be trained in a completely unsupervised manner with MCMC (**Methods**). The estimated parameters were used to calculate PPA and FDR for all genes (**Methods**).

VBASS showed better power than extTADA on simulated data with bulk RNA-seq expression

We tested the performance of VBASS and extTADA on simulated CHD dataset (**Methods**). As expected, both models showed good false discovery control and local false discovery control (Sup Fig. 1). VBASS outperformed extTADA with better recall under same precision level (**Figure 2A**) under sample sizes from 2,645-20,000. Although the difference in power decreases with increasing sample size, VBASS still outperformed extTADA by roughly 10% increase of recall at sample size of 10,000, which is feasible for CHD in the next few years.

To test the power of VBASS with respect to the size of genes, we calculated the recall rate at same significance levels (FDR ≤ 0.05) on both models for genes with different mutation rates. VBASS showed better statistical power especially for genes with higher mutation rates under small sample sizes (**Figure 2B**). As the sample size increases, the power difference of VBASS and extTADA becomes smaller on large genes, while VBASS still outperforms extTADA on medium-mutation-rate genes (**Figure 2B**). Overall, our simulation results showed that VBASS can increase the statistical power for prioritizing disease risk genes by estimating risk prior as a function of expression.

VBASS showed better power than extTADA on simulated data with scRNA-seq expression

We ran VBASS and extTADA separately on the simulation dataset (Method). Both models showed good false discovery control (**Figure 3A**). To test the statistical power of VBASS and extTADA, we plotted the precision-recall curve using the output posterior probabilities from VBASS, extTADA and the real parameters we used in simulation. VBASS outperformed extTADA with higher recall under same precision (**Figure 3B**). Further comparison showed good correlation between the prior value $\bar{\pi}_g$ informed by VBASS and real π_g we used in simulation (**Figure 3C**), indicating that VBASS could reconstruct the prior of being risk through single cell expression data. Moreover, we assessed the association between expression profile x and π via spearman correlation, the result of VBASS is close to real values (**Figure 3D**). Overall, those results showed that our model can not only reach higher statistical power on simulation data set than extTADA but also uncover the association between cell type expression profiles and disease risk.

VBASS identified novel CHD candidate risk genes on published DNV data

We applied VBASS to a CHD data set with DNVs from 2645 trios[13]. We used the mouse embryonic E14.5 heart bulk RNA-seq data to set gene expression rank percentile[6,7]. The estimated distribution of expression rank under null and alternative hypothesis showed most of the risk genes are enriched in rank percentile ≥ 75% (genes with rank percentile ≥ 0.75 are roughly 3 times more likely to be risk than other genes) (**Figure 4A**; **Table 1**), consistent with previous burden analysis of *de novo* variants[7]. With FDR ≤ 0.1, we identified 49 candidate risk genes. In contrast, using the original TADA method, we were able to identify only 40 candidate genes (**Figure 4B-C**, **Table 2**, Table S4). Among the gene that only detected by VBASS, *FLT4* was reported to be a risk gene via combined analysis of *de novo* and inherited variants in the original paper, while *TSC1* and *FBN1* were in their curated CHD gene dataset from literature search[13,38]. *CHD4* was reported to be significantly associated with CHD in a UK CHD cohort of 1891 probands[8], while 3 (*FRYL*, *SETD5*, *KMT2C*) have both LoF and missense variants carriers, 2 (*GANAB*, *KDM5A*) have only missense variants carriers in that cohort. Furthermore, 4 (*CHD4*, *SETD5*, *KMT2C*, *FBN1*) are significantly associated with neurodevelopmental disorders[39], while 11 (*CHD4*, *FRYL*, *GANAB*, *SETD5*, *MINK1*, *ANK3*, *KMT2C*, *IQGAP1*, *TSC1*, *KDM5A*, *FBN1*) have both LoF and missense variants carriers, and 2 (*CAD*, *SLIT3*) have only missense variants carriers in that cohort. Overall, these genes have additional genetic evidence in other cohorts and are plausible candidates. These results indicate that the assumption of VBASS is biologically sound and suggests its higher statistical power even in lower cohort size.

VBASS identified novel ASD candidate risk genes on published DNV data

Previous studies have shown that gene expression in multiple cell types in the brain is associated with ASD risk[17,18,40]. This is in part what motivated the design of VBASS. We obtained ASD DNV data from a recent preprint[15] that combined exome and genome data from four studies (see Methods), and single cell RNA-seq data of human fetal midbrain and prefrontal cortex from two publications[21,22]. We applied VBASS and extTADA to the full ASD data set with 16616 trios. VBASS identified 122 genes with PPA above 0.8 (Table S5). To compare the performance in identification of novel candidate risk genes, we removed the known risk genes used in training and calculate Bayesian FDR of all other genes with VBASS and extTADA (methods). Then we compared the candidate genes identified by VBASS and extTADA at significance level 0.05 and 0.1 (FDR ≤ 0.05 and FDR ≤ 0.1 respectively). At significance level 0.05, VBASS identified 51 genes (Table S6), among which 5 were not identified as candidates by extTADA (Fig. 5A, Table S6). Among the 5 genes, 2 (*DLG4, PAX5*) were reported to be risk genes in SFARI[27] data base (release 2021 Q4) with score of 1 while not in our training gene list.

METTL23 is a transcriptional partner of GABPA and essential for human recognition[41], and disruption of *METTL23* was reported to cause mild autosomal recessive intellectual disability[42]. *ATF4* was reported to have significant altered expression in the middle frontal gyrus of ASD subjects[43]. At significance level 0.1, VBASS identified 75 genes (Table S6), where 6 were not identified by extTADA (**Figure 5A**, Table S6). Among the 6 genes, 2 (*ZMYND8 CASZ1*) were scored 1 in SFARI data base and *CMPK2* was scored 3. *LMTK3* was reported to cause behavioral abnormalities such as locomotor hyperactivity and reduced anxiety in mice knock-out models[44,45]. Furthermore, 7 out of the 11 genes identified only by VBASS (*DLG4, METTL23, SPRY2, LMTK3, PFN2, CASZ1, ZMYND8*) have additional genetic evidence in related cohorts[39]. There were six genes (*CCDC40, FUBP3, PRKAR1B, SIN3A, ITGB5, PMM2*) identified only by extTADA but not by VBASS, likely because of their low detection rates or co-expression strength with other candidates in the single cell datasets. Finally, we studied what are the cell types that associated the most with disease risk. According to spearman correlation analysis, oculomotor / trochlear nucleus (hOMTN), GABAergic neurons (hGaba) and dopaminergic neurons (hDA1) in gestation week 9-10 are more associated with autism risk, while microglia cells and endothelial cells (hEndo) are less associated with autism risk (**Figure 5B**). This observation is consistent with previous evidence of abnormalities in GABAergic neurons and synapses in neurodevelopmental disorders characterized by a shared symptomatology of ASD symptoms[46], while reductions in GABA have been reported in several brain regions in children with ASD[47,48]. There were also evidences that dopaminergic dysfunctions were associate with autistic-like behavior[49,50].

Discussion

In this study, we described VBASS for identification of candidate risk genes by joint analysis of *de novo* variants of cases and gene expression profile of normal samples. The core idea of the method is that prior probability of a gene increase disease risk is a function of expression profile in relevant cell types, and that we can estimate the parameters of the function from the data in an empirical Bayesian framework. For bulk RNA-seq data, we set the function to be a sigmoid function with three parameters. For scRNA-seq data, we use deep neural networks to approximate the function and learn the contribution of cell types jointly with genetic data. Using simulation, we showed that VBASS have accurate error rate control and better statistical power than existing methods under both scenarios.

We applied VBASS to a published CHD DNV data set and estimated that high-expression genes are approximately 3 times more likely to be risk genes than low-expression genes in developing heart. We identified 14 more candidate risk genes, 6 of which have additional support in independent cohorts. We applied VBASS to a published ASD DNV data set and identified 5 and 6 more candidate genes at significance level 0.05 and 0.1 respectively, 8 of them have literature support or additional genetic evidence in neurodevelopmental disorders. Moreover, we showed that gene expression profiles of GABAergic neurons and dopaminergic neurons during gestation week 9-10 are strongly associated with autism risk, indicating their potential roles in neural circuits formation.

VBASS is based on the biological hypothesis that gene expression level in relevant cell or tissue types informs the plausibility of being a disease risk gene. The bulk-RNA seq version is optimized for a single expression profile that is informative of disease risk, such as bulk RNA sequencing data for congenital heart disease. The single cell RNA-seq version is optimized for the conditions in which multiple cell types and time points are associated with disease risk. One alternative approach to improve power based on informative non-genetic data is to calculate p-values for each gene using genetic data and then optimize FDR estimation using non-genetic data as covariates[51-53], While it is a generalizable approach, these methods require p-values to have proper distributions (uniform) under the null. In the analysis of *de novo* or ultra-rare variants, the data is usually too sparse to support a proper distribution of p-values under the null. VBASS does not have this limitation.

A limitation of VBASS is that it only estimates the association of cell types with disease risk. It is not designed to answer questions about whether a certain cell type confers causality in the diseases caused by risk variants. Additionally, the performance of VBASS is partially determined by how well the expression data captures true expression states of genes. In this study, we used average expression of genes in cells within a cell type inferred from single cell data. This approach has limitations in representing rare and transient cellular states. More advanced representation, like RNA velocity[54,55], together with more comprehensive measurements of cell types may improve the model.

Finally, we note the inference part of VBASS is not limited to scRNA-seq data but could be extended to other functional genomics modalities of genes, such as single cell ATAC-seq data or regulator-targets information without much modification of architecture.

Conclusions

We developed VBASS, a new computational method that integrates expression data with Bayesian probabilistic models to improve statistical power of risk gene discovery. It showed proper error rate control and better power than current Bayesian methods in simulation and real datasets. VBASS is freely available for academic use at: https://github.com/ShenLab/VBASS.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availabilities for data and meterials

All data sets (*de novo* variants and gene expression data) were obtained from publications and available from the corresponding publications. VBASS is available on Github: https://github.com/ShenLab/VBASS.

Competing interests

The authors have declared that no conflict of interest exists.

Funding
This work was supported by NIH grants R01GM120609, R03HL147197, and U01HL153009, and Simons Foundation Autism Research Initiative (SIMONS606450).

Author's contributions

Conceptualization, Y.S.; Methodology, G.Z. and Y.S.; Software, G.Z.; Investigation, G.Z., Y.A.C. and Y.S.; Writing – Original Draft, G.Z. and Y.S.; Writing – Review & Editing, G.Z., Y.A.C. and Y.S.; Supervision, Y.S.; Funding Acquisition, Y.S.

Acknowledgements
We would like to thank Dr. Wendy Chung, Dr. David Knowles, Dr. Nicholas Tatonetti, Dr. Haicang Zhang, Dr. Xueya Zhou, Dr. Xiao Fan, Yige Zhao, Joseph Obiajulu, Xi Fu, and members of Shen lab for helpful discussions.

Reference:

1. Rynn, L., Cragan, J. & Correa, A. Update on overall prevalence of major birth defects - Atlanta, Georgia, 1978-2005 (Reprinted from MMWR, vol 57,m pg 1-5, 2008). *Jama-Journal of the American Medical Association* **299**, 756-758 (2008).
2. O'Roak, B.J. *et al.* Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. *Nat Genet* **43**, 585-9 (2011).
3. Neale, B.M. *et al.* Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* **485**, 242-5 (2012).
4. De Rubeis, S. *et al.* Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* **515**, 209-15 (2014).
5. Sanders, S.J. *et al.* Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. *Neuron* **87**, 1215-1233 (2015).
6. Zaidi, S. *et al.* De novo mutations in histone-modifying genes in congenital heart disease. *Nature* **498**, 220-3 (2013).
7. Homsy, J. *et al.* De novo mutations in congenital heart disease with neurodevelopmental and other congenital anomalies. *Science* **350**, 1262-6 (2015).
8. Sifrim, A. *et al.* Distinct genetic architectures for syndromic and nonsyndromic congenital heart defects identified by exome sequencing. *Nat Genet* **48**, 1060-5 (2016).
9. Qi, H. *et al.* De novo variants in congenital diaphragmatic hernia identify MYRF as a new syndrome and reveal genetic overlaps with other developmental disorders. *PLoS Genet* **14**, e1007822 (2018).
10. Satterstrom, F.K. *et al.* Large-Scale Exome Sequencing Study Implicates Both Developmental and Functional Changes in the Neurobiology of Autism. *Cell* **180**, 568-584 e23 (2020).
11. Qiao, L. *et al.* Rare and de novo variants in 827 congenital diaphragmatic hernia probands implicate LONP1 as candidate risk gene. *Am J Hum Genet* **108**, 1964-1980 (2021).
12. He, X. *et al.* Integrated model of de novo and inherited genetic variants yields greater power to identify risk genes. *PLoS Genet* **9**, e1003671 (2013).
13. Jin, S.C. *et al.* Contribution of rare inherited and de novo variants in 2,871 congenital heart disease probands. *Nat Genet* **49**, 1593-1601 (2017).
14. Iossifov, I. *et al.* The contribution of de novo coding mutations to autism spectrum disorder. *Nature* **515**, 216-21 (2014).
15. Zhou, X. *et al.* Integrating <em>de novo</em> and inherited variants in over 42,607 autism cases identifies mutations in new moderate risk genes. *medRxiv*, 2021.10.08.21264256 (2021).
16. Zuk, O. *et al.* Searching for missing heritability: designing rare variant association studies. *Proc Natl Acad Sci U S A* **111**, E455-64 (2014).
17. Zhang, C. & Shen, Y. A Cell Type-Specific Expression Signature Predicts Haploinsufficient Autism-Susceptibility Genes. *Hum Mutat* **38**, 204-215 (2017).
18. Chen, S. *et al.* Dissecting Autism Genetic Risk Using Single-cell RNA-seq Data. *bioRxiv*, 2020.06.15.153031 (2020).
19. Rozenblatt-Rosen, O., Stubbington, M.J.T., Regev, A. & Teichmann, S.A. The Human Cell Atlas: from vision to reality. *Nature* **550**, 451-453 (2017).

20. Tabula Muris, C. *et al.* Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* **562**, 367-372 (2018).
21. La Manno, G. *et al.* Molecular Diversity of Midbrain Development in Mouse, Human, and Stem Cells. *Cell* **167**, 566-580 e19 (2016).
22. Zhong, S. *et al.* A single-cell RNA-seq survey of the developmental landscape of the human prefrontal cortex. *Nature* **555**, 524-528 (2018).
23. Cao, J. *et al.* A human cell atlas of fetal gene expression. *Science* **370**(2020).
24. He, P. *et al.* The changing mouse embryo transcriptome at whole tissue and single-cell resolution. *Nature* **583**, 760-767 (2020).
25. Kingma, D.P. & Welling, M. Auto-Encoding Variational Bayes. arXiv:1312.6114 (2013).
26. Kingma, D.P., Rezende, D.J., Mohamed, S. & Welling, M. Semi-Supervised Learning with Deep Generative Models. arXiv:1406.5298 (2014).
27. Abrahams, B.S. *et al.* SFARI Gene 2.0: a community-driven knowledgebase for the autism spectrum disorders (ASDs). *Mol Autism* **4**, 36 (2013).
28. pfeliciano@simonsfoundation.org, S.C.E.a. & Consortium, S. SPARK: A US Cohort of 50,000 Families to Accelerate Autism Research. *Neuron* **97**, 488-493 (2018).
29. Coe, B.P. *et al.* Neurodevelopmental disease genes implicated by de novo mutation and copy number variation morbidity. *Nat Genet* **51**, 106-116 (2019).
30. Buxbaum, J.D. *et al.* The autism sequencing consortium: large-scale, high-throughput sequencing in autism spectrum disorders. *Neuron* **76**, 1052-6 (2012).
31. RK, C.Y. *et al.* Whole genome sequencing resource identifies 18 new candidate genes for autism spectrum disorder. *Nat Neurosci* **20**, 602-611 (2017).
32. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38**, e164 (2010).
33. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol* **17**, 122 (2016).
34. Ioannidis, N.M. *et al.* REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *Am J Hum Genet* **99**, 877-885 (2016).
35. Samocha, K.E. *et al.* A framework for the interpretation of de novo mutation in human disease. *Nat Genet* **46**, 944-50 (2014).
36. Nguyen, H.T. *et al.* Integrated Bayesian analysis of rare exonic variants to identify risk genes for schizophrenia and neurodevelopmental disorders. *Genome Med* **9**, 114 (2017).
37. Kingma, D.P. & Ba, J. Adam: A Method for Stochastic Optimization. arXiv:1412.6980 (2014).
38. Hinton, R.B. *et al.* Cardiovascular manifestations of tuberous sclerosis complex and summary of the revised diagnostic criteria and surveillance and management recommendations from the International Tuberous Sclerosis Consensus Group. *J Am Heart Assoc* **3**, e001493 (2014).
39. Kaplanis, J. *et al.* Evidence for 28 genetic disorders discovered by combining healthcare and research data. *Nature* **586**, 757-762 (2020).
40. Willsey, A.J. *et al.* Coexpression networks implicate human midfetal deep cortical projection neurons in the pathogenesis of autism. *Cell* **155**, 997-1007 (2013).
41. Reiff, R.E. *et al.* METTL23, a transcriptional partner of GABPA, is essential for human cognition. *Hum Mol Genet* **23**, 3456-66 (2014).

42. Bernkopf, M. *et al.* Disruption of the methyltransferase-like 23 gene METTL23 causes mild autosomal recessive intellectual disability. *Hum Mol Genet* **23**, 4015-23 (2014).
43. Crider, A., Ahmed, A.O. & Pillai, A. Altered Expression of Endoplasmic Reticulum Stress-Related Genes in the Middle Frontal Cortex of Subjects with Autism Spectrum Disorder. *Mol Neuropsychiatry* **3**, 85-91 (2017).
44. Takahashi, M. *et al.* Hyperactive and impulsive behaviors of LMTK1 knockout mice. *Sci Rep* **10**, 15461 (2020).
45. Inoue, T. *et al.* LMTK3 deficiency causes pronounced locomotor hyperactivity and impairs endocytic trafficking. *J Neurosci* **34**, 5927-37 (2014).
46. Coghlan, S. *et al.* GABA system dysfunction in autism and related disorders: from synapse to symptoms. *Neurosci Biobehav Rev* **36**, 2044-55 (2012).
47. Rojas, D.C., Singel, D., Steinmetz, S., Hepburn, S. & Brown, M.S. Decreased left perisylvian GABA concentration in children with autism and unaffected siblings. *Neuroimage* **86**, 28-34 (2014).
48. Puts, N.A.J. *et al.* Reduced GABA and altered somatosensory function in children with autism spectrum disorder. *Autism Res* **10**, 608-619 (2017).
49. Chao, O.Y. *et al.* Altered dopaminergic pathways and therapeutic effects of intranasal dopamine in two distinct mouse models of autism. *Mol Brain* **13**, 111 (2020).
50. Kosillo, P. & Bateup, H.S. Dopaminergic Dysregulation in Syndromic Autism Spectrum Disorders: Insights From Genetic Mouse Models. *Front Neural Circuits* **15**, 700968 (2021).
51. Ignatiadis, N., Klaus, B., Zaugg, J.B. & Huber, W. Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nat Methods* (2016).
52. Zhang, M.J., Xia, F. & Zou, J. Fast and covariate-adaptive method amplifies detection power in large-scale multiple hypothesis testing. *Nat Commun* **10**, 3433 (2019).
53. Yurko, R., G'Sell, M., Roeder, K. & Devlin, B. A selective inference approach for false discovery rate control using multiomics covariates yields insights into disease risk. *Proc Natl Acad Sci U S A* **117**, 15028-15035 (2020).
54. La Manno, G. *et al.* RNA velocity of single cells. *Nature* **560**, 494-498 (2018).
55. Bergen, V., Lange, M., Peidli, S., Wolf, F.A. & Theis, F.J. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nature Biotechnology* **38**(2020).

Figures

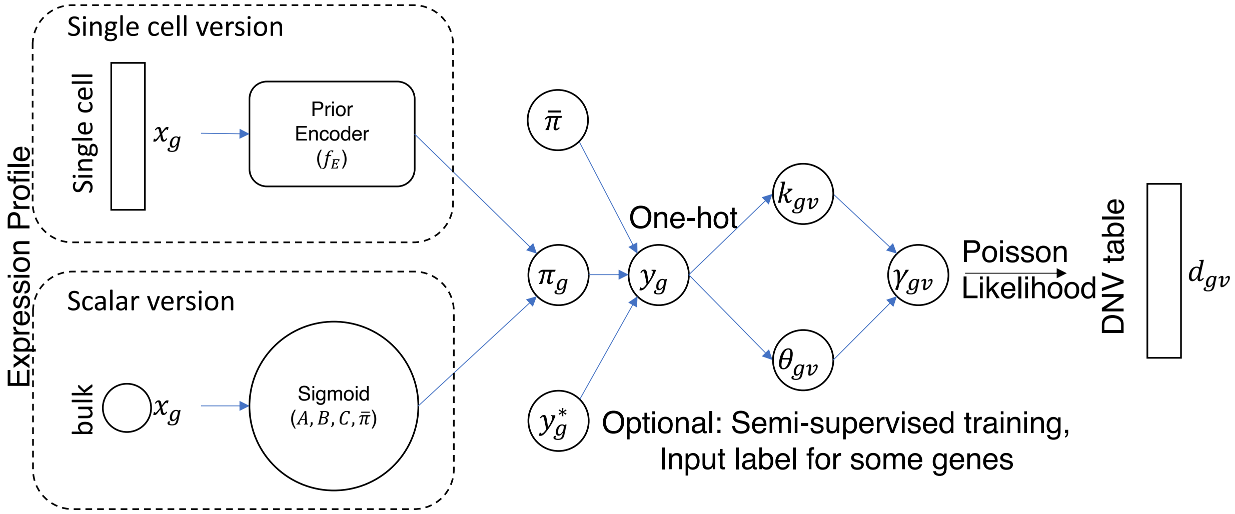


Figure 1. Model architecture. The architecture of VBASS. The input could be either a vector of single cell expression profile or a scalar of bulk expression profile. For the vector input, f_E is a neural network to inference the gene specific parameter π_g, while for the scalar input, it could be simplified to a sigmoid function with four parameters, $A, B, C, \bar{\pi}$. π_g will parameterize the distributions of y_g. This distribution will be penalized by a Bernoulli prior via KL penalty term. This model also takes predefined labels as input, where y_g is given by one-hot encoding of the labels and the Bernoulli KL penalty is replaced with a cross-entropy loss on the real label. k_{gv}, θ_{gv} are two random variables conditioned on y_g that reconstruct the parameters of Gamma-Poisson distribution for d_{gv}.

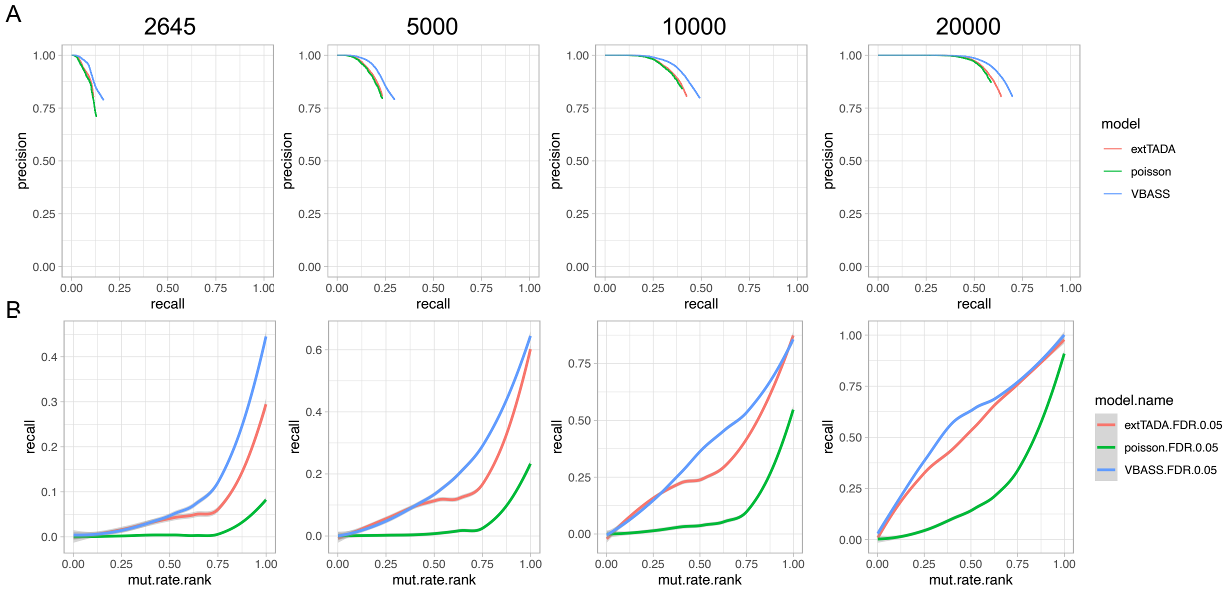


Figure 2. Performance comparison of scalar version VBASS and extTADA on simulation. A) Precision-recall in two models, only show the part with FDR ≤ 0.2 for extTADA and VBASS, only show the part with FDR ≤ 0.01 for Poisson test. B) Comparison of recall (y-axis) for genes sets with different mutation rates (x-axis)

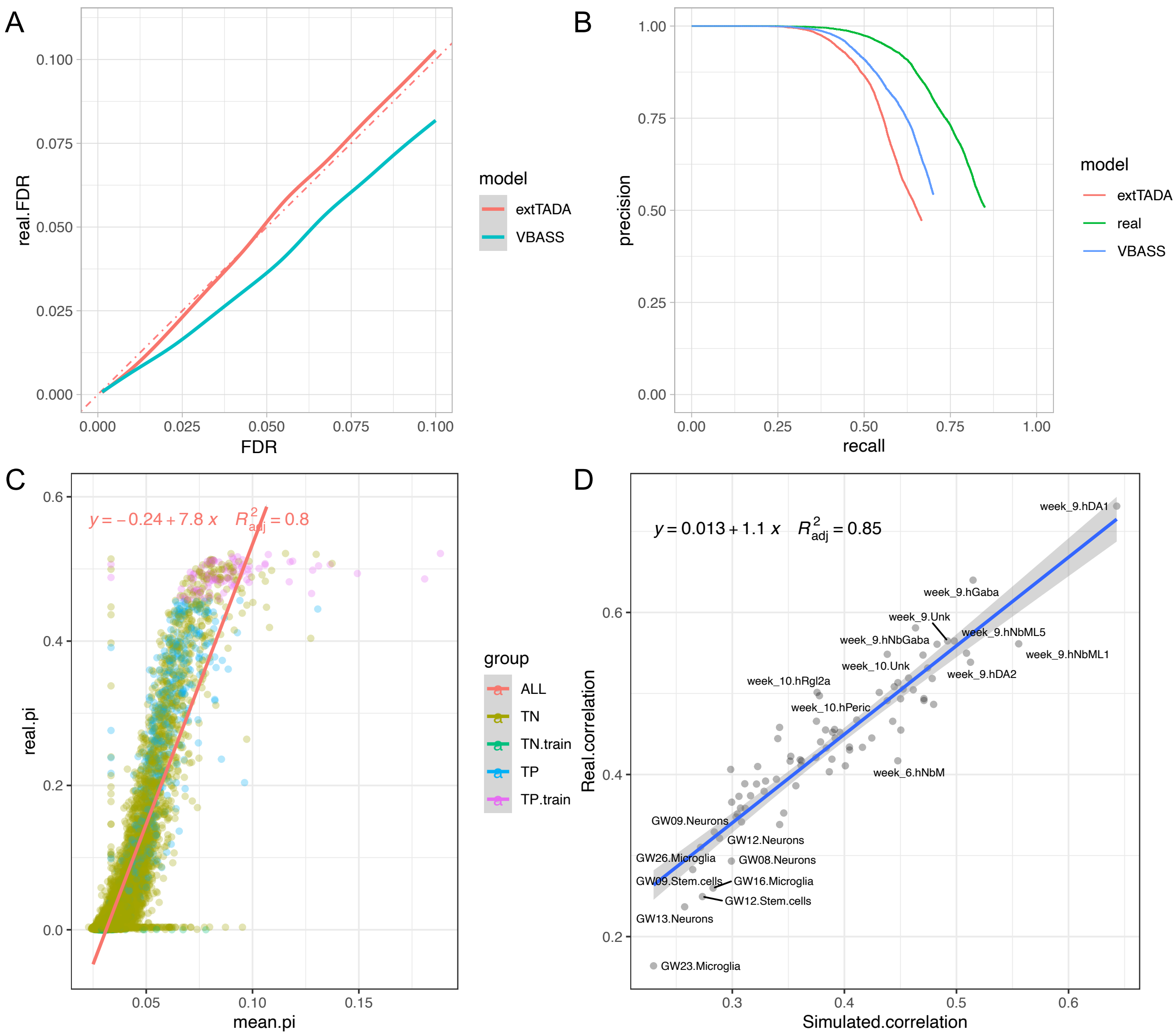


Figure 3. Performance of single cell version VBASS on simulation data. A) Plot of true false discovery rate (real.FDR, y axis) at different FDR cutoff (x axis) estimated by extTADA and VBASS. B) Comparison of precision recall for extTADA and VBASS, only shown for the part with FDR ≤ 0.5. C) Scatter plot of disease risk prior (π) that we assigned in simulation (y-axis) and informed by VBASS (x-axis). Genes were colored by labels and whether used in semi-supervised training, where TN and TP correspond to true negative and true positive, respectively. D) Comparison of correlation between real disease risk prior and cell type expression (y-axis) versus correlation between VBASS informed prior and cell type expression (x-axis). Each dot represents a cell type.

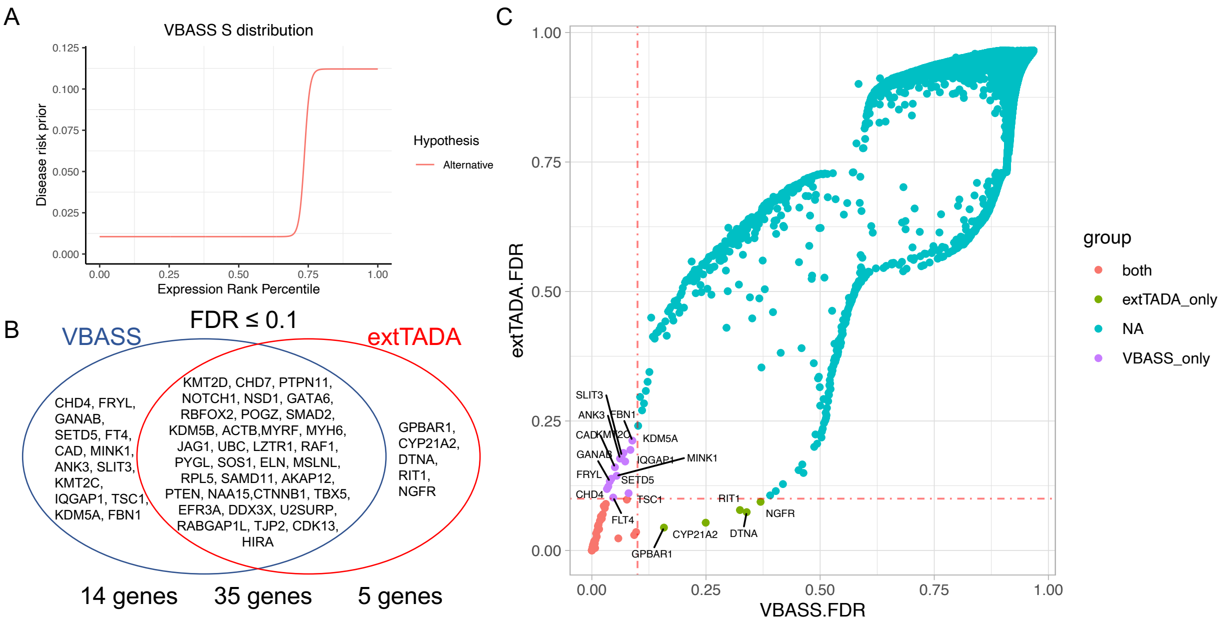


Figure 4. Performance comparison of VBASS and extTADA on CHD data. A). Function of disease risk prior on expression rank percentile estimated by VBASS. B). Genes identified by VBASS and extTADA at significance level 0.1. C). FDR of genes in extTADA (y-axis) and VBASS (x-axis), genes were colored by significance in both models (red), only in VBASS (purple) or only in extTADA (green) at significance level 0.1 (FDR ≤ 0.1).

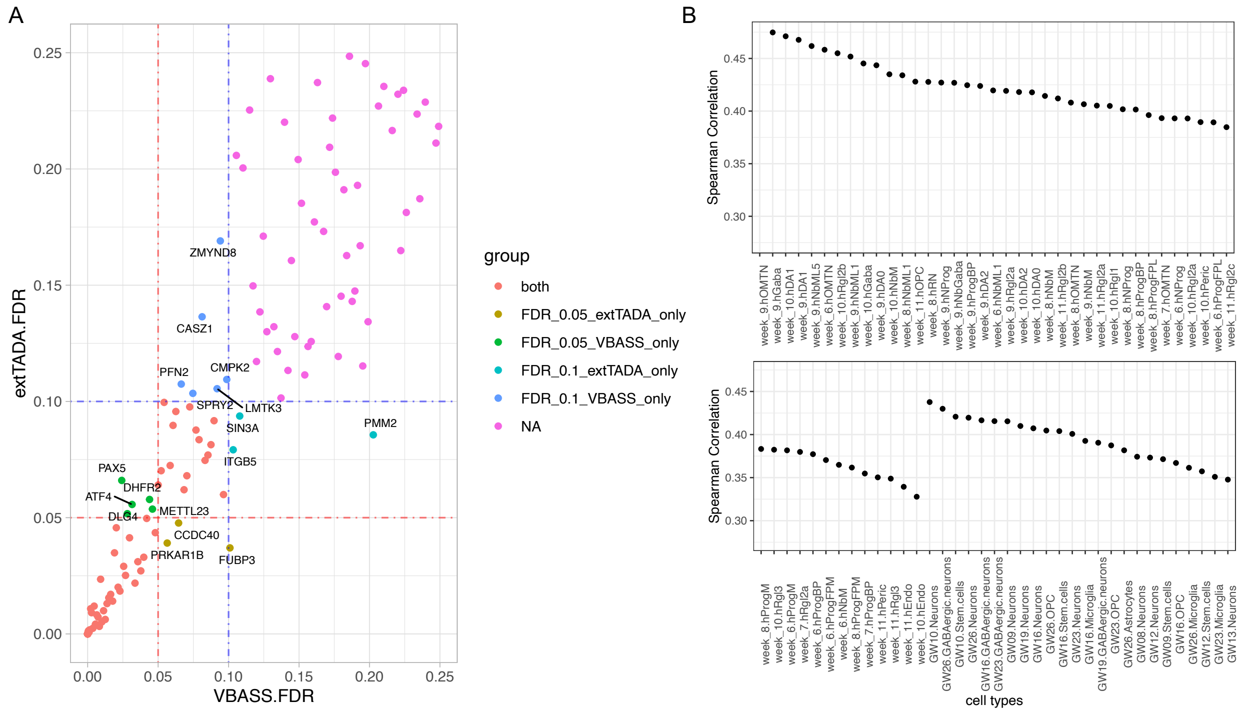


Figure 5. Performance comparison of VBASS and extTADA on ASD data. A). FDR of genes in extTADA (y-axis) or VBASS (x-axis). B). Spearman correlation between cell type expression and disease risk prior (π). The cell types from two single cell data sets were separated and ordered by correlation with π respectively.

Tables

Table 1. Estimated VBASS parameters in CHD data. Mean, posterior mean; sd, standard error; 2.5% and 97.5%, confidence interval; n_eff, effective sample number in MCMC; Rhat, convergence diagnostic in MCMC.

	mean	sd	2.50%	97.50%	n_eff	Rhat
$\boldsymbol{\pi_0}$	0.04	0.01	0.03	0.05	1845.42	1.00
$\boldsymbol{A}$	104.15	87.67	20.01	351.78	2263.73	1.00
$\boldsymbol{B}$	0.74	0.02	0.71	0.78	2136.99	1.00
$\boldsymbol{C}$	0.28	0.10	0.09	0.47	2093.73	1.00
$\boldsymbol{\bar{\gamma}_{LGD}}$	19.95	5.43	10.32	31.87	2745.30	1.00
$\boldsymbol{\bar{\gamma}_{Dmis}}$	11.79	3.60	5.81	19.36	3013.96	1.00
$\boldsymbol{\bar{\beta}_{LGD}}$	0.84	0.02	0.82	0.89	2193.48	1.00
$\boldsymbol{\bar{\beta}_{Dmis}}$	0.90	0.07	0.83	1.07	2144.47	1.00

Table 2. Genes identified by VBASS but not extTADA. dn_LGD, de novo LGD variants; dn_Dmis, de novo Dmis variants.

Gene Symbol	dn_LGD	dn_Dmis	VBASS FDR	Expression Rank	extTADA FDR
CHD4	0	3	0.033	0.990	0.119
FRYL	2	0	0.036	0.837	0.123
GANAB	1	1	0.039	0.934	0.133
SETD5	1	1	0.043	0.949	0.138
FLT4	2	0	0.047	0.734	0.102
CAD	0	3	0.051	0.853	0.161
MINK1	0	2	0.054	0.875	0.144
ANK3	2	0	0.062	0.948	0.177
SLIT3	1	1	0.066	0.860	0.183
KMT2C	1	2	0.069	0.792	0.188
IQGAP1	0	2	0.073	0.856	0.172
TSC1	1	1	0.080	0.728	0.110
KDM5A	0	2	0.084	0.859	0.194
FBN1	0	3	0.089	0.928	0.212

Supplemental materials

Supplementary Table 1. Labels of genes for VBASS in semi-supervised training.

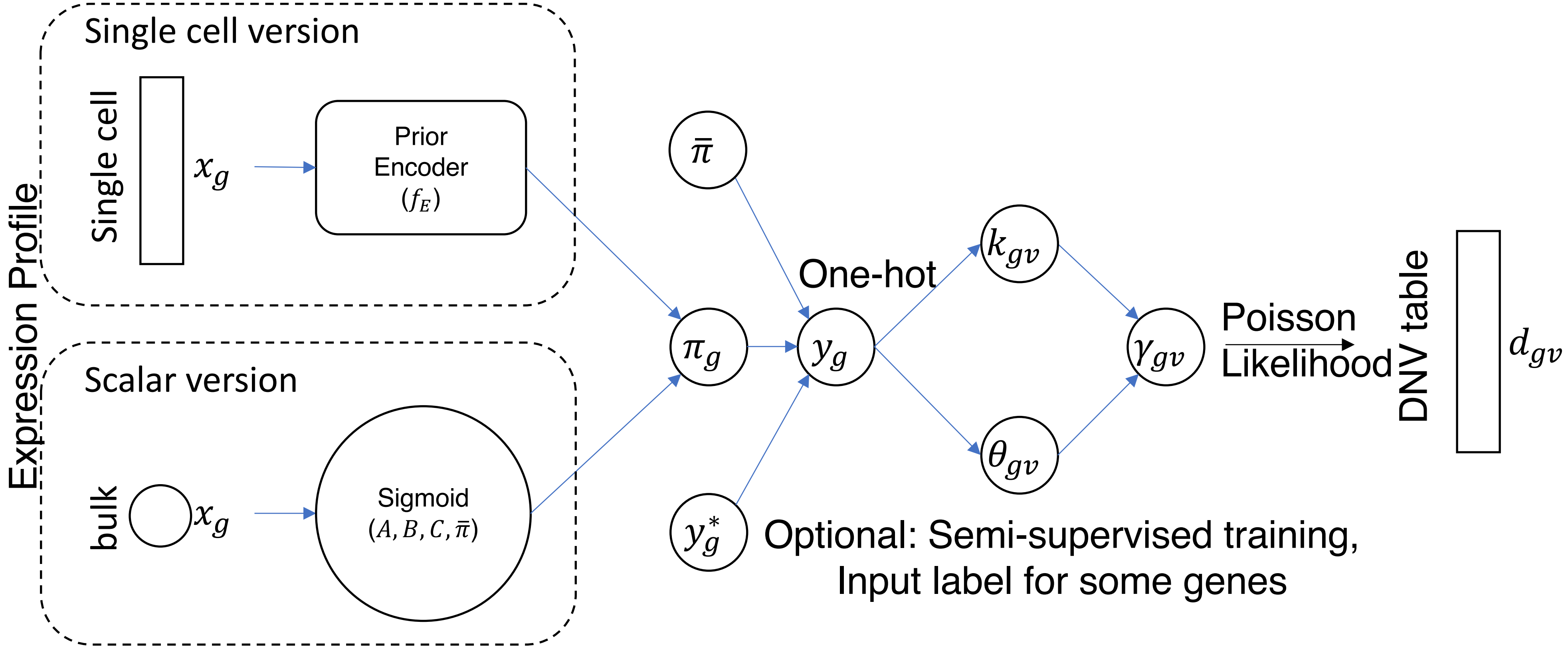
Supplementary Table 2. De novo variants of 2645 CHD trios in Jin et al 2017.

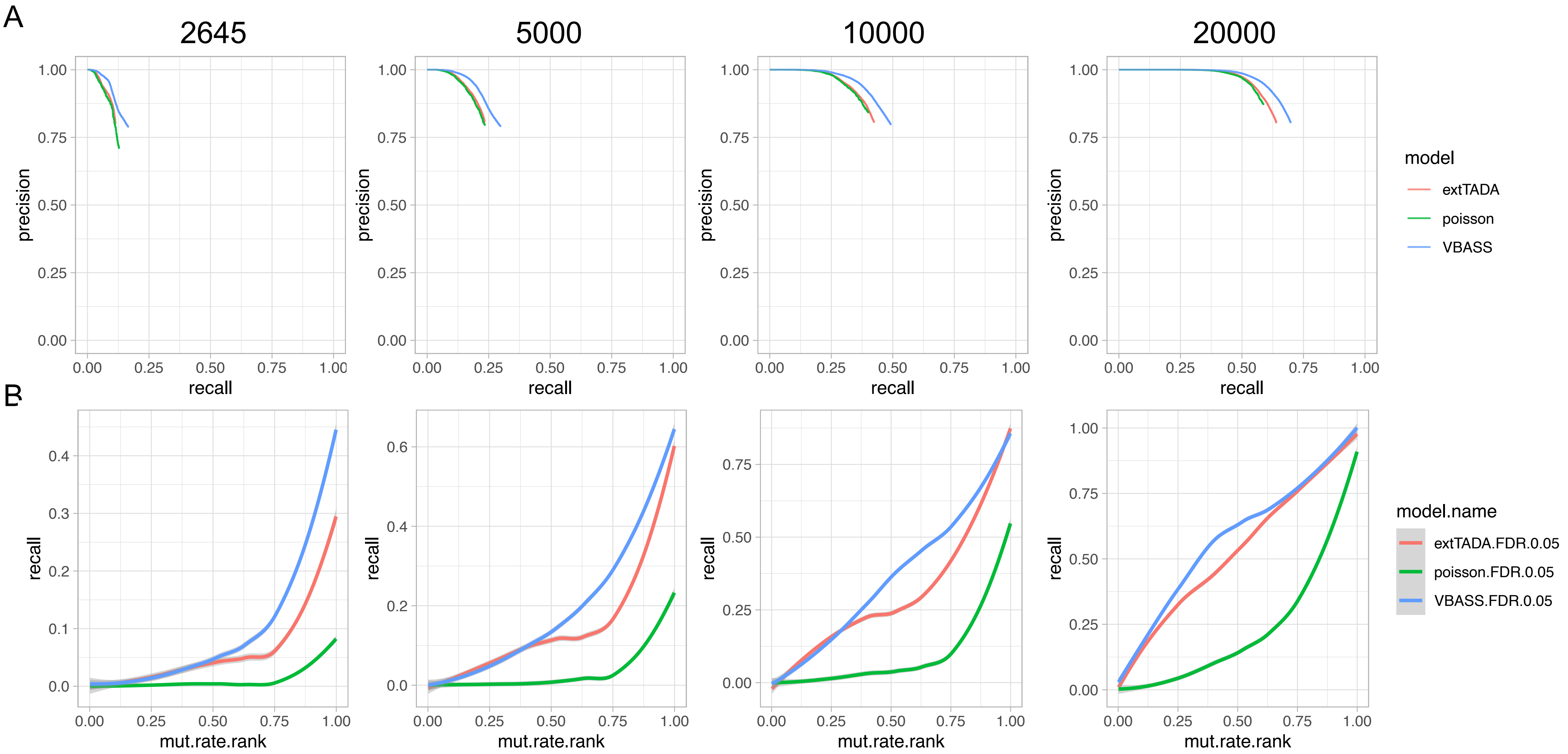
Supplementary Table 3. De novo variants of 16616 ASD trios in Zhou et al 2021.

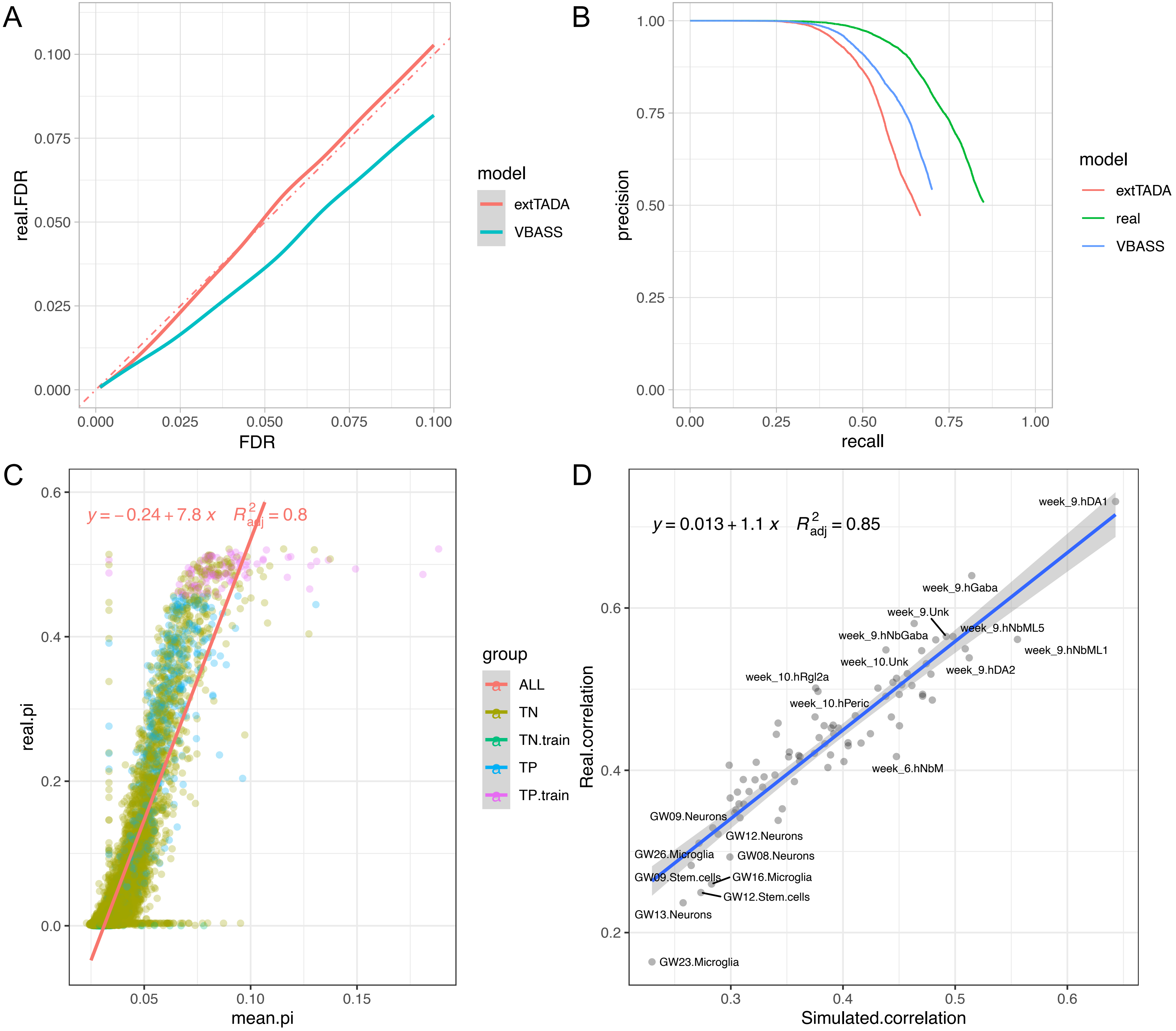
Supplementary Table 4. Posterior probabilities of all genes calculated in CHD cohort by VBASS and extTADA.

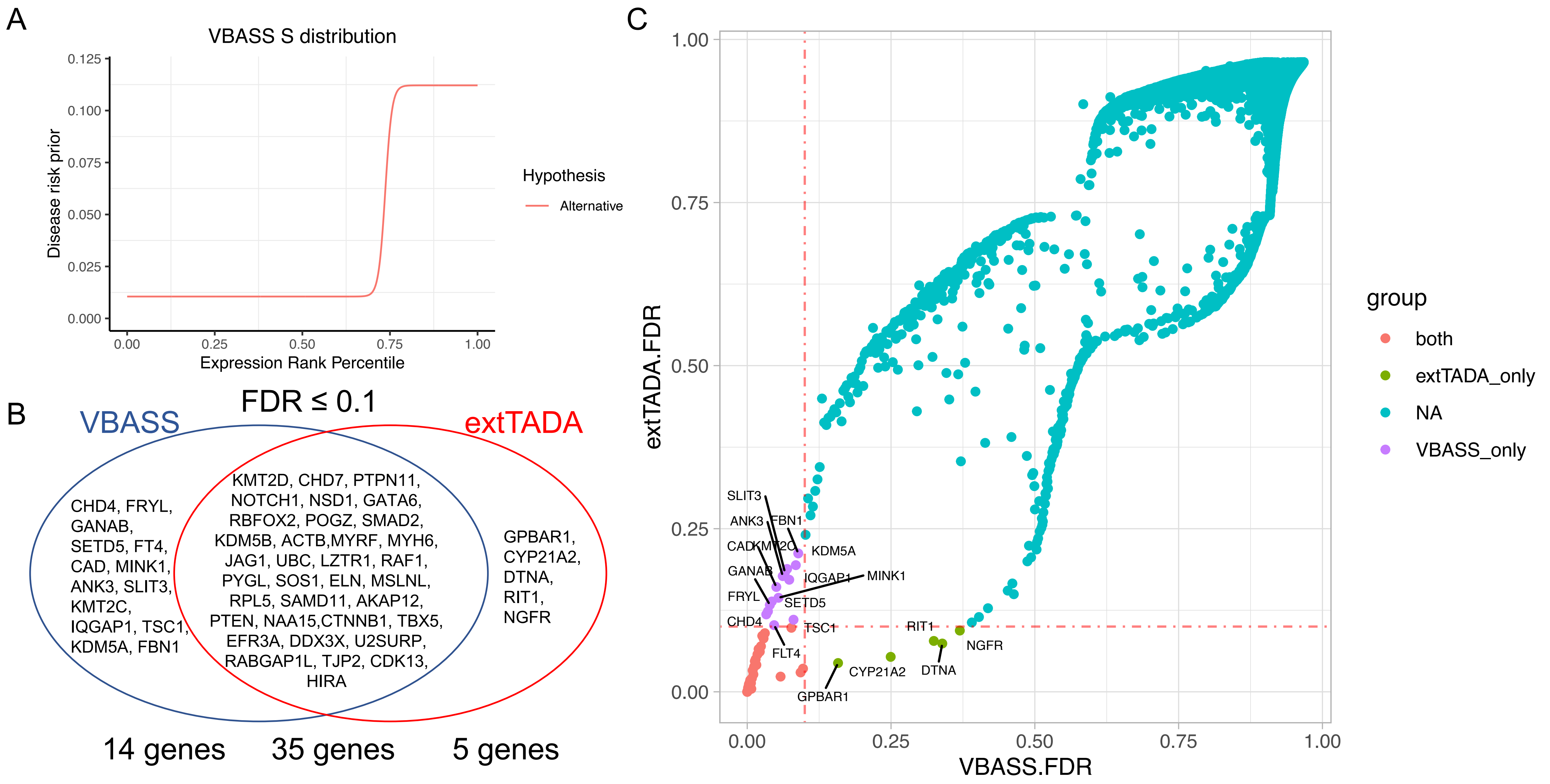
Supplementary Table 5. Posterior probabilities of all genes calculated in ASD cohort by VBASS and extTADA.

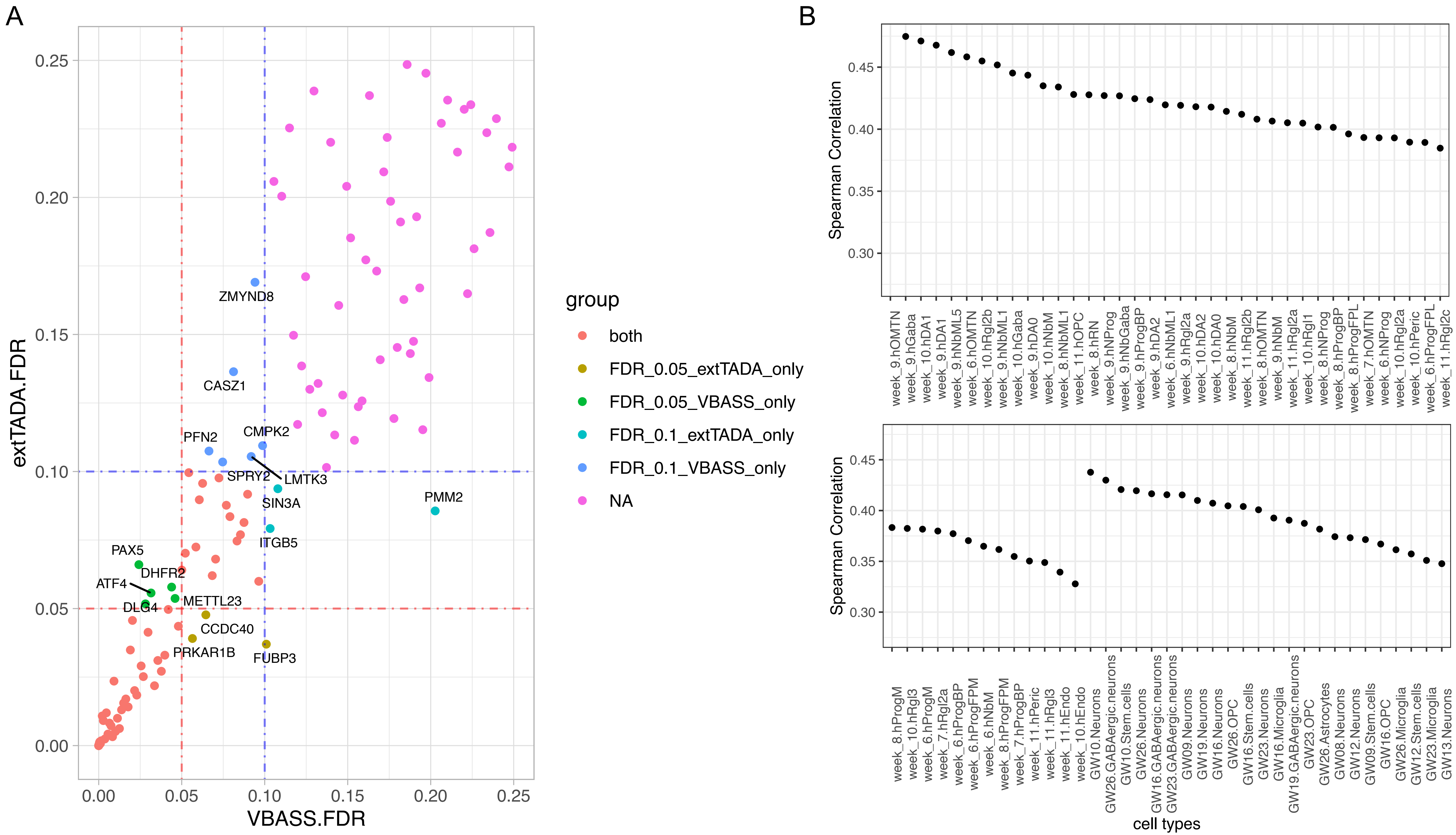
Supplementary Table 6. Posterior probabilities of all genes calculated in ASD cohort by VBASS and extTADA. Removed positive training genes when calculating FDR.


Expression Profile
Single cell version
Single cell
x_g
Prior Encoder (f_E)
Scalar version
bulk
x_g
Sigmoid ($A, B, C, \bar{\pi}$)
$\bar{\pi}$
π_g
One-hot
y_g
y_g^*
k_{gv}
θ_{gv}
γ_{gv}
Poisson Likelihood
DNV table
d_{gv}
Optional: Semi-supervised training, Input label for some genes


A
2645
5000
10000
20000
precision
recall
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
model
extTADA
poisson
VBASS
B
recall
mut.rate.rank
0.0
0.1
0.2
0.3
0.4
0.6
0.00
0.25
0.50
0.75
1.00
model.name
extTADA.FDR.0.05
poisson.FDR.0.05
VBASS.FDR.0.05


A
real.FDR
0.000
0.025
0.050
0.075
0.100
FDR
0.000
0.025
0.050
0.075
0.100
model
extTADA
VBASS
B
precision
0.00
0.25
0.50
0.75
1.00
recall
0.00
0.25
0.50
0.75
1.00
model
extTADA
real
VBASS
C
$y = -0.24 + 7.8\,x \quad R^2_{adj} = 0.8$
real.pi
0.0
0.2
0.4
0.6
mean.pi
0.05
0.10
0.15
group
ALL
TN
TN.train
TP
TP.train
D
$y = 0.013 + 1.1\,x \quad R^2_{adj} = 0.85$
Real.correlation
0.2
0.4
0.6
Simulated.correlation
0.3
0.4
0.5
0.6
week_9.hDA1
week_9.hGaba
week_9.Unk
week_9.hNbML5
week_9.hNbGaba
week_9.hNbML1
week_10.Unk
week_9.hDA2
week_10.hRgl2a
week_10.hPeric
week_6.hNbM
GW09.Neurons
GW12.Neurons
GW26.Microglia
GW08.Neurons
GW09.Stem.cells
GW16.Microglia
GW12.Stem.cells
GW13.Neurons
GW23.Microglia


A
VBASS S distribution
Disease risk prior
0.125
0.100
0.075
0.050
0.025
0.000
0.00
0.25
0.50
0.75
1.00
Expression Rank Percentile
Hypothesis
Alternative
B
FDR ≤ 0.1
VBASS
extTADA
CHD4, FRYL, GANAB, SETD5, FT4, CAD, MINK1, ANK3, SLIT3, KMT2C, IQGAP1, TSC1, KDM5A, FBN1
KMT2D, CHD7, PTPN11, NOTCH1, NSD1, GATA6, RBFOX2, POGZ, SMAD2, KDM5B, ACTB,MYRF, MYH6, JAG1, UBC, LZTR1, RAF1, PYGL, SOS1, ELN, MSLNL, RPL5, SAMD11, AKAP12, PTEN, NAA15,CTNNB1, TBX5, EFR3A, DDX3X, U2SURP, RABGAP1L, TJP2, CDK13, HIRA
GPBAR1, CYP21A2, DTNA, RIT1, NGFR
14 genes
35 genes
5 genes
C
extTADA.FDR
1.00
0.75
0.50
0.25
0.00
VBASS.FDR
0.00
0.25
0.50
0.75
1.00
SLIT3
ANK3
FBN1
CADKMT2C
KDM5A
GANAB
IQGAP1
MINK1
FRYL
SETD5
CHD4
TSC1
RIT1
NGFR
FLT4
CYP21A2
DTNA
GPBAR1
group
both
extTADA_only
NA
VBASS_only


A
extTADA.FDR
VBASS.FDR
0.00
0.05
0.10
0.15
0.20
0.25
ZMYND8
CASZ1
PFN2
CMPK2
SPRY2
LMTK3
SIN3A
PMM2
ITGB5
PAX5
DHFR2
ATF4
DLG4
METTL23
CCDC40
PRKAR1B
FUBP3
group
both
FDR_0.05_extTADA_only
FDR_0.05_VBASS_only
FDR_0.1_extTADA_only
FDR_0.1_VBASS_only
NA
B
Spearman Correlation
0.30
0.35
0.40
0.45
week_9.hOMTN
week_9.hGaba
week_10.hDA1
week_9.hDA1
week_9.hNbML5
week_6.hOMTN
week_10.hRgl2b
week_9.hNbML1
week_10.hGaba
week_9.hDA0
week_10.hNbM
week_8.hNbML1
week_11.hOPC
week_8.hRN
week_9.hNProg
week_9.hNbGaba
week_9.hProgBP
week_9.hDA2
week_6.hNbML1
week_9.hRgl2a
week_10.hDA2
week_10.hDA0
week_8.hNbM
week_11.hRgl2b
week_8.hOMTN
week_9.hNbM
week_11.hRgl2a
week_10.hRgl1
week_8.hNProg
week_8.hProgBP
week_8.hProgFPL
week_7.hOMTN
week_6.hNProg
week_10.hRgl2a
week_10.hPeric
week_6.hProgFPL
week_11.hRgl2c
Spearman Correlation
0.30
0.35
0.40
0.45
week_8.hProgM
week_10.hRgl3
week_6.hProgM
week_7.hRgl2a
week_6.hProgBP
week_6.hProgFPM
week_6.hNbM
week_8.hProgFPM
week_7.hProgBP
week_11.hPeric
week_11.hRgl3
week_11.hEndo
week_10.hEndo
GW10.Neurons
GW26.GABAergic.neurons
GW10.Stem.cells
GW26.Neurons
GW16.GABAergic.neurons
GW23.GABAergic.neurons
GW09.Neurons
GW19.Neurons
GW16.Neurons
GW26.OPC
GW16.Stem.cells
GW23.Neurons
GW16.Microglia
GW19.GABAergic.neurons
GW23.OPC
GW26.Astrocytes
GW08.Neurons
GW12.Neurons
GW09.Stem.cells
GW16.OPC
GW26.Microglia
GW12.Stem.cells
GW23.Microglia
GW13.Neurons
cell types