RESEARCH

Open Access

A Bayesian framework to integrate multi-level genome-scale data for Autism risk gene prioritization

Ying Ji[1,2], Rui Chen[1,2], Quan Wang[1,2], Qiang Wei[1,2], Ran Tao[2,3*†] and Bingshan Li[1,2*†]

*Correspondence: r.tao@vumc.org; bingshan.li@vanderbilt.edu

†Ran Tao and Bingshan Li jointly supervised the study, and are corresponding authors of this study.

[1] Department of Molecular Physiology and Biophysics, Vanderbilt University, Nashville, TN 37212, USA

[2] Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, TN 37212, USA

Full list of author information is available at the end of the article

Abstract

Background: Autism spectrum disorder (ASD) is a group of complex neurodevelopment disorders with a strong genetic basis. Large scale sequencing studies have identified over one hundred ASD risk genes. Nevertheless, the vast majority of ASD risk genes remain to be discovered, as it is estimated that more than 1000 genes are likely to be involved in ASD risk. Prioritization of risk genes is an effective strategy to increase the power of identifying novel risk genes in genetics studies of ASD. As ASD risk genes are likely to exhibit distinct properties from multiple angles, we reason that integrating multiple levels of genomic data is a powerful approach to pinpoint genuine ASD risk genes.

Results: We present BNScore, a Bayesian model selection framework to probabilistically prioritize ASD risk genes through explicitly integrating evidence from sequencing-identified ASD genes, biological annotations, and gene functional network. We demonstrate the validity of our approach and its improved performance over existing methods by examining the resulting top candidate ASD risk genes against sets of high-confidence benchmark genes and large-scale ASD genome-wide association studies. We assess the tissue-, cell type- and development stage-specific expression properties of top prioritized genes, and find strong expression specificity in brain tissues, striatal medium spiny neurons, and fetal developmental stages.

Conclusions: In summary, we show that by integrating sequencing findings, functional annotation profiles, and gene-gene functional network, our proposed BNScore provides competitive performance compared to current state-of-the-art methods in prioritizing ASD genes. Our method offers a general and flexible strategy to risk gene prioritization that can potentially be applied to other complex traits as well.

Keywords: Gene prioritization, Bayesian model selection, ASD risk genes

Background

Genetics plays an important role in the etiology of Autism spectrum disorder (ASD). Dozens of ASD risk genes have been identified from whole exome sequencing (WES) studies (e.g., de novo and inherited mutations) [1–3], but the vast majority of ASD risk genes remains unknown, as it has been estimated that more than 1000 genes are involved in risk of ASD [4]. It is challenging to identify ASD risk genes due to its broad spectrum of genetic architectures with multiple biological processes involved [5]. WES or whole genome-sequencing (WGS) studies of parent-offspring trios has been successful in identifying ASD risk genes via de novo mutations. However, the power of such approaches is inevitably low given the rarity of de novo mutations. Computational approaches provide cost-effective alternatives to effectively nominate ASD risk genes [6].

Various computational methods have been developed for ASD risk gene identification, most of which employ diverse biological evidence to prioritize risk genes [4, 7, 8]. Conceptually, a candidate gene is often scored higher if the gene is similar or closer to known ASD risk genes (referred to as "seed genes" hereafter) based on the biological evidence. In practice, those approaches start by bringing together a "gold-standard" seed gene set, then train a statistical model using seed genes along with their relevant biological evidence, and finally rank all genes based on the predicted scores from the trained model. Two key ingredients to facilitate successful modeling are (1) a robust seed gene set comprising true ASD risk genes and (2) relevant biological evidence that are representative of these genes.

The biological evidence for prioritization mainly pertains to genetic sequence properties, functional annotation, and network information [9, 10]. A few recent studies are based on one or two types of the aforementioned evidence. He et al. [4] prioritized genes through genetic sequence properties alone (e.g., multiple occurrences of mutations in unrelated patients) and didn't consider other evidence that are important to ASD risk. Functional annotation based methods [11, 12] rank genes according to the similarity between candidate and seed genes' annotation profiles. They tend to bias towards well-annotated genes and are less effective for genome-wide prediction [13]. Network-based approaches [7, 8, 14, 15] rely on network proximity between candidate and seed genes, thus are less biased to well-annotated genes. Instead, the results may bias towards highly connected genes. Recently, an increasing proportion of network-based approaches are framing the prioritization problem as a classification problem (i.e., classify genes into ASD risk genes versus non-risk genes) to be solved by machine-learning algorithms [7, 8]. Apart from traditional machine learning approaches, deep learning has also been rapidly gaining popularity. Graph neural network (GNN) can directly analyze data structured as graphs, such as biological networks. Zhang et al. [15] recent applied a GNN classifier to prioritize ASD genes using the human molecular interaction network input for training and reported to outperform other commonly used machine learning algorithms. Apart from efforts based on network only, Lin et al. [16] integrated network evidence with other biological evidence using machine learning classifiers for risk gene prioritization. These machine-learning based methods have discovered novel ASD genes, but their inherent "black box" nature limits the interpretability of the final results.

Herein, we present BNScore, a novel Bayesian model selection approach for ASD risk gene prioritization through explicitly integrating three major types of biological evidence: (1) seed genes derived from ASD sequencing studies; (2) multiple lines of gene-level functional annotations; and (3) distance to known ASD risk genes in a biological network. The framework is flexible in that it can readily include additional features relevant to ASD to further increase prediction accuracy. In addition, the Bayesian set-up renders a clear interpretation of the prediction scores (i.e., Bayesian posterior odds of being a risk gene versus not) for each gene across the genome. We demonstrate the validity of our approach and its improved performance over existing methods by examining the resulting top candidate ASD risk genes against sets of high-confidence benchmark genes and large scale ASD genome-wide association studies (GWAS). We study the brain spatiotemporal gene expression specificity of identified top candidate genes to implicate tissues, cell types, and development stages in the etiology of ASD.

Methods

We frame the task of finding risk genes as a Bayesian model selection problem. For each gene, we select between models M_0, not a ASD risk gene and M_1, a risk gene. Let $\boldsymbol{\theta_j}$ denote the parameters associated with model M_j $(j = 1, 0)$ and let $\boldsymbol{D} = (D_1, D_2, ..., D_p)^{\mathrm{T}}$ denote the p-dimensional functional annotation data for each gene. We formulate the posterior odds of a gene being a risk gene as

$$\begin{aligned}\frac{P(M_1|\boldsymbol{D})}{P(M_0|\boldsymbol{D})} &= \frac{P(M_1)P(\boldsymbol{D}|M_1)}{P(M_0)P(\boldsymbol{D}|M_0)}\\ &= \frac{P(M_1)}{P(M_0)}\frac{\int p(\boldsymbol{D}|\boldsymbol{\theta_1}, M_1)p(\boldsymbol{\theta_1}|M_1)d\boldsymbol{\theta_1}}{\int p(\boldsymbol{D}|\boldsymbol{\theta_0}, M_0)p(\boldsymbol{\theta_0}|M_0)d\boldsymbol{\theta_0}},\end{aligned}$$

where $\frac{P(M_1)}{P(M_0)}$ and $\frac{P(\boldsymbol{D}|M_1)}{P(\boldsymbol{D}|M_0)}$ are the prior odds and Bayes factor of being a risk gene, respectively. For each gene, we specify $\frac{P(M_1)}{P(M_0)}$ based on its average distance to seed genes in a gene-gene functional network denoted by N_S, based on the rationale that disease-associated genes are assumed to be closer to each other than random pairs in the network [5, 8, 11]. We assume that the annotations in $\boldsymbol{D}$ are independent of each other under both M_0 and M_1 and compute gene-level Bayes factor $\frac{P(\boldsymbol{D}|M_1)}{P(\boldsymbol{D}|M_0)}$ by taking the product of the Bayes factors from individual annotations, i.e., $\frac{P(\boldsymbol{D}|M_1)}{P(\boldsymbol{D}|M_0)} = \prod_{l=1}^{p} \frac{P(D_l|M_1)}{P(D_l|M_0)}$.

In our framework, seed genes and "background" genes are essential to derive Bayes factors and prior odds. We use 65 genes identified from a large exome sequencing study as seed genes [3]. We randomly select 500 genes across the genome (excluding the 65 seed genes) and regard them as background genes. We acknowledge that these background genes are not strictly non-ASD genes. However, there is no gold standard for non-ASD risk genes and it is reasonable to assume that the vast majority of genes across the genome are true non-ASD genes. There may be a few true positive genes ended up in our background gene set, in which situation the resulting inference could be slightly conservative. Nevertheless, we believe that this strategy is more robust than using an unreliable "gold-standard" background gene set. We choose 500 background genes by trial and error to balance between two considerations: a smaller set may be insufficiently representative of the genome background and a larger set can be too heterogeneous to be useful [17].

Prior odds of a gene being an ASD risk gene

We assume that the prior probability of a gene being a risk gene is determined by two factors: (1) the overall fraction of risk genes in the genome and (2) the average distance of this gene to seed genes in a gene-gene functional network. The rationale is that the closer the two genes are in the network, the higher chance that they have similar functions. We generate the prior odds from

$$\frac{P(M_1)}{P(M_0)} = \frac{1000}{18000 - 1000} P(N_s),$$

where $\frac{1000}{18{,}000-1000}$ is the assumed constant ratio of risk versus non-risk genes in the genome, as it's estimated that around 1000 genes in 18000 genes are ASD risk genes [4]. $P(N_s)$ is the average distance between the current gene and all seed genes, which is calculated based on Gene Ontology (GO) [18, 19]. Specifically, we first build a network connecting all pairs of genes based on the number and strength of GO terms shared by each gene pair. In this network, the distance between any gene pair is proportional to the log likelihood ratio of the two genes sharing the same GO annotations versus not [14].

We then construct a transition matrix from the network and apply the random walk with restart algorithm [20] to derive the reaching probabilities between any pair of genes.

Finally, we calculate $P(N_s)$ as the average reaching probabilities of the current gene to all seed genes (see Additional file 1: Section A of Supplementary Notes for details).

Bayes factor of a gene being an ASD risk gene

To reflect each gene's strength of ASD association from a collection of functional annotations, we first identify ASD related biological processes and then summarize the ensemble evidence in a Bayes factor. We consider two forms of functional annotations: (1) binary annotation: presence/absence in biological processes previously implicated in ASD (e.g., genes involved in the developmental processes, which have been reported to be important in the pathogenesis of ASD [21]); and (2) continuous annotation: gene-level metrics (e.g., probability of being loss-of-function (LoF) intolerant (pLI) score [22]). We model binary and continuous annotations using Beta-Bernoulli and Normal-Inverse Gamma distributions, respectively [23]. We specify the prior distributions $p(\boldsymbol{\theta_1}|M_1)$ and $p(\boldsymbol{\theta_0}|M_0)$ via parametric Empirical Bayes approaches using seed and background genes (see Additional file 1: Section B of Supplementary Notes for details).

ASD related functional annotations

We initially collect 61 ASD related binary and continuous annotations from literature and then remove redundant or irrelevant annotations using seed and background genes. Specifically, we use Fisher's exact test to identify binary annotations enriched for seed genes, and we use t-test to identify continuous annotations with significant differences between seed and background genes. The selected annotations consist of (1) biological processes implicated in ASD, e.g., genes encoding chromatin modifiers, (2) important regulatory targets, e.g., targets of FMRP, which is a polyribosome-associated RNA binding protein that plays important roles in synaptic function and neuronal plasticity, and (3) generic gene-level metrics, e.g., the pLI score; see Additional file 1: Table S1 for a complete list of annotations.

Computing Bayes factor based on a binary annotation

We assume that a binary annotation D_l follows a Bernoulli distribution $Bernoulli(\theta_{lj})$, where θ_{lj} represents the fraction of disease-associated genes with this annotation under model M_j $(j = 0, 1)$. We assume the prior distribution of θ_{lj} to be $Beta(\alpha_{lj}, \beta_{lj})$, where α_{lj} and β_{lj} are hyperparameters. Suppose we have n seed genes, k of which possess this binary annotation, i.e., $\sum_{g=1}^{n} D_{lg} = k$. It can be shown that the marginal distribution of D_l under model M_j is $p(D_l|\alpha_{lj}, \beta_{lj}) = \int p(D_l|\theta_{lj})p(\theta_{lj}|\alpha_{lj}, \beta_{lj})d\theta_{lj} = \frac{B(\alpha_{lj}+k, n-k+\beta_{lj})}{B(\alpha_{lj}, \beta_{lj})}$, where $B(\cdot,\cdot)$ is the beta function. We obtain moment estimators of hyperparameters $\widetilde{\alpha}_{lj}$ and $\widetilde{\beta}_{lj}$ under M_j using the seed and background genes (see Additional file 1: Section B.1 of Supplementary Notes for details). Then, for each candidate gene (i.e., any gene not in the seed and background gene sets), we determine which model gives a better fit using the Bayes factor $\frac{p(D_l|M_1)}{p(D_l|M_0)} = \frac{p(D_l|\widetilde{\alpha}_{l1}, \widetilde{\beta}_{l1})}{p(D_l|\widetilde{\alpha}_{l0}, \widetilde{\beta}_{l0})}$.

Computing Bayes factor based on a continuous annotation

We assume that a continuous annotation D_l follows a normal distribution $N(\mu_{lj}, \theta_{lj})$, where μ_{lj} and θ_{lj} are the mean and variance, respectively, under model M_j $(j = 0, 1)$. We assume a Normal-Inverse Gamma prior for μ_{lj} and θ_{lj}, i.e., $\mu_{lj}|\theta_{lj} \sim N(\mu_{0lj}, \frac{\theta_{lj}}{\kappa_{lj}})$; $\theta_{lj} \sim IG(\upsilon_{lj}/2, \upsilon_{lj}\sigma_{lj}^2/2)$. Let $\boldsymbol{\eta_{lj}} = (\mu_{0lj}, \kappa_{lj}, \upsilon_{lj}, \sigma_{lj}^2)^{\mathrm{T}}$ denote the hyperparameters. The marginal distribution of D_l under model M_j is $p(D_l|\boldsymbol{\eta_{lj}}) = \int p(D_l|\mu_{lj}, \theta_{lj})p(\mu_{lj}|\mu_{0lj}, \theta_{lj}, \kappa_{lj})p(\theta_{lj}|\upsilon_{lj}, \sigma_{lj}^2)d\mu_{lj}d\theta_{lj}$, which can be shown to be a non-standardized t distribution $t_{\upsilon_{lj}}(\mu_{0lj}, \sigma_{lj}^2(1 + \kappa_{lj})/\kappa_{lj})$ [24]. We obtain moment estimators of hyperparameters $\widetilde{\boldsymbol{\eta}}_{lj}$ under M_j using the seed and background genes (see Additional file 1: Section B.2 of Supplementary Notes for details). Then, for each candidate gene, we determine which model gives a better fit using the Bayes factor $\frac{p(D_l|M_1)}{p(D_l|M_0)} = \frac{p(D_l|\widetilde{\boldsymbol{\eta}}_{l1})}{p(D_l|\widetilde{\boldsymbol{\eta}}_{l0})}$.

Comparing with existing methods for ASD gene prioritization

We compare our method with three state-of-the-art methods developed by Krishnan et al. [8] (referred to as "2016Krishnan"), Duda et al. [7] (referred to as "2018Duda"), Lin et al. [16] (referred to as "2020Lin"), and Zhang et al. [15] (referred to as "2020Zhang"). These methods used machine learning algorithms with different types of evidence: 2016Krishnan utilized a brain-specific functional network reflecting brain-specific expression and biological processes; 2018Duda utilized publicly available tissue-specific microarray, protein interaction, and phenotype annotation data sets; 2020Lin utilized spatiotemporal gene expression patterns in human brain, gene-level constraint metrics, and other gene variation features; 2020Zhang utilized a human molecular interaction network based on literature of physical protein interactions experimentally documented. These methods used different seed and background gene sets for training but they all include the 65 seed genes described previously. We use their gene prioritization scores directly rather than retraining the models. For fair comparison, we depleted the training genes and only evaluate the testing genes for each method using the external benchmarks. Specifically, we obtain the 2016Krishnan, 2018Duda, 2020Lin, and 2020Zhang scores from Supplementary Table 3 of Krishnan et al. [8], Supplemental Table 1 of Duda et al. [7], and Supplemental Table 3 of Lin et al. [16], Github repository [25] by Zhang et al. [15].

Measure of gene expression specificity

We use the specificity index (*SI* and *pSI*) defined by Dougherty et al. [26] to measure expression specificity of candidate genes under various biological conditions (i.e., tissue, cell type, brain region, and developmental stage). Suppose there are m potential conditions. We want to calculate the *SI* for gene g under the first condition compared to the other $m-1$ conditions. Let $E_{1,g}$ denote the expression of gene g under the first condition. We compute a fold-change value for gene g to measure its relative expression under the first condition to the kth condition and obtain the rank of this fold change value $R_{1/k,g}$ relative to other genes. Then, we define *SI* for gene g as the average rank of $R_{1/k,g}$ under all conditions, i.e., $SI_{g,1} = \sum_{k=2}^{m} R_{1/k,g}/(m-1)$. Raw *SI* scores are not directly comparable across conditions due to the differences in number of genes expressed under each condition, so *pSI*, a permutation based p value for each *SI* is computed by randomly shuffling expression values and computing *SI* to determine the probability of observing a *SI* value less than or equal to a predefined threshold of 0.05 in the permutated distribution. The *pSI* value is used to assess gene expression specification under certain conditions. The code used to calculate *SI* and *pSI* was obtained from Dougherty lab website [27].

Results

Benchmarking using sequencing-identified novel ASD genes

We obtained 102 ASD genes identified in a recently published large exome sequencing study [28] (referred to as the “2020 study”). Among these 102 ASD genes, 65 had been previously identified in Sanders et al. [3] (referred to as the “2015 study”). These two studies enable us to conduct a “time-lapse” data experiment [29]. That is, we prioritized ASD risk genes based on the 65 seed genes identified in the 2015 study, and then evaluated the top candidate genes against those identified in the 2020 study but not in the 2015 study (referred to as the “ASD2020” gene set).

We employed two strategies to evaluate our model performance: 1) we performed gene set enrichment analysis to assess whether the top candidate genes are significantly enriched in the ASD2020 gene set; 2) we calculated area under the curve (AUC) of receiver-operating characteristic (ROC) curves (ROC-AUC) and precision-recall curves (PR-AUC). The first strategy evaluates a binary classification of risk versus non-risk genes and ignores the relative ranking of genes; the second strategy takes the ranking of genes into account and should render a more robust and comprehensive evaluation.

The ROC and PR curves in Fig. 1 show that our approach, BNScore achieves the best prediction accuracy in the ASD2020 gene set. In particular, BNScore achieves an ROC-AUC value of 0.91, higher than the other methods (with ROC-AUC values of $0.78-0.88$). The improvement over the second best method, 2020Lin, is statistically significant (one-sided Delong’s test p value $= 0.026$). Similarly, BNScore achieves a PR-AUC value of 0.68, a sizable improvement over the other methods (with PR-AUC values of $0.49-0.59$).

We tested the enrichment of the top candidate genes in the ASD2020 gene set, using the rest genes in the genome as background. Figure 2 shows that the top candidate genes predicted by BNScore are substantially more enriched in the ASD2020 gene set than the

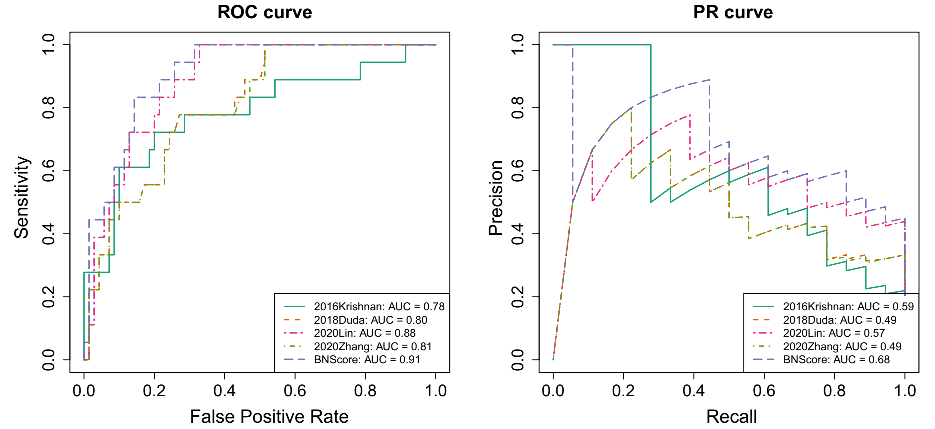

Fig. 1 ROC and PR curves of the four methods in the ASD2020 gene set

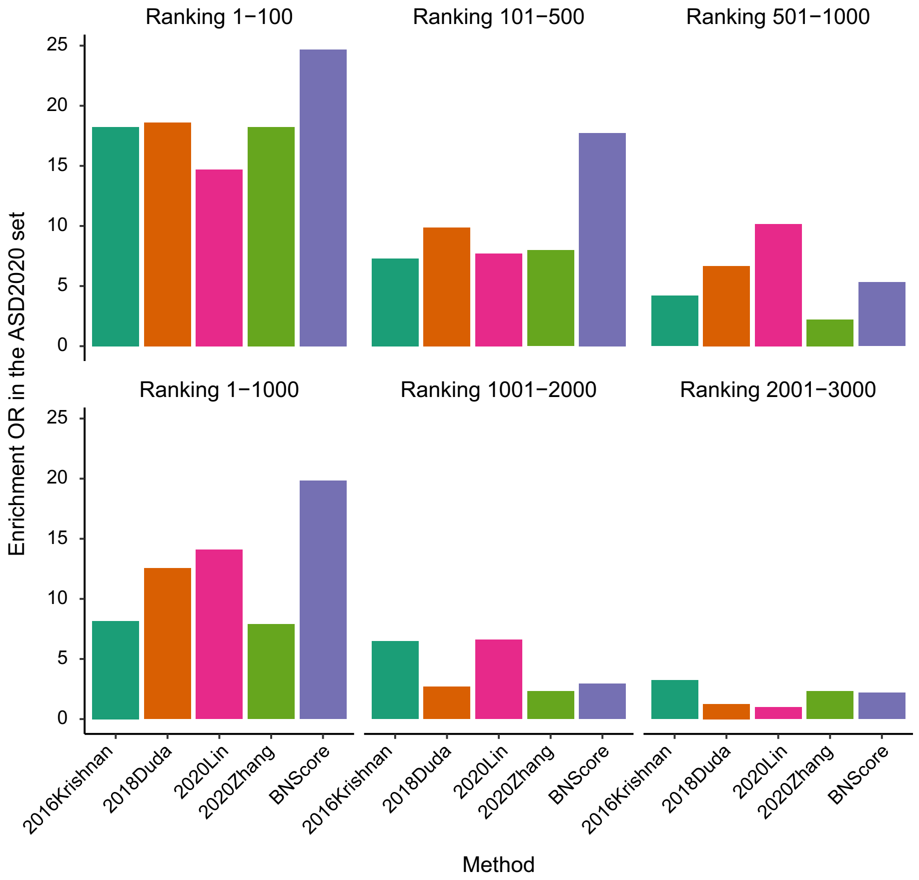

Fig. 2 Enrichment of top candidate genes in the ASD2020 gene set

top candidated genes predicted by the other three methods. For example, the enrichment odds ratio (OR) is 24.7 (p value $= 4.77 \times 10^{-8}$) for the top 100 genes predicted by BNScore, compared to 18.4 (p value $= 1.02 \times 10^{-4}$), 14.7 (p value $= 2.41 \times 10^{-4}$), 14.1 (p value = 2.65×10^{-4}), and 18.23 (p value = 8.60×10^{-4}) for 2018Dula, 2020Lin, 2016Krishnan, and 2020Zhang respectively. The enrichment OR is 19.8 for the top 1000 genes predicted by BNScore, while the ORs are less than 15 for the other three methods. We note that genes with ranking greater than 1000 show much lower enrichment for ASD2020 genes. This is the case for all methods, indicating that most ASD risk genes are likely concentrated in the top 1000 prioritized genes, consistent with the estimate of 1000 ASD risk genes by He et al. [4]. Therefore, we focused on the top 1000 genes as the primary candidate gene sets for the rest of the study.

Benchmarking using SFARI gene lists

We generated benchmark gene sets from the SFARI database [30], which contains genes linked to ASD from a variety of evidence sources and curated into several categories by experts. For each method, we excluded seed genes overlapped with the benchmark SFARI gene sets before evaluation. We formed three tiers of gold-standard gene sets according to the SFARI classification criteria: genes classified as "high confidence" by SFARI were designated as tier 1 evidence (T1); genes classified as "strong candidate" were designated as tier 2 evidence (T2); and genes of "suggestive evidence" were designated as tier 3 evidence (T3). We employed similar strategies as in Sect. to compare the performance of our proposed method with existing ones.

The ROC and PR curves in Fig. 3 show that our proposed BNScore achieved the best accuracy in T1 SFARI gene set. We also tested the top 1000 candidate genes predicted by all methods for enrichment in the SFARI gene lists. Figure 4 shows that the top candidate genes predicted by BNScore are more enriched for T1 genes than the top candidate genes predicted by the other methods. The enrichment OR is 26.9 (p-value = 1.37×10^{-65}) for BNScore, compared to 22.9 (p value = 1.50×10^{-45}, 13.9 (p value = 1.28×10^{-26}), 9.6 (p value = 3.82×10^{-14}), and 11.7 (p value = 1.01×10^{-26}) for 2020Lin, 2018Duda, 2020Zhang, and 2016Krishnan, respectively.

The AUCs for the ROC and PR curves for all methods are lower in the T2 and T3 gene sets compared to their counterparts in the T1 set (Additional file 1: Fig. S1), and the enrichment ORs are much smaller in the T2 and T3 sets compared to those in the T1

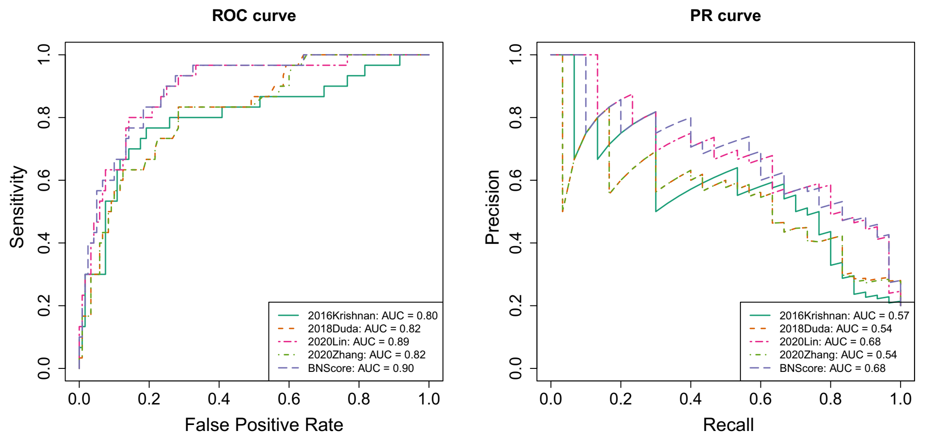

Fig. 3 ROC and PR curves of the four methods in the SFARI T1 gene set

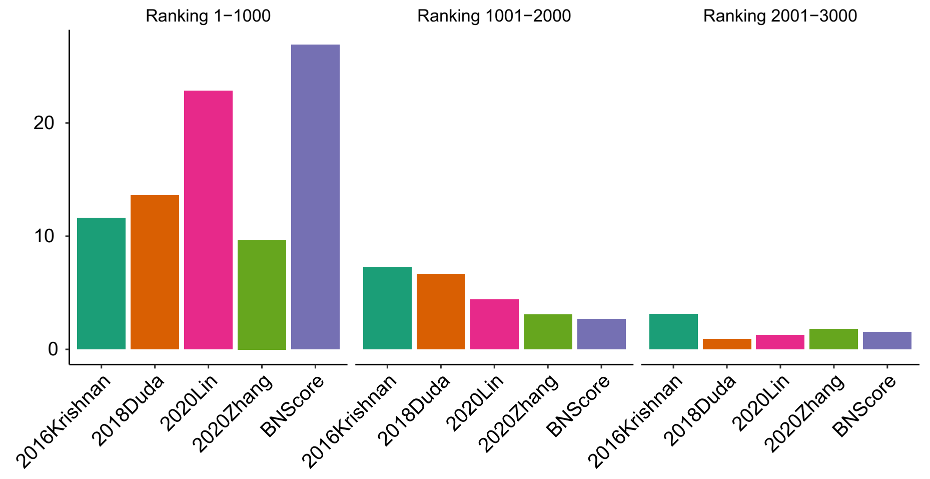

Fig. 4 Enrichment of top candidate genes in the SFARI T1 gene set

set (Additional file 1: Fig. S2). Furthermore, we did not observe significant enrichment in any SFARI gene set for genes ranked greater than 2000 by any method. This suggests that the T2 and T3 gene sets may contain a larger proportion of non-ASD genes than the T1 set. Consequently, the relative performance of the methods in T2 and T3 sets may not be as trustworthy as that in the T1 set.

Benchmarking using ASD GWAS

We used partitioned linkage disequilibrium score regression (LDSC) [31, 32] to assess common SNP heritability enrichment near or within the top candidate genes using summary statistics from the most recent ASD GWAS [33]. Partitioned LDSC is a method to estimate the proportion of genome-wide SNP-heritability attributable to a SNP set, referred to as an “annotation” [32], while taking into account all other annotations. We annotated SNPs that are within 10 kb to the transcription start sites of the top candidate genes [34], and then used partitioned LDSC to evaluate whether these SNPs have enriched ASD heritability. Template files and code to construct annotations were adapted from the LDSC Github repository [35]. We restricted the analysis to Hapmap3 SNPs according to the recommendations from the LDSC authors. Figure 5 shows that the top candidate genes predicted by BNScore, 2016Krishnan, and 2020Lin are significantly enriched for ASD heritability. For the top 1000 genes, the enrichment OR is 7.3 (p value $= 0.0068$) for BNScore, higher than that of 2020Lin (OR $= 6.1$, p value $= 0.036$) and 2016Krishnan (OR $= 6.9$, p value $= 0.025$). We did not include the top candidate genes predicted by 2018Duda or 2020Zhang in Fig. 5 because they were not significantly enriched for ASD heritability (e.g., p value $= 0.53$, p value $= 0.58$ for 2018Duda, 2020Zhang top 1000 genes, respectively).

Examining expression specificity of top candidate genes

We explored the tissue, cell type, and brain developmental stage specificity of the prioritized ASD risk genes to further dissect the ASD etiology. To this end, we calculated the gene expression specificity with respect to tissues, cell types, and brain developmental

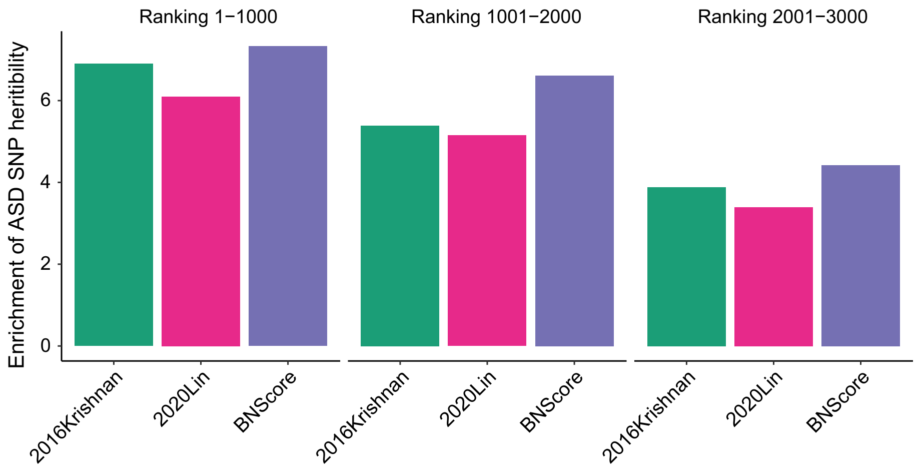

Fig. 5 Enrichment of ASD heritability for top candidate genes

stages based on transcriptome data using the specificity index (*pSI*) developed by Dougherty et al. [26]. We tested our top 1000 candidate ASD risk genes for enrichment in each tissue's, cell type's, or brain developmental stage's specific gene list using Fisher's exact test. We used the Benjamini-Hochberg (BH) procedure to control the false discovery rate [36].

To evaluate tissue specificity, we used the Genotype-Tissue Expression (GTEx) dataset [37]. For tissues with multiple replicates, we averaged reads per kilobase of transcript per million mapped reads values before use. For each tissue, we defined a list of specifically expressed genes by selecting genes with $pSI < 0.05$ (the smaller the more specific). As shown in Fig. 6, we found the brain tissues to be the only category of tissues significantly enriched for the top candidate genes (BH corrected $p - \text{value} = 1.42 \times 10^{-10}$).

To study cell type specificity, we used the mice datasets published by Xu et al. [38]. For each cell type, we defined a list of specifically expressed genes by selecting genes with $pSI < 0.05$. Figure 7 shows the over-representation of candidate genes in striatal medium spiny neurons and retina specific genes. Defects in the striatum have previously been found to specifically contribute to the motor, social, and communication impairments seen in ASD patients [39, 40], while retina has been used as an accessible window to understand brain wiring and functions as it uses and produces most neurotransmitters found in the brain [41].

To study spatiotemporal expression patterns in human brain, we used the Brainspan dataset [42], which was condensed into six major regional divisions across ten developmental stages. For each brain region and developmental stage, we defined a list of specifically expressed genes by selecting genes with $pSI < 0.05$. Figure 8 shows strong enrichment signals for our candidate ASD risk genes in early and mid-fetal stage genes across all brain regions. We observed only a few enrichment signals in the later development stages, e.g., in cerebellum during mid-late childhood and in cortex during young adulthood. These findings are consistent with the reported heterogeneity of ASD, as abruptions from many brain regions across many developmental stages may all contribute to the onset of ASD [8, 43].

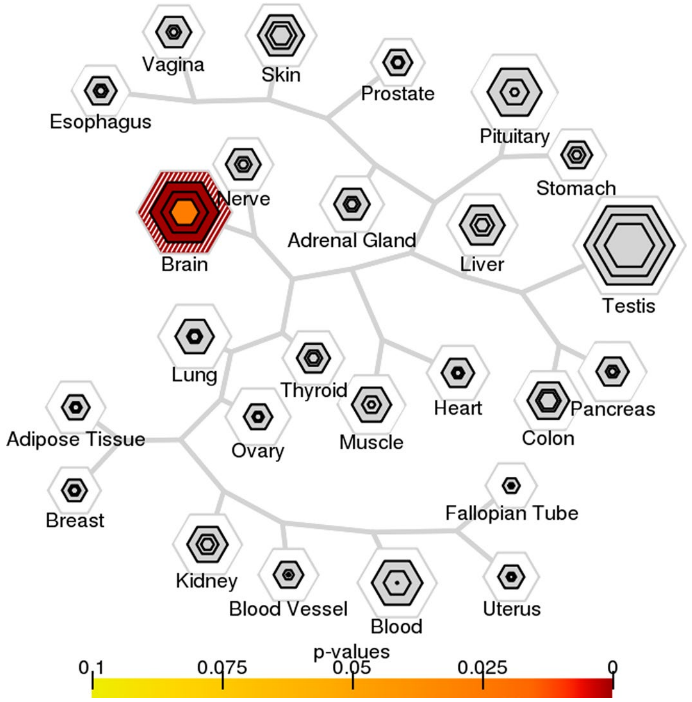

Fig. 6 Tissue expression specificity of top candidate ASD genes. The results are reported in bulleye plots with the size of the bullseye scaled to the number of enriched genes and color coded by Fisher's exact test *p*-values. Brain tissue is the only category of tissues significantly enriched for the top candidate genes (BH corrected *p* value = 1.42×10^{-10})

Discussion

We present BNScore, a Bayesian model selection based framework to facilitate genome-wide ASD gene discovery. The Bayesian modeling frameword has advantages in interpretability of final results compared to hypothesis-testing approaches and machine learning algorithms. Our approach is flexible in integrating multiple types of biological evidence. Currently, it integrates sequencing study results, diverse functional annotations, and network information to obtain genome-wide prediction of ASD risk genes. It is straightforward to incorporate new lines of evidence as they become available in the future.

Our prediction is validated by three benchmark datasets not used in the training process: (1) a recently published exome sequencing study [28], (2) genes from the SFARI database, and (3) a recently published ASD GWAS study [33]. Our approach outperforms the existing methods in most situations, pinpointing 1000 top candidate genes with high confidence. We observe that our top candidate genes are specifically expressed in brain tissues, in striatal medium spiny neurons and retina, and in early developmental stages across brain regions, offering hypotheses for further validation of the implicated tissue, cell types, and developmental stages.

The performance gain for BNScore relative to the other approaches may be attributed to the integration of a variety types of biological evidence, including sequencing study results, diverse functional annotations, and network information to prioritize ASD risk genes. The data integration allows us to take advantage of the complementary

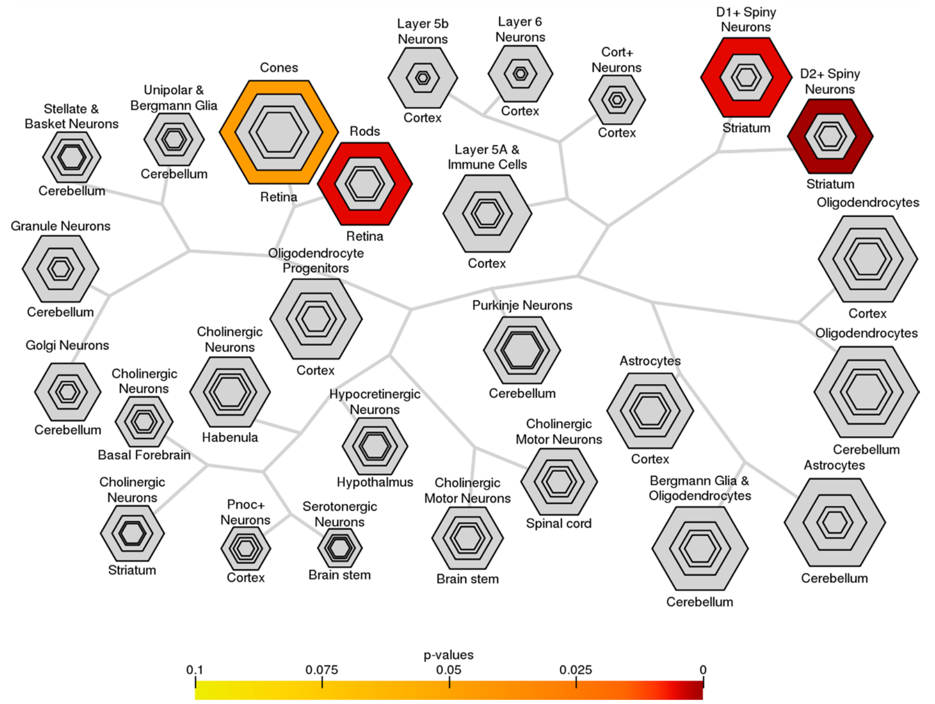

Fig. 7 Cell type expression specificity of top candidate ASD genes. The results are reported in bulleye plots with the size of the bullseye scaled to the number of enriched genes and color coded by Fisher's exact test *p* values. Striatal medium spiny neurons (BH corrected *p* value $= 2.25 \times 10^{-4}$) and retina (BH corrected *p* value $= 0.05 \times 10^{-1}$) are significantly enriched for top candidate genes

information contained in them to improve the prediction of ASD risk genes. Also, genes with multiple sources of evidence pointing to them might be more likely to be true risk genes [44]. In fact, the enrichment ORs of the top 1000 genes ranked using network distance only, Bayes factor based on binary annotations only, Bayes factor based on continuous annotations only are 3.34 (p value $= 1.9 \times 10^{-2}$), 13.09 (p value $= 3.26 \times 10^{-17}$), 17.28 (p value $= 6.16 \times 10^{-22}$), respectively, which are all smaller than the OR of the final BNScore model (OR $= 19.83$, p value $= 1.82 \times 10^{-24}$) for ASD2020 genes.

There are other types of evidence that may be useful for ASD risk gene prioritization. For example, epigenomics data that have been implicated in ASD risk genes may provide additional important evidence [45] , and phenome networks may provide more evidence for gene-gene connections [46]. Our current analysis is limited to protein-coding genes. Recent studies have also shown that long non-coding RNAs (lncRNA), which are important regulators of gene expression, could also be prioritized for ASD risk through transferring knowledge from protein-coding genes [47]. Given the flexibility of our framework, these aspects could be readily explored in future studies and might contribute to further improvement in prediction accuracy.

We assumed parametric models in the Bayes factor calculations because of their ease of implementation and computational efficiency. These models work well in our empirical datasets. In the future, one may encounter other types of annotations that cannot be appropriately represented by these simple models. In this situation, one may consider more sophisticated semiparametric or nonparametric models. In

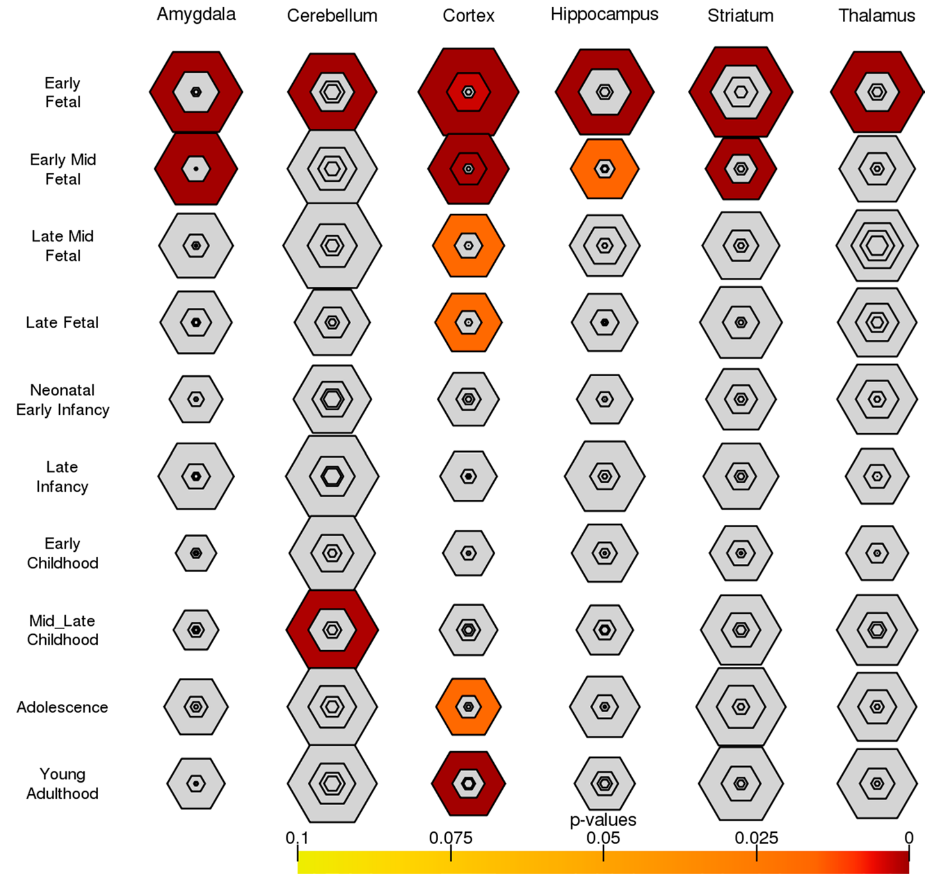

Fig. 8 Brain region and developmental stage expression specificity of top candidate ASD genes. The results are reported in bulleye plots with the size of the bullseye scaled to the number of enriched genes and color coded by Fisher's exact test *p*-values. Strong enrichment signals for top candidate genes are observed across all brain regions in early and mid-fetal stage, and in cerebellum and cortex during the later development stages

general, the advantage of Bayesian modeling lies in its ability to incorporate prior probabilities for the data generating mechanism, isolate the effect of each feature, and identify parameters that are interpretable and of special interest (e.g. θ_{ij} for binary annotation in our approach). As a contrast, machine learning models usually does not attempt to isolate the effect of any single variable and usually does not model the data generating process but instead attempt to learn from the dataset through intractable processes. With that being said, we acknowledge that there are scenarios where machine learning methods can be preferable, e.g., in the presence of complicated nonlinear interactions, when the dataset is huge, or when overall prediction is the only goal and there is no need to succinctly describe the impact of any one variable.

The performance of our proposed BNScore, as well as any other method for ASD gene prioritization, ultimately depends on the seed ASD genes, which may not be representative of the full spectrum of ASD risk genes. In other words, our approach is more powerful to identify new ASD candidate genes similar to "known" disease genes. Therefore, we interpret our results with caution: we can implicate candidate genes but we are not confident to exclude genes that are not similar to seed genes, since such genes may lead to diseases through entirely unexpected mechanisms yet represented in the seed genes.

Like other gene prioritization approaches, our approach is based on the available incomplete annotation data sources, which themselves incorporate false positive/negative annotations and bias studies of human genome. The prioritized genes offer relevant hypotheses to researchers to further investigate.

Conclusion

Overall, owing to the benefits from integrating sequencing findings, functional annotation profiles, and gene-gene functional network, our approach provides competitive performance compared to current state-of-the-art methods when validated in benchmark datasets. The Bayesian setup provides easily interpretable results. With the expansion of both genomic data and epigenomic data in the future, the identification of risk genes could be further improved by expanding our framework to include more annotations. Although designed for ASD, we note that this approach can be extended to other complex traits. It is our hope that this framework can offer prioritized risk genes to researchers to facilitate the identification of disease risk genes.

Abbreviations

ASD: Autism spectrum disorder; ROC-AUC: Area under the receiver-operating characteristic (ROC) curve; PR-AUC: Area under the precision-recall curve; SFARI: Simons foundation autism research initiative; OR: Odds ratio; SI: Specificity index; LDSC: Linkage disequilibrium score regression.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-022-04616-y.

Additional file 1: Supplementary Notes, Tables and Figures.

Acknowledgements

This work was conducted in part using the resources of the Advanced Computing Center for Research and Education at Vanderbilt University, Nashville, TN.

Authors' contributions

BL designed the study, YJ did all analyses. RC, QWang and QWei contributed data collection and interpretation of the results. RT provided advice on statistical modeling. BL and RT jointly supervised the study. YJ drafted the manuscript, RT and BL revised the manuscript. All authors read and approved the final manuscript.

Funding

This work has been supported by National Institute of Health (NIH) Grants U01HG009086 and R01HL151152.

Availability of data and materials

The datasets generated and code are available in the GitHub repository https://github.com/yingji15/ASD_public.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Department of Molecular Physiology and Biophysics, Vanderbilt University, Nashville, TN 37212, USA. [2]Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, TN 37212, USA. [3]Department of Biostatistics, Vanderbilt University, Nashville, TN 37212, USA.

Received: 23 May 2021 Accepted: 15 February 2022
Published online: 22 April 2022

References

1. De Rubeis S, He X, Goldberg AP, Poultney CS, Samocha K, Cicek AE, Kou Y, Liu L, Fromer M, Walker S, Singh T, Klei L, Kosmicki J, Shih-Chen F, Aleksic B, Biscaldi M, Bolton PF, Brownfeld JM, Cai J, Campbell NG, Carracedo A, Chahrour MH, Chiocchetti AG, Coon H, Crawford EL, Curran SR, Dawson G, Duketis E, Fernandez BA, Gallagher L, Geller E, Guter SJ, Hill RS, Ionita-Laza J, JimenzGonzalez P, Kilpinen H, Klauck SM, Kolevzon A, Lee I, Lei I, Lei J, Lehtimki T, Lin CF, Maayan A, Marshall CR, McInnes AL, Neale B, Owen MJ, Ozaki N, Parellada M, Parr JR, Purcell S, Puura K, Rajagopalan D, Rehnstrm K, Reichenberg A, Sabo A, Sachse M, Sanders SJ, Schafer C, Schulte-Rther M, Skuse D, Stevens C, Szatmari P, Tammimies K, Valladares O, Voran A, Li-San W, Weiss LA, Willsey AJ, Yu TW, Yuen RK, Cook EH, Freitag CM, Gill M, Hultman CM, Lehner T, Palotie A, Schellenberg GD, Sklar P, State MW, Sutcliffe JS, Walsh CA, Scherer SW, Zwick ME, Barett JC, Cutler DJ, Roeder K, Devlin B, Daly MJ, Buxbaum JD, Akawi N, Al-Turki S, Ambridge K, Barrett J, Barrett D, Bayzetinova T, Carter N, Clayton S, Coomber E, Firth H, Fitzgerald T, Fitzpatrick D, Gererty S, Gribble S, Hurles M, Jones P, Jones W, King D, Krishnappa N, Mason L, McRae J, Michael P, Middleton A, Miller R, Morley K, Parthiban V, Prigmore E, Rajan D, Sifrim A, Tivery A, van Kogelenberg M, Wright C, Adli M, Al-Awadi S, Al-Gazali L, Allub Z, Al-Saad S, Al-Saffar M, Ataman B, Balkhy S, Barkovich AJ, Barry BJ, Bastaki L, Bauman M, Ben-Omran T, Braverman NE, Chahrour MH, Chang BS, Chaudhry HR, Coulter M, Gama AM, Daoud A, Eapen V, Felie JM, Gabriel SB, Gascon GG, Greenberg ME, Hanson E, Harmin DA, Hashmi A, Herguner S, Hill RS, Hisama FM, Jiralerspong S, Joseph RM, Khalil S, Khuri-Bulos N, Kwaja O, Kwan BY, LeClair E, Lim ET, Markianos K, Martin M, Masri A, Meyer B, Mochida GH, Morrow EM, Mukaddes NM, Nasir RH, Niaz S, Okarmura-Ikeda K, Oner O, Parlow JN, Poduri A, Rajab A, Rappaport L, Rodriguez J, Schmitz-Abe K, Shen Y, Stevens CR, Stoler JM, Sunu CM, Tan WH, Taniguchi H, Teebi A, Walsh CA, Ware J, Wu BL, Yoo SY, Yu T, Anney R, Ayub M, Bailey A, Baird G, Barrett J, Blackwood D, Bolton P, Breen G, Collier D, Cormican P, Craddock N, Crooks L, Curran S, Danecek P, Durbin R, Gallagher L, Green J, Gurling H, Holt R, Joyce C, LeCouteur A, Lee I, Lnnqvist J, McCarthy S, McGuffin P, McIntosh A, McQuillen A, Merkangas A, Monaco A, Muddyman D, O'Donovan M, Owen M, Palotie A, Parr J, Paunio T, Pietilainen O, Rehnstrm K, Skuse D, Stalker J, StClair D, Suvisaari J, Williams H. Synaptic, transcriptional and chromatin genes disrupted in autism. Nature. 2014;515(7526):209–15.
2. Iossifov I, O'Roak BJ, Sanders SJ, Ronemus M, Krumm N, Levy D, Stessman HA, Witherspoon KT, Vives L, Patterson KE, Smith JD, Paeper B, Nickerson DA, Dea J, Dong S, Gonzalez LE, Mandell JD, Mane SM, Murtha MT, Sullivan CA, Walker MF, Waqar Z, Wei L, Willsey AJ, Yamrom B, Lee YH, Grabowska E, Dalkic E, Wang Z, Marks S, Andrews P, Leotta A, Kendall J, Hakker I, Rosenbaum J, Ma B, Rodgers L, Troge J, Narzisi G, Yoon S, Schatz MC, Ye K, McCombie WR, Shendure J, Eichler EE, State MW, Wigler M. The contribution of de novo coding mutations to autism spectrum disorder. Nature. 2014;515(7526):216–21.
3. Sanders SJ, He X, Willsey AJ, Ercan-Sencicek AG, Samocha KE, Cicek AE, Murtha MT, Bal VH, Bishop SL, Dong S, Goldberg AP, Jinlu C, Keaney JF, Klei L, Mandell JD, Moreno-De-Luca D, Poultney CS, Robinson EB, Smith L, Solli-Nowlan T, Su MY, Teran NA, Walker MF, Werling DM, Beaudet AL, Cantor RM, Fombonne E, Geschwind DH, Grice DE, Lord C, Lowe JK, Mane SM, Martin DM, Morrow EM, Talkowski ME, Sutcliffe JS, Walsh CA, Yu TW, Ledbetter DH, Martin CL, Cook EH, Buxbaum JD, Daly MJ, Devlin B, Roeder K, State MW. Insights into autism spectrum disorder genomic architecture and biology from 71 risk loci. Neuron. 2015;87(6):1215–33.
4. He X, Sanders SJ, Liu L, De Rubeis S, Lim ET, Sutcliffe JS, Schellenberg GD, Gibbs RA, Daly MJ, Buxbaum JD, et al. Integrated model of de novo and inherited genetic variants yields greater power to identify risk genes. PLoS Genet. 2013;9(8):1003671.
5. Iakoucheva LM, Muotri AR, Sebat J. Getting to the cores of autism. Cell. 2019;178(6):1287–98.
6. Ignatiadis N, Klaus B, Zaugg JB, Huber W. Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. Nat Methods. 2016;13(7):577–80.
7. Duda M, Zhang H, Li H-D, Wall DP, Burmeister M, Guan Y. Brain-specific functional relationship networks inform autism spectrum disorder gene prediction. Transl Psych. 2018;8(1):1–9.
8. Krishnan A, Zhang R, Yao V, Theesfeld CL, Wong AK, Tadych A, Volfovsky N, Packer A, Lash A, Troyanskaya OG. Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder. Nat Neurosci. 2016;19(11):1454–62.
9. Doncheva NT, Kacprowski T, Albrecht M. Recent approaches to the prioritization of candidate disease genes. Wiley Interdiscip Rev Syst Biol Med. 2012;4(5):429–42.
10. Aerts S, Lambrechts D, Maity S, Van Loo P, Coessens B, De Smet F, Tranchevent LC, De Moor B, Marynen P, Hassan B, Carmeliet P, Moreau Y. Gene prioritization through genomic data fusion. Nat Biotechnol. 2006;24(5):537–44.
11. Liu L, Lei J, Sanders SJ, Willsey AJ, Kou Y, Cicek AE, Klei L, Lu C, He X, Li M, Muhle RA, Ma'ayan A, Noonan JP, Sestan N, McFadden KA, State MW, Buxbaum JD, Devlin B, Roeder K. DAWN: a framework to identify autism genes and subnetworks using gene expression and genetics. Mol Autism. 2014;5(1):22.
12. Chen J, Xu H, Aronow BJ, Jegga AG. Improved human disease candidate gene prioritization using mouse phenotype. BMC Bioinform. 2007;8(1):392.
13. Chen J, Bardes EE, Aronow BJ, Jegga AG. Toppgene suite for gene list enrichment analysis and candidate gene prioritization. Nucleic Acids Res. 2009;37(Suppl 2):305–11.
14. Gilman SR, Iossifov I, Levy D, Ronemus M, Wigler M, Vitkup D. Rare de novo variants associated with autism implicate a large functional network of genes involved in formation and function of synapses. Neuron. 2011;70(5):898–907.
15. Zhang Y, Chen Y, Hu T. Panda: prioritization of autism-genes using network-based deep-learning approach. Genet Epidemiol. 2020;44(4):382–94.
16. Lin Y, Afshar S, Rajadhyaksha AM, Potash JB, Han S. A machine learning approach to predicting autism risk genes: validation of known genes and discovery of new candidates. Front Genet. 2020;11:389.

17. Moreau Y, Tranchevent L-C. Computational tools for prioritizing candidate genes: boosting disease gene discovery. Nat Rev Genet. 2012;13(8):523–36.
18. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. Gene ontology: tool for the unification of biology. Nat Genet. 2000;25(1):25–9.
19. Consortium G.O. The gene ontology resource: 20 years and still going strong. Nucleic Acids Res. 2019;47(D1):330–8.
20. Tong H, Faloutsos C, Pan J-Y. Random walk with restart: fast solutions and applications. Knowl Inf Syst. 2008;14(3):327–46.
21. Parikshak NN, Luo R, Zhang A, Won H, Lowe JK, Chandran V, Horvath S, Geschwind DH. Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. Cell. 2013;155(5):1008–21.
22. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016;536(7616):285–91.
23. Casella G. An introduction to empirical Bayes data analysis. Am Stat. 1985;39(2):83–7.
24. Zhang Y-Y, Rong T-Z, Li M-M. The empirical Bayes estimators of the mean and variance parameters of the normal distribution with a conjugate normal-inverse-gamma prior by the moment method and the MLE method. Commun Stat Theory Methods. 2019;48(9):2286–304.
25. Zhang Y. PANDA Github repository. 2019. https://github.com/MIB-Lab/PANDA/tree/master/panda Accessed 11 May 2021.
26. Dougherty JD, Schmidt EF, Nakajima M, Heintz N. Analytical approaches to RNA profiling data for the identification of genes enriched in specific cells. Nucleic Acids Res. 2010;38(13):4218–30.
27. Dougherty JD. PSI package. 2016. http://genetics.wustl.edu/jdlab/psi_package/. Accessed 11 Sep 2020.
28. Satterstrom FK, Kosmicki JA, Wang J, Breen MS, De Rubeis S, An JY, Peng M, Collins R, Grove J, Klei L, Stevens C, Reichert J, Mulhern MS, Artomov M, Gerges S, Sheppard B, Xu X, Bhaduri A, Norman U, Brand H, Schwartz G, Nguyen R, Guerrero EE, Dias C, Betancur C, Cook EH, Gallagher L, Gill M, Sutcliffe JS, Thurm A, Zwick ME, Brglum AD, State MW, Cicek AE, Talkowski ME, Cutler DJ, Devlin B, Sanders SJ, Roeder K, Daly MJ, Buxbaum JD, Aleksic B, Anney R, Barbosa M, Bishop S, Brusco A, Bybjerg-Grauholm J, Carracedo A, Chan MCY, Chiocchetti AG, Chung BHY, Coon H, Cuccaro ML, Curr A, DallaBernardina B, Doan R, Domenici E, Dong S, Fallerini C, Fernndez-Prieto M, Ferrero GB, Freitag CM, Fromer M, Gargus JJ, Geschwind D, Giorgio E, Gonzlez-Peas J, Guter S, Halpern D, Hansen-Kiss E, He X, Herman GE, Hertz-Picciotto I, Hougaard DM, Hultman CM, Ionita-Laza I, Jacob S, Jamison J, Jugessur A, Kaartinen M, Knudsen GP, Kolevzon A, Kushima I, Lee SL, Lehtimki T, Lim ET, Lintas C, Lipkin WI, Lopergolo D, Lopes F, Ludena Y, Maciel P, Magnus P, Mahjani B, Maltman N, Manoach DS, Meiri G, Menashe I, Miller J, Minshew N, Montenegro EMS, Moreira D, Morrow EM, Mors O, Mortensen PB, Mosconi M, Muglia P, Neale BM, Nordentoft M, Ozaki N, Palotie A, Parellada M, Passos-Bueno MR, Pericak-Vance M, Persico AM, Pessah I, Puura K, Reichenberg A, Renieri A, Riberi E, Robinson EB, Samocha KE, Sandin S, Santangelo SL, Schellenberg G, Scherer SW, Schlitt S, Schmidt R, Schmitt L, Silva IMW, Singh T, Siper PM, Smith M, Soares G, Stoltenberg C, Suren P, Susser E, Sweeney J, Szatmari P, Tang L, Tassone F, Teufel K, Trabetti E, Trelles MDP, Walsh CA, Weiss LA, Werge T, Werling DM, Wigdor EM, Wilkinson E, Willsey AJ, Yu TW, Yu MHC, Yuen R, Zachi E, Agerbo E, Als TD, Appadurai V, Bkvad-Hansen M, Belliveau R, Buil A, Carey CE, Cerrato F, Chambert K, Churchhouse C, Dalsgaard S, Demontis D, Dumont A, Goldstein J, Hansen CS, Hauberg ME, Hollegaard MV, Howrigan DP, Huang H, Maller J, Martin AR, Martin J, Mattheisen M, Moran J, Pallesen J, Palmer DS, Pedersen CB, Pedersen MG, Poterba T, Poulsen JB, Ripke S, Schork AJ, Thompson WK, Turley P, Walters RK. Large-scale exome sequencing study implicates both developmental and functional changes in the neurobiology of autism. Cell. 2020;180(3):568–84.
29. Cáceres JJ, Paccanaro A. Disease gene prediction for molecularly uncharacterized diseases. PLoS Comput Biol. 2019;15(7):1007078.
30. Initiative SFAR. SFARI Gene scoring module. 2020. https://gene.sfari.org/database/gene-scoring/. Accessed 10 Sep 2020.
31. Bulik-Sullivan B, Finucane HK, Anttila V, Gusev A, Day FR, Loh P-R, Duncan L, Perry JR, Patterson N, Robinson EB, et al. An atlas of genetic correlations across human diseases and traits. Nat Genet. 2015;47(11):1236.
32. Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, Loh P-R, Anttila V, Xu H, Zang C, Farh K, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat Genet. 2015;47(11):1228.
33. Grove J, Ripke S, Als TD, Mattheisen M, Walters RK, Won H, Pallesen J, Agerbo E, Andreassen OA, Anney R, et al. Identification of common genetic risk variants for autism spectrum disorder. Nat Genet. 2019;51(3):431–44.
34. Liu X, Finucane HK, Gusev A, Bhatia G, Gazal S, O'Connor L, Bulik-Sullivan B, Wright FA, Sullivan PF, Neale BM, et al. Functional architectures of local and distal regulation of gene expression in multiple human tissues. Am J Hum Genet. 2017;100(4):605–16.
35. Brendan Bulik-Sullivan HF. LDSC Github repository (2015). https://github.com/bulik/ldsc. Accessed 10 Sep 2020.
36. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B (Methodol). 1995;57(1):289–300.
37. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, Hasz R, Walters G, Garcia F, Young N, et al. The genotype-tissue expression (GTEX) project. Nat Genet. 2013;45(6):580–5.
38. Xu X, Wells AB, O'Brien DR, Nehorai A, Dougherty JD. Cell type-specific expression analysis to identify putative cellular mechanisms for neurogenetic disorders. J Neurosci. 2014;34(4):1420–31.
39. Nickl-Jockschat T, Habel U, Maria Michel T, Manning J, Laird AR, Fox PT, Schneider F, Eickhoff SB. Brain structure anomalies in autism spectrum disorder-a meta-analysis of VBM studies using anatomic likelihood estimation. Hum Brain Map. 2012;33(6):1470–89.
40. Fuccillo MV. Striatal circuits as a common node for autism pathophysiology. Front Neurosci. 2016;10:27.
41. London A, Benhar I, Schwartz M. The retina as a window to the brain-from eye research to CNS disorders. Nat Rev Neurol. 2013;9(1):44.
42. Sunkin SM, Ng L, Lau C, Dolbeare T, Gilbert TL, Thompson CL, Hawrylycz M, Dang C. Allen brain atlas: an integrated spatio-temporal portal for exploring the central nervous system. Nucleic Acids Res. 2012;41(D1):996–1008.
43. Dinstein I, Heeger DJ, Behrmann M. Neural variability: friend or foe? Trends Cognit Sci. 2015;19(6):322–8.

44. Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D. Methods of integrating data to uncover genotype-phenotype interactions. Nat Rev Genet. 2015;16(2):85–97.
45. Schanen NC. Epigenetics of autism spectrum disorders. Hum Mol Genet. 2006;15(Suppl 2):138–50.
46. McDiarmid TA, Belmadani M, Liang J, Meili F, Mathews EA, Mullen GP, Hendi A, Wong W-R, Rand JB, Mizumoto K, et al. Systematic phenomics analysis of autism-associated genes reveals parallel networks underlying reversible impairments in habituation. Proc Natl Acad Sci. 2020;117(1):656–67.
47. Wang J, Wang L. Prediction and prioritization of autism-associated long non-coding RNAS using gene expression and sequence features. BMC Bioinform. 2020;21(1):1–15.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.