scientific reports

Check for updates

OPEN

"Guilt by association" is not competitive with genetic association for identifying autism risk genes

Margot Gunning[1,2,3] & Paul Pavlidis[1,2,4]✉

Discovering genes involved in complex human genetic disorders is a major challenge. Many have suggested that machine learning (ML) algorithms using gene networks can be used to supplement traditional genetic association-based approaches to predict or prioritize disease genes. However, questions have been raised about the utility of ML methods for this type of task due to biases within the data, and poor real-world performance. Using autism spectrum disorder (ASD) as a test case, we sought to investigate the question: can machine learning aid in the discovery of disease genes? We collected 13 published ASD gene prioritization studies and evaluated their performance using known and novel high-confidence ASD genes. We also investigated their biases towards generic gene annotations, like number of association publications. We found that ML methods which do not incorporate genetics information have limited utility for prioritization of ASD risk genes. These studies perform at a comparable level to generic measures of likelihood for the involvement of genes in any condition, and do not out-perform genetic association studies. Future efforts to discover disease genes should be focused on developing and validating statistical models for genetic association, specifically for association between rare variants and disease, rather than developing complex machine learning methods using complex heterogeneous biological data with unknown reliability.

Elucidating the genetic architecture of complex human disorders and diseases is currently a major challenge in medical research. Identifying genes involved in disease is often a time consuming and expensive process, so many researchers have been attracted to the idea of using predictions generated by machine learning (ML) algorithms[1–4]. However, the effectiveness of ML approaches, in contrast to traditional genetic association, is unclear.

Algorithms used in gene function or disease prioritization tasks generally operate on a principle called guilt by association (GBA) (Gillis and Pavlidis[6]; Lanckriet et al. [7]), which postulates that genes with "associations" are more likely to be "guilty" of sharing functions. Associations can be sourced from multiple data types, such as gene expression, physical or genetic interactions, and protein sequence similarity. There are many ways these data types can be integrated into a machine learning method, and depending on the data types and algorithm, the associations among genes may be implicitly or explicitly represented as a network in which both direct and indirect associations can be used for inference.

Previous work from our group has shown that applications of ML to gene function prediction are highly influenced by biases in the underlying data[5,6,8]. For example, protein interactions are often biased toward well studied genes, which often have high numbers of associated functional annotations ("multifunctional"). Furthermore, annotations and number of associations can be correlated, and this turns out to be a driver of GBA behavior: GBA tends to ascribe new functions to genes which are highly connected within the network rather than learning additional, novel information from the connection patterns[6,8]. The implication of this "multifunctionality bias" is that GBA can seem to work in cross-validation settings, while providing predictions with little specific value. As an extreme illustration of this phenomenon, a million-edge network of gene associations can be reduced to 23 associations while not substantially impacting GBA performance[5]. For these and other reasons, the real-life

[1]Michael Smith Laboratories, University of British Columbia, Vancouver, BC V6T 1Z4, Canada. [2]Department of Psychiatry, University of British Columbia, Vancouver, BC V6T 1Z4, Canada. [3]Graduate Program in Bioinformatics, University of British Columbia, Vancouver, BC V6T 1Z4, Canada. [4]Djavad Mowafaghian Centre for Brain Health, University of British Columbia, Vancouver, BC V6T 1Z4, Canada. ✉email: paul@msl.ubc.ca

Name	Genetic	GBA ML	Generic	Method and citation
Princeton		✓		Evidence-weighted support vector machine classifier[1]
FRN		✓		Evidence-weighted random forest classifier[9]
DAMAGES		✓		Logistic regression classifier[22]
RF_Lin		✓		Random forest classifier. Does not provide scores for training labels[10]
PANDA		✓		Graph neural network classifier[4]
forecASD	✓	✓		Ensemble stacked random forest classifier[11]
DAWN	✓	✓		Cluster analysis with co-expression and TADA data. Does not provide scores for all protein-coding genes in genome[23]
DeRubeis	✓			TADA on de novo, inherited, and case–control LoF and missense variants[18]
Sanders	✓			TADA on de novo, inherited, and case–control LoF and missense variants, and small deletions[21]
iHart	✓			TADA on de novo, inherited, and case–control LoF and missense variants, and small deletions[19]
Satterstrom	✓			TADA on de novo and case–control LoF and missense variants with pLI and "missense badness score" in framework[24]
Spark	✓			TADA on de novo LoF and missense variants. Does not provide genome-wide scores[20]
Iossifov	✓			ASD-specific likely gene disruptive score[25]
ExAC pLI			✓	Probability loss of function intolerance score based on approximately 60,000 exomes[26]
gnomAD pLI			✓	Probability loss of function intolerance score based on approximately 120,000 exomes[27]
o/e LoF			✓	Observed/expected loss of function score from gnomAD[26]

Table 1. Summary of the ASD gene prioritization studies and generic methods for disease gene prioritization we used.

performance of GBA methods can be questioned. Focusing on the disease gene identification case, we are not aware of any instances where GBA has been responsible for a *bona fide* disease gene identification.

Recently there has been interest in a specific use case for GBA-based ML: predicting genes responsible for genetic risk of autism spectrum disorder (ASD). Multiple GBA-based ML studies have been produced with claims of providing greater insight into the genetic etiology of ASD[1,4,9–11]. ASD is a neurodevelopmental disorder with a genetically heterogenous etiology[12]. Currently, much ASD research is aimed at identifying very rare, highly penetrant de novo variants in ASD probands because this class of variation has been found to impart a large proportion of risk[13–16]. While statistical methods for evaluating rare variants are still a topic of active research, genetic association underpins all ASD risk gene identification to date. In this context, ML methods have a challenge for acceptance by geneticists. Assessing the quality of ML-based ASD gene predictions is essential to provide realistic estimates of performance or complementarity to genetic approaches. However, we are unaware of attempts to compare them directly to genetic association studies, and assess their real-world applicability. In this paper we aimed to assess the reliability and usability of guilt by association machine learning approaches for ASD gene prioritization.

Methods

ASD gene sets. We compiled two ASD genes sets for algorithm evaluation (Table 1). We used the SFARI-Gene 2.0[17] database as a source of well known, high-confidence ASD genes. SFARIGene collects information on ASD genetic risk factors and genes, and is manually curated by MindSpec, Incorporated. Genes are categorized by amount and quality of evidence for associated with ASD, and assigned a score ranging from 1 (high confidence) to 3 (suggestive evidence), or S (syndromic). We used the 144 genes from SFARI category 1 currently considered to be high-confidence ASD genes (Feb 2020) as our SFARI-HC gene set. Many of SFARI-HC genes were initially identified by the genetic association studies of De Rubeis et al.[18] and Sanders et al.[17]. Different subsets of these high-confidence genes had been used for training of the GBA ML algorithms discussed below. We complied a second high-confidence ASD risk gene set recently identified in three large-scale Transmission and De Novo Association Analysis (TADA) studies: Ruzzo et al.[19] (iHart), Feliciano et al.[20] (Spark), and Satterstrom et al.[21]. These three studies were built based on the background of the original TADA genetic association studies, De Rubeis et al.[18] and Sanders et al.[17]. We refer to this set as "novel-HC" to reflect that most of the genes on this list were not used in the training of the GBA ML algorithms, largely because they were identified after the publication of the ML methods. We considered evaluation using the novel-HC genes a "testing scenario" because the ultimate use case of the machine learning algorithms is to highly prioritize and predict novel ASD genes.

ASD gene prioritization studies and generic measures of disease gene likelihood. We considered 13 ASD gene prioritization studies (Table 1). Each study scored genes based on the authors' assessment of their probability of contributing to ASD risk. All studies also provided lists of genes they considered to be high-confidence ASD risk gene candidates based on a thresholding of their rankings. We obtained these scores from the supplemental tables of the publications. We also evaluated three measures of constraint against loss-of-function (LoF) variation because they can be thought of as generic measures of disease gene likelihood (see below for descriptions).

We mapped gene symbols and Entrez gene identification numbers provided by each study to NCBI official gene symbols and Entrez gene identification numbers, and kept only protein-coding genes[28]. We used the mean score when a gene was listed more than once in a study. We ranked the scores from each study so that 1.0 was the highest possible score, indicating higher assessed likelihood of being involved in ASD, and 0.0 was the lowest possible score. The probability loss of function scores (pLI) from ExAC and gnomAD were already in the proper scale, with higher scores indicating genes likely to have high constraint against loss of function (LoF) variation. The scale of the observed/expected LoF score is opposite to the pLI scale and does not range from 0 to 1. We ranked genes based on o/e LoF score from lowest to highest. Lastly, for protein-coding genes not assessed in each GBA ML and GA study, we set the prediction or association score to be 0.0, or in the case of the o/e LoF score, the highest observed value of 2.0. Studies are organized into four categories based on the approach they used. Below we provide a brief description of each data source; see Supplemental Materials for more information.

Genetic association studies. The studies described below are among the most important in terms of identifying what are generally considered high-confidence ASD genes[18–20,29]. We included them in this study primarily to help establish a baseline to which the GBA ML approaches described in subsequent sections can be compared. Many of these studies are based on the TADA approach.

DeRubeis[18] used whole-exome sequencing (WES) data from approximately 13,000 samples from trios and case-controls to identify de novo and inherited LoF variants, and de novo likely damaging missense variants (Mis3 by PolyPhen2). They used a TADA analysis to identify 33 ASD risk genes at FDR < 0.1. Samples from the Autism Sequencing Consortium (ASC), from Simons Simplex Consortium (SSC) (O'Roak et al.[15]; Sanders et al.[16]; Iossifov et al.[13]), and other cohorts were used. Association scores were provided for 18,735 genes.

Sanders[21] used WES data from approximately 17,000 samples from trios and case-controls to identify de novo and inherited LoF variants, de novo likely damaging missense variants (Mis3 by PolyPhen2), and small de novo deletions. They employed a TADA analysis to identify 65 ASD risk genes at FDR < 0.1. They sequenced roughly 2,500 SSC families in addition to using SSC samples from Levy et al.[30], Iossifov et al.[31] and Dong et al.[32], and ASC samples from De Rubeis et al.[18], and samples from Pinto et al.[33], among others. Association scores were provided for 18,665 genes.

iHart[19] used whole-genome sequencing (WGS) data from 2,308 individuals from 493 multiplex Autism Genetic Resource Exchange (AGRE) families to identify de novo and inherited LoF variants and de novo likely damaging missense variants (Mis3 by PolyPhen2). They used their data and the Sanders data, and the Sanders TADA model to identify 69 ASD risk genes with FDR < 0.1, including 16 novel findings. Association scores were provided for 18,472 genes.

Spark[20] was the pilot study for the Simons Powering Autism Research for Knowledge (SPARK) project. They identified inherited and de novo likely damaging missense mutations (CADD ≥ 25) in 465 SPARK trios. They combined their de novo variants with de novo variants from 4,773 other simplex ASD trios from the ASC (De Rubeis et al.[18]) and SSC (Iossifov et al.[31]; Krumm et al.[34]), among other sources, for a TADA analysis. They identified 67 genes with FDR < 0.1, with 13 novel findings. They provided scores for the 2,249 genes found to have additional variation in SPARK families[20].

Satterstrom (Satterstrom et al.[21]) is the most recent and largest-scale genetic association study, with over 30,000 samples. They used samples from the SSC (Iossifov et al.[13]; Iossifov et al.[31]; O'Roak et al.[15]; Sanders et al.[16]), the ASC (De Rubeis et al.[18] and others), others from the AGRE and many other cohorts around the world. They used WES to identify de novo and case–control LoF, and de novo missense mutations (predicted by MPC, the "missense, PolyPhen-2, constraint score"), and employed TADA analysis to identify 102 ASD risk genes at FDR < 0.1. They considered 31 significant genes to be novel findings. Association scores were provided for 17,484 genes. Importantly, Satterstrom et al. modified the TADA method from the studies mentioned above by using the pLI score from ExAC and the MPC score to estimate the priors for the relative risk of LoF and missense variant classes.

Iossifov[25] computed a "Likely Gene-Disruptive" (LGD) score based on recurrence of LGD variants, the difference in frequency of LGD variants between ASD probands and unaffected siblings (ascertainment differential), and the load of LGD variation in ASD probands. They used data from WES of 2,471 families from the SSC (Iossifov et al.[31]), and exome variants from approximately 6,000 controls from the Exome Variant Server[25]. The theory behind the LGD score is similar to the TADA test and to generic measures of constraint against LoF and missense variation because they use recurrence of variants across multiple samples and models of expected LGD variation in a typical gene to increase power to find disease genes[25,26,29]. They provided scores for 23,953 genes, and identified their top 239 genes as likely ASD risk gene candidates[25].

GBA ML studies. Studies in this class do not use information from ASD genetic association studies, but they use machine learning algorithms to distinguish ASD from non-ASD risk genes using other types of non-genetics data.

Princeton[1] is an evidence weighted support vector machine (SVM) built on a functional interaction network made from human gene expression, protein–protein interaction, regulatory, and genetic and chemical perturbation data. For training they used 594 ASD genes, and 1,189 manually curated non-mental health associated genes as positives and negatives, respectively. The positive ASD genes were given one of three weights (1.0, 0.5, 0.25) based on strength of evidence of association with ASD. Krishnan et al. provided likelihood rankings for 25,825 genes, and identified their top decile as likely ASD risk gene candidates.

FRN[9] is a random forest classifier built on an evidence-weighted functional interaction network of human, mouse and rat brain gene expression, protein–protein interaction, protein docking and phenotype annotation data. They used 143 high-confidence ASD genes from SFARI and the Sanders publication above as positive

training genes, and 1,176 of the of the 1,189 Princeton non-mental health associated genes as negative training genes. They provided likelihood rankings for 21,114 genes, and identified their top decile as likely ASD risk gene candidates.

DAMAGES[22] used a combination of regularized and logistic regression using cell-type specific gene expression data and measures of constraint against LoF and missense variation from ExAC. First, they created a DAMAGES (D) score using principal component analysis (PCA) and regression analysis on gene-expression profiles of 24 mouse central nervous system cell types in 6 regions. They created profiles of 145 genes found to have de novo LoF variants in ASD probands and unaffected siblings from multiple cohorts as training samples. Next, using logistic regression, they combined the D score with ExAC measures to create an ensemble (E) score. They used 36 genes with 2 or more de novo LoF variants in ASD probands, and 156 genes with 1 or more de novo LoF variants in sibling controls from multiple cohorts as positive and negative training genes, respectively. They provided likelihood rankings for 15,881 human genes, and identified their top 117 genes as likely ASD risk gene candidates.

RF_Lin[10] is a random forest classifier. They built an evidence-weighted network of BrainSpan co-expression and protein–protein interaction data, and extracted network features such as hubness and centrality[35]. The features of their classifier included their network association matrix, selected network features, and gene-level constraint measures from ExAC. They used the positive and negative training labels employed by FRN described above. They provided likelihood rankings for 17,099 genes, and identified their top decile as likely ASD risk gene candidates. They did not provide scores for their training genes.

PANDA[4] used a network-based deep-learning approach to prioritize autism genes. They built a human molecular interaction network from protein–protein interaction data from multiple sources, and used a training set of 760 ASD genes from SFARI Gene 2.0 and OMIM weighted by confidence of association with ASD (1.0, 0.75, 0.5). They provided likelihood rankings for 23,472 genes, and defined an "autism subnetwork" made up of 2,346 genes (approximately top decile).

Genetics-GBA Hybrid ML studies. The studies in this section used a combination of ASD-specific genetic association information (e.g., from the studies listed above) along with other features to build their models. Information from the two classes of features are integrated prior to training a machine learning algorithm to distinguish ASD from non-ASD risk genes, using high-confidence ASD genes from genetic association studies as their positive training set.

DAWN[23] built a co-expression network from BrainSpan data of the prefrontal and motor-somatosensory neocortex at 10–24 weeks post-conception, and overlaid association statistics from a TADA analysis[35]. Using unsupervised model-based clustering (Weighted Gene Co-expression Network Analysis) and a hidden Markov random field, they modeled the correlation of genetic association scores across the co-expression network to identify highly correlated nodes, or "network ASD genes." Following a false discovery rate estimation procedure, they identified 127 likely ASD risk gene candidates from 10,233 genes.

forecASD[11] is a stacked random forest ensemble classifier using BrainSpan[35] gene expression data, STRING[36] protein–protein interaction data, and genome-wide results from Princeton, DAWN, DAMAGES, Sanders and DeRubeis studies described above. They used 76 SFARI high-confidence genes and 1,000 randomly selected non-SFARI genes as positive and negative training examples, respectively. They provided likelihood rankings for 17,957 genes, and identified their top decile of genes as likely ASD risk gene candidates.

Generic measures of disease gene likelihood. The scores in this section were developed without any disease specificity, and measure the depletion of LoF variation within a gene. Therefore, these scores act as generic proxies for the likelihood of a gene to be involved in *any* genetic disease. We downloaded these scores from the gnomADv.2.1.1 database on 2019-07-18[27].

ExAC_pLI measures the probability of a gene to be extremely intolerant of LoF variation. It's scale is ranges from 0 to 1, with genes over 0.9 representing those extremely intolerant to LoF variation and under higher constraint. It was developed based on data from approximately 60,000 exomes. *GnomAD_pLI* is similar but computed from an expanded data set of roughly 120,000 exomes.

oe_LoF measures the deviation of the number of observed LoF variants within a gene to the expected number. This score differs from the above two because its scale is reversed, with scores below 0.35 indicating extreme depletion of LoF variation and higher constraint[27]. This measure was recommended for identifying genes likely to be depleted of LoF variation because it is more interpretable than the pLI (i.e., a score of 0.4 indicates that 40% of the expected LoF variants within a gene have been observed), and better captures intermediate levels of haploinsufficiency. In addition, unlike the pLI, the o/e LoF score reports a confidence interval; the upper 95% confidence bound is recommended as the criterion to be compared to 0.35[27].

Evaluation. We plotted receiver operating characteristic (ROC) curves and precision-recall curves to assess recovery of the novel high-confidence (novel-HC) and SFARI high-confidence (SFARI-HC) ASD gene sets. When evaluating the ability of the scores to rank the novel-HCASD gene set, we removed the SFARI high-confidence ASD genes and other ASD genes used in the training of the ML algorithms from their gene rankings. This was done to ensure that the algorithms were not penalized for performance on ASD genes. The top ranks provided by the studies are their predictions as to the most likely ASD risk candidates. Therefore, the PR curves are the preferred evaluation metric because they are more sensitive to classification errors in the top ranks.

We calculated area under the ROC curve (AUROC) using the "auc" function with the "trapezoid" method from the DescTools R package[37] to account for ties in the rankings. We calculated precision at 20% recall (P20R) of total genes in the ASD gene sets, and precision at 43% recall (P43R) of total genes in the ASD gene sets. Precision at 20% recall was selected as a 'midrange' for display purposes, and has previously been used as a reported

point statistic in function prediction algorithm assessments[38]. The exception is for the pLI scores; as more than 20% of the high-confidence ASD genes have the maximum pLI score of 1.0, we report precision at 43% recall to have a consistent comparison for precision-recall across all studies.

We used 2,500 bootstrapped samples (gene-level) to calculate 95% confidence intervals for AUROC and precision-recall statistics. The bootstrapped samples were stratified, and done with replacement. This means that we sampled from the ASD gene sets and the rest of the scored protein coding genes separately in each of the 2,500 iterations to ensure balanced coverage, and that the same gene could be sampled more than once in each iteration. Therefore, in each bootstrapped sample, we kept only unique genes for evaluation. Studies whose performance measures confidence intervals did not overlap were considered significantly different from each other.

We measured the correlation between the ASD gene likelihood rankings provided by each study, and other metrics of interest using the Spearman correlation coefficient. The other metrics of interest included a multifunctionality rank, node degree of a BioGrid protein–protein interaction network, number of publications, and the SFARI numeric gene score. If a gene did not have a score for other metrics of interest, it was given a value of 0.0 for consistency with ASD gene prioritization studies.

Each method provided a cut off for their set of likely ASD genes, and we calculated the overlap in their top gene sets as their shared number of genes.

forecASD analysis. We obtained code for the forecASD classifier from https://github.com/LeoBman/forecASD, and re-ran it locally[11]. A minor difference from the preprint is the GitHub code uses a different version of the randomForest R package (version 4.6-14 vs 4.6-12 in the preprint)[11,39]. We refit the final ensemble model (03_ensemble_model.R) with different sets of the input features used in final ensemble model: the noClass (noC) model removed features from other classifiers listed above; the noClassPPI (noCP) model eliminated the other classifiers, and the STRING score; noClassPPIBS (noCPB) model eliminated the other classifiers, the STRING score, and the BrainSpan score; the PPIOnly (PPI) model only used the STRING score; and the BrainSpanOnly (BS) model only used the BrainSpan score. Feature importance was measured by mean decrease in accuracy and mean decrease in Gini node impurity. Mean decrease in accuracy is measured by randomly permuting each feature, and measuring the out-of-bag (cross-validation) accuracy of the resulting trees. Mean decrease in Gini measures how well the features can split the data from mixed labelled nodes into pure single class nodes. Brueggeman et al. did not provide code for their feature importance plots; we used "varImpPlot" from the randomForest package[11,39]. When rerunning their provided code, we found that two columns in their metadata had been mislabelled, D (DAMAGES) and D_ens (DAMAGES ensemble), necessitating re-labelling for plotting of feature importance. As for the other methods, we evaluated each model using the two ASD gene sets and with the same metrics described above.

Results

Method outline. We considered 13 ASD gene prioritization studies, and three measures of generic disease gene likelihood for evaluation. Each study provided scores for genes based on the author's assessment of their probability of contributing to ASD risk. We evaluated their ability to prioritize novel high-confidence and known high-confidence ASD genes using ROC and Precision-Recall curves, and 95% confidence intervals of AUROC and precision at 20% recall. Additionally, we looked at how the scores correlated with one another, and with other generic network features such as number of physical interaction partners to assess potential biases.

Systems-based GBA ML methods do not prioritize novel high-confidence ASD genes well compared to other disease gene prioritization methods. The first test we performed was investigating how well the GBA ML studies prioritized novel high-confidence ASD genes which were not used to build their predictions, and comparing their performance to genetics-based and generic approaches for disease gene prioritization (Fig. 1). Because genetic association remains the gold standard method for identifying genetic risk factors, our operating assumption was that in order to be considered a successful method, an ASD-specific GBA ML study should have comparable performance to the genetic association studies alone, and should outperform generic measures of disease gene likelihood. Lastly, the more recent genetic association studies (iHart and Satterstrom) were built up from the DeRubeis and Sanders studies in that they are using overlapping samples, and similar model parameters and variant classes in their TADA analyses. Therefore, we expected that the DeRubeis and Sanders studies would rank the novel-HC ASD genes at lower or borderline significant levels, and that the iHart and Satterstrom studies would show higher rankings of each other's hits.

We found that GBA ML studies had comparable performance to the generic measures of disease gene likelihood, as is shown by their overlapping 95% confidence intervals for precision at 20% recall (i.e., $P20R_{FRN} = 0.69–4.61\%$; $P20R_{ExAC\_pLI} = 0.69–1.92\%$) (Fig. 1D,H, Table 2). While the studies had high AUROC statistics with overlapping 95% confidence intervals, these metrics are somewhat misleading because they are not sensitive to false positive predictions in top rankings, which are most relevant for prioritization studies (Fig. 1B,F; Table 2). This finding suggests limited utility of GBA ML studies for ASD gene prioritization: use of a simple non-ASD specific measure constraint against LoF variation has comparable performance to complex ML approaches (Fig. 1A–H; Table 2).

The best performing GBA ML method was the hybrid genetics-GBA method forecASD ($P20R_{forecASD} = 4.63–11.21\%$), which had similar levels of performance to the genetic association studies developed before the iHart, Spark and Satterstrom studies (i.e., $P20R_{Sanders} = 3.49–10.54\%$) (Fig. 1C,D,K,L; Table 2). The other hybrid method, DAWN, has similar performance to other GBA ML studies, but this may be in part because they only provide predictions scores for roughly 10,000 genes in the genome (Fig. 1A–D, Table 2).

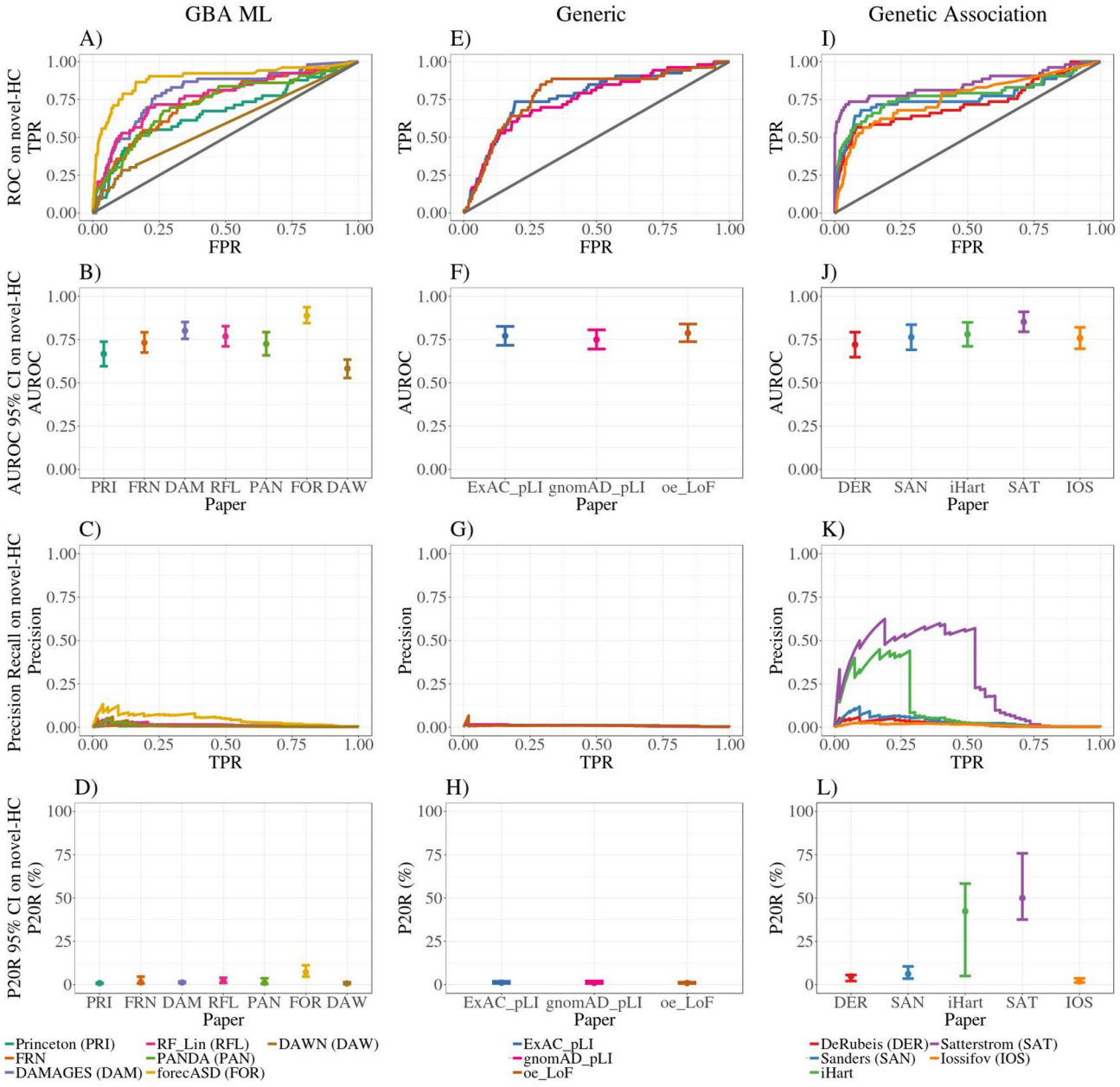

Figure 1. ROC, Precision-recall and summary statistics on novel-HC genes. Novel-HC genes were discovered by new TADA studies (iHart, Spark and Satterstrom), and most were not used in training of GBA ML studies. GBA ML studies have comparable performance to generic measures of disease gene likelihood (LoF constraint measures), with high AUROC (**A**,**B**,**E**,**F**), but low precision at 20% recall (**C**,**D**,**G**,**H**). GBA ML methods incorporating genetics information, particularly forecASD, have significantly better performance. Note that DAWN does not provide likelihood estimates for all protein-coding genes in the genome. Genetic association studies also show high AUROC (**I**,**J**). Previous TADA studies (DER, SAN) show moderate performance while the newer TADA studies are not performing at 100% precision (**L**). 95% confidence intervals were created from 2500 stratified bootstrap samples (**B**,**D**,**F**,**H**,**J**,**L**). *TPR* true positive rate,*FPR* false positive rate,*AUROC* area under the receiver operator curve, *P20R* precision at 20% recall.

It is important to note that the Satterstrom and iHart studies do not have particularly high performance by these benchmarks, despite being, in effect, a comparison of genetics findings to updated genetics findings (Fig. 1I–L; Table 2). In other words, the two recent TADA-based studies do not agree on what genes are significantly associated with ASD. Additionally, the previous TADA studies have some performance (i.e., $P20R_{Sanders} = 3.49–10.54\%$), which would suggest that they were able to identify some of the novel genes at marginal levels of significance, and with the accumulation of more data, these genes became significant in the newer studies (Fig. 1I–L; Table 2).

Paper	AUROC	P20R (%)	P43R (%)
Princeton	0.67 (0.60, 0.74)	0.88 (0.56, 1.00)	0.85 (0.47, 1.16)
FRN	0.73 (0.67, 0.79)	1.28 (0.69, 4.61)	0.87 (0.55, 1.33)
DAMAGES	0.80 (0.75, 0.85)	1.25 (0.87, 1.79)	1.40 (0.76, 1.90)
RF_Lin	0.77 (0.71, 0.83)	2.12 (0.96, 3.94)	1.49 (0.84, 1.87)
PANDA	0.72 (0.66, 0.79)	1.08 (0.44, 3.69)	0.68 (0.43, 0.96)
forecASD	0.89 (0.84, 0.94)	7.13 (4.63, 11.21)	5.56 (3.24, 9.38)
DAWN	0.58 (0.53, 0.63)	0.69 (0.37, 1.37)	0.53 (0.35, 0.72)[a]
ExAC_pLI	0.77 (0.72, 0.83)	1.14 (0.69, 1.92)	1.11 (0.78, 1.46)
gnomAD_pLI	0.75 (0.7, 0.81)	1.05 (0.70, 2.08)	1.15 (0.72, 1.39)
oe_LoF	0.79 (0.74, 0.84)	0.96 (0.64, 1.45)	1.10 (0.81, 1.37)
DeRubeis	0.72 (0.65, 0.79)	4.73 (2.08, 5.60)	2.37 (1.44, 3.77)
Sanders	0.76 (0.69, 0.84)	6.20 (3.49, 10.54)	2.80 (1.68, 6.33)
iHart	0.78 (0.71, 0.85)	42.35 (5.00, 58.33)	4.07 (1.69, 7.61)
Satterstrom	0.85 (0.79, 0.91)	50.00 (37.57, 75.79)	54.13 (24.41, 69.91)
Iossifov	0.76 (0.70, 0.82)	1.82 (1.24, 3.65)	1.83 (1.10, 2.68)

Table 2. Summary statistics on novel-HC genes. *AUROC* area under the receiver operator characteristic curve, *P20R* precision at 20% recall, *P43R* precision at 43% recall. Values in parentheses are the upper and lower 95% confidence interval bounds. [a]For P43R indicates a tie at recall of 20/43% of gene set.

GBA ML methods do not predict high-confidence ASD genes. We analyzed how well the GBA ML studies recovered SFARI high-confidence genes, many of which were used in the training of the ML algorithms, and compared the results to other methods for disease gene prioritization (Fig. 2). The genes in the SFARI-HC set were discovered by different genetic association studies, many of which were first identified by the DeRubeis and Sanders studies. Given the relationship between all the TADA studies, we expected the original and newer genetic association studies to highly prioritize SFARI high-confidence genes. The systems-based GBA ML studies used different subsets of SFARI high-confidence genes, and other ASD associated genes, during training. Therefore, we would expect that these studies should also highly prioritize SFARI-HC genes. We note that this is not a pure test of training performance because not all SFARI-HC genes were used during the training step. However, because the methods were developed at different times using different training gene sets, we opted for a consistent evaluation gene set across methods.

Our findings from this set of analyses parallel what we found for the novel-HC gene set. Mainly, we found that the GBA ML studies have comparable performance to the generic measures of disease gene likelihood with overlapping 95% confidence intervals for precision at 20% recall (i.e., $P20R_{FRN} = 5.33–12.06\%$; $P20R_{ExAC\_pLI} = 9.68–14.08\%$) (Fig. 2C,D,G,H; Table 3). RF_Lin did not provide predictions for their training genes, which partially explains its poorer performance relative to other studies ($AUROC_{RF\_Lin} = 0.44–0.55$; $P20R_{RF\_Lin} = 3.49–7.08\%$) (Fig. 2A–D; Table 3). Again we found that the genetics-GBA method forecASD had the best performance of the GBA ML studies with similar performance to genetic association studies (i.e., $P20R_{forecASD} = 38.23–77.32\%$; $P20R_{Sanders} = 53.74–95.00\%$; $P20R_{Satterstrom} = 65.48–100.00\%$) (Fig. 2C,D,K,L; Table 3). As per our expectations, the genetic association studies performed well in this training performance test (Fig. 2I–L; Table 3). These results show that systems-based GBA ML studies are providing little ASD-specific information above that provided by the generic measures of constraint against LoF variation (Fig. 2A–H; Table 3). Once again, these findings highlight the limited utility of the systems-based GBA ML studies for prioritizing ASD risk genes.

Low agreement between ML and genetic association. As previously discussed, GBA postulates that genes with shared associations are more likely to have shared functions or be involved in the same diseases. However, predictions can be driven by underlying multifunctionality bias whereby new functions are ascribed to genes that are well characterized because they are highly studied, and have a high number of association annotations[6,8]. In other words, we hypothesized that GBA methods using heterogeneous biological networks biased towards well-studied genes would tend to rank generically "disease-related" genes highly simply because they are well studied. Furthermore, because this ranking is not ASD-specific, it cannot readily identify novel and specific relationships. On the other hand, methods which do not recapitulate these generic rankings may perform badly because the main source of apparent performance of GBA methods is their ability to prioritize well studied genes ("multifunctionality bias" as per Gillis and Pavlidis).

We compared the genetic association and GBA ML scores to generic network features and generic gene annotations (Fig. 3). Our results show that some of the GBA ML studies are indeed biased. For example, the genetics-GBA study, forecASD, has moderate correlation with physical node degree ($R_{Spearman} = 0.34$) and number of publications ($R_{Spearman} = 0.34$), as do DAMAGES, RF_Lin and PANDA (Fig. 3). In the work of Gillis and Pavlidis, correlations of this magnitude were sufficient to explain a large fraction of predictive performance. In contrast, Princeton and FRN did not appear to show bias (i.e. $R_{S:FRN,pnd} = 0.16$, $R_{S:FRN,numPubs} = -0.03$). Furthermore, as expected the TADA analyses show little to no agreement with these generic features (i.e. $R_{S:iHart,pnd} = 0.05$,

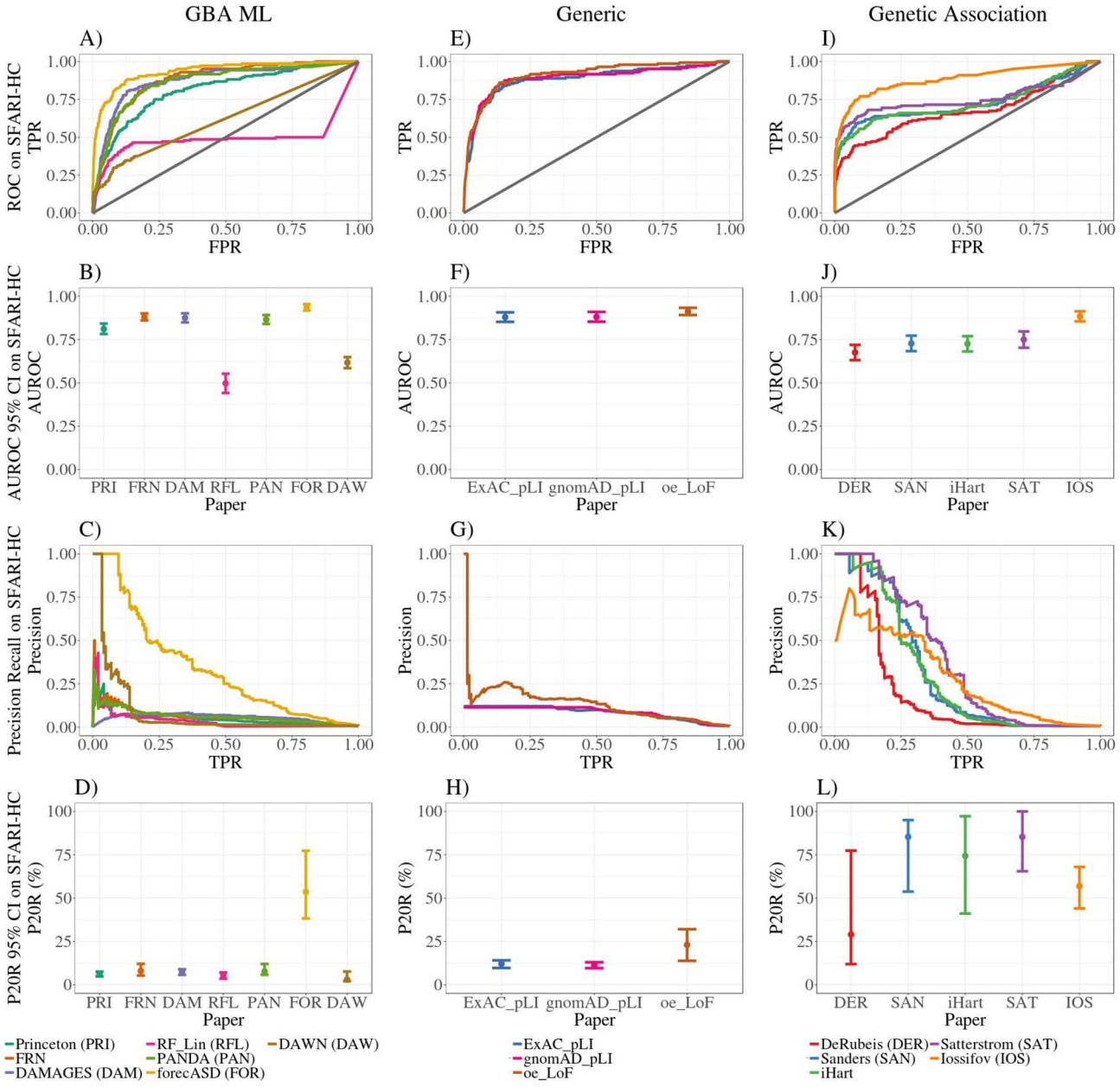

Figure 2. ROC, precision-recall and summary statistics for SFARI-HC genes. Many SFARI-HC genes were initially discovered by early TADA studies (DER, SAN), and used in training of GBA ML studies. Thus, this evaluation acts as a control experiment. Systems-based GBA ML studies have comparable performance to generic measures of disease gene likelihood (LoF constraint measures), with high AUROC (**A**,**B**,**E**,**F**), but low precision at 20% recall (**C**,**D**,**G**,**H**) on SFARI-HC genes. The GBA ML method with genetics information, forecASD, had significantly better performance compared to other GBA ML methods. Note that DAWN does not provide estimates for all protein-coding genes in the genome, and RF_Lin does not provide estimates for their training genes, which partially explain their poorer performance. Genetic association studies show high AUROC (**I**,**J**) and high precision at 20% recall (**K**,**L**). 95% confidence intervals were created from 2500 stratified bootstrap samples (**B**,**D**,**F**,**H**,**J**,**L**). *TPR* true positive rate, *FPR* false positive rate, *AUROC* area under the receiver operator curve, *P20R* precision at 20% recall.

$R_{S:iHart,numPubs} = 0.05$). These findings offer some explanation for the poor performance of the systems-based GBA ML studies when tested on novel genes. Methods which are not biased towards well studied genes, such as Princeton and FRN, may be performing poorly because there is no bias to drive apparent performance[6,8]. On the other hand, studies which are biased towards well studied genes, such as RF_Lin and PANDA, may be performing poorly because GBA is assigning new functions to highly connected genes in the network, and not learning ASD-specific information[6,8]. However, further work is required to delineate how multifunctionality is affecting

Study	AUROC	P20R (%)	P43R (%)
Princeton	0.81 (0.78, 0.84)	6.14 (4.79, 7.58)	4.36 (3.49, 6.02)
FRN	0.88 (0.86, 0.90)	8.00 (5.33, 12.06)[a]	5.79 (4.53, 6.37)
DAMAGES	0.87 (0.85, 0.90)	7.36 (5.70, 8.90)	7.13 (5.78, 8.78)
RF_Lin	0.50 (0.44, 0.55)	5.59 (3.49, 7.08)	2.58 (0.53, 3.82)
PANDA	0.87 (0.84, 0.89)	7.67 (5.76, 11.99)	5.78 (4.79, 6.80)
forecASD	0.94 (0.92, 0.95)	53.56 (38.23, 77.32)	31.80 (23.85, 40.22)
DAWN	0.62 (0.59, 0.65)	3.00 (2.07, 7.62)	1.70 (1.39, 2.02)[a]
ExAC_pLI	0.88 (0.85, 0.91)	12.06 (9.68, 14.08)[a]	9.57 (8.31, 12.07)[a]
gnomAD_pLI	0.88 (0.85, 0.91)	11.42 (9.59, 12.96)[a]	11.42 (9.29, 12.96)[a]
oe_LoF	0.91 (0.89, 0.93)	23.02 (13.80, 32.05)	16.30 (12.29, 18.80)[a]
DeRubeis	0.68 (0.63, 0.72)	29.00 (11.89, 77.38)	4.42 (1.68, 6.34)
Sanders	0.73 (0.68, 0.77)	85.29 (53.74, 95.00)	12.45 (6.68, 20.99)
iHart	0.73 (0.68, 0.77)	74.36 (41.06, 97.22)	14.28 (5.68, 28.05)
Satterstrom	0.75 (0.70, 0.80)	85.29 (65.48, 100.00)	31.51 (16.37, 55.71)
Iossifov	0.88 (0.86, 0.91)	56.85 (43.99, 68.00)[a]	31.00 (18.04, 44.85)

Table 3. Summary statistics on SFARI-HC genes. Headings are per Table 2, and [a]means a tie at recall of 20/43% of gene set.

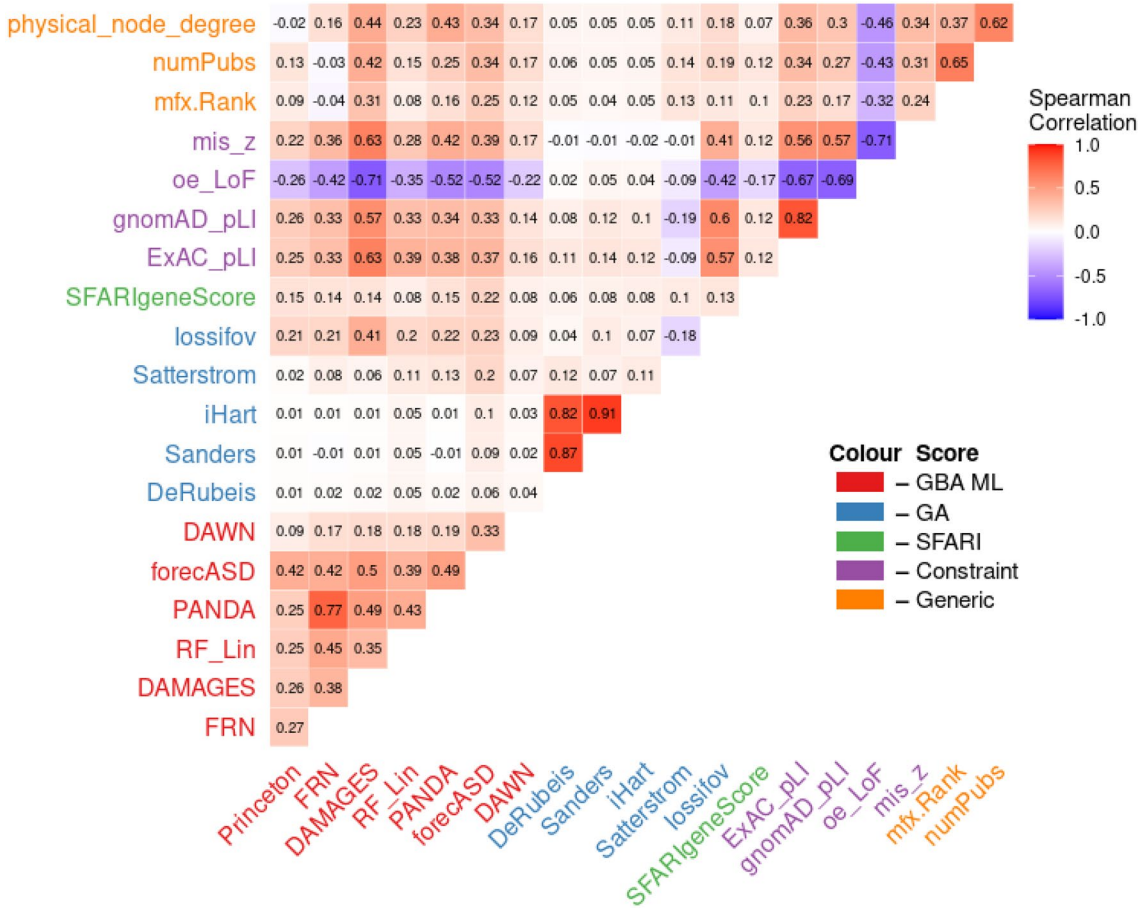

Figure 3. Correlations among gene rankings. Values are Spearman correlations. Notable patterns include low correlations between genetic association methods, ML methods and other network features such as node degree and publication number; increased correlation between select ML methods and other network features; low correlation between Satterstrom score and pLI despite its incorporation in the statistical framework; low correlation between SFARI gene score and generic gene annotations.

	Princeton	FRN	DAMAGES	RF_Lin	forecASD	DAWN	DeRubeis	Sanders	iHart	Satterstrom	Iossifov	exac_pLI	gnomad_pLI	oe_LoF	SFARI-HC	NOVEL-HC
Princeton	***2467***	1014	70	842	***831***	38	16	28	34	55	108	***1045***	1026	997	***83***	29
FRN		**2111**	74	985	842	42	20	35	36	62	121	1093	1023	1031	99	27
DAMAGES			**117**	89	90	9	4	7	7	12	25	116	115	116	16	0
RF_Lin				**2089**	854	30	6	5	8	43	129	1436	1335	1378	59	33
forecASD					***1803***	63	33	65	65	89	187	1109	1044	1052	***118***	43
DAWN						**127**	11	17	16	19	27	50	53	55	20	2
DeRubeis							**33**	26	23	21	24	26	24	25	23	0
Sanders								**65**	52	39	45	46	44	44	39	0
iHart									**69**	36	40	45	41	42	35	16
Satterstrom										**102**	53	89	81	83	52	32
Iossifov											**239**	204	198	203	66	6
exac_pLI												***3220***	2475	2477	***121***	37
gnomad_pLI													**3046**	2838	125	35
oe_LoF														**2957**	124	37
SFARIHC															***144***	7
NOVELHC																**60**

Table 4. Overlap of top ranked ASD genes from each study. Numbers on the diagonal represent the number of ASD genes predicted. Values highlighted in bold italics are discussed in the main text.

each study. Lastly, high agreement between generic measures of constraint and systems-based GBA ML studies further suggests that their predictions are generic, and not specific to ASD (i.e. $R_{S:forecASD,ExAC\_pLI} = 0.37$) (Fig. 3).

The lack of agreement between Satterstrom and the other TADA analyses further highlights the non-equivalence between the genetic association studies and the need for TADA model validation (i.e. $R_{S:Satterstrom,\ DeRubeis} = 0.12$) (Fig. 3). Satterstrom directly incorporates ExAC pLI into its TADA model, however, it displays little correlation with pLI ($R_{S:ExAC\_pLI} = -0.09$), and low to moderate agreement with other generic network features (i.e. $R_{S:Satterstrom,numPubs} = 0.14$). While it is possible that using pLI incorporated some generic disease gene bias into Satterstrom, the direct effects of pLI on the score are likely complex and non-linear due to TADA's approach of collapsing multiple pieces of information to derive the per-gene association scores[24,29]. Therefore, Spearman correlation may not adequately capture the relationship.

Iossifov is the genetic association study with the highest agreement with generic gene annotations. Notably, it has high correlation with pLI ($R_{S:ExAC\_pLI} = 0.60$). Iossifov is the most similar to pLI in its construction: both scores attempt to quantity the deviation of the observed number of LoF variants from an expectation of LoF variation derived from complex models incorporating rates synonymous variation, among many other factors[25–27]. The Iossifov score is ASD-specific because they incorporate an estimate of the number of causal ASD genes, and the observed load of LoF variation in ASD probands, whereas the LoF constraint scores were developed without any disease specificity[25].

Lastly, the presence of some agreement between the SFARI gene score and generic measures of constraint and generic network features further demonstrate that high-confidence ASD genes have a relationship with constraint scores in that many confirmed ASD genes are constrained against LoF variation (pLI > 0.9), and that they are likely well-studied genes (Fig. 3). As genes are associated with disease, they become more studied, and they usually collect a high number of functional and physical annotations. While these annotations may be biologically relevant, they can impact GBA ML studies in a negative way by increasing the effects of multifunctionality, as discussed below.

Overlap in the subset of genes identified as likely ASD candidate risk genes. We next examined whether overlap among top ranked genes may still exist despite low overall correlation (Table 4; Fig. 3). For example, while forecASD and Princeton share 831 genes in their top rankings, forecASD is able to recover 118/144 SFARI high-confidence genes from a potential 1,803 compared to the 83/144 recovered from a potential 2,467 by Princeton (Table 4, bold italics highlights). Likewise, Princeton and ExAC pLI share 1,045 genes in their top rankings, but ExAC pLI captures 121/144 from a potential 3,220 (Table 4, bold italics highlights). This again shows that the systems-based ML studies are not performing as well as those with ASD-specific genetics information, and that they are providing little ASD-specificity above that provided by the generic measures of constraint.

We noted that multiple genes identified in previous TADA analyses are no longer statistically significantly associated with ASD in Satterstrom (i.e., only 36 of iHart's significant findings are in Satterstrom's 102) (Table 4, Supplementary Table S1), and that the TADA analyses only share 17 genes in their top findings (Supplementary Table S2). There are seven genes found by recent TADA studies, which at the time of their publications, were considered novel findings; however, they are now considered to be SFARI-HC genes (Supplementary Table S3). The differences in overlap of top findings between the TADA analyses further highlights that the differences between the underlying models need to be investigated more closely.

Feature	BrainSpan score	STRING score	Other classifiers	De Rubeis	Sanders
Version					
forecASD/Redo	✓	✓	✓	✓	✓
noClass (NoC)	✓	✓		✓	✓
noClassPPI (NoCP)	✓			✓	✓
noClassPPIBS (NoCPB)				✓	✓
PPIOnly (PPI)		✓			
BrainSpanOnly (BS)	✓				

Table 5. Features included in the different forecASD analyses.

Feature importance in forecASD algorithm. While forecASD had significantly better performance with SFARI and novel high-confidence ASD risk genes compared to other systems-based GBA ML studies, it is performing with low precision ($P20R_{forecASD}$ = 4.63–11.21%) (Tables 2, 3). If a GBA ML method is to be considered successful, it must be able to generalize to new data, and highly rank true positives. To understand the driving force behind forecASD performance, we examined its performance using training feature sets made up of different combinations of the original features used in the model (Table 5).

Other classifiers included as features in this analysis were DAWN, Princeton and DAMAGES. The features from the DAWN algorithm were their list of risk ASD genes (rASD), and network score and minimum FDR. The features from the DAMAGES algorithm were the D and Ensemble scores. There were four FDR values from Sanders: tada_asc + ssc + del (most thorough with exome data and small de novo dels), tada_asc + ssc (both sources of exome data), tada_asc (ASC exome only), and tada_ssc (SSC exome only). The per-gene Bayes Factor from DeRubeis was also included (TADA_BF).

We evaluated the forecASD models on both novel-HC and SFARI-HC gene sets (Fig. 4). In both evaluations, we found that forecASD versions incorporating genetics information had significantly better performance than the versions using only protein–protein interaction or BrainSpan gene-expression data (Fig. 4; Supplementary Tables S4, S5). Notably, we found that the forecASD version incorporating only genetics data (noClassPPIBS) had overlapping 95% confidence intervals for precision at 20% recall of novel and SFARI-HC genes with the full forecASD version (i.e. $P20R_{novelHC:forecASD}$ = 4.63–11.21%; $P20R_{novelHC:noClassPPIBS}$ = 2.88–9.73%) (Fig. 4; Supplementary Tables S4, S5). These results further show that forecASD performance is driven by genetic association data (Fig. 5; we note that in the forecASD preprint, STRING was considered the most informative feature, but we were unable to reproduce this result with their code despite reproducing their classification results; we believe it is an error). Taken with our other findings, the implication is that supplementing ASD-specific genetics data with heterogeneous biological data is likely not useful for disease gene discovery, especially when considering the unknown reliability and biases within the data.

Discussion

Our investigation has shown that GBA ML methods that do not use ASD genetics information have limited utility. This appears to be because non-genetic association data provides little to no useful information above that provided by generic measures of disease gene likelihood. This finding likely has implications for other attempts to prioritize genes for complex human genetic diseases: using heterogeneous biological network data likely has diminishing returns due to poor real-world performance and biases.

Non-equivalence of genetic association studies. A complication of our study was that the ASD genetic association studies agree poorly, even when analyzing heavily overlapping sets of subjects. For example, the recent work of Satterstrom et al., fails to replicate many of the genes considered significant ASD risk genes reported by De Rubies et al., despite using essentially all the data from De Rubies et al. The reason for this is not clear. One possibility is that many of the genes reported by De Rubies et al. were false positives uncovered by Satterstrom having more data. Arguing against this, all of the genes identified by De Rubies et al. were considered high-confidence ASD genes by SFARI Gene at the time of our analysis. Another likely culprit is that the methods for detecting statistical association of very rare de novo variants with phenotypes were changed substantially in Satterstrom et al.[21]. We note that iHart uses the same TADA model as Sanders, but with an increased number of samples from multiplex families, and Satterstrom and iHart show the highest agreement in ranking (R_S = 0.92) and overlap of significant genes (52/80). For this reason we consider it likely that the incorporation of pLI and MPC in Satterstrom et al. has a larger impact on the results than changes to the underlying data. Regardless, it is a caveat of our study that there is apparently no universally trustable gold standard set of ASD genes. The impact of this on the interpretation of our study is limited, because as we show, the set of genes used for evaluation does not change the performance outcomes substantially.

ML methods are comparable to generic measures of LoF constraint. Proposed use cases of the GBA ML studies include prediction and/or prioritization of ASD risk genes, framing WES/WGS results for further exploration in resequencing or mechanistic studies, and/or uncovering new and delineating possible pathways implicated in ASD etiology[1,9–11,22,23]. Overall, for GBA ML study to be considered successful in identifying novel ASD genes, it should highly prioritize known ASD genes and provide additional, specific and unbiased

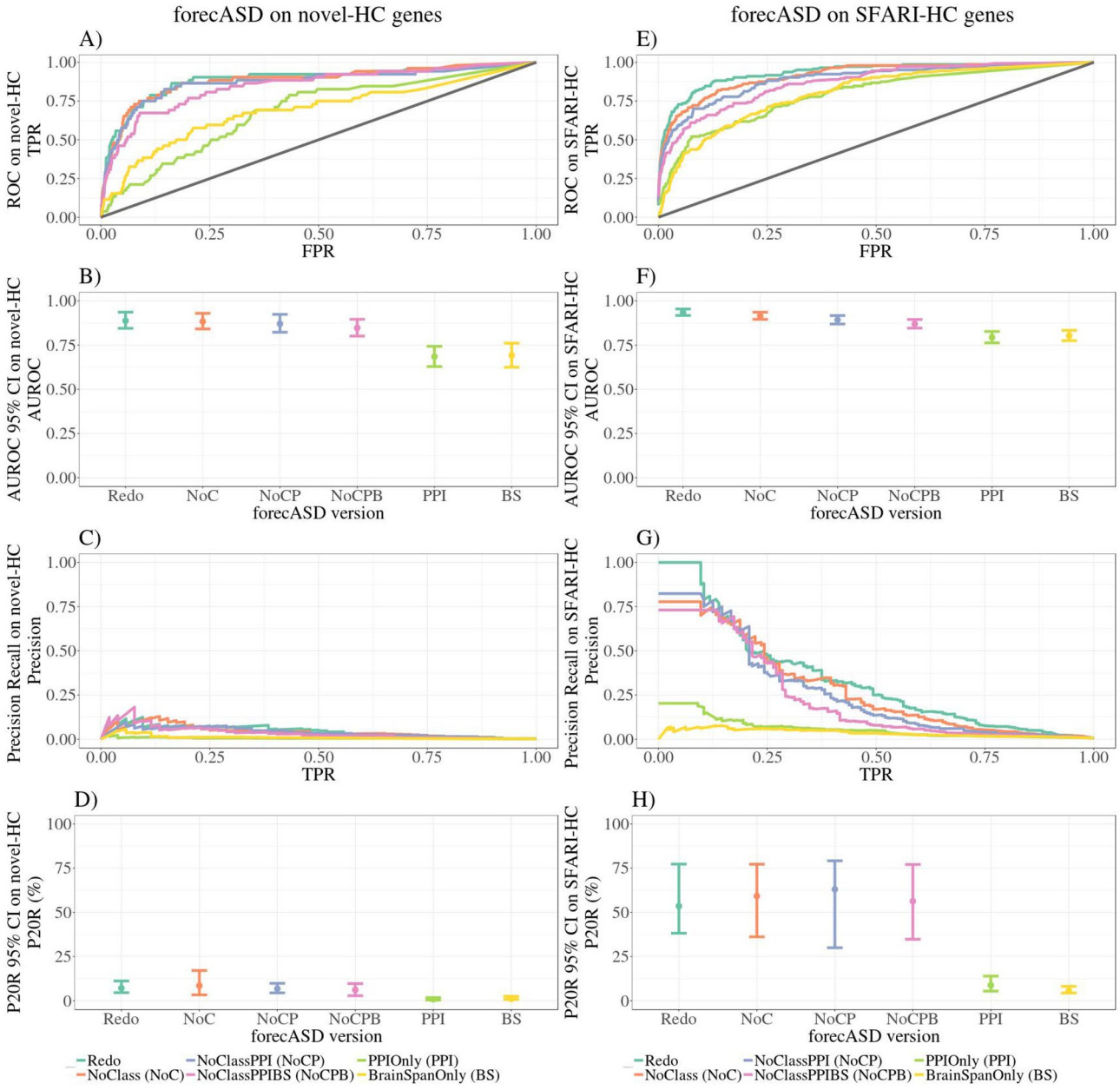

Figure 4. ROC, precision-recall and summary statistics for forecASD versions on novel-HC and SFARI-HC genes. Versions without genetics information, PPIOnly and BrainSpanOnly, show significantly worse performance in both tests. 95% confidence intervals were created from 2500 stratified bootstrap samples. *TPR* true positive rate, *FPR* false positive rate, *AUROC* area under the receiver operator curve, *P20R* precision at 20% recall.

predictions above that which could be obtained from generic measures of constraint. We have shown that the systems-based ML studies failed to do so.

As discussed previously, we expect that methods employing GBA would tend to rank generically "disease-related" genes highly because they are well studied, highly annotated and highly connected within networks. Thus, methods biased towards generic rankings, such as PANDA and DAMAGES, likely struggle to identify novel and disease-specific relationships (Fig. 3). Conversely, GBA methods which are not biased towards generic rankings, such as Princeton and FRN, may perform badly because the main source of apparent performance of GBA methods is their ability to prioritize well studied genes ("multifunctionality bias" as per Gillis and Pavlidis)(Fig. 3). While we found that, overall, the system-based GBA studies perform with low precision, we also found that two studies, Princeton and FRN, are not biased towards well studied, highly annotated, highly connected genes (Tables 3, 4, Fig. 3). These two studies built complex functional interaction networks from multiple data types, including protein–protein interaction and gene expression data. They used Gene Ontology annotations to define "gold standards" of functional relationships and Bayesian frameworks for weighting and data integration[1,9,40,41]. Their poor performance could be due to their GO functional categorization not aligning well with the multiple

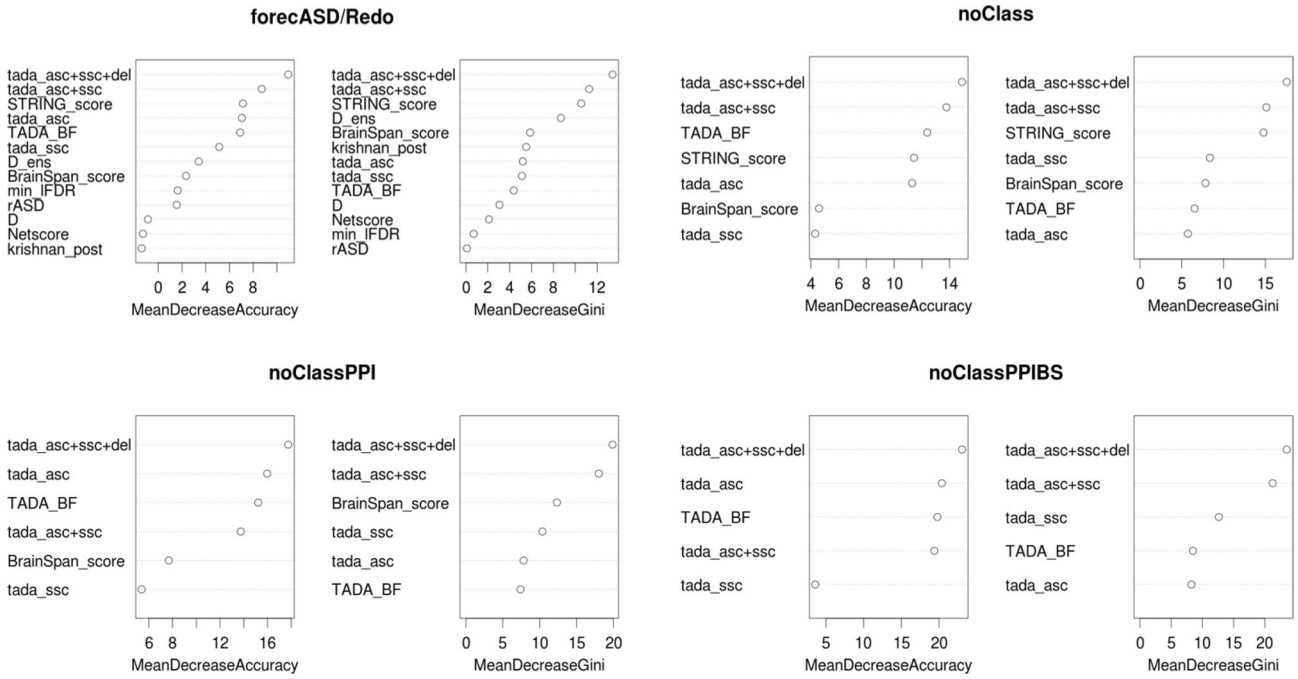

Figure 5. Genetics features drive forecASD performance. The most comprehensive score from the Sanders, tada_asc + ssc + del, was ranked as the most important feature for discerning ASD from non-ASD training genes in each version, followed by other TADA-based statistics.

biological data types and/or not providing useful ASD-specific information. However, it is much more likely that these studies do not perform well because due to the effects, or lack thereof, of multifunctionality bias[6,8].

While further investigation into each study is required to delineate how multifunctionality bias is affecting their performance, a consistent finding across the GBA ML studies was high agreement between the studies and generic measures of constraint against LoF variation (Fig. 3). We have confirmed that measures of constraint against LoF variation are able to identify ASD genes, albeit with low precision, and that they agree with generic network features and annotations (Tables 3, 4; Fig. 3). Many previous studies have found ASD genes, particularly those with high numbers of recurrent de novo variants, to be enriched for genes under high evolutionary constraint, and LoF constraint has previously been reported to be positively correlated with the number of physical interaction partners[18,19,24,26,27]. From this, we can confirm that measures of constraint against LoF variation measure generic susceptibility to disease, and that high constraint does not automatically guarantee a particular disease status, necessitating incorporation with data specific to the disease at hand to increase precision[26,27,42]. Furthermore, while these measures are also correlated with numbers of interaction partners, functions and publications, they may point towards more biologically relevant information, such as the ability of a gene to influence different phenotypic traits, rather than number of connection partners based on network structure ("hubness")[6,8].

The implication of this analysis is that supplementing ASD-specific genetics information with measures of constraint may provide a more fruitful avenue forward compared to creating GBA ML methods using biased biological networks. We can see this already being done by the Satterstrom TADA analysis by their incorporation of the pLI and MPC into the method in attempts to provide more detailed information about variant classes with higher burden in ASD probands[24].

In summary, our results demonstrate that despite using complex data and sophisticated algorithms, ASD GBA ML methods fail to outperform generic measures of disease gene likelihood such as pLI. We suspect this is likely to generalize to the study of other genetic disorders.

Code availability

Code and publicly available raw data re-analyzed in this work are available at https://github.com/margotgunning/ASDMachineLearning.

Received: 22 January 2021; Accepted: 16 July 2021
Published online: 05 August 2021

References

1. Krishnan, A. *et al.* Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder. *Nat. Neurosci.* **19**, 1454–1462 (2016).
2. Lee, I., Blom, U. M., Wang, P. I., Shim, J. E. & Marcotte, E. M. Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome Res.* **21**, 1109–1121 (2011).
3. Moreau, Y. & Tranchevent, L.-C. Computational tools for prioritizing candidate genes: Boosting disease gene discovery. *Nat. Rev. Genet.* **13**, 523–536 (2012).

4. Zhang, Y., Chen, Y. & Hu, T. PANDA: Prioritization of autism-genes using network-based deep-learning approach. *Genet. Epidemiol.* https://doi.org/10.1002/gepi.22282 (2020).
5. Gillis, J. & Pavlidis, P. "Guilt by association" is the exception rather than the rule in gene networks. *PLoS Comput. Biol.* **8**, e1002444 (2012).
6. Gillis, J. & Pavlidis, P. The impact of multifunctional genes on 'guilt by association' analysis. *PloS One* **6**, e17258 (2011).
7. Lanckriet, G.R.G. *et al.* Kernel-based data fusion and its application to protein function prediction in yeast. *Pac Symp Biocomput.* 300–311 (2004).
8. Pavlidis, P. & Gillis, J. Progress and challenges in the computational prediction of gene function using networks. *F1000Research* **1**, 14 (2012).
9. Duda, M. *et al.* Brain-specific functional relationship networks inform autism spectrum disorder gene prediction. *Transl. Psychiatry* **8**, 1–9 (2018).
10. Lin, Y., Rajadhyaksha, A. M., Potash, J. B. & Han, S. A machine learning approach to predicting autism risk genes: Validation of known genes and discovery of new candidates. *bioRxiv* https://doi.org/10.1101/463547 (2018).
11. Brueggeman, L., Koomar, T. & Michaelson, J. J. Forecasting risk gene discovery in autism with machine learning and genome-scale data. *Sci. Rep.* **10**, 1–11 (2020).
12. de la Torre-Ubieta, L., Won, H., Stein, J. L. & Geschwind, D. H. Advancing the understanding of autism disease mechanisms through genetics. *Nat. Med.* **22**, 345–361 (2016).
13. Iossifov, I. *et al.* De novo gene disruptions in children on the autistic spectrum. *Neuron* **74**, 285–299 (2012).
14. Neale, B. M. *et al.* Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* **485**, 242–245 (2012).
15. O'Roak, B. J. *et al.* Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature* **485**, 246–250 (2012).
16. Sanders, S. J. *et al.* De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485**, 237–241 (2012).
17. Abrahams, B. S. *et al.* SFARI Gene 2.0: A community-driven knowledgebase for the autism spectrum disorders (ASDs). *Mol. Autism* **4**, 36 (2013).
18. De Rubeis, S. *et al.* Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* **515,** 209–215 (2014).
19. Ruzzo, E. K. *et al.* Inherited and de novo genetic risk for autism impacts shared networks. *Cell* **178**, 850-866.e26 (2019).
20. Feliciano, P. *et al.* Exome sequencing of 457 autism families recruited online provides evidence for novel ASD genes. *npj Genom. Med.* **4,** 1–14(2019).
21. Sanders, S. J. *et al.* Insights into autism spectrum disorder genomic architecture and biology from 71 risk loci. *Neuron* **87**, 1215–1233 (2015).
22. Zhang, C. & Shen, Y. A cell type-specific expression signature predicts haploinsufficient autism-susceptibility genes. *Hum. Mutat.* **38**, 204–215 (2017).
23. Liu, L. *et al.* DAWN: A framework to identify autism genes and subnetworks using gene expression and genetics. *Mol. Autism* **5**, 22 (2014).
24. Satterstrom, F. K. *et al.* Large-scale exome sequencing study implicates both developmental and functional changes in the neurobiology of autism. *Cell* **20**, 20 (2020).
25. Iossifov, I. *et al.* Low load for disruptive mutations in autism genes and their biased transmission. *Proc. Natl. Acad. Sci.* **112**, E5600–E5607 (2015).
26. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
27. Karczewski, K. J. *et al.* Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *bioRxiv* https://doi.org/10.1101/531210 (2019).
28. Sayers, E. W. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* https://doi.org/10.1093/nar/gkz899 (2019).
29. 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).
30. Levy, D. *et al.* Rare de novo and transmitted copy-number variation in autistic spectrum disorders. *Neuron* **70,** 886–897 (2011).
31. Iossifov, I. *et al.* The contribution of de novo coding mutations to autism spectrum disorder. *Nature* **515**, 216–221 (2014).
32. Dong, S. *et al.* De novo insertions and deletions of predominantly paternal origin are associated with autism spectrum disorder. *Cell Rep.* **9,** 16–23 (2014).
33. Pinto, D. *et al.* Convergence of genes and cellular pathways dysregulated in autism spectrum disorders. *The American Journal of Human Genetics* **94,** 677–694 (2014).
34. Krumm, N. *et al.* Excess of rare inherited truncating mutations in autism. *Nat. Genet.* **47**, 582–588 (2015).
35. Sunkin, S. M. *et al.* Allen Brain Atlas: An integrated spatio-temporal portal for exploring the central nervous system. *Nucleic Acids Res.* **41**, D996–D1008 (2013).
36. Szklarczyk, D. *et al.* STRING v11: Protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**, D607–D613 (2019).
37. Andri Signorell et al. *DescTools: Tools for Descriptive Statistics.* (2019).
38. Peña-Castillo, L. *et al.* A critical assessment of Mus musculus gene function prediction using integrated genomic evidence. *Genome Biol.* **9**, S2 (2008).
39. Liaw, A. & Wiener, M. Classification and regression by randomforest. *R News* **2**, 18–22 (2002).
40. Greene, C. S. *et al.* Understanding multicellular function and disease with human tissue-specific networks. *Nat. Genet.* **47**, 569–576 (2015).
41. The Gene Ontology Consortium. The gene ontology resource: 20 years and still GOing strong. *Nucleic Acids Res.* **47**, D330–D338 (2019).
42. Cummings, B. B. *et al.* Transcript expression-aware annotation improves rare variant discovery and interpretation. *bioRxiv* https://doi.org/10.1101/554444 (2019).

Author contributions

M.G. performed research. P.P. provided supervision. Both authors contributed to the study design and wrote the manuscript.

Funding

This work was supported by the SFARI Foundation and an NSERC Discovery Grant. MG was supported in part by a scholarship from the UBC Bioinformatics Graduate Program via the NSERC CREATE program in High-Dimensional Biology and a CGS-M scholarship.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-95321-y.

Correspondence and requests for materials should be addressed to P.P.

Reprints and permissions information is available at www.nature.com/reprints.

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