frontiers in Psychiatry

ORIGINAL RESEARCH
published: 12 May 2021
doi: 10.3389/fpsyt.2021.637022

Identifying Subgroups of Patients With Autism by Gene Expression Profiles Using Machine Learning Algorithms

Ping-I Lin[1,2]*, *Mohammad Ali Moni*[1], *Susan Shur-Fen Gau*[3] *and Valsamma Eapen*[1,2]

[1] *School of Psychiatry, The University of New South Wales, Sydney, NSW, Australia,* [2] *South Western Sydney Local Health District, Liverpool, NSW, Australia,* [3] *Department of Psychiatry, National Taiwan University Hospital and College of Medicine, Taipei, Taiwan*

Objectives: The identification of subgroups of autism spectrum disorder (ASD) may partially remedy the problems of clinical heterogeneity to facilitate the improvement of clinical management. The current study aims to use machine learning algorithms to analyze microarray data to identify clusters with relatively homogeneous clinical features.

Methods: The whole-genome gene expression microarray data were used to predict communication quotient (SCQ) scores against all probes to select differential expression regions (DERs). Gene set enrichment analysis was performed for DERs with a fold-change >2 to identify hub pathways that play a role in the severity of social communication deficits inherent to ASD. We then used two machine learning methods, random forest classification (RF) and support vector machine (SVM), to identify two clusters using DERs. Finally, we evaluated how accurately the clusters predicted language impairment.

Results: A total of 191 DERs were initially identified, and 54 of them with a fold-change >2 were selected for the pathway analysis. Cholesterol biosynthesis and metabolisms pathways appear to act as hubs that connect other trait-associated pathways to influence the severity of social communication deficits inherent to ASD. Both RF and SVM algorithms can yield a classification accuracy level >90% when all 191 DERs were analyzed. The ASD subtypes defined by the presence of language impairment, a strong indicator for prognosis, can be predicted by transcriptomic profiles associated with social communication deficits and cholesterol biosynthesis and metabolism.

Conclusion: The results suggest that both RF and SVM are acceptable options for machine learning algorithms to identify AD subgroups characterized by clinical homogeneity related to prognosis.

Keywords: autism spectrum disorder, genomics, social cognition, language, machine learning

OPEN ACCESS

Edited by:
Claudio Toma,
Severo Ochoa Molecular Biology
Center (CSIC-UAM), Spain

Reviewed by:
Richard Woodman,
Flinders University, Australia
Noèlia Fernàndez-Castillo,
Centre for Biomedical Network
Research (CIBER), Spain

****Correspondence:***
Ping-I Lin
daniel.lin@unsw.edu.au

Specialty section:
This article was submitted to
Behavioral and Psychiatric Genetics,
a section of the journal
Frontiers in Psychiatry

Received: *02 December 2020*
Accepted: *13 April 2021*
Published: *12 May 2021*

Citation:
Lin P-I, Moni MA, Gau SS-F and
Eapen V (2021) Identifying Subgroups
of Patients With Autism by Gene
Expression Profiles Using Machine
Learning Algorithms.
Front. Psychiatry 12:637022.
doi: 10.3389/fpsyt.2021.637022

INTRODUCTION

Clinical heterogeneity is a norm rather than an exception in autism spectrum disorder (ASD), a complex neurodevelopmental disorder characterized by social communication deficits and stereotyped behaviors. Heterogeneous clinical features pose great challenges for diagnostics for ASD, such that children who receive a diagnosis of ASD have a range of vastly different presentations, trajectories, and outcomes. Further, the diagnostic criteria for ASD have been continuously revised through different editions of the Diagnostic and Statistical Manual for Mental Disorders (DSM), particularly the substantial changes in the 5th edition (DSM 5) where the wide range of clinical presentations have been brought together under a single ASD diagnostic entity (1). The current diagnostic system lacks an evidence-based approach and we urgently require a scientific approach to understanding which interventions are likely to be the most effective for which child with ASD (2). Accumulating evidence has shown that no pharmaceutical treatments have thus far been conclusively found to substantially reduce core symptoms of ASD (3). This may be partially attributable to the fact that most clinical trials did not take clinical heterogeneity into account and hence treatment effects remain equivocal. Variable clinical presentations may reflect different biological pathways. The identification of biomarkers for etiological pathways may hence hold the key to unraveling mechanisms underlying the variation in clinical presentations (4), which in turn may pave the way for personalized medicine in ASD.

The goal of identifying biomarkers for clinical homogeneity is to tackle challenges arising from clinical heterogeneity for research on either etiologies or treatments of ASD. One of the most extensively studied biomarkers for ASD is genetic factors. There are two different strategies to evaluate genetic markers for clinical heterogeneity: bottom-up and top-down approaches. The bottom-up approach is to define a priori subgroups using phenotypic information under the premise that some genetic loci are more likely to contribute to susceptibility to disease in a certain subgroup(s). Therefore, stratifying the population by a clinical marker (e.g., age of onset) will allow investigators to detect genetic association effects that are larger in certain subgroups. The top-down approach, on the other hand, is based on the premise that certain genetic markers can be used to distinguish subgroups, each of which is characterized by relatively homogeneous phenotypic profiles underscored by similar biological pathways—which imply similar therapeutic targets. Many of the earlier genome-wide linkage or association studies that aimed to unravel genetic underpinnings of clinical heterogeneity chose the second approach, which is to identify genetic markers associated with the phenotype defined by strict diagnostic criteria of ASD (5–7). Using the data from the Autism Diagnostic Interview-Revised (ADI-R) (8), Autism Diagnostic Observation Schedule (ADOS) (9), Vineland Adaptive Behavior Scales (VABS) (10), head circumferences, and ages at assessment as classifying variables, Veatch et al. identified clinically similar subgroups of individuals with ASD and found that the genotypes were more similar within subgroups compared to the whole population—the proportion of the total genetic variance contained in a subpopulation was 0.17 (11). However, this approach has not yielded highly replicable and clinically meaningful findings that can lead to conclusively validated etiological factors yet (12). Furthermore, another genome-wide association study of 2,576 families with ASD probands did not discover any genetic loci that exert a larger effect on the disease risk in subpopulations defined by the diagnosis, IQ, and symptom profiles; heritability estimates were also found to be similar in subpopulations to the whole population (13). Results from different groups show that an increased number of gene-truncating variants (highly pathogenic variants) may exert a considerable impact on IQ in ASD patients (14, 15); and higher burden of this pool of variants in ASD patients correlates with lower IQ scores. These studies showed that genomic approaches are able to identify genetic loci exerting larger effect on disease risk or associated with clinical outcomes, although genetic loci must be considered in an additive manner.

The top-down approach often starts with a few selected genetic loci associated with the disease. Despite fruitful findings from genome-wide and candidate gene-based association studies, few genetic loci can be used to improve accuracy in diagnostics or optimize treatment effects of therapeutics for ASD. Nevertheless, several genetic markers are found to be useful for classifying patients with ASD into relatively homogeneous subgroups. For example, Bruining et al. reported prominently higher symptom homogeneity in both the ASD group with 22q11 deletions and ASD group with Klinefelter Syndrome (KS), compared to the heterogeneous ASD sample (16). Transcriptomic profiles have also been used to identify genetic markers to classify individuals with ASD. Hu and Lai used the gene expression data to identify a subset of the "classifier" genes, which resulted in an overall class prediction accuracy of nearly 82%, ∼90% sensitivity, and 75% specificity (17). These results seem to demonstrate the value of the top-down approach.

Determining subgroups of ASD is challenging mainly because of the complexity of biological factors and clinical heterogeneity inherent to ASD. To tackle these challenges, one of the solutions is to implement state-of-the-art statistical methods that can efficiently parse through high-dimensionality data, such as machine learning (ML) algorithms, to differentiate subgroups with meaningful etiological, diagnostic, or therapeutic implications (18). Previous evidence suggests that ML algorithms can be used to reduce the number of items from standardized ASD assessment tools to make the assessment more efficient (19) and predict clinical outcomes with ASD phenotypic clusters and genetic data of copy number variations (20). The ML algorithms appear to be useful to identify phenotypic clusters as ASD subgroups that can predict clinical outcomes (21). In the current study, we attempted to implement the ML algorithms in the context of the bottom-up approach, which is to identify clusters using genomic information, and then explore the relationship between the genomic clusters and clinical features of ASD.

METHODS

Data Collection

The goal of the current study is to evaluate whether transcriptomic profiles correlated with clinical severity levels of ASD—which were measured with social communication questionnaire (SCQ) (22), can classify patients into two subgroups defined on the basis of language (i.e., the subgroup with language impairment vs. the subgroup without language impairment). The language function is considered as a strong predictor for cognitive ability and adaptive skills in children with ASD (23), and its variation within ASD patients is influenced by genetic factors (24–26). The presence of language impairment was defined as the total score (verbal) >10 in the section of Qualitative Abnormalities in Communication in Autism Diagnostic Interview-Revised (ADI-R) (27). A total of 31 children diagnosed with ASD were recruited in the current study. The clinical diagnoses were made by Gau, a board-certified child psychiatrist, and confirmed by the ADI-R interview with the parents. The Chinese version of the ADI-R been approved by the Western Psychological Services in May 2007 (28) mRNA was extracted from lymphoblastoid cell lines (LCL) of all participants. The microarray experiment was performed at the Core Laboratory of National Taiwan University Hospital in Taiwan, using the Affymetrix Human Genome U133 Plus 2.0 Array (Affymetrix Inc., Santa Clara, CA, USA). The experimental procedures followed the protocols provided by the manufacturer. The study was conducted with the ethical approval by the Institutional Research Board at National Taiwan University Hospital in Taiwan.

Statistical Methods

Transcriptome-Wide Association Analysis

We evaluated the integrity of 28S and 18S rRNA by electrophoresis of 2 mg of total RNA in 1.2% agarose gel containing 2.2 M formaldehyde and in a running buffer containing 0.2 M of MOPS (pH 7.0), 20 mM of sodium acetate and 10 mM of EDTA (pH 8.0). The A260/A280 ratio was used to measure the quality of RNA. The ratio between 1.9 and 2.1 was considered good quality. The intensity files of all the subjects were input into the computer program GAP: Generalized Association Plots (29, 30) for quality control using visualization and descriptive statistics. We used the Robust Multi-array Analysis (RMA) method to normalize the data (31). In order to filter out probe sets with low variations and to reduce the impact of multiple comparisons, we kept only the 1,000 probe sets with the largest standard deviations. We searched for differential expression regions (DERs) by prioritizing the gene expression levels associated with the clinical severity indicated by SCQ scores, we used the generalized linear model to screen for probes across the whole genome with mRNA levels associated with the SCQ scores with unadjusted $p < 0.00001$. All original intensity ratio data were transformed into logarithmic 2 values after being normalized. We controlled for the batch effect by adjusting for the batch as a binary covariate since there were two batches. These probes constitute the primary source of predictors to determine ASD subgroups.

Gene Ontology and Pathway Analysis

The DERs with a fold-change >2 were selected for the gene ontology and pathway analysis to evaluate the biological relevance and functional pathways of the significant genes. We have incorporated the KEGG (32), WikiPathways (33), BioCarta (34), and Reactome (35) pathway database for the cell signaling pathways. We have also considered the GO Biological Process (2018) database for gene ontological analysis (36). The GO terms and pathways enriched by the list of genes were identified using the hypergeometric analyses with an adjusted $P \leq 0.05$ was considered as statistically significant.

Gene Over-Representation Analysis

Then we used the webtool at ConsensusPathDB (http://cpdb.molgen.mpg.de/) to identify pathway-pathway interaction network (CPDB analysis) (37). The analysis criteria included: (1) one-next neighbors for the radius with $p < 0.01$, (2) pathway-based sets at least two overlapped genes and $p < 0.01$, and (3) gene ontology level 2 categories with $p < 0.01$. The results from the second approach helped visualize the possible "hub" pathway from the top 10 networks associated with the candidate genes.

We chose two machine learning (ML) algorithms to evaluate the clustering results: random forest classification and support vector machine algorithms. The presence of language impairment was considered as a dichotomous clinical outcome to determine classification errors. We chose the first ML algorithm proposed by Shi and Horvath (38). We used the Random Forest classification (RF) algorithm in an unsupervised mode to generate a proximity matrix. The gene expression data were analyzed using RF using two different approaches for comparison. The first approach is to reduce data dimensionality using principal component analysis to identify principal component (PC) scores for each subject. The top 10 PCs were selected to calculate the proximity matrix that provides a rough estimate of the distance between samples based on the proportion of times the samples end up in the same leaf node. The proximity matrix values were then converted to a dissimilarity matrix to classify the sample into two subgroups using partitioning around medoid (PAM) (39). The second approach is to use the information of all 191 probes with gene expression levels significantly associated with SCQ scores to generate the RF proximity matrix. Similarly, the RF proximity matrix was used to classify the sample into two subgroups using the PAM clustering analysis (39) to classify the patients into two clusters to determine the final cluster assignment. The RF-PAM clustering analysis could allow us to evaluate the classification error by calculating the frequency of patients with language impairment in the cluster, in which the majority of patients had no language impairment, and vice versa.

We further chose Support Vector Machine (SVM) as the second ML algorithm to classify the patients into two subgroups (40). To reduce data dimensionality, we implemented principal component analysis to identify the principal component (PC) scores for each subject. The data of PC scores were split in a 7:3 ratio—in other words, 70% of the data was used for training the model and the remaining 30% was for testing the model. Estimating the C (Cost) parameter to classify

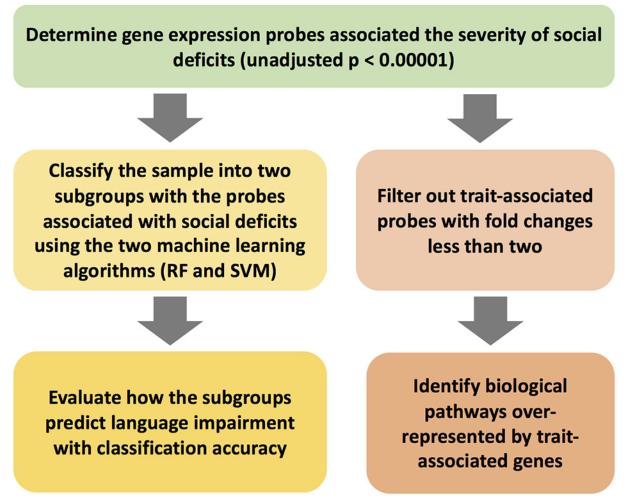
Determine gene expression probes associated the severity of social deficits (unadjusted p < 0.00001)
Classify the sample into two subgroups with the probes associated with social deficits using the two machine learning algorithms (RF and SVM)
Filter out trait-associated probes with fold changes less than two
Evaluate how the subgroups predict language impairment with classification accuracy
Identify biological pathways over-represented by trait-associated genes

FIGURE 1 | The workflow of the study scheme.

TABLE 1 | Clinical features of the patients in the current study.

	Language impairment (51.3%)	No language impairment (48.7%)	Relationship with language impairment*
Age	9.00 (*SD*: 2.52)	8.91 (*SD*: 3.99)	$P > 0.05$
ADIR-BV	17.83 (*SD*: 3.27)	8.55 (*SD*: 1.13)	$P < 0.0001$
ADIR-BN	8.92 (*SD*: 2.71)	3.64 (*SD*: 1.43)	$P < 0.0001$
SCQ	22.19 (*SD*: 4.84)	11.47 (*SD*: 4.84)	$P < 0.0001$
VIQ	82.08 (*SD*: 20.77)	111.91 (*SD*: 10.12)	$P = 0.0003$
PIQ	90.83 (*SD*: 15.74)	101.36 (*SD*: 15.34)	$P > 0.05$
SRS	89.61 (*SD*: 16.12)	79.55 (*SD*: 27.99)	$P > 0.05$

ADIR-BV, Autism Diagnostic Interview–Revised, Qualitative Abnormalities in Communication, Total Verbal score. ADIR-BN, Autism Diagnostic Interview–Revised, Qualitative Abnormalities in Communication, Total Non-Verbal score. SCQ, Social Communication Questionnaire score; VIQ, verbal IQ; PIQ, performance IQ; SRS, Social Responsiveness Scale score.

**The student's t-test was performed to evaluate whether the the two subgroups classified by the presence of language impairment had different values in each continuous variable.*

the data was performed using SVM with the linear kernel function. The choice of kernel function was made based on the recommendation from a prior study that using microarray data to predict the diagnosis of colon cancer—which concludes that linear kernel function leads to a lower prediction error than the RBF, quadratic, and polynomial kernel functions (41). The prediction accuracy and Kappa value estimated when the C value was held constant at 1. The Kappa value was calculated using the formula $(p_o - p_e)/(1\text{-}p_e)$, where p_o and p_e denote the observed agreement and expected agreement for classification, respectively. We further used the confusion matrix, which contains the number of correct and incorrect predictions summarized with count values and broken down by each class, to predict the prediction accuracy of the SVM model. The accuracy is calculated as (TP + TN)/(TP+TN+FP+FN), where TP and TN refer to true positives and true negatives, respectively; FP and FN refer to false positives and false negatives, respectively. These two measures (i.e., accuracy and Kappa value) were chose to evaluate the SVM performance as recommended by previous studies (42, 43). The Kappa statistics could lead to a biased performance estimate in unbalanced situations (44), which is not the characteristic of the current sample. The SVM analysis was performed using the R package "*caret*" (45).

RESULTS

The workflow of the current project is shown in **Figure 1**. The clinical features of the 31 subjects are summarized in **Table 1**. The group with language impairment and the group without language impairment has significant differences in clinical features associated with both social communication function and verbal IQ scores.

The transcriptomic association study reveals 191 probes that were statistically significantly associated with SCQ scores with a $p < 0.00001$. The batch effect seemingly did not affect the association results (**Supplementary Figure 1**). We selected 54 of them with a fold-change >2 for the pathway analysis. Differentially expressed 54 genes with logarithmic fold changes

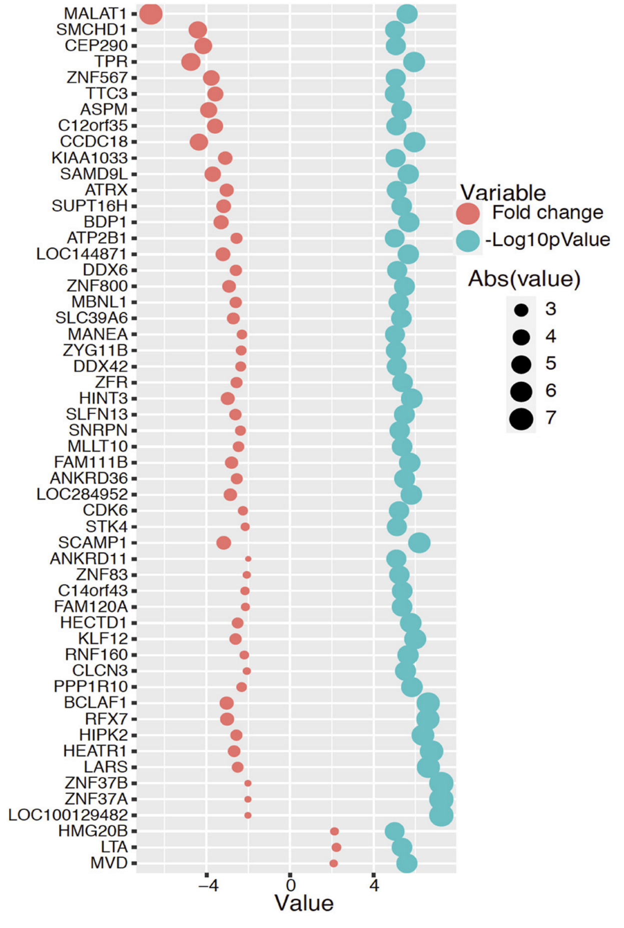

FIGURE 2 | Differentially expressed 54 genes with fold changes and –logarithmic 10 adjusted *p*-values. The red circle represents logarithmic fold change and the blue color circle represents –logarithmic 10 adjusted *p*-value for each significant gene.

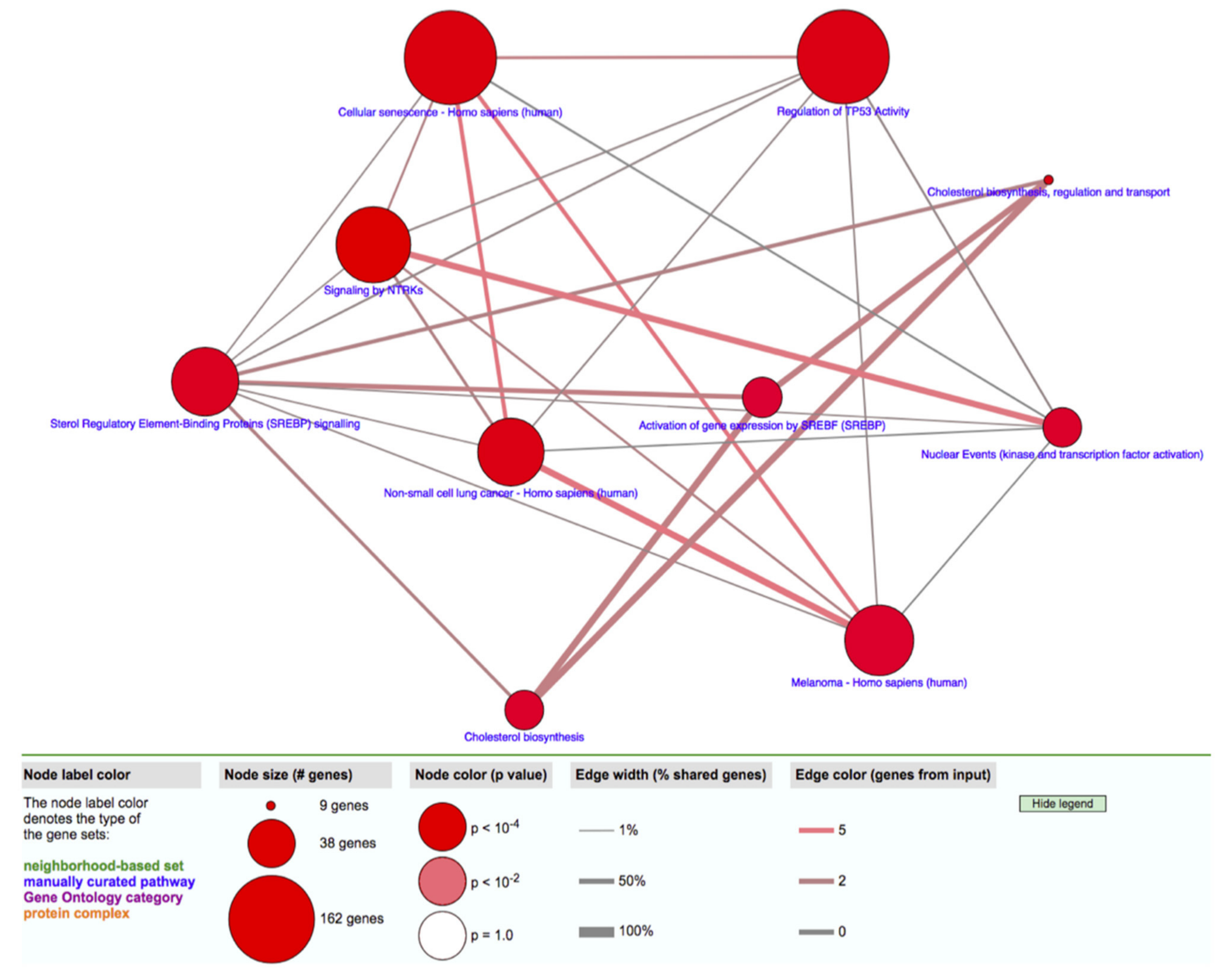
Cellular senescence - Homo sapiens (human)
Regulation of TP53 Activity
Cholesterol biosynthesis, regulation and transport
Signaling by NTRKs
Sterol Regulatory Element-Binding Proteins (SREBP) signalling
Activation of gene expression by SREBF (SREBP)
Nuclear Events (kinase and transcription factor activation)
Non-small cell lung cancer - Homo sapiens (human)
Melanoma - Homo sapiens (human)
Cholesterol biosynthesis
Node label color
The node label color denotes the type of the gene sets:
neighborhood-based set
manually curated pathway
Gene Ontology category
protein complex
Node size (# genes)
9 genes
38 genes
162 genes
Node color (p value)
p < 10-4
p < 10-2
p = 1.0
Edge width (% shared genes)
1%
50%
100%
Edge color (genes from input)
5
2
0
Hide legend

FIGURE 3 | Gene network analysis. The relationship among pathways enriched with candidate genes with expression levels associated with SCQ scores is shown.

and –logarithmic 10 adjusted *p*-values are listed in **Figure 2**. Only three pathways were found to be over-represented by these 54 genes with adjusted $p < 0.05$: cholesterol biosynthetic process (GO:0006695), secondary alcohol biosynthetic process (GO:1902653), and regulation of signal transduction by p53 class mediator (GO:1901796). The CPBD analysis shows that Sterol Regulatory Element-Binding Proteins (SREBP) signaling pathway is the pathway connected with 9 of the 10 pathways including cholesterol biosynthetic pathway, so it can be regarded as the "hub" associated with genetic network for ASD (**Figure 3**). This pathway of SREBP focuses on the regulation of lipid metabolism by SREBP.

The RF-PAM analysis identified two clusters (**Figure 4**). The classification accuracy was 67.7% when the top 10 PCs were used to generate the proximity matrix, while the classification accuracy was 96.9% when all 191 probes were used to generate the proximity matrix. The SVM analysis based on the top 10 PC scores shows that the clustering results reached classification accuracy at 93.3% (95% CI 68.1–99.8%) and no-information rate (i.e., the largest proportion of the observed classes) at 53.3% ($p = 0.0011$). Other parameters relevant to prediction performance include Kappa value = 0.86, sensitivity = 0.86, specificity = 1.00, and balanced accuracy = 0.93. The SVM analysis using the information of all probes with differential gene expressions associated with SCQ scores yielded a slightly higher classification accuracy than the SVM analysis based on the top 10 PC scores. The classification accuracy at 99.9% (95% CI 78.2–100%) and no-information rate (i.e., the largest proportion of the observed classes) at 53.3% ($p = 8.035 \times 10^{-5}$) were achieved when 191 probes were analyzed. This classification accuracy can be demonstrated in gene expression level distributions stratified by language impairment (**Supplementary Figure 2**). The SVM clustering results are shown in **Figure 5**. The results suggest that the first two principal components could identify support vectors that fell in the area with better prediction confidence (**Figure 5A**), compared with the results predicted by individual

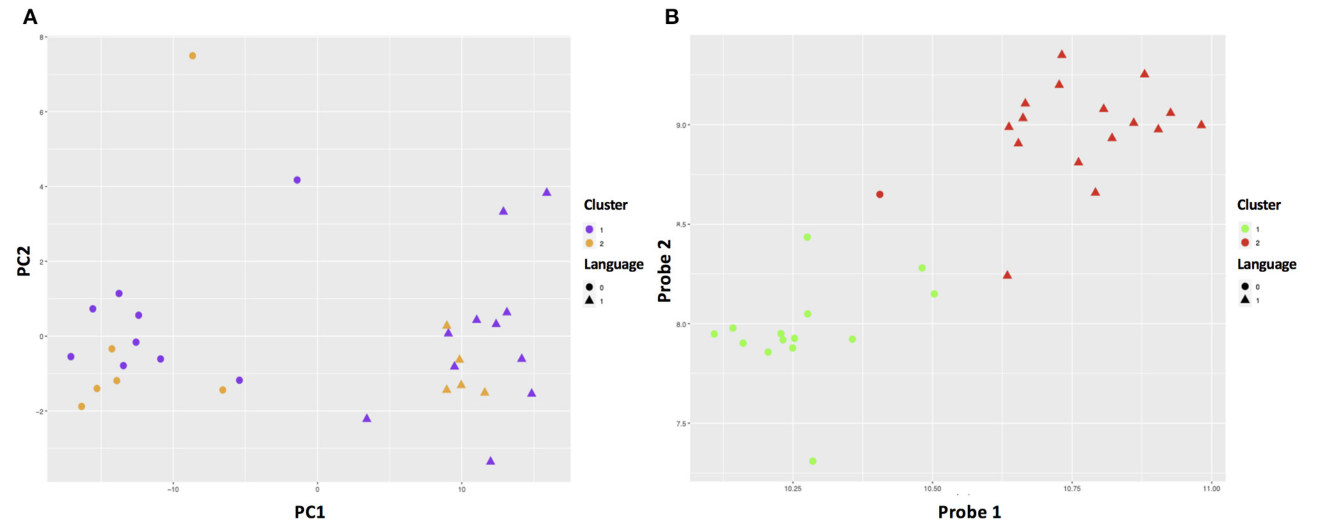

FIGURE 4 | ASD subgroups identified using RF and PAM clustering algorithms. Dim1 and Dim2 correspond to principal components 1 and 2, respectively. **(A,B)** The results based on the top 10 principal components (PCs) and the 191 probes, respectively. We used the first two predictors to make the plots to demonstrate how different approaches classified the sample.

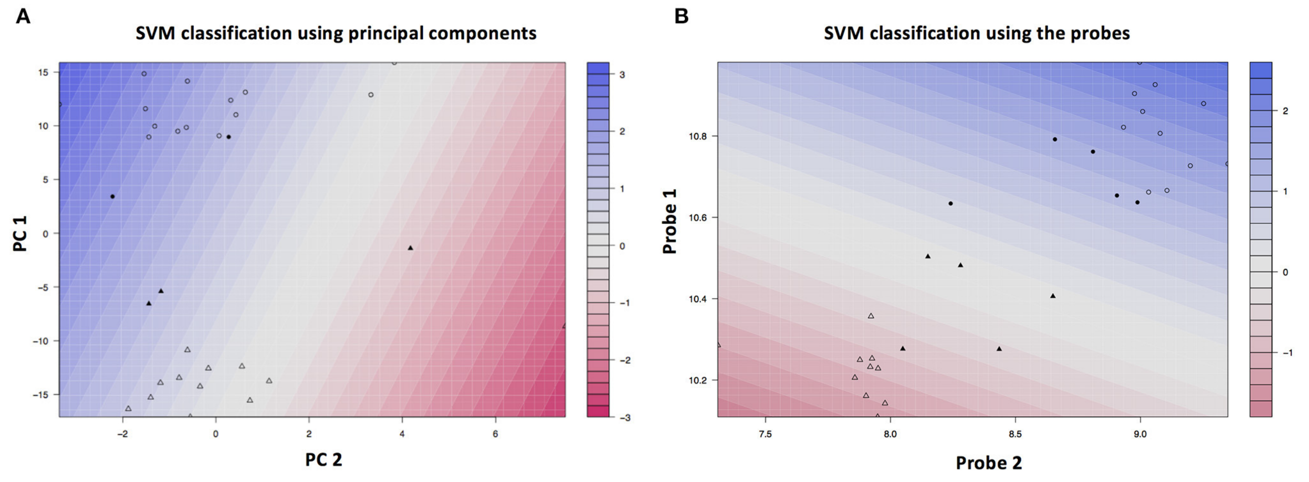

FIGURE 5 | SVM clustering results based on the top PCs. **(A)** Shows the color gradient that indicates how confidently a new point would be classified based on its features. PC1 and PC2 represent the first and second principal components, respectively. **(B)** Shows the color gradient that indicates how confidently a new point would be classified based on its features when predictors were based on all SCQ-associated probes. Probe 1 and probe 2 represent the first and second probes, respectively. The solid symbols indicate the support vectors and the hollow circles indicates other subjects. The circles and triangles represent the first and second subgroups, respectively.

probes (**Figure 5B**). The predicting performance of the RF-PAM and SVM algorithms is listed in **Table 2**.

DISCUSSIONS

We conducted a proof-of-concept study to demonstrate how transcriptomic data from a small sample could provide useful biomarkers to classify ASD subgroups. The selection of the predictors was based on DERs associated with SCQ scores, which indicate the variation in severity levels of social communication deficits, a hallmark clinical feature of ASD. The DER with strongest evidence for the association with social deficits in our sample is matched with the HEATR1 gene (HEAT Repeat Containing 1). The HEART1 gene is associated with schizophrenia (46). The HEATR1 gene abnormalities in the brain during the embryonic stage has been reported in zebrafish (47). The candidate genes that harbor these DERs suggest several genetic pathways that modulate the variation in social communication functions. Among these pathways, the pathway

TABLE 2 | Predicting performance of two machine learning algorithms.

Method	Predictors	Prediction accuracy
RF-PAM	191 probes	96.90%
RF-PAM	10 PC*	67.70%
SVM	191 probes	99.90%
SVM	10 PC*	93.30%

**Principal component.*

of cholesterol biosynthesis/metabolism and sterol regulatory element-binding proteins (SREBP) pathway—cholesterol metabolism appear to act as hubs that connect other top SCQ-associated pathways. Particularly, the SREBP pathway shares most genes with other SCQ-associated pathways. These two pathways are related to lipid metabolism. Cholesterol synthesis and uptake are tightly modulated at the transcriptional level through negative feedback control, which is regulated by SREBPs (48). The relationship between lipid metabolism and brain functions has been well-documented. A growing body of evidence has indicated that cholesterol metabolism plays a key role in synaptic functions (49–51). Dysregulated cholesterol metabolism has been extensively documented in ASD (51–58). A recent study implemented a personalized medicine approach combining healthcare claims, electronic health records, familial whole-exome sequences, and neurodevelopmental gene expression patterns, and identified an ASD subtype characterized by dyslipidemia (59). There are certainly several other genetic pathways involved in molecular mechanisms underlying social communication deficits. Nevertheless, our results indicate that cholesterol synthesis/metabolism pathways act as hubs that connect most other biological pathways, which suggest that the genomic functional changes associated with lipid metabolism may moderate other genomic changes, such as the p53 signaling pathway, that regulate social communication functions.

Using the DERs as biomarkers, we clustered the sample into two subgroups using two different ML algorithms. Both the RF-PAM and SVM analyses yielded similar levels of classification accuracy when all 191 markers were utilized. However, compared to the analysis using the RF-PAM algorithm, the analysis using the SVM algorithm seemed to be more robust when we performed dimension reduction for all the 191 markers with the PCA method. The RF algorithm is applicable when there are more predictors than observations, relatively insensitive to the noise (e.g., a large number of irrelevant genes), and does not rely on excessive fine-tuning of parameters (60). RF algorithm is more robust to small sample size as the SVM algorithm (61, 62). However, Brown et al. found that SVM outperforms other techniques that include Fisher's linear discriminant, Parzen window, and tow decision tree learners when using gene expression data to predict clinical outcomes (63). Additionally, Statnikov et al. conducted a comprehensive comparison of RF and SVM using microarray data for 22 diagnostic and prognostic datasets and concluded that SVM is superior to RF in terms of classification accuracy (64). Although the purpose of this study is not to comprehensively evaluate which ML algorithm outperforms the other ML algorithm, our results seem to lend some support to the robustness of the SVM algorithm. Nevertheless, the RF algorithm is at least as robust as the SVM algorithm when the dimension of input variables is not substantially reduced.

One of the major limitation of the current study is the small sample size. Nevertheless, some machine learning algorithm, such as SVM, can handle a small sample with a large number of features. Additionally, model overfitting may arise due to a lack of another independent sample for validation. Furthermore, unknown confounders may cause spurious associations between the phenotype and genomic markers. However, the goal of this proof-of-concept study is prediction of subtypes rather than the identification of etiologies. Therefore, confounders would not yield a substantial impact on prediction results (65).

The clinical and etiological heterogeneity in ASD has meant that there is considerable variability in treatment outcomes across different interventions and between individuals receiving the same intervention. Hence the traditional diagnostic and "one size fits all" approach to ASD intervention needs improvement. Further, we currently do not have a sufficient understanding of "what would work for whom," thereby limiting opportunities for maximizing outcomes for children and their families with economic ramifications for broader society. In this context, ML algorithms have been found to be useful in predicting diagnostic accuracy in ASD with neuroimaging data (66). Further, one recent study used Gaussian Mixture Models and Hierarchical Agglomerative Clustering, which provide a statistical framework for learning latent cluster memberships to determine ASD subgroups with differentiated treatment responses (67). Our findings that using ML algorithms, children could be classified into two groups based on the presence of language impairment, offers promise for unraveling clinically meaningful subgroups in ASD. This, in turn, can be used for predicting likely responsiveness (and non-responsiveness) to specific treatment pathways. This "precision" approach to assessment and intervention will ensure that resources for appropriate intervention and supports are allocated in an evidence-based manner. This is critical as without timely recognition of the variability in the clinical presentation, neurocognitive level of functioning, and psychosocial circumstances coupled with individualized intervention, children and their families may miss key opportunities of brain plasticity available in the critical early years. ML techniques as utilized in this study offer a viable solution to address this by allowing matching interventions and supports that are tailored to the individual profile and needs.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found at: https://figshare.com/articles/dataset/Autism_gene_expression_data/14251328.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Research Ethics Committee of the National Taiwan University Hospital. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

P-IL and MM carried out the statistical analysis. P-IL and VE conceived of the study and drafted the manuscript. SG participated in the study design and coordination. All authors read and approved the final manuscript.

FUNDING

The genetic data collection and analysis were supported by grants from the Ministry of Science and Technology (NSC 99-3112-B-002-036), Taiwan, and National Taiwan University Hospital (NCTRC201114), Taiwan, awarded to SG.

ACKNOWLEDGMENTS

The authors would like to thank the subjects who participated in this study and facility support at National Taiwan University Hospital.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyt.2021.637022/full#supplementary-material

Supplementary Figure 1 | The evaluation of potential batch effect due to the microarrays timing. **(A)** The kernel density distributions of gene expression levels of the two batches are shown. **(B)** Time 1 and time 2 indicate the association test results that adjusted for the time (i.e., batch) vs. the results without adjusting for the time.

Supplementary Figure 2 | Randomly selected four probes associated with SCQ scores stratified by the presence of language impairment. The red and blue curves represent the group without language impairment and the group with language impairment, respectively.

REFERENCES

1. American Psychiatric Association (2013). *Diagnostic and statistical manual of mental disorders (5th ed.)*. Arlington: American Psychiatric Association. p. 31–2. doi: 10.1176/appi.books.9780890425596
2. Eapen V. Genetic basis of autism: is there a way forward? *Curr Opin Psychiatry.* (2011) 24:226–36. doi: 10.1097/YCO.0b013e328345927e
3. Bowers K, Lin P-I, Erickson C. Pharmacogenomic medicine in autism: challenges and opportunities. *Pediatr Drugs.* (2015) 17:115–24. doi: 10.1007/s40272-014-0106-0
4. McPartland JC, Bernier RA, Jeste SS, Dawson G, Nelson CA, Chawarska K, et al. The autism biomarkers consortium for clinical trials (ABC-CT): scientific context, study design, and progress toward biomarker qualification. *Front Integr Neurosci.* (2020) 14:16. doi: 10.3389/fnint.2020.00016
5. Anney R, Klei L, Pinto D, Regan R, Conroy J, Magalhaes TR, et al. A genome-wide scan for common alleles affecting risk for autism. *Hum Mol Genet.* (2010) 19:4072–82. doi: 10.1093/hmg/ddq307
6. Yonan AL, Alarcón M, Cheng R, Magnusson PKE, Spence SJ, Palmer AA, et al. A genomewide screen of 345 families for autism-susceptibility loci. *Am J Hum Genet.* (2003) 73:886–97. doi: 10.1086/378778
7. Liu J, Nyholt DR, Magnussen P, Parano E, Pavone P, Geschwind D, et al. A genomewide screen for autism susceptibility loci. *Am J Hum Genet.* (2001) 69:327–40. doi: 10.1086/321980
8. Lord C, Rutter M, Le Couteur A. Autism Diagnostic Interview-Revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *J Autism Dev Disord.* (1994) 24:659–85.
9. Gotham K, Pickles A, Lord C. Standardizing ADOS scores for a measure of severity in autism spectrum disorders. *J Autism Dev Disord.* (2009) 39:693–705. doi: 10.1007/s10803-008-0674-3
10. Icabone DG. Vineland adaptive behavior scales. *Diagnostique.* (1999) 24:257–73. doi: 10.1177/153450849902401-423
11. Veatch OJ, Veenstra-Vanderweele J, Potter M, Pericak-Vance MA, Haines JL. Genetically meaningful phenotypic subgroups in autism spectrum disorders. *Genes Brain Behav.* (2014) 13:276–85. doi: 10.1111/gbb.12117
12. Anney R, Klei L, Pinto D, Almeida J, Bacchelli E, Baird G, et al. Individual common variants exert weak effects on the risk for autism spectrum disorders. *Hum Mol Genet.* (2012) 21:4781–92. doi: 10.1093/hmg/dds301
13. Chaste P, Klei L, Sanders SJ, Hus V, Murtha MT, Lowe JK, et al. A genome-wide association study of autism using the Simons simplex collection: does reducing phenotypic heterogeneity in autism increase genetic homogeneity? *Biol Psychiatry.* (2015) 77:775–84. doi: 10.1016/j.biopsych.2014.09.017
14. Torrico B, Shaw AD, Mosca R, Vivó-Luque N, Hervás A, Fernàndez-Castillo N, et al. Truncating variant burden in high-functioning autism and pleiotropic effects of LRP1 across psychiatric phenotypes. *J Psychiatry Neurosci.* (2019) 44:350–9. doi: 10.1503/jpn.180184
15. Chiang AH, Chang J, Wang J, Vitkup D. Exons as units of phenotypic impact for truncating mutations in autism. *Mol Psychiatry* (2020) 25:1–11. doi: 10.1038/s41380-020-00876-3
16. Bruining H, de Sonneville L, Swaab H, de Jonge M, Kas M, van Engeland H, et al. Dissecting the clinical heterogeneity of autism spectrum disorders through defined genotypes. *PLoS ONE.* (2010) 5:e10887. doi: 10.1371/journal.pone.0010887
17. Hu VW, Lai Y. Developing a Predictive Gene Classifier for Autism Spectrum Disorders Based upon Differential Gene Expression Profiles of Phenotypic Subgroups. *N Am J Med Sci (Boston)* (2013) 6:1–18. doi: 10.7156/najms.2013.0603107
18. Mottron L, Bzdok D. Autism spectrum heterogeneity: fact or artifact? *Mol Psychiatry.* (2020) 25:3178–85. doi: 10.1038/s41380-020-0748-y
19. Küpper C, Stroth S, Wolff N, Hauck F, Kliewer N, Schad-Hansjosten T, et al. Identifying predictive features of autism spectrum disorders in a clinical sample of adolescents and adults using machine learning. *Sci Rep.* (2020) 10:4805. doi: 10.1038/s41598-020-61607-w
20. Asif M, Martiniano HFMC, Marques AR, Santos JX, Vilela J, Rasga C, et al. Identification of biological mechanisms underlying a multidimensional ASD phenotype using machine learning. *Transl Psychiatry.* (2020) 10:43. doi: 10.1038/s41398-020-0721-1
21. Akter T, Shahriare Satu M, Khan MI, Ali MH, Uddin S, Lio P, et al. Machine learning-based models for early stage detection of autism spectrum disorders. *IEEE Access.* (2019) 7:166509–27. doi: 10.1109/ACCESS.2019.2952609
22. Schanding Jr. GT, Nowell KP, Goin-Kochel RP. Utility of the social communication questionnaire-current and social responsiveness scale as teacher-report screening tools for autism spectrum disorders. *J Autism Dev Disord.* (2012) 42:1705–16. doi: 10.1007/s10803-011-1412-9
23. Mayo J, Chlebowski C, Fein DA, Eigsti IM. Age of first words predicts cognitive ability and adaptive skills in children with ASD. *J Autism Dev Disord.* (2013) 43:253–64. doi: 10.1007/s10803-012-1558-0

24. Lin PI, Kuo PH, Chen CH, Wu JY, Gau SSF, Wu YY, et al. Runs of homozygosity associated with speech delay in autism in a taiwanese Han population: evidence for the recessive model. *PLoS ONE.* (2013) 8:e72056. doi: 10.1371/journal.pone.0072056
25. Lin PI, Chien YL, Wu YY, Chen CH, Gau SSF, Huang YS, et al. The WNT2 gene polymorphism associated with speech delay inherent to autism. *Res Dev Disabil.* (2012) 33:1533–40. doi: 10.1016/j.ridd.2012.03.004
26. Eicher JD, Gruen JR. Language impairment and dyslexia genes influence language skills in children with autism spectrum disorders. *Autism Res.* (2015) 8:229–34. doi: 10.1002/aur.1436
27. Lord C, Rutter M, Le Couteur A. Autism Diagnostic Interview-Revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *J Autism Dev Disord.* (1994) 24:659–85. doi: 10.1007/BF02172145
28. Gau SSF, Lee CM, Lai MC, Chiu YN, Huang YF, Kao J Der, et al. Psychometric properties of the Chinese version of the social communication questionnaire. *Res Autism Spectr Disord.* (2011) 5:809–18. doi: 10.1016/j.rasd.2010.09.010
29. Chen CH. Generalized association plots: information visualization via iteratively generated correlation matrices. *Stat Sin.* (2002) 12:7–29.
30. Wu HM, Tien YJ, Chen CH. GAP: a graphical environment for matrix visualization and cluster analysis. *Comput Stat Data Anal.* (2010) 54:767–78. doi: 10.1016/j.csda.2008.09.029
31. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res.* (2003) 31:e15. doi: 10.1093/nar/gng015
32. Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* (2000) 28:27–30. doi: 10.1093/nar/28.1.27
33. Slenter DN, Kutmon M, Hanspers K, Riutta A, Windsor J, Nunes N, et al. WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Res.* (2018) 46:D661–7. doi: 10.1093/nar/gkx1064
34. Nishimura D. BioCarta. *Biotech Softw Internet Rep.* (2001) 117–20. doi: 10.1089/152791601750294344
35. Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P, et al. The reactome pathway knowledgebase. *Nucleic Acids Res.* (2018) 44:D481–7. doi: 10.1093/nar/gkx1132
36. Carbon S, Douglass E, Dunn N, Good B, Harris NL, Lewis SE, et al. The gene ontology resource: 20 years and still GOing strong. *Nucleic Acids Res.* (2019) 47:D330–8. doi: 10.1093/nar/gky1055
37. Kamburov A, Stelzl U, Lehrach H, Herwig R. The ConsensusPathDB interaction database: 2013 Update. *Nucleic Acids Res.* (2013) 41:D793–800. doi: 10.1093/nar/gks1055
38. Shi T, Horvath S. Unsupervised learning with random forest predictors. *J Comput Graph Stat.* (2006) 15:118–38. doi: 10.1198/106186006X94072
39. Kaufman L, Rousseeuw PJ. *Partitioning Around Medoids (Program PAM), in Finding Groups in Data: An Introduction to Cluster Analysis*. Hoboken: Wiley (2008). doi: 10.1002/9780470316801.ch2
40. Cortes C, Vapnik V. Support-vector networks. *Mach Learn.* (1995) 20:273–97. doi: 10.1023/A:1022627411411
41. Devi Arockia Vanitha C, Devaraj D, Venkatesulu M. Gene expression data classification using Support Vector Machine and mutual information-based gene selection. *Procedia Comput Sci.* (2014) 47:13–21. doi: 10.1016/j.procs.2015.03.178
42. Soleymani A, Pennekamp F, Petchey OL, Weibel R. Developing and integrating advanced movement features improves automated classification of ciliate species. *PLoS ONE.* (2015) 11:e0145345. doi: 10.1371/journal.pone.0145345
43. Verda D, Parodi S, Ferrari E, Muselli M. Analyzing gene expression data for pediatric and adult cancer diagnosis using logic learning machine and standard supervised methods. *BMC Bioinformatics.* (2019) 20:390. doi: 10.1186/s12859-019-2953-8
44. Delgado R, Tibau XA. Why Cohen's Kappa should be avoided as performance measure in classification. *PLoS ONE.* (2019) 14:e0222916. doi: 10.1371/journal.pone.0222916
45. Kuhn M. caret Package. *J Stat Softw.* (2008) 28:1–26.
46. Roussos P, Guennewig B, Kaczorowski DC, Barry G, Brennand KJ. Activity-dependent changes in gene expression in schizophrenia human-induced pluripotent stem cell neurons. *JAMA Psychiatry.* (2016) 73:1180–8. doi: 10.1001/jamapsychiatry.2016.2575
47. Azuma M, Toyama R, Laver E, Dawid IB. Perturbation of rRNA synthesis in the bap28 mutation leads to apoptosis mediated by p53 in the zebrafish central nervous system. *J Biol Chem.* (2006) 281:13309–16. doi: 10.1074/jbc.M601892200
48. Sato R. Sterol metabolism and SREBP activation. *Arch Biochem Biophys.* (2010) 501:177–81. doi: 10.1016/j.abb.2010.06.004
49. Paul SM, Doherty JJ, Robichaud AJ, Belfort GM, Chow BY, Hammond RS, et al. The major brain cholesterol metabolite 24(S)-hydroxycholesterol is a potent allosteric modulator of N-Methyl-D-Aspartate receptors. *J Neurosci.* (2013) 33:17290–300. doi: 10.1523/JNEUROSCI.2619-13.2013
50. Wang H. Lipid rafts: a signaling platform linking cholesterol metabolism to synaptic deficits in autism spectrum disorders. *Front Behav Neurosci.* (2014) 8:104. doi: 10.3389/fnbeh.2014.00104
51. Petrov AM, Kasimov MR, Zefirov AL. Cholesterol in the pathogenesis of alzheimer's, parkinson's diseases and autism: link to synaptic dysfunction. *Acta Naturae.* (2017) 9:26–37. doi: 10.32607/20758251-2017-9-1-26-37
52. Tamiji J, Crawford DA. The neurobiology of lipid metabolism in autism spectrum disorders. *NeuroSignals.* (2011) 18:98–112. doi: 10.1159/000323189
53. Gillberg C, Fernell E, Kočovská E, Minnis H, Bourgeron T, Thompson L, et al. The role of cholesterol metabolism and various steroid abnormalities in autism spectrum disorders: a hypothesis paper. *Autism Res.* (2017) 10:1022–44. doi: 10.1002/aur.1777
54. Richardson AJ, Ross MA. Fatty acid metabolism in neurodevelopmental disorder: a new perspective on associations between attention-deficit/hyperactivity disorder, dyslexia, dyspraxia and the autistic spectrum. *Prostaglandins Leukot Essent Fat Acids.* (2000) 63:1–9. doi: 10.1054/plef.2000.0184
55. Aneja A, Tierney E. Autism: the role of cholesterol in treatment. *Int Rev Psychiatry.* (2008) 20:165–70. doi: 10.1080/09540260801889062
56. Cartocci V, Catallo M, Tempestilli M, Segatto M, Pfrieger FW, Bronzuoli MR, et al. Altered brain cholesterol/isoprenoid metabolism in a rat model of autism spectrum disorders. *Neuroscience.* (2018) 372:27–37. doi: 10.1016/j.neuroscience.2017.12.053
57. Esparham AE, Smith T, Belmont JM, Haden M, Wagner LE, Evans RG, et al. Nutritional and metabolic biomarkers in autism spectrum disorders: an exploratory study. *Integr Med.* (2015) 14:40–53.
58. Tierney E, Bukelis I, Thompson RE, Ahmed K, Aneja A, Kratz L, et al. Abnormalities of cholesterol metabolism in autism spectrum disorders. *Am J Med Genet Part B Neuropsychiatr Genet.* (2006) 141B:666–8. doi: 10.1002/ajmg.b.30368
59. Luo Y, Eran A, Palmer N, Avillach P, Levy-Moonshine A, Szolovits P, et al. A multidimensional precision medicine approach identifies an autism subtype characterized by dyslipidemia. *Nat Med.* (2020) 26:1375–9. doi: 10.1038/s41591-020-1007-0
60. Breiman L. Random forests. *Mach Learn.* (2001) 45:5–32. doi: 10.1023/A:1010933404324
61. Díaz-Uriarte R, Alvarez de Andrés S. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics.* (2006) 7:3. doi: 10.1186/1471-2105-7-3
62. Kim SY. Effects of sample size on robustness and prediction accuracy of a prognostic gene signature. *BMC Bioinformatics.* (2009) 10:147. doi: 10.1186/1471-2105-10-147
63. Brown MPS, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS, et al. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc Natl Acad Sci USA.* (2000) 97:262–7. doi: 10.1073/pnas.97.1.262
64. Statnikov A, Wang L, Aliferis CF. A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. *BMC Bioinformatics.* (2008) 9:319. doi: 10.1186/1471-2105-9-319

65. Van Diepen M, Ramspek CL, Jager KJ, Zoccali C, Dekker FW. Prediction versus aetiology: common pitfalls and how to avoid them. *Nephrol Dial Transplant*. (2017) 32:ii1–5. doi: 10.1093/ndt/gfw459

66. Moon SJ, Hwang J, Kana R, Torous J, Kim JW. Accuracy of machine learning algorithms for the diagnosis of autism spectrum disorder: systematic review and meta-analysis of brain magnetic resonance imaging studies. *J Med Internet Res*. (2019) 6:e14108. doi: 10.2196/14108

67. Stevens E, Dixon DR, Novack MN, Granpeesheh D, Smith T, Linstead E. Identification and analysis of behavioral phenotypes in autism spectrum disorder via unsupervised machine learning. *Int J Med Inform*. (2019) 129:29–36. doi: 10.1016/j.ijmedinf.2019.05.006

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.