Cell type-specific predictive models for prioritizing genes and gene sets associated with autism spectrum disorder

Jinting Guan (jtguan@xmu.edu.cn)
Xiamen University https://orcid.org/0000-0001-9433-5677
Yang Wang
Xiamen University
Yiping Lin
Xiamen University
Qingyang Yin
Xiamen University
Yibo Zhuang
Xiamen YLZ Yihui Technology Co., Ltd
Guoli Ji
Xiamen University

Research

Keywords: autism spectrum disorder, cell type-specific, predictive model, gene set

Posted Date: August 10th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-51727/v1

Abstract

Background Autism spectrum disorder (ASD) is characterized by substantial phenotypic and genetic heterogeneity. Although bulk transcriptomic analyses revealed convergence of disease pathology on common pathways, the brain cell type-specific molecular pathology of ASD is still needed to study. Different gene functions may be dysregulated and causal genes may be distinct among different brain cells in ASD. Gene expression profiling-based machine learning studies can be conducted for the diagnosis of ASD, prioritizing high-confidence gene candidates and promoting the design of effective interventions.

Methods To characterize the cell type heterogeneity of ASD and to take advantage of the potential of gene expression signature as diagnostic biomarkers for ASD, we construct multiple kinds of classification models for ASD based on the recently available human brain nucleus gene expression data of ASD and controls. Firstly, we construct cell type-specific predictive models based on individual genes to screen cell type-specific genes associated with ASD. Then from the view of gene set, we construct cell type-specific gene set-based predictive models to screen cell type-specific gene sets associated with ASD. These two kinds of predictive models can be applied to predict the diagnosis of a given nucleus with known cell type. Lastly, we further construct a multi-label predictive model for predicting the cell type and diagnosis of a given nucleus at the same time.

Results It is found that the functions of genes with predictive power for ASD are not consistent and the top important genes are distinct among different cells, demonstrating the cell type heterogeneity of ASD. Our findings suggest that layer 2/3 and layer 4 excitatory neurons, layer 5/6 cortico-cortical projection neurons, parvalbumin interneurons, and protoplasmic astrocytes are preferentially affected in ASD. Gene *BCYRN1* and *CCK* are prioritized in excitatory neurons, and *HSPA1A* is of note in protoplasmic astrocytes.

Limitations Our study utilized methods of machine learning to identify biomarkers of ASD, while it is more convincing if subsequent experiments could be conducted to validate the results.

Conclusions The results show that it may be feasible to use single cell/nucleus gene expression for ASD detection and the constructed predictive models can promote the diagnosis of ASD. Our analytical pipeline prioritizes ASD-associated cell type-specific genes and gene sets, which may be used as potential biomarkers of ASD.

1. Background

Autism spectrum disorder (ASD) represents a group of neurodevelopmental disorders, characterized by substantial phenotypic and genetic heterogeneity. Genetic studies have identified variants that contribute to the risk of developing ASD [1–7]. However, it remains perplexing how these reported variants lead to the pathogenesis of ASD. A major mode of action is that these genetic variants cause gene expression alternations, direct analysis of gene expression in disease-relevant tissue is thus valuable for understanding the molecular mechanism of ASD. As ASD is believed to result from functional aberrations

within brains, bulk transcriptomic analyses between autistic and normal brains have been applied for identifying aberrant gene expression patterns in ASD [8–10]. However, the brain is a highly heterogeneous organ including different cell types that are highly interconnected. Genes may demonstrate diverse functions across different brain cell types. In ASD, different functions may be dysregulated and causal genes may be distinct among different cells. Although bulk transcriptomic studies revealed convergence of disease pathology on common pathways, the brain cell type-specific molecular pathology of ASD is still needed to study.

Recently, the newly available single-nucleus RNA-sequencing data of ASD [11] makes it possible to study the cell type heterogeneity of ASD directly. The authors identified differentially expressed genes between ASD and control groups in a cell type-specific way and analyzed the functions of the cell type-specific differentially expressed genes to characterize the heterogeneity of dysregulated gene expression patterns among brain cell types in ASD. As genes interact with others, the integrity of disease gene modules instead of individual genes may determine the manifestation of a disease in cells [12, 13]. Therefore, in addition to identifying the individual cell type-specific risk genes, it is essential to identify cell type-specific gene sets/modules associated with ASD.

There have been more and more studies evaluating the effectiveness of machine learning for diagnosing ASD, exploring its genetic underpinnings, and designing effective interventions [14]. These studies were based on different kinds of datasets, such as behavior evaluation based on Autism Diagnostic Observation Schedule (ADOS) [15, 16] and Autism Diagnostic Interview-Revised (ADI-R) [17, 18], brain images for magnetic resonance image (MRI) [19, 20] and electroencephalogram (EEG) [21], and genetic profiles [10, 22–24]. To detect ASD candidate genes, several predictive models were constructed based on gene expression profiling, including the one built using differential expressed genes between ASD and controls based on gene expression microarrays of blood [23], and the one built using aberrant gene expression in ASD based on bulk transcriptomic data of brains [10]. Actually, for identifying ASD risk genes, genetic and genomic studies were usually performed, such as genome-wide association studies, copy number variation studies and whole exome sequencing, while these methods are expensive and time-consuming, and the generated potential candidate genes are numerous and not easy to be validated [22]. Gene screening methods based on machine learning can prioritize genes and identify high-confidence candidates, which may provide new insights for the experimental studies.

In this study, to characterize the cell type heterogeneity of ASD and to take advantage of the potential of gene expression signature being diagnostic biomarkers for ASD, we analyze the human brain nucleus gene expression data of ASD and controls published in [11] and construct multiple kinds of classification models for ASD using the algorithm of partial least squares, identifying cell type-specific genes and gene sets associated with ASD. Firstly, we construct cell type-specific predictive models based on individual genes to screen cell type-specific genes associated with ASD. Then from the view of gene set, we construct cell type-specific gene set-based predictive models to screen cell type-specific gene sets associate with ASD. These two kinds of predictive models can be applied to predict the diagnosis of a given nucleus with known cell type. Lastly, we further construct a multi-label predictive model for

predicting the cell type and diagnosis of a given nucleus at the same time. Our results suggest that it may be feasible to use brain cell/nucleus gene expression for ASD detection and the constructed predictive models can promote the diagnosis of ASD. Our analytical pipeline prioritizes ASD-associated cell type-specific genes and gene sets, promoting to depict the cell type heterogeneity of ASD.

2. Materials And Methods

2.1 Human brain nucleus gene expression data

We used the single-nucleus RNA-seq data published in [11], which includes 104559 nuclei from 41 post-mortem tissue samples from the prefrontal cortex and anterior cingulate cortex of 15 ASD patients and 16 control subjects. The nuclei were divided into 17 cell types, including fibrous astrocytes (AST-FB), protoplasmic astrocytes (AST-PP), endothelial, parvalbumin interneurons (IN-PV), somatostatin interneurons (IN-SST), SV2C interneurons (IN-SV2C), VIP interneurons (IN-VIP), layer 2/3 excitatory neurons (L2/3), layer 4 excitatory neurons (L4), layer 5/6 corticofugal projection neurons (L5/6), layer 5/6 cortico-cortical projection neurons (L5/6-CC), microglia, maturing neurons (Neu-mat), NRGN-expressing neurons (Neu-NRGN-I), NRGN-expressing neurons (Neu-NRGN-II), oligodendrocytes, and OPC. We downloaded the matrices of raw counts from website of autism.cells.ucsc.edu. Then we preprocessed the data with R package of scran [25], including the quality control of nuclei and genes, removing a minority of nuclei from different cell cycle phases, and normalizing the gene expression data. We excluded nuclear and mitochondrial genes downloaded from Human MitoCarta2.0 [26]. Then we applied scran to obtain highly variable genes, which include a total of 12036 genes. We used the expression level of 12036 genes for downstream analyses, which contains 85125 nuclei, including 3655, 7085, 1991, 3719, 4190, 1836, 5621, 12795, 6518, 3402, 4385, 2495, 3532, 589, 1459, 12206, and 9647 nuclei from cell types of AST-FB, AST-PP, endothelial, IN-PV, IN-SST, IN-SV2C, IN-VIP, L2/3, L4, L5/6, L5/6-CC, microglia, Neu-mat, Neu-NRGN-I, Neu-NRGN-II, oligodendrocytes, and OPC respectively.

2.2 Annotated gene sets

A total of 913 ASD candidate genes were downloaded from Simons Foundation Autism Research Initiative (SFARI), which include 119, 144, 219, and 472 genes from categories of S (syndromic), 1 (high confidence), 2 (strong candidate), and 3 (suggestive evidence). For gene set analysis, three kinds of annotated gene sets from Molecular Signatures Database (MSigDB)[27] were used, including H: hallmark gene sets, C2: curated gene sets (containing gene sets from chemical and genetic perturbations, and canonical pathways of Biocarta, KEGG, PID and Reactome), and C5: GO gene sets. By intersecting the genes in gene sets and our analyzed gene expression matrix, we kept 3741 gene sets containing more than 30 overlapping genes.

2.3 The algorithm of partial least squares

Partial least squares (PLS) [28] regression combines features from principal component analysis and multiple regression. It has the ability to address the problem of modelling multicollinearity, noisy, and

even incomplete highly dimensional data [29]. PLS can solve both single- and multi-label classification problems. Partial least squares discriminant analysis (PLS-DA) is a PLS regression, with the dependent variable being categorical.

Supposed X is an $n\times m$ matrix containing n observations of m genes, Y is a $n\times p$ matrix containing n observations of p response variables, then X and Y can be decomposed by:

$$X = TP^T + E, Y = UQ^T + F$$

where T and U are $n\times k$ score matrices (called component scores or latent variables) of X and Y respectively, P and Q are $m\times k$ and $p\times k$ orthogonal loading matrices, and E and F are the residual matrices. The decompositions of X and Y are made so as to maximize the covariance between T and U. Then based on T, P, U, and Q, we can first fit U and T, then the linear relationship between X and Y can be obtained.

2.4 The construction of predictive models

The R package of caret [30] was adopted to construct predictive models based on the algorithm of PLS. Firstly, for each cell type, we extracted the gene expression data of nuclei from the cell type and constructed a cell type-specific predictive model. Secondly, for each cell type and each annotated gene set, we extracted the expression data of nuclei from the cell type in the genes included in the gene set, and constructed a cell type-specific gene set-based predictive model. These two kinds of predictive models can predict the diagnosis of a nucleus with known cell type. Specifically, we split the extracted gene expression data into training set and test set at the ratio of 7:3 using stratified sampling. For the training set, we selected the optimal model by applying 10-fold cross validation for 10 times and tuning over the model hyperparameter (the number of PLS components) with grid search from 1 to 15 with step of one. To evaluate the model performance, the area under the receiver operating characteristic (ROC) curve (denoted as AUC) was used, because this metric can deal well with the problem of label imbalance and not be influenced by the selection of threshold. Then from the optimal model, we obtained the predictive probability of each nucleus being a nucleus from ASD patients. Next, we used R package of pROC [31] to obtain the best threshold on training set and the threshold was used to determine the predictive performances on training set and test set. For each predictive model, we calculated the importance of each gene using the function of *varImp* in caret.

In order to predict the cell type and diagnosis of a given nucleus at the same time, we constructed a multi-label predictive model based on PLS using R package of mlr [32]. For each nucleus, we used 18 labels to describe it, with one label being the diagnosis and the other 17 cell type labels obtained using one-hot encoding. We split the whole gene expression data including all cell types and all genes into training set and test set at the ratio of 7:3 using stratified sampling. Based on the training set, we selected the optimal model by applying 5-fold cross validation for 5 times and tuning over the model hyperparameter with grid search from 1 to 15 with step of one. Hamming loss was used as performance indicator. Then from the optimal model, we obtained the predictive probability of each nucleus belonging to each label. For the

labels of cell types, the predictive cell type of each nucleus was set as the cell type whose predictive probability is the largest. For the diagnosis label, we extracted the predictive probability of training set and applied ROC analysis to obtain the optimal cutoff on training set for determining the predictive diagnosis of each nucleus in training and test sets.

3. Results

3.1. Methodological overview

The overview of our analytical method can be seen in Fig. 1. Firstly, to screen genes associated with ASD in each cell type, we constructed cell type-specific predictive models, which can predict the diagnosis of a nucleus whose cell type is known, using the algorithm of partial least squares (PLS) (**Materials and Methods**). Specifically, for each cell type, we extracted the gene expression data of the nuclei from the cell type and split the data into training and test sets. We selected the optimal model based on the training set, and then obtained the predictive probability of each nucleus being a nucleus from ASD patients. Next, ROC analysis was performed to obtain the best threshold on training set and the threshold was used to determine the predictive performance on training and test sets. To prioritize genes, we calculated the importance of each gene in the cell type-specific predictive model. In addition, in order to use less genes to achieve similar performances, we performed recursive feature elimination with cross-validation (RFECV) to reduce the number of genes used to re-construct cell type-specific predictive models. The optimal genes obtained using RFECV were denoted as RFE genes, which were used for the downstream analyses to depict the cell type heterogeneity of ASD.

Secondly, to screen gene sets associated with ASD in each cell type, we constructed cell type-specific gene set-based predictive models using PLS. Specifically, for each cell type and each gene set, we extracted the expression level of the nuclei from the cell type in the genes included in the considered gene set, and constructed a predictive model. To prioritize gene sets, we ranked gene sets using their predictive performance on the test set, and kept the gene sets whose predictive accuracy (ACC), sensitivity (SN), and specificity (SP) are larger than 70% as cell type-specific gene sets associated with ASD. Besides, for the total genes included in these identified gene sets, we calculated their frequency and averaged importance, and used the genes with top averaged importance to re-construct cell type-specific predictive models.

Lastly, we further constructed a multi-label predictive model using PLS, which can predict the cell type and the diagnosis of a given nuclei at the same time. For the labels of cell types, the predictive cell type of each nucleus was set as the cell type whose predictive probability is the largest. For the diagnosis label, we extracted the predictive probability of training set and applied ROC analysis to obtain the optimal cutoff for determining the predictive diagnosis of each nucleus in training and test sets.

3.2. Cell type-specific genes associated with ASD

For each of the 17 cell types, we first constructed a cell type-specific predictive model using all genes (Table 1, **Supplementary File 1**). To score genes in each cell type, we calculated the importance of genes and ranked the genes (**Supplementary File 2**). Next, in order to use less genes to achieve similar

performances, we used the genes with top 500, 1000, and 1500 importance respectively to re-construct cell type-specific predictive models. We found out that using top 1000 genes made the model performance better than the one using top 500, while approach the one using top 1500 genes (**Supplementary File 1**). Therefore, for each cell type, we applied RFECV to reduce the number of genes to up to 1000 genes and also to obtain the optimal number of genes used to construct a cell type-specific predictive model. The optimal genes obtained using RFECV were denoted as RFE genes. It is noted that the performances on test sets of the cell type-specific predictive models based on RFE genes approach the ones based on all genes (Fig. 2. **A, Supplementary File 1**), hence, we used the RFE genes for the subsequently analyses in this section.

Table 1

The classification performances of cell type-specific predictive models built using all genes. The number of nuclei from ASD and controls are listed. ROC analysis was applied to obtain the AUC and the optimal cutoff point on the training set, then the optimal cutoff was used to determine the predictive accuracy (ACC), sensitivity (SN), and specificity (SP) on the training and test sets.

Cell type (ASD/Control)	Training set				Test set			
	ACC	SN	SP	AUC	ACC	SN	SP	AUC
AST-FB (2033/1622)	0.91	0.92	0.9	0.97	0.72	0.78	0.63	0.79
AST-PP (4749/2336)	0.93	0.93	0.93	0.98	0.84	0.87	0.79	0.90
Endothelial (850/1141)	0.92	0.91	0.92	0.97	0.76	0.70	0.80	0.83
IN-PV (1811/1908)	0.95	0.94	0.96	0.99	0.80	0.77	0.82	0.88
IN-SST (1945/2245)	0.94	0.92	0.95	0.98	0.76	0.70	0.81	0.83
IN-SV2C (990/846)	0.98	0.98	0.97	1.00	0.80	0.83	0.76	0.88
IN-VIP (3098/2523)	0.89	0.88	0.91	0.96	0.79	0.79	0.78	0.86
L2/3 (6962/5833)	0.95	0.95	0.95	0.99	0.89	0.90	0.88	0.96
L4 (3415/3103)	0.93	0.91	0.94	0.98	0.83	0.80	0.87	0.91
L5/6 (1710/1692)	0.93	0.93	0.93	0.98	0.78	0.77	0.80	0.86
L5/6-CC (2279/2106)	0.97	0.98	0.97	1.00	0.85	0.88	0.82	0.93
Microglia (1174/1321)	0.91	0.90	0.93	0.97	0.76	0.73	0.78	0.84
Neu-mat (1853/1679)	0.85	0.82	0.88	0.93	0.75	0.70	0.80	0.83
Neu-NRGN-I (321/268)	0.97	0.99	0.94	0.99	0.69	0.75	0.63	0.74
Neu-NRGN-II (828/631)	0.82	0.86	0.78	0.89	0.63	0.70	0.53	0.68
Oligodendrocytes (4587/7619)	0.83	0.86	0.81	0.91	0.77	0.79	0.75	0.85
OPC (5085/4562)	0.83	0.82	0.84	0.91	0.75	0.74	0.76	0.82

By examining the number of RFE genes in every cell type (Table 2), we found that in several cell types, such as AST-PP, IN-PV, L2/3, L4, and L5/6-CC, there are more RFE genes and the corresponding cell type-specific predictive models have better performances than other cell types (Fig. 2. **A**). This implies that these cell types may be more vulnerable in ASD and more genes may be dysregulated in these cell types. Then, for each cell type, we also applied edgeR [33] to identify differential expressed genes in ASD compared to controls. It can be seen that in the mentioned cell types above, there are indeed more differential expressed genes, which also indicates that these cell types may be mainly affected by ASD. By performing hypergeometric tests, we found that the RFE genes are significantly overlapped with the differential expressed genes identified by edgeR (Table 2). Then we checked if building cell type-specific predictive models using edgeR genes would be better than the ones using RFE genes, while the model performances using RFE genes is better than the ones using edgeR genes (**Supplementary File 1**). This shows that genes that are not identified by edgeR may have predictive power for ASD. Next, we found that there are more SFARI ASD genes overlapped with RFE genes in neuron-related cell types. We also performed overrepresentation tests between RFE genes and SFARI ASD genes, and found that RFE genes are significantly overlapped with ASD genes (Table 2).

Table 2

The overrepresentation tests between RFE genes and differential expressed genes identified by edgeR and SFARI ASD genes. The number of overlapping genes between RFE genes and edgeR genes or ASD genes, the number of total edgeR genes or ASD genes, and the FDR-adjusted hypergeometric test *P*-value are shown. The genes with top five importance are listed, of which edgeR genes are bold, and SFARI ASD genes are underlined.

Cell type	Number of RFE genes	Overlapping genes/edgeR genes (FDR-adjusted *P*-value)	Overlapping genes/ASD genes (FDR-adjusted *P*-value)	Top five important genes
AST-FB	200	120/257 (1.5e-158)	22/299 (5.8e-09)	DPP10, TMSB4X, SPARCL1, ZFP36L1, PCDH9
AST-PP	1000	667/1464 (0.0e + 00)	98/299 (2.2e-34)	PTGDS, HSPA1A, TRPM3, RP11-179A16.1, CIRBP
Endothelial	500	115/146 (3.2e-134)	40/299 (6.3e-11)	HERC2P3, AKAP12, TMSB4X, *RP11-649A16.1*, RPS28
IN-PV	1000	384/695 (4.2e-251)	103/299 (2.7e-38)	AC105402.4, MTATP6P1, CNTNAP3, CIRBP, *ARL17B*
IN-SST	1000	549/1346 (5.9e-291)	104/299 (5.3e-39)	SST, AC105402.4, VGF, HSPA1A, BCYRN1
IN-SV2C	900	345/616 (7.4e-244)	100/299 (9.7e-40)	CCK, BCYRN1, AC105402.4, MEG3, HSPB1
IN-VIP	1000	676/1820 (0.0e + 00)	104/299 (5.3e-39)	HSPA1A, CCK, RPS15, MEG3, RGS12
L2/3	1000	863/4690 (7.8e-230)	107/299 (2.9e-41)	BCYRN1, CCK, CNTNAP2, MEG3, CAMK2N1
L4	1000	715/2477 (1.3e-294)	113/299 (3.7e-46)	BCYRN1, CCK, NCAM2, SLC17A7, MTATP6P1
L5/6	900	467/1069 (4.0e-281)	98/299 (2.7e-38)	BCYRN1, AC105402.4, MTATP6P1, ATP1B1, SLC17A7
L5/6-CC	1000	701/3183 (7.1e-202)	114/299 (7.5e-47)	BCYRN1, CCK, AC105402.4, *RP11-750B16.1*, MT-RNR2

Cell type	Number of RFE genes	Overlapping genes/edgeR genes (FDR-adjusted *P*-value)	Overlapping genes/ASD genes (FDR-adjusted *P*-value)	Top five important genes
Microglia	200	74/106 (4.9e-112)	20/299 (1.4e-07)	FKBP5, TMSB4X, NEAT1, SLC1A3, CHN2
Neu-mat	900	351/476 (1.7e-312)	116/299 (2.9e-53)	AC105402.4, *XIST*, CAMK2N1, MEG3, ROBO2
Neu-NRGN-I	100	2/2 (6.8e-05)	12/299 (7.0e-06)	*RP11-750B16.1, PTMA, NRGN, GNAO1*, TSPAN7
Neu-NRGN-II	100	7/8 (1.9e-14)	6/299 (3.8e-02)	PRNP, NRGN, STMN1, RP11-750B16.1 *PLP1*
Oligodendrocytes	600	410/1420 (9.4e-253)	57/299 (9.5e-19)	PTGDS, NRXN1, CNDP1, ABCA2, CREB5
OPC	900	528/1413 (1.9e-285)	102/299 (2.9e-41)	GPC5, TMSB4X, HSPH1, CNTNAP2, OLIG1

For each cell type-specific predictive model built based on RFE genes, we calculated the importance of each RFE gene (**Supplementary File 3**). Table 2 lists the top RFE genes in each cell type. Figure 2. **B** also demonstrates the expression of top three RFE genes in ASD and control groups for the representative cell types, including AST-PP, endothelial, IN-PV, L2/3, microglia, oligodendrocytes, and OPC. The top genes among different cell types are distinct, implying the cell type heterogeneity of ASD. However, some top genes appearing in several cell types are of note. For instance, gene *BCYRN1* (brain cytoplasmic RNA 1, a long non-coding RNA) has the largest importance in all excitatory neurons, including L2/3, L4, L5/6, and L5/6-CC. Gene *BCYRN1* is involved in the regulation of synaptogenesis, and there have been several literatures linking *BCYRN1* and Alzheimer's disease, a neurological disease [34, 35], which implies the possible association between *BCYRN1* and ASD. Besides, *BCYRN1* has been prioritized in a blood-based gene expression study of ASD [36].

To further characterize the cell type heterogeneity of ASD, we compared the RFE genes across different cell types. We preformed gene ontology analyses using R package of clusterProfiler [37], with background genes set as the genes in the analyzed gene expression matrix. The functions of cell type-specific RFE genes are not consistent among different cell types (**Supplementary File 4**). For instance, in IN-PV, the enriched GO terms include neuron projection, axon, somatodendritic compartment, and cell part morphogenesis, while in L2/3 the top GO terms are associated with ribosome, cotranslational protein targeting to membrane, and protein localization to endoplasmic reticulum (Fig. 2. **B**).

3.3. Cell type-specific gene sets associated with ASD

In addition to screening individual genes associated with ASD, we also constructed cell type-specific gene set-based predictive models to screen ASD-related gene sets. For each cell type and each gene set, we extracted the expression level of the nuclei from the cell type in the genes included in the considered gene set, and constructed a predictive model (**Materials and Methods**). We retained the gene sets whose ACC, SN, and SP on test set are larger than 70%, and there are 5, 1, 88, 15, 137 gene sets identified in cell types of AST-PP, IN-PV, L2/3, L4, and L5/6-CC respectively (**Supplementary File 5**). Figure 3. **A** shows the top five gene sets in each of these five cell types and the performances of corresponding cell type-specific gene set-based predictive models. For AST-PP, the top ASD-associated gene sets include *REACTOME_DISEASE, GO_REGULATION_ OF_CELL_POPULATION_PROLIFERATION, GO_POSITIVE_REGULATION_OF_CATALYTIC_ ACTIVITY, GO_SIGNALING_RECEPTOR_BINDING* and *GO_ENZYME_LINKED_RECEPTOR_ PROTEIN_SIGNALING_PATHWAY*. For other neuron cell types, the ASD-associated gene sets are mostly related to cell junction, synapse, neuron projection, neurogenesis, neuron differentiation, and cell projection organization. By checking the top important genes in each cell type-specific gene set, we found that several genes appear in the majority of the gene sets, for example, gene *HSPA1A* [heat shock protein family A (*HSP70*) member 1A] shows up in all AST-PP specific ASD-associated gene sets (Fig. 3. **B**). Therefore, for each cell type, we analyzed the frequency of each gene included in the identified gene sets, and calculated the averaged importance of genes (**Supplementary File 5**). Figure 3. **C** shows the genes with top five averaged importance in each cell type. Gene *HSPA1A* is noted in AST-PP. Actually, heat shock proteins play a central role in the development of neurological disorders, of which *HSP70* family has been shown its functions [38], and *HSPA1A*, a member of *HSP70* family, has already been associated with ASD [39]. As to gene *CCK* (cholecystokinin), it is prioritized in excitatory neurons, which a kind of gut peptide hormones. Gut peptide hormones have been found across different brain regions and many of them are involved with ASD-related deficits [40].

Next, based on the genes with averaged importance > 10% in corresponding cell types, we re-constructed a cell type-specific predictive model for each of these five cell types. It is noted that their predictive performances are even better that the ones of the cell type-specific gene set-based predictive models (Fig. 3. **D**). We checked the functions of these genes (**Supplementary File 6**), and found their functions are distinct, especially among AST-PP, IN-PV and excitatory neurons (Fig. 3. **E**). In AST-PP, the top genes are associated with the functions of enzyme linked receptor protein signaling pathway, transmembrane receptor protein tyrosine kinase signaling pathway, positive regulation of phosphorus and phosphate metabolic process, and cellular component morphogenesis. In IN-PV, the top genes are related to synaptic and postsynaptic membrane, cation channel complex, and neuron projection. As to the cell types of excitatory neurons, the top genes are associated with ribosome, SRP-dependent cotranslational protein targeting to membrane, nuclear-transcribed mRNA catabolic process, nonsense-mediated decay, and protein targeting to ER.

3.4. A multi-label classification model predicting cell type and diagnosis

To predict the cell type and diagnosis of a given nucleus at the same time, we applied PLS to construct a multi-label predictive model (**Materials and Methods**). We split the whole gene expression data to training set and test set. For the diagnosis label, we extracted the predictive probability of training set and applied ROC analysis to obtain the optimal cutoff for determining the predictive diagnosis of each nucleus in training and test sets. For the cell type labels, the predictive cell type of each nucleus was set as the cell type whose predictive probability is the largest. The Hamming loss of the multi-label predictive model is 0.02, and the accuracy achieves 72.8% with 92.7% accuracy for cell type labels and 78.5% accuracy for diagnosis label. Then we examined the predictive performance of the model in each cell type. For each cell type, Fig. 4 illustrates the proportion of the number of nuclei predicted as each cell type to the total number of nuclei, the proportion of correct and incorrect predictions for the label of diagnosis, the proportion of correct predictions for all labels in the test set. It can be seen that for most cell types, the predictive cell types are correct, except for AST-FB and Neu-mat. Because AST-FB and AST-PP are cell clusters of astrocytes and they may have similar gene expression patterns, a part of nuclei from AST-FB is predicted as AST-PP. As to Neu-mat, more than 40% nuclei were predicted as L2/3, which may indicate the gene expression patterns between Neu-mat and L2/3 are similar. For most cell types, the predictive accuracy of diagnosis label is larger than 70%, and the top highest accuracy values appear in L2/3, L5/6-CC, IN-SV2C, L4 and AST-PP, showing these cell types may be more vulnerable in ASD.

4. Discussion

Genetic studies have identified variants associated with ASD, while the causal variants and the specific cell types in which the disease-risk variants may be active are unclear. Genes may demonstrate diverse functions across different brain cell types. Different functions may be dysregulated and causal genes may be distinct among different brain cells in ASD. Recently, the newly available single-nucleus RNA-sequencing data of ASD [11] makes it possible to study the cell type heterogeneity of ASD directly. The authors identified differentially expressed genes between ASD and controls in a cell type-specific way, and found that the top differentially expressed (DE) neuronal genes were identified in L2/3 and IN-VIP, and the top genes differentially expressed in non-neuronal cell types were identified in AST-PP and microglia. The relative changes of DE genes in L2/3 and microglia were the most predictive of clinical severity of ASD patients and the cell types that are recurrently affected across multiple patients included L2/3 and L5/6-CC. They concluded that synaptic signaling of upper-layer excitatory neurons and the molecular state of microglia are preferentially affected in ASD, and the dysregulation of specific groups of genes in cortico-cortical projection neurons correlates with clinical severity of ASD.

Actually, except for genetic and genomic studies, gene prioritization studies [10, 22–24] can be applied to detect ASD risk genes, which can help to identify high-confidence gene candidates. In this study, to characterize the cell type heterogeneity of ASD and to identify cell type-specific genes and gene sets associated with ASD, we constructed multiple kinds of predictive models based on the human brain nucleus gene expression data of ASD and controls [11]. By constructing cell type-specific predictive models based on individual genes, we found that AST-PP, IN-PV, L2/3, L4, and L5/6-CC may be more vulnerable in ASD. They have more RFE genes and the corresponding cell type-specific predictive models

have better performances. Actually, they have more differential expressed genes identified by edgeR and more SFARI ASD genes. These indicate more genes may be dysregulated in these cell types and these cell types may be mainly affected by ASD. The functions of genes with predictive power for ASD are not consistent and the top important genes are distinct among different cell types, implying the cell type heterogeneity of ASD. However, some genes appearing as top important genes in several cell types are of note. For instance, gene *BCYRN1* has the largest importance in all excitatory neurons, including L2/3, L4, L5/6, and L5/6-CC. Gene *BCYRN1* is involved in the regulation of synaptogenesis, and there have been several literatures linking *BCYRN1* and Alzheimer's disease, a neurological disease [34, 35], which implies the possible association between *BCYRN1* and ASD. Besides, *BCYRN1* has been prioritized in a blood-based gene expression study of ASD [36].

As genes interact with others, the integrity of disease gene modules instead of individual genes may determine the manifestation of a disease in cells [12, 13]. Therefore, in addition to identifying the individual cell type-specific risk genes, it is valuable to identify cell type-specific gene sets/modules associated with ASD. From the view of gene set, by constructing cell type-specific gene set-based predictive models, we also noted cell types of AST-PP, IN-PV, L2/3, L4, and L5/6-CC. The identified gene sets specific to these cell types are not consistent. For AST-PP, the ASD-associated gene sets include *REACTOME_DISEASE, GO_REGULATION_OF_CELL_ POPULATION_PROLIFERATION, GO_POSITIVE_REGULATION_OF_CATALYTIC_ACTIVITY, GO_SIGNALING_RECEPTOR_BINDING* and *GO_ENZYME_LINKED_RECEPTOR_PROTEIN_ SIGNALING_PATHWAY*. For other four neuronal cell types, the ASD-associated gene sets are mostly related to cell junction, synapse, neuron projection, neurogenesis, neuron differentiation, and cell projection organization. We found gene *HSPA1A* appears as the most important gene in all AST-PP specific ASD-associated gene sets. Actually, heat shock proteins play a central role in the development of neurological disorders, of which *HSP70* family has been shown its functions [38], and *HSPA1A*, a member of *HSP70* family, has already been associated with ASD [39]. Gene *CCK* is prioritized in L2/3, L4, and L5/6-CC, which a kind of gut peptide hormones. Gut peptide hormones have been found across different brain regions and many of them are involved with ASD-related deficits [40].

Overall, we found that the functions of genes with predictive power for ASD are not consistent and the top important genes are distinct among different cell types, depicting the cell type heterogeneity of ASD. The findings suggest that L2/3, L4, L5/6-CC, AST-PP, and IN-PV are mainly affected in ASD. The results show that it may be feasible to use single cell/nucleus gene expression for ASD detection and the constructed predictive models can promote the diagnosis of ASD. Our method prioritizes ASD-associated cell type-specific genes and gene sets, which may be used as potential biomarkers of ASD, promoting the design of effective interventions.

5. Limitations

In this study, we analyzed the newly available human brain nucleus gene expression data of ASD and controls and constructed cell type-specific predictive models for ASD, identifying ASD-associated cell

type-specific genes and gene sets. We constructed the models by splitting the data into training, validation and test sets, while it is more completed to use another independent dataset to test the models. However, currently there is only one human brain cell/nucleus RNA-seq dataset of ASD, which is the data we analyzed in this study. Besides, our study utilized methods of machine learning to detect biomarkers of ASD, while it is more convincing if subsequent experiments could be conducted to validate the results.

6. Conclusions

To screen cell type-specific genes and gene sets associated with ASD, we constructed cell type-specific classification models, which can predict the diagnosis of a given nucleus with known cell type, based on the human brain nucleus gene expression data of ASD and controls. We further constructed a multi-label classification model for predicting the cell type and diagnosis of a given nucleus at the same time. The findings suggest that layer 2/3 and layer 4 excitatory neurons, layer 5/6 cortico-cortical projection neurons, parvalbumin interneurons, and protoplasmic astrocytes are preferentially affected in ASD. The functions of genes with predictive power for ASD are not consistent and the top important genes are distinct among different cell types, depicting the cell type heterogeneity of ASD. Our results show that it may be feasible to use single cell/nucleus gene expression for ASD detection and the constructed predictive models can promote the diagnosis of ASD. Our analytical pipeline prioritizes ASD-associated cell type-specific genes and gene sets, which may be used as potential biomarkers of ASD. This study provides new insights for understanding the cell type-specific molecular pathology of ASD.

Abbreviations

ASD: autism spectrum disorder

AST-FB: fibrous astrocytes

AST-PP protoplasmic astrocytes

IN-PV: parvalbumin interneurons

IN-SST: somatostatin interneurons

IN-SV2C: SV2C interneurons

IN-VIP: VIP interneurons

L2/3: layer 2/3 excitatory neurons

L4: layer 4 excitatory neurons

L5/6: layer 5/6 corticofugal projection neurons

L5/6-CC: layer 5/6 cortico-cortical projection neurons

Neu-mat: maturing neurons

Neu-NRGN-I: NRGN-expressing neurons

Neu-NRGN-II: NRGN-expressing neurons

PLS: partial least squares

ROC: receiver operating characteristic

AUC: area under ROC curve

RFECV: recursive feature elimination with cross-validation

SN: sensitivity

SP: specificity

ACC: accuracy

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The analyzed datasets and codes in this study are available from the authors upon request.

Funding

This work has been supported by the National Natural Science Foundation of China (Nos. 61803320, and 61573296), and the Fundamental Research Funds for the Central Universities in China (Xiamen University: 202010384099), and the fund with Xiamen YLZ Yihui Technology Co., Ltd (XDHT2020131A).

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

JG conceived the study. YW, YL and QY wrote the codes and analyzed the data. JG, YZ and GJ interpreted the results. JG and YW wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements

Not applicable.

References

1. De Rubeis S, He X, Goldberg AP, Poultney CS, Samocha K, Cicek AE, et al. Synaptic, transcriptional and chromatin genes disrupted in autism. Nature. 2014;515(7526):209–15.
2. Turner TN, Hormozdiari F, Duyzend MH, McClymont SA, Hook PW, Iossifov I, et al. Genome Sequencing of Autism-Affected Families Reveals Disruption of Putative Noncoding Regulatory DNA. Am J Hum Genet. 2016;98(1):58–74.
3. O'Roak BJ, Vives L, Girirajan S, Karakoc E, Krumm N, Coe BP, et al. Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. Nature. 2012;485(7397):246–50.
4. Sanders SJ, Murtha MT, Gupta AR, Murdoch JD, Raubeson MJ, Willsey AJ, et al. De novo mutations revealed by whole-exome sequencing are strongly associated with autism. Nature. 2012;485(7397):237–41.
5. Neale BM, Kou Y, Liu L, Ma'ayan A, Samocha KE, Sabo A, et al. Patterns and rates of exonic de novo mutations in autism spectrum disorders. Nature. 2012;485(7397):242–5.
6. Iossifov I, Ronemus M, Levy D, Wang Z, Hakker I, Rosenbaum J, et al. De novo gene disruptions in children on the autistic spectrum. Neuron. 2012;74(2):285–99.
7. Gaugler T, Klei L, Sanders SJ, Bodea CA, Goldberg AP, Lee AB, et al. Most genetic risk for autism resides with common variation. Nat Genet. 2014;46(8):881–5.
8. Voineagu I, Wang X, Johnston P, Lowe JK, Tian Y, Horvath S, et al. Transcriptomic analysis of autistic brain reveals convergent molecular pathology. Nature. 2011;474(7351):380–4.
9. Gupta S, Ellis SE, Ashar FN, Moes A, Bader JS, Zhan J, et al. Transcriptome analysis reveals dysregulation of innate immune response genes and neuronal activity-dependent genes in autism. Nature communications. 2014;5:5748.

10. Guan J, Yang E, Yang J, Zeng Y, Ji G, Cai JJ. Exploiting aberrant mRNA expression in autism for gene discovery and diagnosis. Hum Genet. 2016;135(7):797–811.
11. Velmeshev D, Schirmer L, Jung D, Haeussler M, Perez Y, Mayer S, et al. Single-cell genomics identifies cell type–specific molecular changes in autism. Science. 2019;364(6441):685–9.
12. Kitsak M, Sharma A, Menche J, Guney E, Ghiassian SD, Loscalzo J, et al. Tissue Specificity of Human Disease Module. Sci Rep. 2016;6(1):35241.
13. Mohammadi S, Davila-Velderrain J, Kellis M. Reconstruction of Cell-type-Specific Interactomes at Single-Cell Resolution. Cell Systems. 2019;9(6):559 – 68.e4.
14. Hyde KK, Novack MN, LaHaye N, Parlett-Pelleriti C, Anden R, Dixon DR, et al. Applications of Supervised Machine Learning in Autism Spectrum Disorder Research: a Review. Review Journal of Autism Developmental Disorders. 2019;6(2):128–46.
15. Levy S, Duda M, Haber N, Wall DP. Sparsifying machine learning models identify stable subsets of predictive features for behavioral detection of autism. Molecular autism. 2017;8:65-.
16. Duda M, Kosmicki JA, Wall DP. Testing the accuracy of an observation-based classifier for rapid detection of autism risk. Translational Psychiatry. 2014;4(8):e424-e.
17. Duda M, Daniels J, Wall DP. Clinical Evaluation of a Novel and Mobile Autism Risk Assessment. J Autism Dev Disord. 2016;46(6):1953–61.
18. Wall DP, Dally R, Luyster R, Jung J-Y, DeLuca TF. Use of Artificial Intelligence to Shorten the Behavioral Diagnosis of Autism. PloS one. 2012;7(8):e43855.
19. Heinsfeld AS, Franco AR, Craddock RC, Buchweitz A, Meneguzzi F. Identification of autism spectrum disorder using deep learning and the ABIDE dataset. NeuroImage: Clinical. 2018;17:16–23.
20. Chen R, Jiao Y, Herskovits EH. Structural MRI in Autism Spectrum Disorder. Pediatr Res. 2011;69(8):63–8.
21. Bosl WJ, Tager-Flusberg H, Nelson CA. EEG Analytics for Early Detection of Autism Spectrum Disorder: A data-driven approach. Sci Rep. 2018;8(1):6828.
22. Cogill S, Wang L. Support vector machine model of developmental brain gene expression data for prioritization of Autism risk gene candidates. Bioinformatics. 2016;32(23):3611–8.
23. Kong SW, Collins CD, Shimizu-Motohashi Y, Holm IA, Campbell MG, Lee IH, et al. Characteristics and predictive value of blood transcriptome signature in males with autism spectrum disorders. PloS one. 2012;7(12):e49475.
24. Oh DH, Kim IB, Kim SH, Ahn DH. Predicting Autism Spectrum Disorder Using Blood-based Gene Expression Signatures and Machine Learning. Clin Psychopharmacol Neurosci. 2017;15(1):47–52.
25. Lun ATL, McCarthy DJ, Marioni JC. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. F1000Research. 2016;5:2122.
26. Calvo SE, Clauser KR, Mootha VK. MitoCarta2.0: an updated inventory of mammalian mitochondrial proteins. Nucleic acids research. 2016;44(D1):D1251-D7.

27. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdottir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. Bioinformatics. 2011;27(12):1739–40.
28. Wold H. Estimation of principal components and related models by iterative least squares. Multivariate analysis. 1966:391–420.
29. Boulesteix A-L, Strimmer K. Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. Brief Bioinform. 2006;8(1):32–44.
30. Kuhn M. Building predictive models in R using the caret package. Journal of statistical software. 2008;28(5):1–26.
31. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, et al. pROC: an open-source package for R and S + to analyze and compare ROC curves. 2011;12(1):1–8.
32. Bischl B, Lang M, Kotthoff L, Schiffner J, Richter J, Studerus E, et al. mlr: Machine Learning in R. 2016;17(1):5938–42.
33. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139–40.
34. Wan P, Su W, Zhuo Y. The Role of Long Noncoding RNAs in Neurodegenerative Diseases. Mol Neurobiol. 2017;54(3):2012–21.
35. Hu G, Niu F, Humburg BA, Liao K, Bendi VS, Callen S, et al. Molecular mechanisms of long noncoding RNAs and their role in disease pathogenesis. 2018;9(26).
36. Ivanov HY, Stoyanova VK, Popov NT, Bosheva M, Vachev TIJB. Blood-based gene expression in children with autism spectrum disorder. 2015;17:e8966.
37. Yu G, Wang L-G, Han Y, He Q-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS. 2012;16(5):284–7.
38. Turturici G, Sconzo G, Geraci F. Hsp70 and its molecular role in nervous system diseases. Biochem Res Int. 2011;2011:618127.
39. Lin M, Zhao D, Hrabovsky A, Pedrosa E, Zheng D, Lachman HM. Heat Shock Alters the Expression of Schizophrenia and Autism Candidate Genes in an Induced Pluripotent Stem Cell Model of the Human Telencephalon. PloS one. 2014;9(4):e94968.
40. Qi X-R, Zhang L. The Potential Role of Gut Peptide Hormones in Autism Spectrum Disorder. Frontiers in cellular neuroscience. 2020;14:73.

Description Of Supplementary Materials

Supplementary File 1. The performances of cell type-specific predictive models built based on all genes, top 500, 1000 and 1500 important genes, RFE genes and edgeR genes.

Supplementary File 2. The calculated gene importance in each cell type-specific predictive models built based on all genes.

Supplementary File 3. The calculated gene importance in each cell type-specific predictive models built based on RFE genes.

Supplementary File 4. The enriched GO terms with cell type-specific RFE genes.

Supplementary File 5. The identified ASD-associated cell type-specific gene sets along with their top five important genes and predictive performances. For these gene sets, the frequency and averaged importance of each gene included in the gene sets are listed.

Supplementary File 6. The enriched GO terms for genes with averaged importance > 10% included in the identified ASD-associated cell type-specific gene sets.

Figures

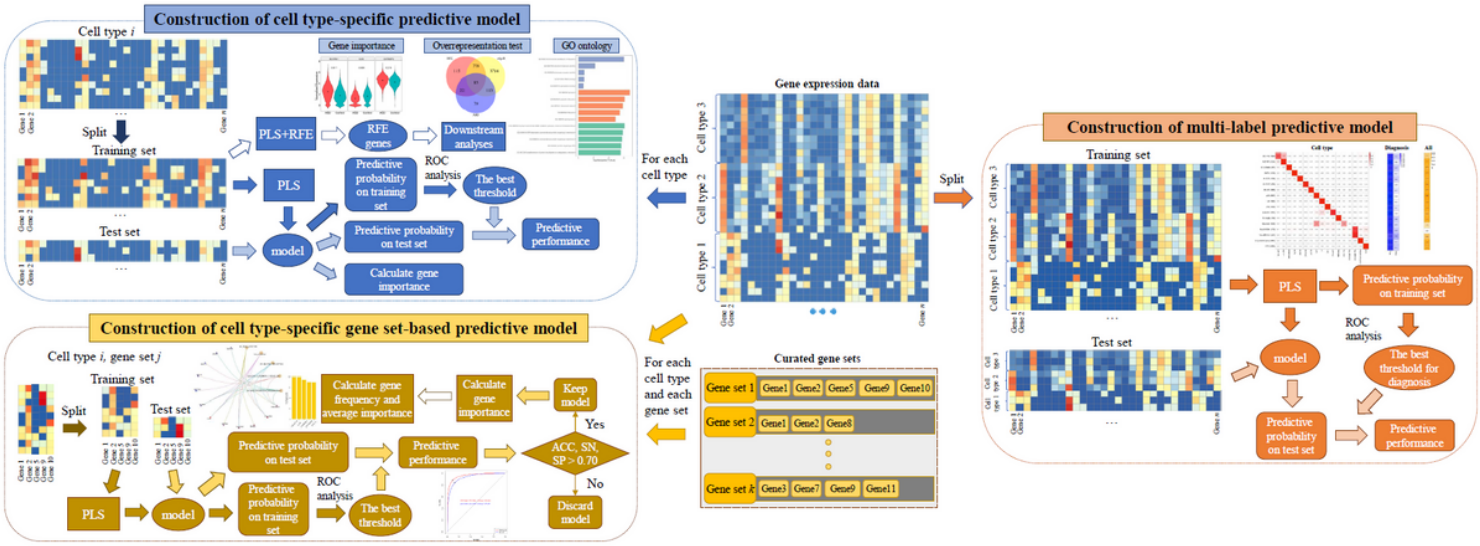

Figure 1

The methodological overview.

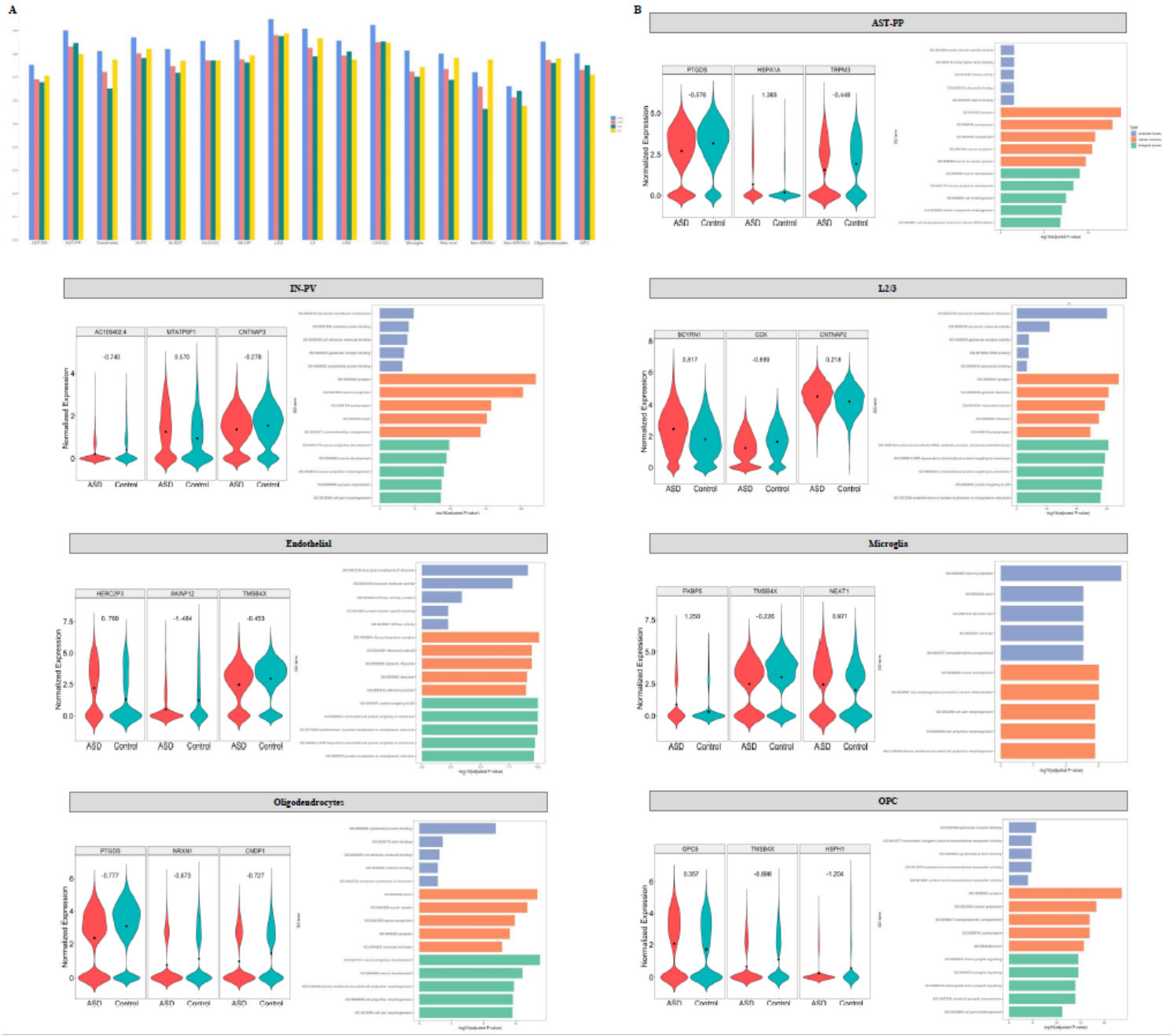

Figure 2

A) The classification performance on test set of cell type-specific predictive models built using RFE genes. ROC analysis was applied to obtain the AUC and the optimal cutoff point on the training set, then the optimal cutoff was used to determine the predictive accuracy (ACC), sensitivity (SN), and specificity (SP) on the test set. For the cell types of AST-PP, endothelial, IN-PV, L2/3, microglia, oligodendrocytes, and OPC, (B) the expression of top three important genes in ASD and control groups are shown along with the top enriched GO terms with the RFE genes.

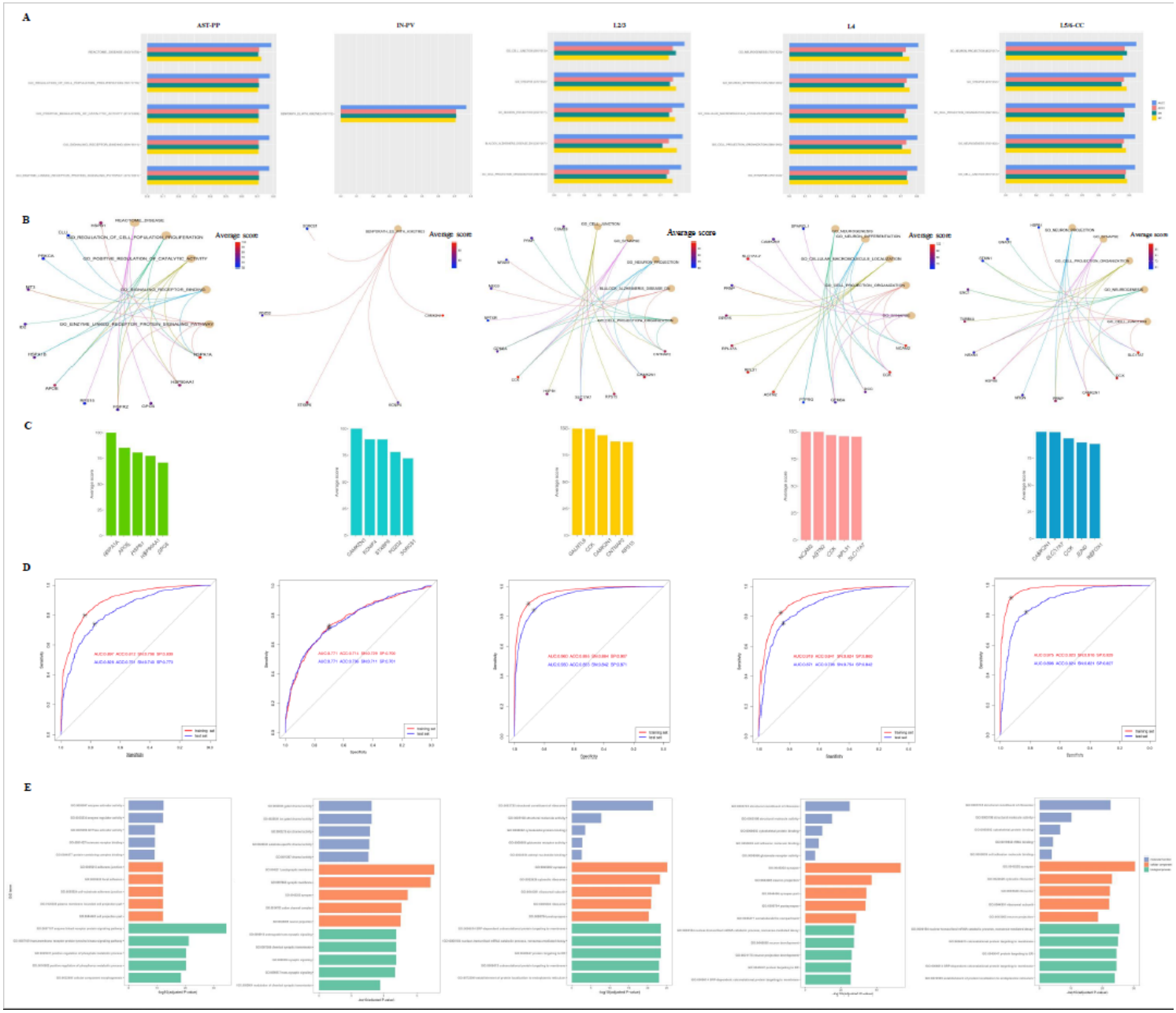

Figure 3

(A) The identified top five gene sets associated with ASD by constructing cell type-specific gene set-based predictive models. The number of overlapping genes between the gene expression data and the gene set, the total number of genes in the gene set, and the performances of corresponding cell type-specific gene set-based predictive models are shown. For each cell type, (B) illustrates the top five gene sets and the genes with top five importance in each gene set, and (C) plots the genes with top averaged importance. (D) The performances of predictive models built using genes with averaged importance > 10% and (E) the enriched GO terms with these genes.

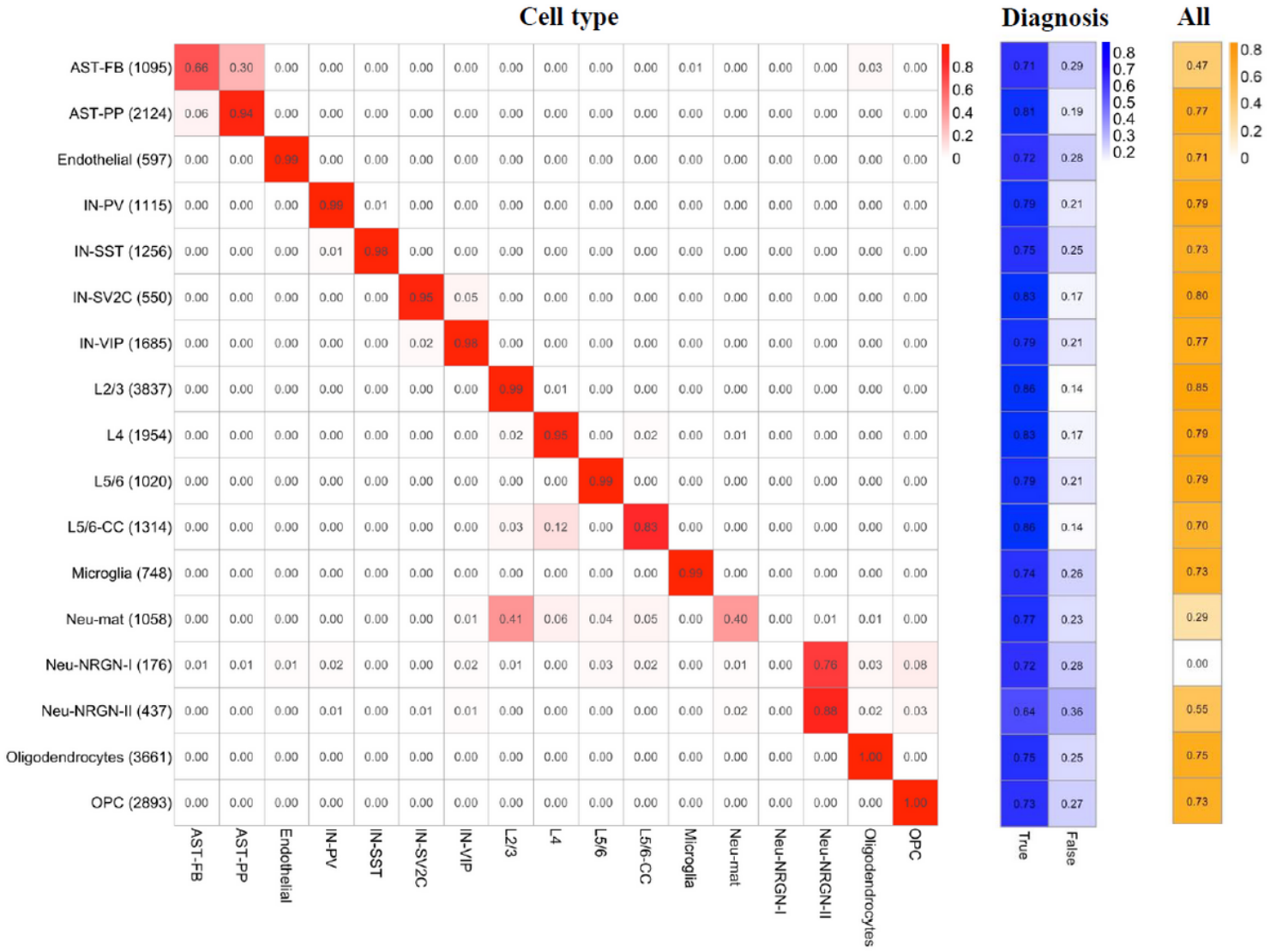

Figure 4

For each cell type, (left) the proportion of the number of nuclei predicted as each cell type to the total number of nuclei, (middle) the proportion of correct and incorrect predictions for the label of diagnosis, (right) the proportion of correct predictions for all labels in the test set.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryFile1.Modelperformance.xlsx
- SupplementaryFile2.Genescores.xlsx
- SupplementaryFile3.RFEgenescores.xlsx
- SupplementaryFile4.GOanalysisofRFEgenes.xlsx
- SupplementaryFile5.ASDassociatedGenesets.xlsx

- SupplementaryFile6.GOtopaveragedscoregenes.xlsx