Sparse representation based patient-specific disease prediction and treatment for autism spectrum disorder

Dongbai Liu[1,2], Megan Herceg[3], Hongbao Cao[4,5], Fuquan Zhang[6*]

[1]Department of Neurology, The First People's Hospital Affiliated to Soochow University, Suzhou, Jiangsu Province, 215006, China; [2]Department of Neurology, Jiangyin People's hospital, Jiangyin, Jiangsu Province, 214400, China; [3]School of Systems Biology, George Mason University, Manassas, VA, 20110, USA;[4]Department of Genomics Research, R&D Solutions, Elsevier Inc., Rockville, MD, 20852, USA; [5]Unit on Statistical Genomics, NIMH/NIH, Bethesda, 20852, USA; [6]Wuxi Mental Health Center, Nanjing Medical University, Wuxi, Jiangsu Province, 214151, China

∗ Corresponding to:
Dr. Fuquan Zhang
Wuxi Mental Health Center, Nanjing Medical University
156 Qianrong Road, Wuxi, Jiangsu Province, 214151, China
E-mails: zhangfq@njmu.edu.cn
Tel: +86 510 83219310, Fax: +86 510 83012201

Running Head: Computational analysis for ASD.

Abstract

We proposed a sparse representation based variable selection (SRVS) approach with the purpose to assist personalized diagnosis for autism spectrum disorder (ASD). The approach was applied in 4 independent gene expression datasets. The SRVS method identified a unique gene vector for each patient group, leading to significantly higher classification accuracy compared to randomly selected genes (82.82%, 97.22%, 100%, and 62.41%; p-values: 0, 0, 0.0014, and 0.0014). The SRVS method also outperformed the ANOVA-based gene selection in terms of classification ratio. Our results suggest that SRVS together with literature data could serve as an effective method for personalized biomarker selection.

Keywords: Autism spectrum disorder; sparse representation; variable selection; precision medicine

1. Introduction

Autism spectrum disorder (ASD) is a common, highly heritable neurodevelopmental condition typically diagnosed in early childhood, which can last throughout a person's life [3]. People with ASD have substantial challenges in social interaction, communication, and learning. These patients display restrictive, repetitive behaviors, interests, and activities [3]. Despite nearly 50 years elapsed since Leo Kanner's seminal description of "Autistic Disturbances of Affective Contact" in a pediatric population [19], the neurological manifestations and causes of ASD are not fully understood. Previous studies found that ASD has an estimated heritability of around 50% [32]. In addition, twin studies suggested that genetic factors play a role in the etiology of autism [30, 31]. Recently, researchers found multiple genes involved in the etiology of autism[22, 14]. As such, significant researches investigating the genetic causes of ASD have been conducted, targeting earlier diagnosis and novel treatment development for the disease.

Hundreds of genes/proteins have been linked to ASD, reflecting its heterogeneity. Mutations of some genes have been frequently reported as risk factors for ASD, such as SLC6A4 and OXT [8, 29]. However, these genes may also be associated with multiple other diseases. For example, genetic variance of SLC6A4 was suggested to represent a risk factor for eating disorders (ED) [6], and its elevated activity has been associated with obsessive compulsive disorder (OCD) [40]. The lack of specificity of these genes decreases their utility in the diagnosis and treatment of ASD. On the other hand, many genes only appear in a small portion of ASD cases, such as AGER, BCHE and APBA2 [2, 4, 10]. Research is ongoing; there are dozens of novel genes associated with ASD that are identified each year. For instance, AKT1, ARC and ARHGAP32 were newly reported as ASD risk genes in 2016 [1, 26, 39]. These genes have not been identified in many other ASD studies, which may be due to the specificity of genome variations between subjects [23]. Therefore, early diagnosis and prediction of ASD may require multiple genes as biomarkers. Moreover, subject specificity should also be considered for the treatment.

To address this issue, this study first developed an ASD genetic database (ASD_042017), curating all ASD target genes available within the Pathway Studio (PS; http://pathwaystudio.com/), which possesses the largest databases in the field [21]. The disease prediction capability of the curated genes within ASD_042017 were tested using multiple independent gene expression datasets with an ASD case/control classification approach. A sparse representation-based variable selection (SRVS) algorithm was employed to select the best gene vector as a biomarker/feature for the classification. The SRVS algorithm has been previously shown to be effective in genetic and imaging variable

selection [7]. Instead of selecting a specific number of variables, the SRVS method generates a sparse regression weight for each variable, which can be used for variable ranking [7].

Our results confirm the specificity of the genomic variation of different ASD patient groups, and support the hypothesis that for a given ASD patient group, there may exist a gene vector from the curated ASD target genes that possesses significant predication power to separate ASD cases from heathy controls.

2. Methods and Materials

2.1 Development and analysis of ASD_042017

Figure 1 presents the database schema of **ASD_042017**. The database contains 529 genes, 41 drugs/small molecules, 106 diseases, and 91 pathways. Also included is the information from 2,098 and 267 supporting references for ASD-Gene and ASD-Drug relations, respectively. For each relation, there are one or more supporting references. The current ASD_042017 database has been deposited into '**Bioinformatics Database**' (http://database.gousinfo.com). It is scalable and will be updated monthly or upon request. To note, ASD presents multiple subtypes, including essential and complex autism. The 529 target genes were collected for all these ASD subtypes, which were intended to present a comprehensive target gene pool for the feature/gene selection for different patient groups.

The 91 pathways (**ASD_042017→Related Pathways**) were acquired using Pathway Enrichment Analysis (PEA) module of Pathway Studio (www.pathwaystudio.com). The 529 ASD genes were significantly enriched within these pathways (p-value<3e-08; q=0.001 for FDR). The 106 diseases (**ASD_042017→Related Diseases**) were identified using the Sub-Network Enrichment Analysis (SNEA) module in Pathway Studio (http://pathwaystudio.gousinfo.com/SNEA.pdf). The 529 ASD target genes present significant overlap with the genes linked to each of these 106 diseases (Fisher's Exact Test p-value<2.5e-33; q=0.001 for FDR).

The Gene-Gene Interaction (GGI) Network (**ASD_042017→GGI Network**) was generated based on the pathways enriched. Two genes were identified as connected if they shared one or more pathways, and the number of shared pathways dictated the weight of the edge. 100 potential drugs (**ASD_042017→Potential Drugs**) were identified using the SNEA module in Pathway Studio. The drugs/small molecules were significantly related to the ASD target genes, and those that have not been identified in clinical trial (96 out of 100) could represent potential drug candidates for ASD. These drugs demonstrated significant overlap with the drugs/small molecules related to ASD directly (**ASD_042017→Related drugs**; p-value = 1.96e-46).

Put Figure 1 about here.

2.2 SRVS for gene vector selection

The SRVS was used to rank the 529 ASD target genes according to a given experiment dataset. For each gene, a sparse weight was assigned by SRVS. The gene vector composed of the top n genes from SRVS yields the genetic marker for an ASD case/control group, where n is the number of genes corresponding to the maximum classification ratio (CR) as defined in Eq. (1).

$$classification\ ratio\ (CR) = \frac{\#correcly\ classified\ subjects}{\#total\ subjects} \quad (1)$$

In general, a sparse representation model is presented as Eq. (2)

$$y = X\delta + \varepsilon, \quad (2)$$

where $y \in R^{n\times 1}$ is the observation vector; $X \in R^{n\times p}$, $p \gg n$ are measurements of the data, and $\varepsilon \in R^{n\times 1}$ is the measurement error caused by noise. The goal is to reconstruct the unknown vector $\delta \in R^{p\times 1}$ based on *y* and X.

To best approximate y, by choosing a small number of non-zero entries of δ for the model given by Eq. (2), we consider the following L_p minimization problem (P0):

$$(P0) \quad min\|\delta\|_p \ \ subject\ to\ \ \|y - X\delta\|_2 \le \varepsilon \quad (3)$$

where $\|*\|_p$ is the L_p norm, and $p \in [0,1]$. The following algorithm is designed to solve the minimization problem (P0) given by Eq. (3) and detect the columns of X relevant to y.

SRVS Algorithm

1. Initialize $\delta^{(0)} = 0$;
2. For the Step l, randomly choose k columns from $X = \{x_1, \ldots, x_p\} \in R^{n\times p}$ to construct a $n \times k$ sub-matrix denoted as $X_l \in R^{n\times k}$; and mark the selected columns' indexes as $I_l \in R^{1\times k}$;
3. Solve the following L_p minimization problem to find the optimal sparse solution $\delta_l \in R^{k\times 1}$:

$$min\|\delta_l\|_p \quad \text{s.t.} \quad \|y - X_l\delta_l\|_2 \le \varepsilon \quad (4)$$

4. Update $\delta^{(l)} \in R^{p\times 1}$ with δ_l: $\delta^{(l)}(I_l) = \delta^{(l-1)}(I_l) + \delta_l$; where $\delta^{(l)}(I_l)$ and $\delta^{(l-1)}(I_l)$ denote the I_l th entries in $\delta^{(l)}$ and $\delta^{(l-1)}$ respectively;
5. If $\left\|\delta^{(l)}/l - \delta^{(l-1)}/(l-1)\right\|_2 > \alpha$, where α is a predefined constant, update $l = l + 1$, and go to Step 2. Otherwise, set $\delta = \delta^{(l)}/l$. The non-zero entries in δ correspond to the column vectors selected.

In Step 3, there are many proposed methods for solving the L_p minimization problem, such as the Homotopy method [13] for $p = 1$, and the orthogonal matching pursuit (OMP) algorithm [9] for $p = 0$.

2.3 Gene expression data

The present study employed four RNA gene expression datasets to evaluate the classification performance using ASD target genes, including GSE18123, GSE38322, GSE7329 and GSE37772. These datasets were selected from Illumina BaseSpace Correlation Engine (http://www.illumina.com) and are publicly available at NCBI Gene Expression Omnibus (www.ncbi.nlm.nih.gov/geo/). The data selection criteria were as follows: 1) the data were from human experiments, 2) the data were from RNA expression datasets, and 3) the study designs were ASD case vs. healthy control. Expression data of healthy controls and ASD patients were extracted from each dataset and used

for ASD case/control classification. The genes in each dataset were limited to ASD target genes curated within the database **ASD_042017**. The key statistics of the 4 datasets are summarized in Table 1.

Put Table 1 about here.

The gene expression profile of the 4 gene expression datasets were also included in **ASD_042017→GSE18123**, **GSE38322**, **GSE7329,** and **GSE37772**. Within each dataset, the SRVS generated weights (SRVSScore) and ANOVA generated p-value score (PValueScore; Log p-values: -10*log(p-value)) were also presented. The p-value for each gene is generated from the one-way ANOVA of the case/control comparison using the corresponding expression data. A SRVSScore or a PValueScore represents the significance of a gene to the dataset according to SRVS or ANOVA methods, respectively.

2.4 ASD case/control classification

To identify the best gene vector and the corresponding classification accuracy, expressed as a classification ratio (CR), the ASD target genes were first ranked by SRVSScore in descending order. Then, a Euclidean distance-based multivariate classification [20] was performed for each dataset, followed by a leave-one-out (LOO) cross-validation. In each run of LOO, gene expression data of one subject was used for testing and the rest for training. The inputs of the classifier were the top n (n=1, 2 …) genes, such that the CRs of using these genes could be identified. A permutation of 5,000 runs was then conducted to test the hypothesis that a randomly selected gene set of the same size will reach equal or greater CR. The gene vector that generates the highest CR will be the best gene vector to select for the dataset.

Following the same process, the best gene vector was identified for each dataset by ANOVA analysis. For comparison purposes, a CR baseline was also generated using randomly selected sets of n (n=1, 2 …) genes. For each point in the CR baseline, the value was the mean of 300 CRs of genes randomly selected from the full dataset.

3. Results

3.1 Target genes from ASD_042017

Pathway enrichment analysis identified 410 out of 529 ASD target genes as significantly enriched within multiple ASD implicated pathways. Fig. 2 presents a gene interaction network using these genes (**ASD_042017→Related Genes**). A connection between two genes indicates that these two genes share one or more pathways (**ASD_042017→Related Pathways**). The adjacency matrix of Fig. 2 is presented in **ASD_042017→GGI Network**.

Put Figure 2 about here.

3.2 ASD case/control classification

Fig. 3 presents the classification results. The maximum CRs are presented at the position of corresponding number of genes. Table 2 summarizes the results of LOO cross validation of the two gene ranking methods on all four datasets, where the maximum CRs, the corresponding number of top genes, and permutation p-values of both methods are provided.

Put Table 2 about here.

Put Figure 3 about here.

Fig. 3 establishes that, compared to the CRs generated by randomly selected gene sets, the genes selected from ASD target genes by both SRVSScore and PValueScore can lead to significant higher classification accuracies. Note that the highest CRs were acquired using only the top genes with highest SRVSScore/PValuesSocre, (see Fig. 3 and Table 2), however adding more genes with lower score may not necessarily improve the classification accuracy. These results suggested the validity of both SRVS method and ANOVA method. Moreover, it was noted that the SRVSScore outperformed PValueScore in terms of CR in all datasets.

Table 2 and Fig. 4 also show that, for each dataset, the top genes selected by both methods were significantly different. For the SRVS method, the percentage of unique genes selected for the four datasets ranged from 60% to 90.48% (Table 2→**# Unique genes**). For the ANOVA method, the four datasets showed 100% unique genes (Table 2→**# Selected Genes**). These results suggest group specificity of the genome variation between the patients within these four datasets.

It is also worth mentioning that the optimum gene markers for different datasets as determined by SRVS and ANOVA were very different. As shown in Table 2, the genes selected by the SRVS method presented an overlap of less than 5% with that of ANOVA for all the 4 datasets (Table 2→ **Overlap genes of two methods (%)**). The results suggested that SRVS performs differently and more effectively than ANOVA does.

Put Figure 4 about here.

4 Discussion

In the last several decades, numerous studies have been conducted to seek the genetic cause of ASD, with hundreds of risk genes identified. Many of these genes were used as drug targets for ASD and most them play roles within ASD implicated genetic pathways. Results from these previous studies laid solid groundwork for the understanding of ASD genetic pathogenesis, which facilitates further study in the field. Of note, only a few genes have been commonly detected for ASD cases (e.g., gene SHANK3). However, most ASD genes present mutations or aberrant activities for multiple disorders, such as BDNF [15]. Dysregulated BDNF has been documented in ASD [37], and extensive work has implicated it in neurodegenerative disorders, including Alzheimer's, Huntington's disease, Parkinson's disease, and diabetic retinopathy [41, 27]. These results reflect the heterogeneity of neurotrophins, and partially explain the large size of the ASD risk gene pool curated through previous studies.

While novel ASD genes are being actively discovered with continuous genetic and genomic studies being performed, far fewer studies are conducted to test the validity of the existing reported ASD risk genes, as a whole, for their diagnostic and predictive capabilities with respect to ASD. We hypothesized that, if the current ASD gene pool is sufficient to cover most genes underlying the genetic pathogeneses of ASD, then for a given ASD patient group, there exists a sub-gene-set from the ASD gene pool that possesses significance in classification/prediction of ASD

patients from controls.

To test our hypothesis, we first conducted a comprehensive literature review using a data mining approach on 2,098 scientific articles, which identified 529 ASD target genes. Relations between these genes and ASD include Clinical Trial, Regulation, Quantitative Change, Biomarker, Genetic Change, Cell Expression, and State Change. The relationships were defined by ResNet relation database (http://pathwaystudio.gousinfo.com/ResNetDatabase.html). The 529 ASD genes were deposited into a genetic database, ASD_042017 (**ASD_042017→Related Genes**), which is available at '**Bioinformatics Database**' (http://database.gousinfo.com/). For each ASD-Gene relation, there is one or more supporting references. The titles and relevant sentences from which the relations were identified are presented at **ASD_042017→Ref for ASD Related Genes**. Within ASD_042017, there are also 91 pathways (**ASD_042017→Related Pathways**), 106 disease-subnetworks (**ASD_042017→ Related Diseases**), and 100 potential drugs/small molecules (**ASD_042017→Potential Drugs**) where these genes were significantly enriched.

Pathway enrichment analysis showed that, most of these genes (410/529) were significantly enriched within multiple genetic pathways that associated with ASD (p-value<3e-08; q=0.001 for FDR). For instance, there are 245 genes significantly enriched within 24 neuro system pathways (p-value< 2.7e-008; q=0.001 for FDR) [11,16,28,36], including dendrite (GO: 0030425; p-value=9.3e-039); postsynaptic membrane (GO: 0045211; p-value=2.2e-034); synapse (GO: 0045202; p-value=2.2e-034); neuronal cell body (GO: 0043025; p-value=5e-033); chemical synaptic transmission (GO: 0007268; p-value=1.1e-029); axon (GO: 0030424; p-value=1.5e-027); neuron projection (GO: 0043005; p-value=2.8e-025); nervous system development (GO: 0007399; p-value=2.1e-024); postsynaptic density (GO: 0097481; p-value=4.5e-024); dendritic spine (GO: 0043197; p-value=5.5e-021). There were also 109 genes enriched within 8 pathways/gene sets related to brain function development (p-value<1.7e-008) [17, 34] and 69 genes enriched within 8 behavior pathways (GO: 0002226; p-value= 1.3e-008) [5, 38].

Disease sub-network analysis SNEA (http://pathwaystudio.gousinfo.com/SNEA.pdf) showed that 484 of 529 genes significantly overlapped with the risk genes for 106 diseases (p-value< 2.4e-033; q=0.001 for FDR). Many of these 106 diseases were related to ASD, including schizophrenia [24], bipolar disorder [25], and Alzheimer's disease [20]. More results from the SNEA can be identified at **ASD_042017→Related Diseases**.

Within ASD_042017, there were 41 known ASD drugs (**ASD_042017→Related Drugs**) that have been through clinical trials and demonstrated effectiveness in treating ASD. Among these 41 drugs, four drugs were overlapped with the top 100 potential drugs (**ASD_042017→Potential Drugs**) whose gene subnetworks were significantly enriched with the 529 ASD genes. Additionally, many of the 529 ASD genes were target genes of known ASD drugs. For instance, a recent study showed that aripiprazole is effective for the treatment of irritability for children with ASD [35]. This may be explained by the fact that aripiprazole induces downregulation of extracellular signal-regulated kinases (ERK2) [18], while hypofunctional ERK2 was suggested to induce ASD [12, 33]. All these results suggested that the 529 ASD genes were linked to ASD and may therefore possess classification/prediction power for the disorder. However, due to the heterogeneity of ASD and the specificity of human genome variation [18], the significance of using these genes as markers for early diagnosis and personalized treatment requires additional testing.

To address this issue, ASD case/control classifications were conducted on four independent gene expression datasets. Two algorithms, SRVS and ANOVA, were used for gene selection from the 529 gene ASD group. The basic logic for gene selection is that, for a given ASD patient group, mutations of these 529 genes will not be present in every patient, and therefore are not effective as biomarkers for all patients.

Compared to randomly selected genes, those selected by both SRVS and ANOVA generated significantly higher prediction power (permutation p-value<0.0014 for SRVS and <0.0018 for ANOVA) and classification accuracy (SRVS vs. ANOVA: 82.82% vs. 76.77%, 97.22% vs. 91.67, 100% vs. 96.67% and 62.41% vs. 58.77%), as shown in Table 2. These results indicated that, for a given dataset, there exists a gene vector from the 529 gene ASD pool that could be used as biomarker vector for the diagnosis and prognosis of the disease. It should also be noted that SRVS outperforms ANOVA in terms of CR on the four datasets tested. This suggests the effectiveness of the SRVS method for feature selection.

Cross analysis on the gene selection results showed that optimum biomarkers are dataset specific, as displayed in Table 2 and Fig. 4. These results reflect the specificity of the genomic variations of different subjects [23], and highlight the necessity of genomic variable selection in the diagnosis and treatment of ASD. From Table 1 it can be seen that the disease status and tissue for experiment of each study were different. Despite of that, the 529 ASD target gene pool together with the SRVS algorithm managed to achieve significantly better classification ratio. These results suggest that our approach is applicable for different ASD subtypes. However, more experiments on datasets of specific subtypes (e.g., datasets only with essential or complex Autism) are needed to confirm the results from this study.

5 Conclusion

Our results support the validity of the literature based ASD risk genes as genetic biomarkers for the early diagnosis and personalized treatment of ASD. Integrating these genes for pathway analysis, disease-subnetwork analysis, druggability analysis, and gene-gene interaction analysis could help elucidate ASD pathogenesis and inform novel drug development. Moreover, SRVS is an effective method in genomic feature selection, which could help in patient-specific diagnosis and treatment.

Acknowledgement

There was no specific grant or any funding from others.

Conflict of interests

The authors declare that they have no conflicts of interests to disclose.

References

1. Alhowikan AM. Activity-Regulated Cytoskeleton-Associated Protein Dysfunction May Contribute to Memory Disorder and Earlier Detection of Autism Spectrum Disorders. Med Princ Pract. **25**(2016), pp.350-354.
2. Babatz TD, Kumar RA, and Sudi J Copy number and sequence variants implicate APBA2 as an autism candidate gene. Autism Res. **2**(2009), pp.359-364.
3. Belmonte MK, Cook EH Jr, Anderson GM, and Rubenstein JL, Autism as a disorder of neural information processing: Directions for research and targets for therapy. Mol Psychiatry. **9**(2004), pp.646-663.
4. Boso M, Emanuele E, and Minoretti P Alterations of circulating endogenous secretory RAGE and S100A9 levels indicating dysfunction of the AGE-RAGE axis in autism. Neurosci Lett. **410**(2006), pp.169-73, 2006.
5. Çaku A, Pellerin D, and Bouvier P, Effect of lovastatin on behavior in children and adults with fragile X syndrome: an open-label study. Am. J. Med. Genet. A, **164A**(2014), pp. 2834-2842, 2014.
6. Calati R, De Ronchi D, Bellini M, and Serretti A. The 5-HTTLPR polymorphism and eating disorders: a meta-analysis. Int J Eat Disord. **44**(3), pp.191-199, 2011.
7. Cao H, Duan J, and Lin D, Sparse representation based biomarker selection for schizophrenia with integrated analysis of fMRI and SNPs. Neuroimage. **102** Pt 1(2014), pp.220-228.
8. Chmielewski W, Beste C. Action control processes in autism spectrum disorder - Insights from a neurobiological and neuroanatomical perspective, Prog Neurobiol. **124(2015)**, pp.49-83, 2015.
9. Davis G, Mallat S, and Avellaneda M. Greedy adaptive approximation. J Constr Approx. **13**(1997), pp.57-98.
10. De Jaco A, Kovarik Z, and Comoletti D, A single mutation near the C-terminus in alpha/beta hydrolase fold protein family causes a defect in protein processing. Chem Biol Interact. **157-158(2005)**, pp.371-372.
11. De Rubeis S, He X, and Goldberg AP, Synaptic, transcriptional and chromatin genes disrupted in autism. Nature. **515**(2014), pp. 209-215.
12. Di Benedetto B, Rupprecht R. Targeting glia cells: novel perspectives for the treatment of neuropsychiatric diseases. Curr Neuropharmacol. **11**(2013), pp.171-185.
13. Donoho DL and Tsaig Y. Fast solution of L1-norm minimization problems when the solution may be sparse. IEEE Transs on Information Theory. 54 (2008), pp. 4789-4812.
14. Freitag CM, Staal W, and Klauck SM, Genetics of autistic disorders: review and clinical implications. Eur Child Adolesc Psychiatry. **19**(3), pp.169-178, 2010.
15. Gadow KD, Roohi J, and Devincent CJ, Association of COMT (Val158Met) and BDNF (Val66Met) gene polymorphisms with anxiety, ADHD and tics in children with autism spectrum disorder. J. Autism Dev. Disord. **39**(2009), pp.1542-1551.
16. Gilman SR, Iossifov I, and Levy D, Rare De novo variants associated with autism implicate a large functional network of genes involved in formation and function of synapses. Neuron. **70**(2011), pp.898-907.
17. Goorden SMI, Van Woerden GM, and Van Der Weerd L, Cognitive deficits in Tsc1+/- mice in the absence of cerebral lesions and seizures. Ann. Neurol. **62**(2007), pp.648-655.
18. Ishii D, Matsuzawa D, and Kanahara N, Effects of aripiprazole on MK-801-induced prepulse inhibition deficits and mitogen-activated protein kinase signal transduction pathway. Neurosci Lett. **471**(2010), pp.53-57.
19. Kanner L. Autistic disturbances of affective contact. Acta Paedopsychiatr. **35**(1968), pp.100-36.
20. Khan SA, Khan SA, Narendra AR, and Mushtaq G. Alzheimer's Disease and Autistic Spectrum Disorder: Is there any Association? CNS Neurol Disord Drug Targets. **15**(2016), pp.390-402.
21. Lorenzi PL, Claerhout S, and Mills GB, A curated census of autophagy-modulating proteins and small molecules: candidate targets for cancer therapy. Autophagy. **10**(2014), pp.1316-1326.
22. Losh M, Sullivan PF, Trembath D, Piven J. Current developments in the genetics of autism: from phenome to

genome. J Neuropathol Exp Neurol. **67**(2008), pp.829-37.
23. Lu YF, Goldstein DB, and Angrist M, Personalized medicine and human genetic diversity. Cold Spring Harb Perspect Med. **4**(2014), pp.a008581.
24. Meyer U, Feldon J, and Dammann O. Schizophrenia and autism: both shared and disorder-specific pathogenesis via perinatal inflammation? Pediatr Res. **69**(2011), pp.26R-33R, 2011.
25. Munesue T, Ono Y, and Mutoh K, High prevalence of bipolar disorder comorbidity in adolescents and young adults with high-functioning autism spectrum disorder: a preliminary study of 44 outpatients. J Affect Disord. **111**(2008), pp.170-175.
26. Nakamura T, Arima-Yoshida F, and Sakaue F, PX-RICS-deficient mice mimic autism spectrum disorder in Jacobsen syndrome through impaired GABAA receptor trafficking. Nat Commun. **7(2016)**, pp.10861.
27. Ola MS, Nawaz MI, Khan HA and Alhomida AS. Neurodegeneration and neuroprotection in diabetic retinopathy. Int J Mol Sci. **14**(2013), pp.2559-2572.
28. Park H and Poo MM. Neurotrophin regulation of neural circuit development and function. Nat. Rev. Neurosci. **14**(2013), pp. 7-23.
29. Parker KJ, Garner JP, and Libove RA, Plasma oxytocin concentrations and OXTR polymorphisms predict social impairments in children with and without autism spectrum disorder, Proc Natl Acad Sci U S A. **111**(2014), pp.12258-12263, 2014.
30. Rutter M, Genetic studies of autism: From the 1970s into the millennium. Journal of Abnormal Child Psychology. **28**(2000), pp.3-14.
31. Rutter M, Macdonald H and Le Couteur A, Genetic factors in child psychiatric disorders: II. Empirical findings. Journal of Child Psychology and Psychiatry. **31**(1990), pp.39-83.
32. Sandin S, Lichtenstein P, Kuja-Halkola R, The familial risk of autism. JAMA. **311**(2014), pp.1770-1777.
33. Satoh Y, Endo S, and Nakata T, ERK2 contributes to the control of social behaviors in mice. J Neurosci. **31**(2011), pp.11953-11967.
34. Satoh Y, Kobayashi Y, and Takeuchi A, Deletion of ERK1 and ERK2 in the CNS causes cortical abnormalities and neonatal lethality: Erk1 deficiency enhances the impairment of neurogenesis in Erk2-deficient mice. J. Neurosci. **31**(2011), pp.1149-1155.
35. Stigler KA. Psychopharmacologic management of serious behavioral disturbance in ASD. Child Adolesc Psychiatr Clin N Am. **23**(2014), pp.73-82.
36. Sweatt JD. Mitogen-activated protein kinases in synaptic plasticity and memory. Curr. Opin. Neurobiol. **14**(2004), pp. 311-317.
37. Tsai SJ. Is autism caused by early hyperactivity of brain-derived neurotrophic factor? Med Hypotheses. **65**(2005), pp.79-82, 2005;.
38. Wang X, Snape M, and Klann E, Activation of the extracellular signal-regulated kinase pathway contributes to the behavioral deficit of fragile x-syndrome. J. Neurochem. **121**(2012), pp. 672-679.
39. Zhang J, Zhang JX, and Zhang QL. PI3K/AKT/mTOR-mediated autophagy in the development of autism spectrum disorder. Brain Res Bull. 1**25(2016)**, pp.152-158.
40. Zhu L, Li G, and Choi SR, An improved preparation of [18F]FPBM: a potential serotonin transporter (SERT) imaging agent. Nucl Med Biol. **40**(2013), pp.974-979, 2013.
41. Zuccato C and Cattaneo E. Brain-derived neurotrophic factor in neurodegenerative diseases. Nat Rev Neurol. **5**(2009), pp.311-22.

Table 1 Key Statistics of 4 Gene Expression Datasets

NCBI GEO ID	GSE18123	GSE38322	GSE7329	GSE37772
# ASDs/Controls	66/33	18/18	15/15	233/206
#genes from ASD_042017	508	507	495	488
Disease Status	autism vs. Control	Idiopathic autistic vs. Control	Autism due to a fragile X mutation (FMR1-FM), or a 15q11-q13 duplication (dup(15q)) vs. Control	Autism probans vs. unaffected siblings
Tissue	Peripheral blood	Brain regions	Lymphoblastoid cells	Peripheral blood

Table 2 LOO cross validation and permutation results

	GSE18123 (case/control:66/33)		**GSE38322 (case/control:18/18)**		**GSE7329 (case/control:15/15)**		**GSE37772 (case/control:233/206)**	
	SRVS	ANOVA	SRVS	ANOVA	SRVS	ANOVA	SRVS	ANOVA
Max CRs	82.82	76.77	97.22	91.67	100	96.67	62.41	58.77
# Selected Genes	25	1	40	2	21	4	5	1
p-value	~0	~0	~ 0	~0	0.0014	0.0004	0.0014	0.0018
Unique genes from all datasets (%)	72% (18/25)	100% (1/1)	80% (32/40)	100% (2/2)	90.48% (19/21)	100% (4/4)	60% (3/5)	100% (1/1)
Overlap genes of two methods (%)	0% (0/25)	0% (0/1)	0% (0/40)	0% (0/2)	4.76% (1/21)	25% (1/4)	0% (0/5)	0% (0/1)

Figure Legends

Fig. 1 ASD genetic database schema

Fig. 2 The Gene-Gene Interaction Network composed of the 410 out of 529 ASD target genes from ASD_042017. The weight of the edge between two node/genes represents the number of pathways shared by the two genes; The larger the size of a node, the higher the number of pathways (**ASD_042017→Related Pathways**) including the gene; The brighter the color, the higher the Fisher's centrality of the gene (number of other genes connected). The adjacency matrix is presented in **ASD_042017→GGI Network.**

Fig. 3 Comparison of different metrics through a LOO cross validation. Genes were ranked in ascending order according to SRVSScore or PValueScore for SRVS method or ANOVA, respectively.

Fig. 4 Venn diagram comparing the top genes selected for different datasets using two methods. (a) Using SRVS methods. (b) Using ANOVA analysis.