bioRxiv preprint doi: https://doi.org/10.1101/029488; this version posted October 19, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Aberrant gene expression in autism

Jinting Guan[1,2], Ence Yang[2,3], Jizhou Yang[2], Gang Wang[2], Yong Zeng[1,2], Guoli Ji[1,4*], James J. Cai[2,5*]

[1]Department of Automation, Xiamen University, Xiamen, Fujian, China.

[2]Department of Veterinary Integrative Biosciences, Texas A&M University, College Station, Texas, USA.

[3]Institute for Systems Biomedicine, School of Basic Medical Sciences, Peking University Health Science Center, Beijing 100191, China.

[4]Innovation Center for Cell Signaling Network, Xiamen University, Xiamen, Fujian, China.

[5]Interdisciplinary Program in Genetics, Texas A&M University, College Station, Texas, USA.

[*]Co-senior authors: G Ji (glji@xmu.edu.cn) and JJ Cai (jcai@tamu.edu)

Keywords: autism spectrum disorder, gene expression, multivariate analysis.

Abstract

Autism spectrum disorder (ASD) is a complex neurodevelopmental condition characterized by phenotypic and genetic heterogeneity. Analyzing gene expression in autistic brains may reveal specific patterns of gene activation and repression associated with the disorder. For this, we developed a new analysis method called aberrant gene expression analysis, based on a multivariate distance measure for outlier detection, to identify aberrantly expressed genes. In a two-group setting, our method detects the discrepancies in gene expression dispersion between populations. Using this method, we re-visited RNA sequencing data previously generated from post-mortem brain tissues of 47 autistic and 57 control samples and identified a number of sets of genes with increased expression variability in autistics. In other words, the expression dispersion of these genes in autistics is more pronounced than that in controls. Many of those genes such as those expressed in synapse and those involved in neuropeptide binding are known to be implicated in ASD. We also found that many co-expressed gene modules present among non-autistic controls disappear among autistics, due to the aberrant gene expression that exclusively affects ASD patients. For the diagnostic purposes, we used a greedy algorithm to globally search for classifier gene sets, for which the expression in peripheral blood samples of autistics is maximally deviant from that of non-autistic controls. The aberrant gene expression respecting the classifier gene sets is more pronounced and specific for ASD samples, allowing the distinguishing of ASD from non-ASD samples. These results suggest that aberrant gene expression has the potential to be used as biomarkers for ASD. Taken together, we have developed a new gene expression analysis method based on a multivariate, dispersion-specific measure, which is powerful in detecting increased gene expression variability functionally

bioRxiv preprint doi: https://doi.org/10.1101/029488; this version posted October 19, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

associated with ASD. We conclude that detecting aberrant gene expression and identifying the underlying genes represent a new discovery and diagnostic strategy for ASD and other genetically heterogeneous disorders.

Author Summary

There is substantial phenotypic and genetic heterogeneity in autism, which complicates studies seeking to identify genetic factors that contribute to the disorder. However, despite this complexity, study designs have focused on using group differences, e.g., in gene expression, between autistic cases and controls to identify the genetic effects. The problem is that, by their nature, group difference-based methods, such as differential expression (DE) analysis, blur or collapse the heterogeneity within autistics, which, in fact, characterizes this spectrum disorder. For instance, DE analysis identifies genes expressed differentially between groups, but, by its design, the method tends to identify genes expressed uniformly or less variably across individuals within groups. By ignoring genes with variable expression in autistic individuals, an important axis of genetic heterogeneity contributing to gene expression variability among affected individuals has been overlooked. We have developed a new analysis method called aberrant gene expression analysis, aimed to identify genes with significant changes in expression variance between diseased and non-diseased samples. This is in sharp contrast to the purpose of conventional DE analysis method, which aims to identify genes with significant changes in expression mean. Using this new method, we detected increased gene expression variability and identified candidate genes with functions relevant to autism.

Introduction

Autism spectrum disorder (ASD, [OMIM 209850]) is a complex neurodevelopmental condition characterized by substantial phenotypic and genetic heterogeneity [1-4]. Both genetic and environmental factors contribute to the increased risk of developing ASD [5, 6]. Recent years have seen major advances in the understanding of the genetic, neurobiological and developmental underpinnings of ASD [7-9]. Genetic studies, especially genome-wide association studies (GWAS), have identified many single-nucleotide variants (SNVs) and copy number variants (CNVs) associated with ASD susceptibility [10-12]. However, it remains difficult to identify the actual causal genes underlying these associations. SNVs that produce association signals in identified loci often fall into intergenic regions, while CNVs often extend across multiple variants or genes, both of which confound the identification of causal genes. Also, there are opposing views on the relative contribution of rare versus common variants to ASD susceptibility. Some studies suggest that low-frequency variants bring a greater impact on the risk for ASD [13-16], while other studies suggest that common variants form a dominating source of the risk [17, 18]. Against this background of complexity, several studies demonstrate the use of gene expression information—measuring mRNA abundance of individual genes, coupled with other genetic approaches, allows for novel insights in understanding ASD [19-21]. Analyzing gene expression and sequence data facilitates revealing the impact of regulatory genetic variants on the gene itself and the indirect consequences on the expression of other genes [22].

To this end, we introduce a novel, gene expression analysis method for identifying ASD-implicated genes. Our working hypothesis is that ASD is associated with aberrant gene expression caused by the promiscuous multigene activation and repression. Indeed, we show that many gene sets that contain genes known to be implicated in ASD tend to be expressed more aberrantly in autistic individuals. Encouraged by these findings, a global search was conducted for unique combinations of genes for ASD diagnosis. These genes are more aberrantly expressed in blood samples of autistics than non-autistics. We use a greedy algorithm to solve the combinatorial problem of global search and identify three gene sets, each containing five genes, which can be used as classifier gene sets to discern gene expression

bioRxiv preprint doi: https://doi.org/10.1101/029488; this version posted October 19, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

specific to patients with ASD. Based on the three gene sets, a diagnostic test for ASD achieves high sensitivity and high specificity. Altogether, our results refine the relationships between gene function and gene expression dispersion among autistic individuals, providing new insights into the genetic, molecular mechanisms underlying the dysregulated gene expression in ASD. Our results also lay out the foundation for the utilization of gene expression as biomarkers for early diagnosis of ASD.

Results

Many biological processes underlying human diseases are accompanied by changes in gene expression in corresponding tissues [23]. ASD is not an exception. Previous studies have detected specific gene expression changes in ASD, concerning genes involved in the synaptic formation, transcriptional regulation, chromatin remodeling, or inflammation and immune response [21, 24]. These analyses mostly focused on departures across the average expression between the case and control groups, without considering or much less focusing on alternative patterns of departure such as those characterized by heterogeneous dispersion. The goal of present study is to detect the difference in heterogeneous multigene expression dispersion between autistic and non-autistic samples.

Overview of aberrant gene expression analysis

We have previously developed a multivariate method, namely aberrant gene expression analysis, to measure the level of multigene expression dispersion in the general population [25]. This analysis method uses Mahalanobis distance (MD) to quantify the dissimilarity in multigene expression patterns between individuals [26]. MD is an appropriate measure because it accounts for the covariance between expression levels of multiple genes. The aberrant gene expression analysis can be used to identify genes more likely to be aberrantly expressed among individuals. It can also be used to identify individual outliers whose expression for a given gene set differs markedly from most of a population.

Here, we extend the MD-based aberrant gene expression analysis to a two-group setting. We estimated the level of dysregulation in multigene expression among autistics relative to the controls, assuming that the dysregulation is due to a promiscuous gene activation and repression associated with autism. We re-analyzed the gene expression data generated from the post-mortem brain tissues of 47 ASD and 57 control samples [19]. For a given gene set, we computed the MD between gene expression of each ASD individual *i* to the multivariate centroid of the controls, denoted as MD_i. We used the sum of squared MD_i (SSMD) to measure the overall dispersion level for all autistic samples vs. the controls. Using permutation tests, we assessed the significance of gene sets and identified gene sets more likely to be aberrantly expressed among individuals affected with ASD (**Materials and Methods**).

Coordinated gene expression is disrupted in ASD

Our MD-based aberrant gene expression analysis is capable of detecting the signal of expression aberration in different forms, including, e.g., (1) the increased individual-to-individual gene expression variance (i.e., the increased gene expression variability) and (2) the decreased expression correlation between genes. To illustrate the effect of disrupted expression, we use gene sets comprising only two genes to show the aberrant gene expression among autistic individuals manifested as the loss of expression correlation between the two genes. **Fig. 1** shows scatter plots of expression levels between gene pairs. In **Fig. 1A**, the expression of *CORO1A* is positively correlated with that of *SYN2* for the controls (left panel, Pearson correlation test, $P = 3.6\times10^{-10}$). The gene *CORO1A* encodes coronin 1A, an actin binding protein. The gene *SYN2*, which is selectively expressed at nerve terminals in mature neurons, encodes synapsin II, a neuron-specific synaptic vesicle phosphoprotein [27, 28]. Synapsins interact with actin filaments in a phosphorylation-dependent manner [29]. As evident from the

bioRxiv preprint doi: https://doi.org/10.1101/029488; this version posted October 19, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

description of gene functions, the correlated expression between the two genes is crucial for their respective actin-related molecular functions in normal individuals. However, such a crucial correlation becomes less significant in the ASD group (middle panel, $P = 0.07$). To make the contrast, we superimposed the data points for ASD individuals onto those of the controls (right panel). The top 10 ASD samples with the largest MD values are highlighted in red. The data points of these ASD samples are the most remote observations, distributed either far away from or orthogonally against the "cloud of data points" around the population mean, in which most control individuals are located. **Fig. 1B** presents a negative example, in which the correlations in expression levels between two genes, *CX3CR1* and *SELPLG*, are presented in both control and ASD groups ($P = 1.3\times10^{-9}$ and 1.4×10^{-10}, respectively), indicating that the coordinated expression between the two genes is not disrupted in ASD. Altogether, these two examples, one positive and one negative, suggest that aberrant gene expression is not universal. The pattern of aberration may be highly specific with regard to certain gene sets (e.g., that in **Fig. 1A)** but not others (e.g., that in **Fig. 1B**).

Fig. 1. A proof-of-concept example, based on real data [19], showing that (**A**) the correlated expression between *SYN2* and *CORO1A* presents among non-ASD samples but is disrupted among ASD samples, while (**B**) the correlated expression between *CX3CR1* and *SELPLG* presents among both non-ASD and ASD samples. Red stars in (**A**) show the top 10 ASD samples with the largest MD_i.

Gene sets that tend to be aberrantly expressed in ASD

To identify gene sets more likely to be aberrantly expressed in autistics, we calculated SSMD for a number of pre-defined gene sets (**Materials and Methods**). These included the curated gene sets in the molecular signatures database of GSEA [30]. The significance of each gene set was assessed using permutation tests with random gene sets. A total of 18 GSEA curated gene sets were found to produce significantly higher SSMD than random gene sets at a false discovery rate (FDR) of 10%. Functions of these gene sets fall into four major categories, namely, metabolism and biosynthesis, immune or inflammatory response, signaling pathway, and vitamins and supplements (**Table 1**). The relevance of these major functional categories with ASD is supported by respective studies [7, 31-36]. For example, the mTOR signaling pathway (with a full name in Reactome database: Energy dependent regulation of the serine/threonine protein kinase mTOR by LKB1-AMPK [37]) is essential to synaptogenesis; gene products of the pathway regulate dendritic spine morphology in synapses. The dysregulation of this pathway is implicated in ASD [7, 33, 34]. **Table 1** also contains four gene sets with miscellaneous functions, unclassified into any of the four major categories but all implicated in ASD. These genes are involved in: (1) activated point mutants of FGFR2 [38, 39], (2) activation of the AP-1 family of transcription factors [40], (3) inwardly rectifying K^+ channels [41, 42], and (4) G2/M checkpoints [43].

Table 1. GSEA curated gene sets that tend to be aberrantly expressed in ASD. *Number of genes included in our analysis/Number of genes in the gene set. Gene sets mentioned in the main text are shown in italic. The previously known ASD-implicated genes are shown in bold.

To determine individual gene's contribution to the total SSMD of a gene set, we computed ΔSSMD for each gene. The ΔSSMD of a gene is the difference between SSMD values of a gene set before and after excluding the gene from the gene set. Top three genes with the largest ΔSSMD are given for all gene sets in **Table 1**. ASD-implicated genes are highlighted.

To further investigate the relationship between gene function and aberrant gene expression, we grouped genes into gene sets, based on their cellular and molecular functions indicated by gene ontology (GO) terms associated with the gene function descriptions. A total of 36 significant GO terms at an FDR of 10% were identified (**Supplementary Table 1**). These terms are distributed

bioRxiv preprint doi: https://doi.org/10.1101/029488; this version posted October 19, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

in 22 biological processes (BP), 11 molecular functions (MF), and three cellular components (CC) sub-ontologies. The relevant processes include cellular response to stimulus, cellular metabolic process, cell morphogenesis and proliferation, regulation of intracellular transport and organelle organization, and tissue development. A close look at these significant GO terms revealed several that are implicated in ASD, e.g., *neuropeptide receptor activity* (GO: 0008188), *neuropeptide binding* (GO: 0042923), and *inhibitory synapse* (GO: 0060077).

We also used the expression data from non-ASD controls to construct the co-expression networks. The WGCNA method [44] was employed to identify 807 functional gene modules according to the observed co-expression relationships between genes. The size of these modules ranged from 4 to 110 genes. We computed SSMD for gene sets of all modules. Using permutation tests, we assessed the significance and identified 76 significant modules (**Supplementary Table 2**). Many modules contain genes with functions relevant to ASD. For example, module 1 is enriched with genes closely associated with synapse and cell junction while module 5 is enriched with genes involved in regulation of neurogenesis/neuron differentiation. **Fig. 2A** shows the co-expression relationships between genes in modules one and five among control samples. The co-expression patterns in the two modules are absent in ASD samples (**Fig. 2B**). It is striking to observe such complete breakdowns of essential functional modules in ASD cases.

Fig. 2. The breakdown of co-expression network modules in ASD. (**A**) Two example modules are presented as gene interaction subnetworks among non-ASD controls. Edge width is proportional to the value of Pearson's correlation coefficient (ranging 0.5 and 0.8). Node size is proportional to the value of ΔSSMD for each gene. The two modules are enriched with genes whose products are closely associated with synapse or cell junction (top) and genes involved in regulation of neurogenesis or neuron differentiation (bottom), respectively. (**B**) The same sets of genes in the two modules are depicted for ASD samples. The missing of edges is due to the lack of co-expression relationships between genes.

To quantify the module difference between ASD and control groups, we used the function `modulePreservation` in the WGCNA R package [45] to calculate two statistics—medianRank and Zsummary—that measure the level of connectivity preservation between modules constructed using control and ASD samples. Most of the 76 significant modules have a large medianRank and a small Zsummary close to zero (**Supplementary Table 2**), which suggest little or no module preservation across the control and ASD samples. To further demonstrate that the 76 significant modules constructed using control data are robust, we obtained an independent data from brain tissues of 93 non-ASD healthy controls (GEO accession: GSE30453) [46] and used this new independent expression data to re-draw these significant modules. We found that, despite the difference in technical platforms (i.e., RNA sequencing vs. microarray) on which two gene expression data were generated, most co-expression relationships between genes in these modules could be recapitulate using the new independent control data (**Supplementary Fig. 1**). These results suggest that these modules are robust and the co-expression relationships between genes in these modules are biologically important and indispensable for healthy controls.

Next, we tested the correlation between ΔSSMD and two network metrics for nodes, i.e., betweenness centrality and clustering coefficient. We previously showed that disease-causing genes have high betweenness centrality and low clustering coefficient values [47]. However, for genes in these co-expression modules tested, no significant correlation was detected, which suggests ΔSSMD captured statistical features of genes that differ from those captured by the two network metrics.

bioRxiv preprint doi: https://doi.org/10.1101/029488; this version posted October 19, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Finally, we note that ΔSSMD as a single-gene measure may be used to prioritize genes with desired functions. For instance, *CPLX2* is among genes with the largest value of ΔSSMD in the module (**Fig. 2A**). It is likely that the sequences of *CPLX2* regulatory region are more heterogeneous among autistics, or the region contains variants associated with large gene expression variability more common in autistics. In either case, ΔSSMD enables to prioritize gene candidates and pinpoint the genomic regions that are likely to accommodate the potential mutations responsible for the increased gene expression variability. Indeed, *CPLX2* encodes a complexin protein that binds to synaphin as part of the SNAP receptor complex and disrupts it, allowing transmitter release. *CPLX2* has been associated with schizophrenia and attention deficit hyperactivity disorder [48, 49], but not with autism yet. Target sequencing of *CPLX2* regulation region in the autistic samples may discover unknown variants associated with autism risk.

Aberrant gene expression as biomarkers for ASD

We sought to determine whether we could classify patients as having ASD vs. controls solely based on the aberrant gene expression that is more pronounced among autistics. For this purpose, we obtained the gene expression data from the peripheral blood samples of 104 ASD patients and 82 healthy controls (GEO accession: GSE18123 [50]). We split the original data set into two, each containing data of 52 ASD and 41 control samples, and used them as "training" and "test" sets, respectively. With the training set data, we calculated MD_i for autistic samples against the control samples. With the test set data, we calculated MD_i for both autistic and control samples against the control samples of the training set (**Materials and Methods**). That is, we calculated MD_i for all samples against the same set of controls in the training set. Our purpose was to identify a set of genes whose aberrant gene expression could be used to distinguish ASD cases from controls (i.e., MD_i respecting the gene sets for ASD and non-ASD samples differs greatly).

We conducted a global search for such a gene set comprising as few as five genes out of all protein-coding autosomal genes expressed in the whole blood (n = 16,365). No information of any pre-defined gene sets was used. An exact solution for such a search is a combinatorial problem requires $>10^{18}$ SSMD calculations, which was computationally infeasible. Instead, we used a greedy algorithm to search from different starting points for producing several local optimal solutions (**Materials and Methods**). We tested the prediction power of candidate gene sets using the receiver operator characteristic (ROC) curve analysis. The area under the ROC curve (AUC) was used to determine the prediction power of a gene set. Three sets of genes, each containing five genes, were identified to produce high accuracy with balanced sensitivity and specificity values for the tests using both training and test data sets (**Fig. 3**). These three gene sets are: {*FAM120A, HDC, OR13C8, PSAP, RFX8*}, {*HBG1, MOCS3, PDGFA, SERAC1, SLFN12L*}, and {*BHMT2, CCL4L1, CD2, FAM189B, MAK*} (see **Supplementary Table 3** for the corresponding SSMD and ΔSSMD values). All three gene sets achieved greater than 70% sensitivity and greater than 70% specificity in all tests (**Table 2**). Many genes in the three classifier gene sets are associated with ASD but in an indirect manner. Mass spectrometry analysis [51] showed that the protein of *FAM120A* interacts with that of an ASD-implicated gene *CYFIP1*. *HDC* encodes L-histidine decarboxylase, catalyzing the biosynthesis of histamine from histidine. A rare functional mutation in *HDC* has been associated with Tourette's syndrome [52], which is a neuropsychiatric disorder potentially related to ASD [53]. The expression of *PDGFA* was found to be down-regulated in patients affected with the 22q11.2 deletion syndrome, which is associated with high rates of ASD in childhood [54]. Other genes including *SERAC1*, *OR13C8*, *RFX8*, *HBG1*, *SLFN12L*, and *BHMT2* are known to be linked with ASD-associated CNVs and/or rare *de novo* SNVs [13, 14, 55-57].

bioRxiv preprint doi: https://doi.org/10.1101/029488; this version posted October 19, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Fig. 3. ROC curves and dot diagrams of MD_i. (**A**) ROC curves graphs for the three classifier gene sets tested with the training and test data sets. Corresponding AUC values for the training (AUC1) and test (AUC2) data sets are given in the inserts. Red cross indicates the optimal operating point of the ROC curve for the training data set. (**B**) Dot diagrams for training (top) and test (bottom) sets showing the distributions of MD_i calculated with respect to the three classifier gene sets for samples in ASD and control groups. Log-transformed MD_i values are shown. The red vertical lines show the optimal cutoff values determined from the ROC curves tested on training data set.

Table 2. The performances of classifiers based on gene set I, II, III tested on the training and test data sets. True positive (TP), true negative (TN), false positive (FP), false negative (FN), sensitivity (SN), specificity (SP), accuracy (ACC) values are reported.

Discussion

Dispersion-specific measure of gene expression and its application in ASD

ASD is a complex disease involving multiple genetic alterations that result in modifications of many cellular processes. Maladaptive patterns of ASD lead to significantly high gene expression variability among affected individuals. Unitary models of autism brain dysfunction have not adequately addressed conflicting evidences, and efforts to find a single unifying brain dysfunction have led the field away from research to explore individual variation and micro-subgroups. Therefore, it has been suggested that researchers must explore individual variation in brain measures within autism [58, 59]. Yet, previous studies have rarely addressed the issue of increased gene expression variability associated with autism. Among few exceptions [21, 60], the authors wrote: “Autistic subjects display significant phenotypic variability which could be due to an intricate interplay of genetic and environmental factors. Thus, we hypothesized that this phenotypic diversity is due to subject-to-subject variability in gene expression.” [21]. Nevertheless, the *status quo* pertaining to gene expression specific to ASD patients is based on the detection of differential gene expression, i.e., the gene-expression differences between mathematical expectation (i.e., mean) of autistic and non-autistic samples. The major assumption underlying differential expression analysis is: ASD cases have the same or similar gene expression change phenotypes, which makes them as a distinctive cohort have significantly higher or lower expression than the controls. However, this assumption contradicts the fact that ASD has highly heterogeneous genetic causes, and also excludes empirical evidences gathered about uncommon molecular changes causing ASD [13-16].

Our overall strategy for this study was based on the quantitative measures of the departure of multigene expression dispersion between individuals. The profound heterogeneity in ASD underscores the importance of leveraging measures of dispersion in order to capture the specific tendency. Gene expression dispersion has been found associated with gene function and disease or physiological status of individuals [61-64]. Discrepancies in gene expression should not only be characterized by the mean but also by other statistics of interest, such as dispersion parameters. Using this thinking and the proven multivariate approach [25], we have further developed MD-based aberrant gene expression analysis and applied it to ASD. The statistical signal captured is the tendency of being more dispersed in multigene expression among autistics than non-autistics. We have shown that our variability-centered method can recapitulate signals from many genes known to be implicated in ASD. Our method does not depend on the prior knowledge about gene function or the identification of mutations in genes. Thus, it is a tool for discovering and identifying genes previously unknown to be involved in ASD progression.

bioRxiv preprint doi: https://doi.org/10.1101/029488; this version posted October 19, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Aberrant gene expression in co-expression network

We have shown that, when applied to the co-expression network, SSMD can reveal the effects of perturbing genetic networks. SSMD analysis informs us about how ASD distorts expression patterns of biological systems. Disturbed ASD genetic networks have been noticed previously [65-67]. However, most existing network analyses were not designed for directly measuring the level of dysregulation. Instead, information about known ASD genes (e.g., in [65, 67]) or differently expressed genes (e.g., in [66]) were used to prioritize the modules, which would not allow modules contain unknown ASD genes to be prioritized for subsequent analyses. In contrast, our approach allows for a straightforward screening of perturbed network modules and provides the raw material for the identification of genetic regulatory mechanisms involved in the variability of gene transcription.

Genetic variants contribute to aberrant gene expression

Our results have provided unique entry points to investigate further on the genetic basis of aberrant gene expression (e.g., increased gene expression variability) in ASD. When genotype or sequence information, along with their gene expression information, become available for ASD samples, it would be possible to assess the influences of the aggregation of rare mutations, CNVs, as well as common genetic variants on aberrant patterns of gene expression in ASD. We have shown previously that certain alleles of common genetic variants can increase gene expression variability among individuals carrying the alleles [68]. We have also shown that epistasis and decanalization equally contribute to the effect of common variants on variable gene expression [69]. Taking these into account, it would be possible to construct gene expression variability networks and use genetic variants of ASD to predict which parts of the network are more vulnerable to the perturbation from genetic and/or environmental factors.

Aberrant gene expression as biomarkers

Recent years have seen an intensive search for biological markers for ASD. Although a wide range of ASD biomarkers have been proposed, as of yet none has been validated for clinical use [70]. Therefore, there is a critical need for valid biological markers for ASD. Based on the aberrant gene expression analysis shown here, gene sets with few genes can be used as novel biomarkers. Application of our gene-expression candidate biomarkers will allow for higher sensitivity and specificity in a diagnostic screen for ASD. We anticipate that if our gene-expression biomarkers are expanded to use the blood gene expression data derived from other platforms (such as different types of microarrays, RNA sequencing, and qPCR), they will offer a significant advancement in developing a clinical blood test. The success of such gene-expression biomarkers will assist in early and objective diagnosis for ASD.

Caveats and future directions

In a previous study [21], Voineagu et al. used differential expression analysis to discover that significantly more genes (510 vs. 8) expressed differentially between frontal and temporal cortex in brains of non-autistic controls than autistics. The loss of gene expression difference in different regions of brains in autistics may be attributable to the increased gene expression variability in each sub-regional gene expression. That is, the heterogeneity in gene expression between different brain regions in the same individual may introduce another level of gene expression variability. This effect, however, has not be explicitly captured by our aberrant gene expression analysis method used in this study. Furthermore, even particular brain regions are highly heterogeneous because of the mixtures of cell types. Thus, aberrant gene expression patterns might in part indicate different relative proportions of cell types in a sample. With the advent of the single-cell based technologies [71], this level of gene expression heterogeneity may be measured and the problem of heterogeneity of cell types among tissue samples as an important source of variability may be addressed in future studies.

bioRxiv preprint doi: https://doi.org/10.1101/029488; this version posted October 19, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

In conclusion, we have developed a novel, variability-centric gene expression analysis, and applied the method to ASD. This advance showcases the value of development and refinement of systems genomics tools in studying human complex diseases. The aberrantly expressed genes identified in this study will facilitate the identification of ASD-predisposing variation, which may eventually reveal the causes of ASD and enable earlier and more targeted methods for diagnosis and intervention.

Materials and Methods

Gene expression data

Whole transcriptomes of 104 brain tissues (47 ASD and 57 controls) were previously determined using RNA sequencing by Gupta et al [19]. The data had been deposited in the National Database for Autism Research (NDAR) under the accession code NDARCOL0002034. Among these samples, 62, 14 and 28 were tissues from cerebral cortex Brodmann Area 19, anterior prefrontal cortex, and a part of the frontal cortex, respectively, resulting in 47 (32 unique individuals) ASD samples and 57 (40 unique individuals) controls. For this study, the raw data of gene expression was normalized using the conditional quantile normalization [72], and then processed using the algorithm of probabilistic estimation of expression residuals (PEER) [73] to remove technical variation. The PEER residuals were obtained after regressing out covariates (age, gender, brain region, and sample collection site) and accounting for ten possible hidden determinants of expression variation. The expression median across all samples was added back to the PEER residuals to give the final processed gene expression levels. Extremely lowly expressed genes with expression median < 2.5 (empirical cutoff) were excluded. The final data matrix contained the expression levels for 10,127 genes in 104 samples. Principal component analysis for all samples indicated that there was no population stratification regarding the global gene expression profiles (**Supplementary Fig. 2**).

Functional gene sets

The curated GSEA gene sets were obtained from the molecular signatures database (MSigDB v5.0, accessed March 2015) [30]. GO terms associated with gene function were downloaded from BioMart (v0.7, accessed February 2015) [74]. The co-expression networks were built for control samples using WGCNA [44]. The power of 16 was chosen using the scale-free topology criterion; the minimum module size was set to 4, and the minimum height for merging modules was 0.25. The resulting modules were plotted using SBEToolbox [75].

Calculation of robust MD between ASD and control samples

For a given gene set, MD_i is the Mahalanobis distance (MD) [26] from an ASD individual i to the multivariate centroid of control individuals. Conventional MD_i was calculated using the following operation:

$$MD(x_{i\cdot}, x_c) = \sqrt{(x_{i\cdot} - x_c)^T \Psi^{-1} (x_{i\cdot} - x_c)}$$

where $x_{i\cdot}$ is the vector of expression of genes in ASD sample i, x_c is the vector of expression means of genes across all control samples, and ψ is the covariance matrix estimated from the controls. Throughout the analyses of this study, a robust version of MD_i was calculated using the algorithm Minimum Covariance Determinant (MCD) [76]. The MCD algorithm subsamples h observations out of control individuals whose covariance matrix had the smallest covariance determinant. By default, $h = 0.75n$, where n is the total number of control samples. The robust MD_i was then computed with the above equation by replacing x_c with the MCD estimate of location, $\hat{x}_c$, i.e., the expression mean of the h controls, and with the MCD estimate of scattering, $\hat{\Psi}$, i.e., the covariance matrix estimated from the h controls. A Matlab implementation of MCD, available in the function `mcdcov` of LIBRA toolbox [77], was used to perform the

bioRxiv preprint doi: https://doi.org/10.1101/029488; this version posted October 19, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

computation of MCD estimator. For a given gene set, MCD estimates of x_c and ψ were computed as the outputs of `mcdcov`, and re-used for calculating robust MD_i for ASD individuals.

The sum of squared MD_i, $SSMD = \sum_{i=1}^{M} MD_i^2$ was calculated for a given gene set to measure the overall dispersion of *M* autistic individuals. To assess the significance of SSMD of a given gene set, permutation tests were performed with *N* reconstructed gene sets of the same size but randomly selected genes. The *P*-value of permutation test, P_{perm}, was determined by the ratio of $\frac{n}{N}$, where *n* is the number of random gene sets having SSMD greater than that of the tested gene set and *N* is the total number of random gene sets used in permutation tests. For the first round of test, we set $N = 1{,}000$ to obtain a short list of gene sets with $P_{perm} < 0.001$. For the second round of test, we set $N = 10{,}000$ to obtain P_{perm} for gene sets in the short list and the correction for multiple testing was performed by controlling the FDR with the Benjamini-Hochberg method [78]. Gene sets with an FDR<10% were considered to be of significance. Next, to measure the relative contribution of each gene in a gene set to the total SSMD of the gene set, ΔSSMD was calculated. ΔSSMD is the difference between the two SSMD values, which were calculated before and after the gene was excluded from the gene set.

ROC curve analyses

For the analysis of aberrant gene expression as the biomarker for ASD diagnosis, peripheral blood gene expression data measured using the Affymetrix Human Gene 1.0 ST array for 104 ASD and 82 controls was downloaded (GEO accession: GSE18123) [50]. The raw intensity data was processed using the R function `rma(`robust multi-array average expression measure`)` in the Affy package. The expression measure was quantile normalized and log2 transformed. The final data matrix contained the expression level information for 16,365 autosomal genes among 186 samples. We equally split samples into training set and test set, each of which contains half of ASD (i.e., 52 ASD samples) and half of control samples (i.e., 41 controls).

ROC curve analysis was used to estimate the specificity and sensitivity of classification tests, in which gene sets were used as classifiers for ASD and controls. For a given gene set, we first obtained the multivariate centroid of controls in training set ($\boldsymbol{G}_{training}$), and calculated MD_i of each sample *i* (including all ASD and control samples in the training set). Using ROC curve analysis, the threshold corresponding to optimal specificity and sensitivity combination (denoted by *T*) was determined. If MD_i is greater than *T*, the sample *i* was classified as an ASD, otherwise a control. The performance of each classifier gene set for predicting ASD and control samples was tested at different threshold *T* values and the top three best-performed gene sets having largest area under the ROC curve (AUC) were identified. We denote the AUC with respect to the training data set as AUC1. For the top gene sets that produced the best AUC1 values, we then assessed their performances of prediction with the test data set. Again, we used the AUC of ROC curve, denoted by AUC2, to measure the performance of gene sets as classifiers for the test data set. For each classifier gene set, MD_i for all samples of the test set, regardless of the disease status of samples, was calculated against $\boldsymbol{G}_{training}$, i.e., the multivariate centroid of controls in training set. This is because we assumed the disease status of samples of the test set were unknown.

With randomly generated gene sets, we examined AUC2 as a function of the size of gene sets. We found no correlation between the two (**Supplementary Fig. 3**), suggesting that random gene sets have no prediction value. We also noticed that there is a strong positive correlation between the size of gene sets and AUC1 (**Supplementary Fig. 3**), which is simply because that the inclusion of more variables (i.e., expression data from more genes) allows a better fit of gene expression variation in control samples of the training set, resulting in a continuously improved AUC1. We also found a weak but significant positive correlation between AUC1 and AUC2 (**Supplementary Fig. 4**). Based on these preliminary results, we decided to search for

bioRxiv preprint doi: https://doi.org/10.1101/029488; this version posted October 19, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

gene sets that could produce the best AUC1 and test the gene sets with AUC2. We used as few as five genes to make these candidate gene sets to avoid the potential problem associated with overfitting of the control data.

Global search for the classifier gene sets

A greedy algorithm was developed to identify subsets of five genes, in regard to which AUC1 reaches its maximal values. The search is global as possible combinations of all expressed genes were considered; no information of any pre-defined gene sets was used by the algorithm. Starting with all pairs of two-gene combinations, AUC1 values were computed, and the top 1,000 two-gene pairs with maximal AUC1 were retained as seeds for subsequent steps. The idea of the greedy strategy is to make a locally optimal choice at each stage have the hope of finding a global optimum. Thus, the assumption here was that the genes in the two-gene combinations producing the greatest AUC1 (i.e., the locally optimal solution) would be among those in five-gene combinations producing the greatest AUC1 (i.e., the global optimum). For each of the selected two-gene combinations, a new gene that can produce largest SSMD was identified and added to the gene pair to make a three-gene combination. The procedure was repeated until the number of genes reached five. At this stage, an additional procedure was introduced to improve the locally optimal solutions achieved by the greedy heuristic: all distinct three-gene combinations were extracted from five-gene combinations as the new candidate subsets. From these three-gene combinations, new genes were added to get a new round of solutions of five-gene combinations. The newly generated five-gene combinations will be retained if they produced larger AUC1 than older ones. This replacement procedure was repeated until no improvement could be made. Computer code is available from the authors upon request.

Acknowledgements

We thank Shannon Ellis and Dan Arking for sharing data, Oliver Stegle and Tuuli Lappalainen for helping with data normalization, and Steve Horvath for co-expression network analysis. We thank Rae L. Russell for proofreading and editing this paper. We acknowledge the Texas A&M Institute for Genome Sciences and Society (TIGSS, former Whole Systems Genomics Initiative) for providing computing resources and system administration support. This work was supported by the fund of China Scholarship Council to J Guan, and the National Natural Science Foundation of China (Nos. 61174161 and 61201358), the specialized Research Fund for the Doctoral Program of Higher Education of China (No. 20130121130004) and the Fundamental Research Funds for the Central Universities in China (Xiamen University: Nos. 2013121025, 201412G009, and CXB2014007) to G Ji.

References

1. Willsey AJ, State MW. Autism spectrum disorders: from genes to neurobiology. Curr Opin Neurobiol. 2015;30:92-9. doi: 10.1016/j.conb.2014.10.015. PubMed PMID: 25464374.
2. Devlin B, Scherer SW. Genetic architecture in autism spectrum disorder. Curr Opin Genet Dev. 2012;22(3):229-37. doi: 10.1016/j.gde.2012.03.002. PubMed PMID: 22463983.
3. Geschwind DH. Genetics of autism spectrum disorders. Trends in cognitive sciences. 2011;15(9):409-16. doi: 10.1016/j.tics.2011.07.003. PubMed PMID: 21855394; PubMed Central PMCID: PMC3691066.
4. Geschwind DH, State MW. Gene hunting in autism spectrum disorder: on the path to precision medicine. Lancet Neurol. 2015. doi: 10.1016/S1474-4422(15)00044-7. PubMed PMID: 25891009.
5. Sandin S, Lichtenstein P, Kuja-Halkola R, Larsson H, Hultman CM, Reichenberg A. The familial risk of autism. JAMA. 2014;311(17):1770-7. doi: 10.1001/jama.2014.4144. PubMed PMID: 24794370; PubMed Central PMCID: PMC4381277.

bioRxiv preprint doi: https://doi.org/10.1101/029488; this version posted October 19, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

6. Persico AM, Bourgeron T. Searching for ways out of the autism maze: genetic, epigenetic and environmental clues. Trends Neurosci. 2006;29(7):349-58. doi: 10.1016/j.tins.2006.05.010. PubMed PMID: 16808981.
7. Abrahams BS, Geschwind DH. Advances in autism genetics: on the threshold of a new neurobiology. Nature reviews Genetics. 2008;9(5):341-55. doi: 10.1038/nrg2346. PubMed PMID: 18414403; PubMed Central PMCID: PMC2756414.
8. Belmonte MK, Cook EH, Jr., Anderson GM, Rubenstein JL, Greenough WT, Beckel-Mitchener A, et al. Autism as a disorder of neural information processing: directions for research and targets for therapy. Mol Psychiatry. 2004;9(7):646-63. doi: 10.1038/sj.mp.4001499. PubMed PMID: 15037868.
9. Elsabbagh M, Johnson MH. Getting answers from babies about autism. Trends in cognitive sciences. 2010;14(2):81-7. doi: 10.1016/j.tics.2009.12.005. PubMed PMID: 20074996.
10. Weiss LA, Arking DE, Gene Discovery Project of Johns H, the Autism C, Daly MJ, Chakravarti A. A genome-wide linkage and association scan reveals novel loci for autism. Nature. 2009;461(7265):802-8. doi: 10.1038/nature08490. PubMed PMID: 19812673; PubMed Central PMCID: PMC2772655.
11. Wang K, Zhang H, Ma D, Bucan M, Glessner JT, Abrahams BS, et al. Common genetic variants on 5p14.1 associate with autism spectrum disorders. Nature. 2009;459(7246):528-33. doi: 10.1038/nature07999. PubMed PMID: 19404256; PubMed Central PMCID: PMC2943511.
12. Glessner JT, Wang K, Cai G, Korvatska O, Kim CE, Wood S, et al. Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. Nature. 2009;459(7246):569-73. doi: 10.1038/nature07953. PubMed PMID: 19404257; PubMed Central PMCID: PMC2925224.
13. 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. doi: 10.1038/nature11011. PubMed PMID: 22495311; PubMed Central PMCID: PMC3613847.
14. Pinto D, Delaby E, Merico D, Barbosa M, Merikangas A, Klei L, et al. Convergence of genes and cellular pathways dysregulated in autism spectrum disorders. Am J Hum Genet. 2014;94(5):677-94. doi: 10.1016/j.ajhg.2014.03.018. PubMed PMID: 24768552; PubMed Central PMCID: PMC4067558.
15. 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. doi: 10.1038/nature10945. PubMed PMID: 22495306; PubMed Central PMCID: PMC3667984.
16. Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, Walsh T, et al. Strong association of de novo copy number mutations with autism. Science. 2007;316(5823):445-9. doi: 10.1126/science.1138659. PubMed PMID: 17363630; PubMed Central PMCID: PMC2993504.
17. 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. doi: 10.1038/ng.3039. PubMed PMID: 25038753; PubMed Central PMCID: PMC4137411.
18. Klei L, Sanders SJ, Murtha MT, Hus V, Lowe JK, Willsey AJ, et al. Common genetic variants, acting additively, are a major source of risk for autism. Mol Autism. 2012;3(1):9. doi: 10.1186/2040-2392-3-9. PubMed PMID: 23067556; PubMed Central PMCID: PMC3579743.
19. 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. Nat Commun. 2014;5:5748. doi: 10.1038/ncomms6748. PubMed PMID: 25494366; PubMed Central PMCID: PMC4270294.
20. Flint J, Timpson N, Munafo M. Assessing the utility of intermediate phenotypes for genetic mapping of psychiatric disease. Trends Neurosci. 2014;37(12):733-41. doi: 10.1016/j.tins.2014.08.007. PubMed PMID: 25216981.

bioRxiv preprint doi: https://doi.org/10.1101/029488; this version posted October 19, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

21. 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. doi: 10.1038/nature10110. PubMed PMID: 21614001; PubMed Central PMCID: PMC3607626.
22. Iossifov I, O'Roak BJ, Sanders SJ, Ronemus M, Krumm N, Levy D, et al. The contribution of de novo coding mutations to autism spectrum disorder. Nature. 2014;515(7526):216-21. doi: 10.1038/nature13908. PubMed PMID: 25363768; PubMed Central PMCID: PMC4313871.
23. Cookson W, Liang L, Abecasis G, Moffatt M, Lathrop M. Mapping complex disease traits with global gene expression. Nature reviews Genetics. 2009;10(3):184-94. doi: 10.1038/nrg2537. PubMed PMID: 19223927.
24. 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. doi: 10.1038/nature13772. PubMed PMID: 25363760.
25. Zeng Y, Wang G, Yang E, Ji G, Brinkmeyer-Langford CL, Cai JJ. Aberrant gene expression in humans. PLoS genetics. 2015;11(1):e1004942. doi: 10.1371/journal.pgen.1004942. PubMed PMID: 25617623; PubMed Central PMCID: PMC4305293.
26. Mahalanobis PC. On the generalised distance in statistics. Proceedings National Institute of Science, India. 1936;2:49-55. Epub 1. doi: citeulike-article-id:4155812.
27. Corradi A, Fadda M, Piton A, Patry L, Marte A, Rossi P, et al. SYN2 is an autism predisposing gene: loss-of-function mutations alter synaptic vesicle cycling and axon outgrowth. Hum Mol Genet. 2014;23(1):90-103. doi: 10.1093/hmg/ddt401. PubMed PMID: 23956174; PubMed Central PMCID: PMC3857945.
28. Cesca F, Baldelli P, Valtorta F, Benfenati F. The synapsins: key actors of synapse function and plasticity. Progress in neurobiology. 2010;91(4):313-48. doi: 10.1016/j.pneurobio.2010.04.006. PubMed PMID: 20438797.
29. Benfenati F, Valtorta F, Bahler M, Greengard P. Synapsin I, a neuron-specific phosphoprotein interacting with small synaptic vesicles and F-actin. Cell Biol Int Rep. 1989;13(12):1007-21. PubMed PMID: 2517594.
30. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdottir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. Bioinformatics. 2011;27(12):1739-40. doi: 10.1093/bioinformatics/btr260. PubMed PMID: 21546393; PubMed Central PMCID: PMC3106198.
31. Tierney E, Bukelis I, Thompson RE, Ahmed K, Aneja A, Kratz L, et al. Abnormalities of cholesterol metabolism in autism spectrum disorders. American journal of medical genetics Part B, Neuropsychiatric genetics : the official publication of the International Society of Psychiatric Genetics. 2006;141B(6):666-8. doi: 10.1002/ajmg.b.30368. PubMed PMID: 16874769; PubMed Central PMCID: PMC2553243.
32. Chow ML, Pramparo T, Winn ME, Barnes CC, Li HR, Weiss L, et al. Age-dependent brain gene expression and copy number anomalies in autism suggest distinct pathological processes at young versus mature ages. PLoS genetics. 2012;8(3):e1002592. doi: 10.1371/journal.pgen.1002592. PubMed PMID: 22457638; PubMed Central PMCID: PMC3310790.
33. Sawicka K, Zukin RS. Dysregulation of mTOR signaling in neuropsychiatric disorders: therapeutic implications. Neuropsychopharmacology. 2012;37(1):305-6. doi: 10.1038/npp.2011.210. PubMed PMID: 22157871; PubMed Central PMCID: PMC3238083.
34. Lazaro MT, Golshani P. The utility of rodent models of autism spectrum disorders. Curr Opin Neurol. 2015;28(2):103-9. doi: 10.1097/WCO.0000000000000183. PubMed PMID: 25734952.

bioRxiv preprint doi: https://doi.org/10.1101/029488; this version posted October 19, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

35. Frye RE, Huffman LC, Elliott GR. Tetrahydrobiopterin as a novel therapeutic intervention for autism. Neurotherapeutics : the journal of the American Society for Experimental NeuroTherapeutics. 2010;7(3):241-9. doi: 10.1016/j.nurt.2010.05.004. PubMed PMID: 20643376; PubMed Central PMCID: PMC2908599.
36. Klaiman C, Huffman L, Masaki L, Elliott GR. Tetrahydrobiopterin as a treatment for autism spectrum disorders: a double-blind, placebo-controlled trial. Journal of child and adolescent psychopharmacology. 2013;23(5):320-8. doi: 10.1089/cap.2012.0127. PubMed PMID: 23782126.
37. Croft D, Mundo AF, Haw R, Milacic M, Weiser J, Wu G, et al. The Reactome pathway knowledgebase. Nucleic Acids Res. 2014;42(Database issue):D472-7. doi: 10.1093/nar/gkt1102. PubMed PMID: 24243840; PubMed Central PMCID: PMC3965010.
38. Schubert D, Martens GJ, Kolk SM. Molecular underpinnings of prefrontal cortex development in rodents provide insights into the etiology of neurodevelopmental disorders. Mol Psychiatry. 2014. doi: 10.1038/mp.2014.147. PubMed PMID: 25450230.
39. Stevens HE, Smith KM, Maragnoli ME, Fagel D, Borok E, Shanabrough M, et al. Fgfr2 is required for the development of the medial prefrontal cortex and its connections with limbic circuits. The Journal of neuroscience : the official journal of the Society for Neuroscience. 2010;30(16):5590-602. doi: 10.1523/JNEUROSCI.5837-09.2010. PubMed PMID: 20410112; PubMed Central PMCID: PMC2868832.
40. Schaaf CP, Sabo A, Sakai Y, Crosby J, Muzny D, Hawes A, et al. Oligogenic heterozygosity in individuals with high-functioning autism spectrum disorders. Hum Mol Genet. 2011;20(17):3366-75. doi: 10.1093/hmg/ddr243. PubMed PMID: 21624971; PubMed Central PMCID: PMC3153303.
41. Guglielmi L, Servettini I, Caramia M, Catacuzzeno L, Franciolini F, D'Adamo MC, et al. Update on the implication of potassium channels in autism: K(+) channelautism spectrum disorder. Frontiers in cellular neuroscience. 2015;9:34. doi: 10.3389/fncel.2015.00034. PubMed PMID: 25784856; PubMed Central PMCID: PMC4345917.
42. Lee H, Lin MC, Kornblum HI, Papazian DM, Nelson SF. Exome sequencing identifies de novo gain of function missense mutation in KCND2 in identical twins with autism and seizures that slows potassium channel inactivation. Hum Mol Genet. 2014;23(13):3481-9. doi: 10.1093/hmg/ddu056. PubMed PMID: 24501278; PubMed Central PMCID: PMC4049306.
43. Fatemi SH, Folsom TD, Reutiman TJ, Sidwell RW. Viral regulation of aquaporin 4, connexin 43, microcephalin and nucleolin. Schizophr Res. 2008;98(1-3):163-77. doi: 10.1016/j.schres.2007.09.031. PubMed PMID: 17997079; PubMed Central PMCID: PMC2259220.
44. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008;9:559. doi: 10.1186/1471-2105-9-559. PubMed PMID: 19114008; PubMed Central PMCID: PMC2631488.
45. Langfelder P, Luo R, Oldham MC, Horvath S. Is my network module preserved and reproducible? PLoS Comput Biol. 2011;7(1):e1001057. doi: 10.1371/journal.pcbi.1001057. PubMed PMID: 21283776; PubMed Central PMCID: PMCPMC3024255.
46. Heinzen EL, Ge D, Cronin KD, Maia JM, Shianna KV, Gabriel WN, et al. Tissue-specific genetic control of splicing: implications for the study of complex traits. PLoS biology. 2008;6(12):e1. doi: 10.1371/journal.pbio.1000001. PubMed PMID: 19222302; PubMed Central PMCID: PMCPMC2605930.
47. Cai JJ, Borenstein E, Petrov DA. Broker genes in human disease. Genome Biol Evol. 2010;2:815-25. doi: 10.1093/gbe/evq064. PubMed PMID: 20937604; PubMed Central PMCID: PMC2988523.
48. Lee HJ, Song JY, Kim JW, Jin SY, Hong MS, Park JK, et al. Association study of polymorphisms in synaptic vesicle-associated genes, SYN2 and CPLX2, with schizophrenia.

bioRxiv preprint doi: https://doi.org/10.1101/029488; this version posted October 19, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Behav Brain Funct. 2005;1:15. doi: 10.1186/1744-9081-1-15. PubMed PMID: 16131404; PubMed Central PMCID: PMC1215472.
49. Lionel AC, Crosbie J, Barbosa N, Goodale T, Thiruvahindrapuram B, Rickaby J, et al. Rare copy number variation discovery and cross-disorder comparisons identify risk genes for ADHD. Sci Transl Med. 2011;3(95):95ra75. doi: 10.1126/scitranslmed.3002464. PubMed PMID: 21832240.
50. 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. doi: 10.1371/journal.pone.0049475. PubMed PMID: 23227143; PubMed Central PMCID: PMCPMC3515554.
51. De Rubeis S, Pasciuto E, Li KW, Fernandez E, Di Marino D, Buzzi A, et al. CYFIP1 coordinates mRNA translation and cytoskeleton remodeling to ensure proper dendritic spine formation. Neuron. 2013;79(6):1169-82. doi: 10.1016/j.neuron.2013.06.039. PubMed PMID: 24050404; PubMed Central PMCID: PMCPMC3781321.
52. Ercan-Sencicek AG, Stillman AA, Ghosh AK, Bilguvar K, O'Roak BJ, Mason CE, et al. L-histidine decarboxylase and Tourette's syndrome. N Engl J Med. 2010;362(20):1901-8. doi: 10.1056/NEJMoa0907006. PubMed PMID: 20445167; PubMed Central PMCID: PMCPMC2894694.
53. Clarke RA, Lee S, Eapen V. Pathogenetic model for Tourette syndrome delineates overlap with related neurodevelopmental disorders including Autism. Transl Psychiatry. 2012;2:e158. doi: 10.1038/tp.2012.75. PubMed PMID: 22948383; PubMed Central PMCID: PMCPMC3565204.
54. Jalbrzikowski M, Lazaro MT, Gao F, Huang A, Chow C, Geschwind DH, et al. Transcriptome Profiling of Peripheral Blood in 22q11.2 Deletion Syndrome Reveals Functional Pathways Related to Psychosis and Autism Spectrum Disorder. PloS one. 2015;10(7):e0132542. doi: 10.1371/journal.pone.0132542. PubMed PMID: 26201030; PubMed Central PMCID: PMCPMC4511766.
55. Poultney CS, Goldberg AP, Drapeau E, Kou Y, Harony-Nicolas H, Kajiwara Y, et al. Identification of small exonic CNV from whole-exome sequence data and application to autism spectrum disorder. Am J Hum Genet. 2013;93(4):607-19. doi: 10.1016/j.ajhg.2013.09.001. PubMed PMID: 24094742; PubMed Central PMCID: PMCPMC3791269.
56. Krumm N, Turner TN, Baker C, Vives L, Mohajeri K, Witherspoon K, et al. Excess of rare, inherited truncating mutations in autism. Nat Genet. 2015;47(6):582-8. doi: 10.1038/ng.3303. PubMed PMID: 25961944; PubMed Central PMCID: PMCPMC4449286.
57. Prasad A, Merico D, Thiruvahindrapuram B, Wei J, Lionel AC, Sato D, et al. A discovery resource of rare copy number variations in individuals with autism spectrum disorder. G3 (Bethesda). 2012;2(12):1665-85. doi: 10.1534/g3.112.004689. PubMed PMID: 23275889; PubMed Central PMCID: PMCPMC3516488.
58. Waterhouse L, Gillberg C. Why autism must be taken apart. Journal of autism and developmental disorders. 2014;44(7):1788-92. doi: 10.1007/s10803-013-2030-5. PubMed PMID: 24390538.
59. Geschwind DH. Autism: many genes, common pathways? Cell. 2008;135(3):391-5. doi: 10.1016/j.cell.2008.10.016. PubMed PMID: 18984147; PubMed Central PMCID: PMCPMC2756410.
60. Garbett K, Ebert PJ, Mitchell A, Lintas C, Manzi B, Mirnics K, et al. Immune transcriptome alterations in the temporal cortex of subjects with autism. Neurobiol Dis. 2008;30(3):303-11. doi: 10.1016/j.nbd.2008.01.012. PubMed PMID: 18378158; PubMed Central PMCID: PMCPMC2693090.
61. Ecker S, Pancaldi V, Rico D, Valencia A. Higher gene expression variability in the more aggressive subtype of chronic lymphocytic leukemia. Genome Med. 2015;7(1):8. doi:

bioRxiv preprint doi: https://doi.org/10.1101/029488; this version posted October 19, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

10.1186/s13073-014-0125-z. PubMed PMID: 25632304; PubMed Central PMCID: PMC4308895.
62. Somel M, Khaitovich P, Bahn S, Paabo S, Lachmann M. Gene expression becomes heterogeneous with age. Curr Biol. 2006;16(10):R359-60. doi: 10.1016/j.cub.2006.04.024. PubMed PMID: 16713941.
63. Mar JC, Matigian NA, Mackay-Sim A, Mellick GD, Sue CM, Silburn PA, et al. Variance of gene expression identifies altered network constraints in neurological disease. PLoS genetics. 2011;7(8):e1002207. doi: 10.1371/journal.pgen.1002207. PubMed PMID: 21852951; PubMed Central PMCID: PMC3154954.
64. Li JJ, Liu Y, Kim T, Min RQ, Zhang ZL. Gene Expression Variability within and between Human Populations and Implications toward Disease Susceptibility. Plos Computational Biology. 2010;6(8):e1000910. doi: ARTN e1000910

DOI 10.1371/journal.pcbi.1000910. PubMed PMID: WOS:000281389500038.
65. Li J, Shi M, Ma Z, Zhao S, Euskirchen G, Ziskin J, et al. Integrated systems analysis reveals a molecular network underlying autism spectrum disorders. Mol Syst Biol. 2014;10:774. doi: 10.15252/msb.20145487. PubMed PMID: 25549968; PubMed Central PMCID: PMC4300495.
66. Pramparo T, Pierce K, Lombardo MV, Carter Barnes C, Marinero S, Ahrens-Barbeau C, et al. Prediction of autism by translation and immune/inflammation coexpressed genes in toddlers from pediatric community practices. JAMA Psychiatry. 2015;72(4):386-94. doi: 10.1001/jamapsychiatry.2014.3008. PubMed PMID: 25739104.
67. Willsey AJ, Sanders SJ, Li M, Dong S, Tebbenkamp AT, Muhle RA, et al. Coexpression networks implicate human midfetal deep cortical projection neurons in the pathogenesis of autism. Cell. 2013;155(5):997-1007. doi: 10.1016/j.cell.2013.10.020. PubMed PMID: 24267886; PubMed Central PMCID: PMC3995413.
68. Hulse AM, Cai JJ. Genetic variants contribute to gene expression variability in humans. Genetics. 2013;193(1):95-108. doi: 10.1534/genetics.112.146779. PubMed PMID: 23150607; PubMed Central PMCID: PMC3527258.
69. Wang G, Yang E, Brinkmeyer-Langford CL, Cai JJ. Additive, epistatic, and environmental effects through the lens of expression variability QTL in a twin cohort. Genetics. 2014;196(2):413-25. doi: 10.1534/genetics.113.157503. PubMed PMID: 24298061; PubMed Central PMCID: PMC3914615.
70. Walsh P, Elsabbagh M, Bolton P, Singh I. In search of biomarkers for autism: scientific, social and ethical challenges. Nature reviews Neuroscience. 2011;12(10):603-12. doi: 10.1038/nrn3113. PubMed PMID: 21931335.
71. Dey SS, Foley JE, Limsirichai P, Schaffer DV, Arkin AP. Orthogonal control of expression mean and variance by epigenetic features at different genomic loci. Mol Syst Biol. 2015;11(5):806. doi: 10.15252/msb.20145704. PubMed PMID: 25943345.
72. Hansen KD, Irizarry RA, Wu Z. Removing technical variability in RNA-seq data using conditional quantile normalization. Biostatistics. 2012;13(2):204-16. doi: 10.1093/biostatistics/kxr054. PubMed PMID: 22285995; PubMed Central PMCID: PMC3297825.
73. Stegle O, Parts L, Durbin R, Winn J. A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. PLoS Comput Biol. 2010;6(5):e1000770. doi: 10.1371/journal.pcbi.1000770. PubMed PMID: 20463871; PubMed Central PMCID: PMC2865505.
74. Smedley D, Haider S, Durinck S, Pandini L, Provero P, Allen J, et al. The BioMart community portal: an innovative alternative to large, centralized data repositories. Nucleic Acids Res. 2015. doi: 10.1093/nar/gkv350. PubMed PMID: 25897122.

bioRxiv preprint doi: https://doi.org/10.1101/029488; this version posted October 19, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

75. Konganti K, Wang G, Yang E, Cai JJ. SBEToolbox: A Matlab Toolbox for Biological Network Analysis. Evol Bioinform Online. 2013;9:355-62. doi: 10.4137/EBO.S12012. PubMed PMID: 24027418; PubMed Central PMCID: PMC3767578.
76. Rousseeuw PJ, Van Driessen K. A fast algorithm for the minimum covariance determinant estimator. Technometrics. 1999;41(3):212-23. doi: Doi 10.2307/1270566. PubMed PMID: WOS:000081562100004.
77. Verboven S, Hubert M. LIBRA: a MATLAB library for robust analysis. Chemometr Intell Lab. 2005;75(2):127-36. doi: DOI 10.1016/j.chemolab.2004.06.003. PubMed PMID: WOS:000227055000002.
78. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B (Methodological). 1995;57(1):289-300. doi: 10.2307/2346101.

Supplementary Information

Supplementary Table 1. GO term-defined gene sets that tend to be aberrantly expressed in autistic brains. Gene sets contain genes annotated with GO terms of three sub-ontologies: biological process (BP), molecular function (MF), and cellular component (CC).

Supplementary Table 2. WGCNA co-expression network modules containing genes that tend to be aberrantly expressed in autistic brains. Modules are annotated by using the DAVID-defined gene function keyword clusters. Representative genes with the corresponding function are shown in bold. Module preservation statistics calculated using function `modulePreservation` of WGCNA are given.

Supplementary Table 3. Genes in the three classifier gene sets and corresponding SSMD and ΔSSMD values.

Supplementary Fig. 1. Reproducibility of co-expression modules in non-ASD control group and the breakdown of modules in ASD. Ten example modules are shown with two independent data sets from controls, as well as one data set from ASD samples. Edge width is proportional to the Pearson's correlation coefficients (ranging 0.5 and 1). Node size is proportional to ΔSSMD for each gene.

Supplementary Fig. 2. Results of principal component analysis (PCA) showing the first four principal components (from PC1 to PC4). The distributions of 104 samples (57 controls and 47 ASD samples) on PCA spaces defined by PC1 and 2, PC2 and 3, and PC3 and 4 are shown.

Supplementary Fig. 3. Box plot of AUC (area under ROC curve) value against size of classifier gene set. For each size of gene set (from 3 to 15), 100 different random gene sets were constructed and tested on training set and test set for obtaining AUCs. The black and red box plots denote AUC values tested on training set (AUC1) and test set (AUC2) varying with the size of classifier gene set, respectively.

Supplementary Fig. 4. Scatter plot of AUC values tested on test set (AUC2) against AUC values tested on training set (AUC1) for 100 different random classifier 5-gene sets. Red line denote the least-squares line of the scatter plot. The Spearman correlation coefficient between AUC1 and AUC2 is 0.32 (P=1.1×10^{-3}). Inset shows the distribution of the Spearman rank correlation coefficients between AUC1 and AUC2 calculated with 1,000 replicates of such 100 random classifier 5-gene sets.

Tables.

Table 1. GSEA curated gene sets that tend to be aberrantly expressed in ASD. *Number of genes included in our analysis/Number of genes in the gene set. Gene sets mentioned in the main text are shown in italic. The previously known ASD-implicated genes are shown in bold.

GSEA gene set	Number of genes*	Top ΔSSMD gene	Reference
Metabolism and biosynthesis			
KEGG_PENTOSE_PHOSPHATE_PATHWAY	19/27	*H6PD, PRPS2,* ***PFKP***	
KEGG_STEROID_BIOSYNTHESIS	14/17	*SC5DL, NSDHL,* ***DHCR7***	
REACTOME_CHOLESTEROL_BIOSYNTHESIS	20/24	*SQLE, HSD17B7, HMGCR*	[31]
REACTOME_BRANCHED_CHAIN_AMINO_ACID_ CATABOLISM	16/17	*DLD, HIBADH, MCCC2*	
Immune/Inflammatory response			
BIOCARTA_LAIR_PATHWAY	4/17	*SELPLG, C3, ITGB1*	
BIOCARTA_41BB_PATHWAY	12/17	*MAPK8, ATF2, MAPK14*	
REACTOME_IL1_SIGNALING	25/39	*CHUK, RBX1,* ***BTRC***	[32]
REACTOME_REGULATION_OF_IFNA_SIGNALING	6/24	*STAT1, PTPN1, JAK1*	
Signaling pathway			
BIOCARTA_IGF1_PATHWAY	20/21	*JUN, CSNK2A1, ELK1*	
PID_S1P_S1P2_PATHWAY	21/24	*MAPK8, MAPK14, JUN*	
PID_HNF3APATHWAY (FOXA1/HNF3A TF network)	22/44	*NDUFV3,* ***PISD****, FOS*	
REACTOME_ENERGY_DEPENDENT_REGULATION_ OF_MTOR_BY_LKB1_AMPK	15/18	*PRKAA1, CAB39,* ***TSC1***	[7, 33, 34]
Vitamins and supplements			
BIOCARTA_VITCB_PATHWAY	6/11	*SLC2A3, COL4A2, SLC2A1*	
REACTOME_TETRAHYDROBIOPTERIN_BH4_SYNTHESIS_ RECYCLING_SALVAGE_AND_REGULATION	9/13	*GCHFR, PTS, AKT1*	[35, 36]
Miscellaneous			
REACTOME_ACTIVATED_POINT_MUTANTS_OF_FGFR2	4/16	***FGF9****, FGFR2, FGF1*	[38, 39]
REACTOME_ACTIVATION_OF_THE_AP1_FAMILY_OF_ TRANSCRIPTION_FACTORS	10/10	*MAPK14,* ***MAPK3****, ATF2*	[40]
REACTOME_INWARDLY_RECTIFYING_K_CHANNELS	20/31	***KCNJ10****, KCNJ4, GNG4*	[41, 42]
REACTOME_G2_M_CHECKPOINTS	22/45	*MCM2, RFC5, RPA2*	[43]

Table 2. The performances of classifiers based on gene set I, II, III tested on the training and test data sets. True positive (TP), true negative (TN), false positive (FP), false negative (FN), sensitivity (SN), specificity (SP), accuracy (ACC) values are reported.

Expression data set	Training set			Test set		
Number of samples (ASD + control)	93 (52 + 41)			93 (52 + 41)		
Classifier gene set	I	II	III	I	II	III
TP	43	44	45	40	41	37
TN	33	32	32	30	30	30
FP	8	9	9	11	11	11
FN	9	8	7	12	11	15
SN (%)	82.69	84.62	86.54	76.92	78.85	71.15
SP (%)	80.49	78.05	78.05	73.17	73.17	73.17
ACC (%)	81.72	81.72	82.80	75.27	76.34	72.04
SN (%), union of I, II, and III	100			98.08		
SP (%), intersection of I, II, and III	97.56			97.56		

Figures.

Fig. 1.

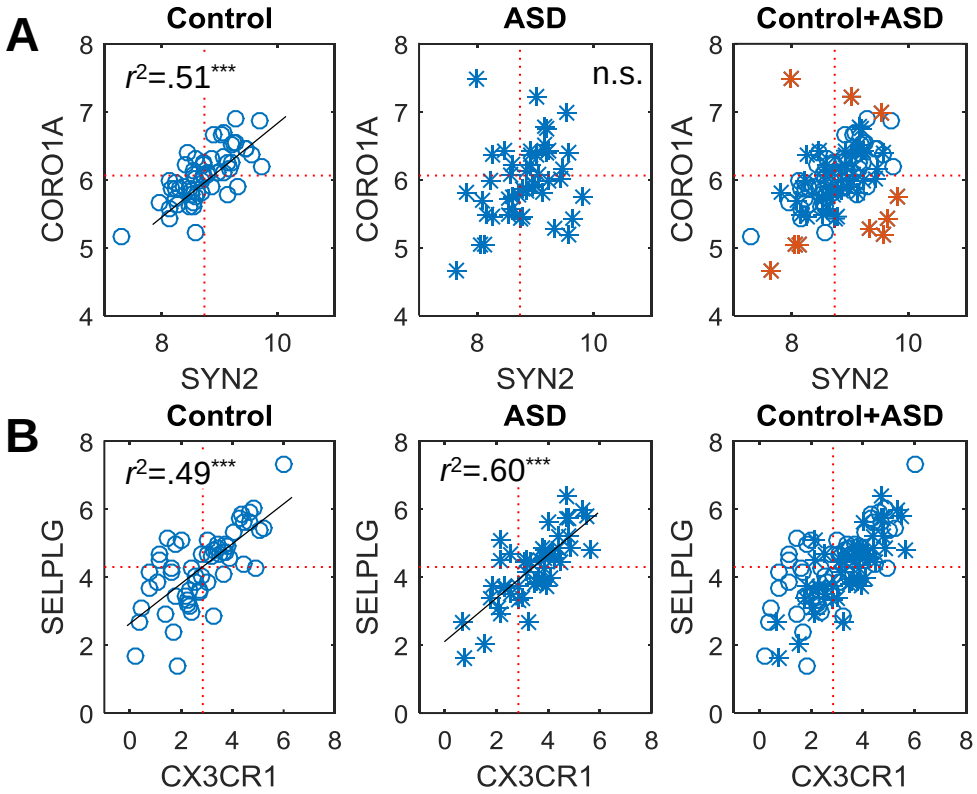


Fig. 2.

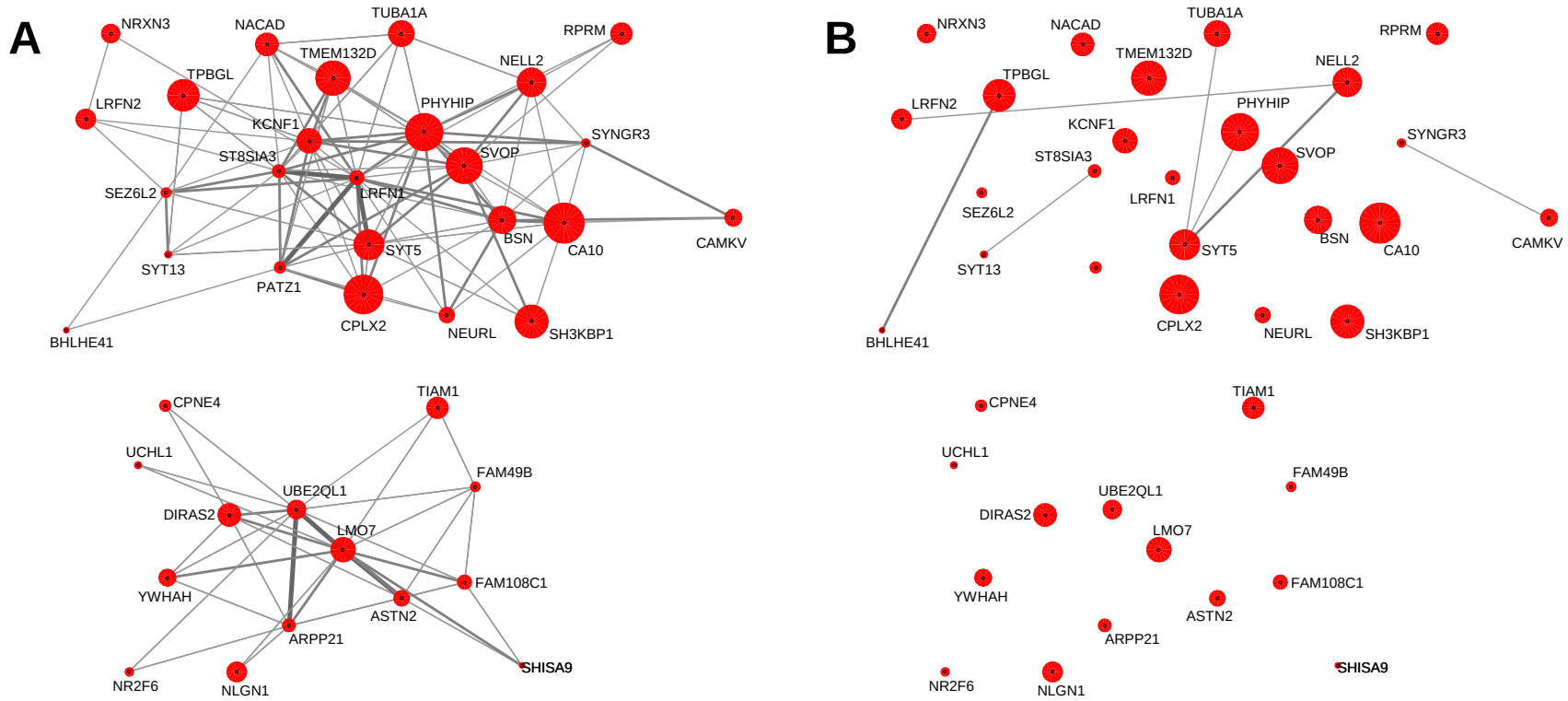
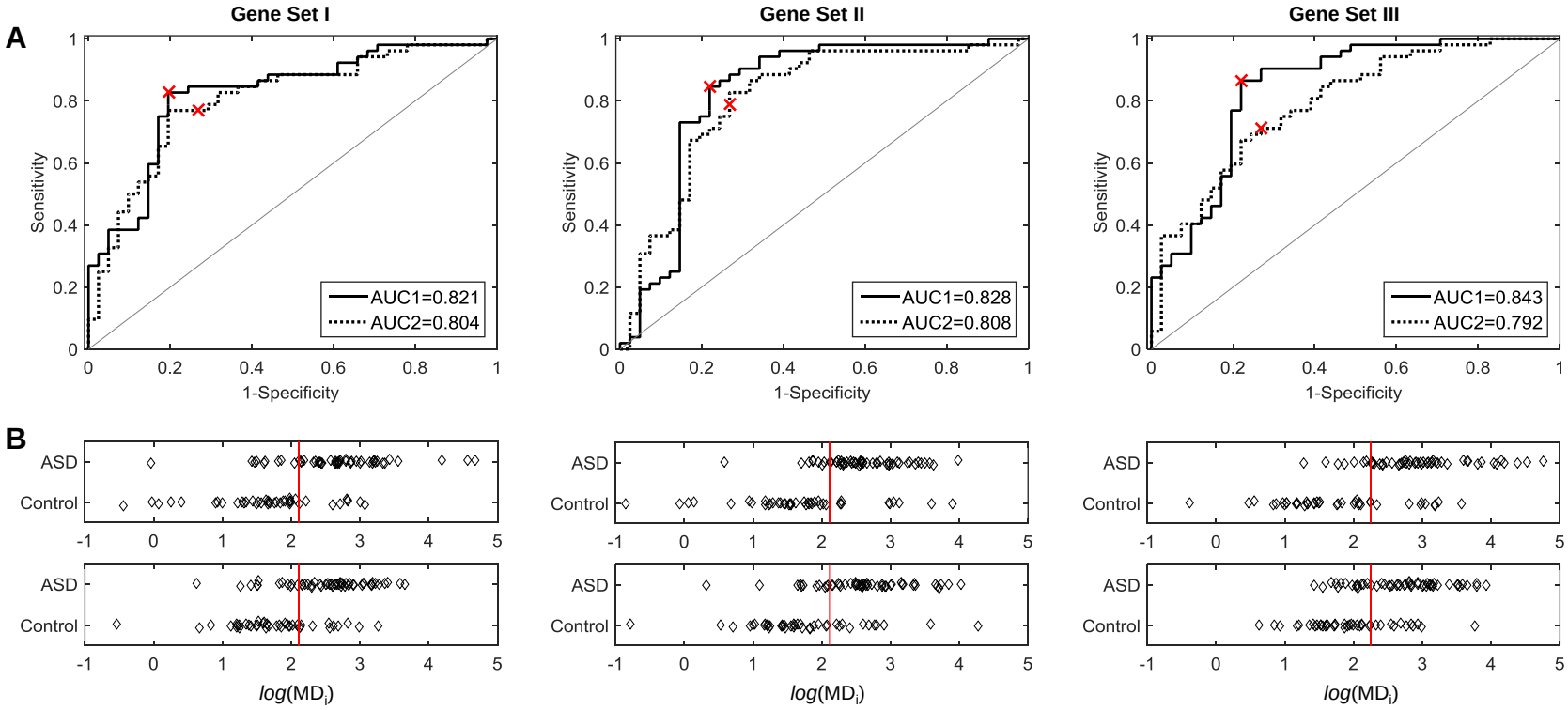

A
NRXN3
NACAD
TUBA1A
RPRM
TMEM132D
NELL2
TPBGL
LRFN2
PHYHIP
KCNF1
SYNGR3
ST8SIA3
SVOP
SEZ6L2
LRFN1
BSN
CA10
CAMKV
SYT5
SYT13
PATZ1
CPLX2
NEURL
SH3KBP1
BHLHE41
B
NRXN3
NACAD
TUBA1A
RPRM
TMEM132D
NELL2
TPBGL
LRFN2
PHYHIP
KCNF1
SYNGR3
ST8SIA3
SVOP
SEZ6L2
LRFN1
BSN
CA10
CAMKV
SYT5
SYT13
CPLX2
NEURL
SH3KBP1
BHLHE41
CPNE4
TIAM1
UCHL1
FAM49B
UBE2QL1
DIRAS2
LMO7
FAM108C1
YWHAH
ASTN2
ARPP21
SHISA9
NR2F6
NLGN1
CPNE4
TIAM1
UCHL1
FAM49B
UBE2QL1
DIRAS2
LMO7
FAM108C1
YWHAH
ASTN2
ARPP21
SHISA9
NR2F6
NLGN1

Fig. 3.


A
Gene Set I
Gene Set II
Gene Set III
Sensitivity
1-Specificity
AUC1=0.821
AUC2=0.804
AUC1=0.828
AUC2=0.808
AUC1=0.843
AUC2=0.792
B
ASD
Control
log(MDi)