Novel correlative analysis identifies multiple genomic variations impacting ASD with macrocephaly

Anthony Wynshaw-Boris (anthony.wynshaw-boris@case.edu)
Case Western Reserve University

chen fu
Justine Ngo
Case Western Reserve University

Shanshan Zhang
Case Western Reserve University

Leina Lu
Case Western Reserve University

Alexancer Miron
Case Western Reserve University

Simon Schafer
The Salk Institute for Biological Studies

Fred Gage
Salk Institute for Biological Studies

Fulai Jin
Frederick Schumacher

Article

Keywords:

Posted Date: September 9th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-2009452/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
Abstract

Autism spectrum disorders (ASD) display both phenotypic and genetic heterogeneity, impeding the understanding of ASD and development of effective means of diagnosis and potential treatments. Genes affected by genomic variations for ASD converge in dozens of gene ontologies (GOs), but the relationship between the variations at the GO level have not been well elucidated. In the current study, multiple types of genomic variations were mapped to GOs and correlations among GOs were measured in ASD and control samples. Several ASD-unique GO correlations were found, suggesting the importance of co-occurrence of genomic variations in genes from different functional categories in ASD etiology. Combined with experimental data, several variations related to WNT signaling, neuron development, synapse morphology/function and organ morphogenesis were found to be important for ASD with macrocephaly, and novel co-occurrence patterns of them in ASD patients were found. Further, we applied this gene ontology correlation analysis method to find genomic variations that contribute to ASD etiology in combination with changes in gene expression and transcription factor binding, providing novel insights into ASD with macrocephaly and a new methodology for the analysis of genomic variation.

Introduction

Autism spectrum disorder (ASD) is a neurodevelopmental disease that consists of social interaction abnormalities and repetitive behaviors. Beyond these two core symptoms, ASD patients may exhibit additional behaviors and comorbidities (1, 2) such as seizures, aggressive behavior, intellectual disability, and brain development abnormalities. Thus, ASD patients manifest substantial phenotypic heterogeneity. About 25% of ASD patients display early brain overgrowth (macrocephaly) as a comorbidity (3). Previous studies demonstrated that this overgrowth begins in mid-gestational fetal development and persists postnatally (3). Although ASD is often diagnosed by 3 years of age based on the presence of core symptoms, the later age of onset makes it difficult to investigate prenatal pathophysiology. Since the brain overgrowth abnormalities precede the behavioral abnormality, understanding the genetic mechanisms of ASD associated with macrocephaly could facilitate an earlier diagnosis and potential therapeutic targets of ASD.

The genetic heterogeneity of ASD is also extensive and has been broadly accepted. Many de novo, rare inherited and common genomic variations have been previously reported (4, 5, 6, 7). ASD-related variations cover multiple categories, including single nucleotide variants (SNVs), insertions-deletions (INDELs), and structural variations (SVs) such as larger deletions or duplications (6, 7, 8, 9). These variations affect more than 700 genes (10) in dozens of gene ontologies (GOs, 11), underscoring the substantial role of genetics in phenotypic heterogeneity. We believe by focusing on ASD with a single phenotypic attribute, namely ASD with macrocephaly, we can minimize the complexity due to ASD phenotypic and genetic heterogeneity, thereby simplifying the discovery of genetic mechanisms responsible for this subset of individuals with ASD.
Traditional GO analysis may not be adequate to identify genes responsible for ASD etiology. First, as a result of population history, evolution or special function, some regions/genes display high rates of variation. For example, genes functionally related to biological/cell adhesion display high rates of evolution in the human genome (12, 13). Variations of these genes can be found in both ASD patients and normal individuals. Second, rare variations, or variations on genes with “house-keeping” functions that may be important for the phenotype, may not cause significant results in GO analysis, as the number of these genes is relatively small.

While previous analyses have concluded that genes dysregulated in different ASD patients vary greatly, they converge on similar sets of pathways (11, 14). However, the combinatorial pattern of converged pathways for genomic variants that are key to ASD etiology remains unclear. It is important to explore the interaction among GOs enriched with genomic variations in ASD. A methodology for the analysis of pathway correlation has recently been developed. ClueGO (15) was developed as a Cytoscape (16) plug-in capable of grouping GO terms into modules based on the similarity among GOs. This tool provided a network visualization of GOs overrepresented in gene lists uncovered in genetic or gene expression analysis. However, the connection between GOs identified by these methods was determined by the similarity of genes in GO terms, not the quantitative measurement of GOs affected by genomic variations.

Further, it is important to integrate the analysis of different types of genetic variations, given that each type accounts for a small proportion of ASD risk (6, 17) Most ASD genomic studies focused on a single type of genomic variation (7, 8, 9). In several recent review articles, the combinatorial effect of SNVs and CNVs were identified (17), but the studies lacked efficient quantitative models. Although most quantitative models were originally designed to integrate variants of a single type, they may be repurposed for the integration of different types of data. For example, DAWN (18, 19) and MAGI (20) were designed to combine de novo mutations into modules with gene expression and protein-protein interaction networks as supporting information. However, these models require transmission score or gene expression data, which could limit the application of these models in ASD research.

Several recent studies for ASD with macrocephaly consistently found that genes in functional groups encompassing “cell adhesion” and “neuron development” were enriched with genomic variations and transcriptionally dysregulated genes (21, 22, 23). Among ASD patients lacking macrocephaly as a comorbidity, several genes in close proximity to regions identified by GWAS of ASD were also in these functional groups (24, 25). These results suggested that genes in these functional groups are important for ASD, although they are not limited to the subgroup of individuals with ASD and macrocephaly. On the other hand, several studies, using mouse and human iPSC models, demonstrated that WNT pathway activity was dysregulated in ASD with macrocephaly, whereas GO analysis for genes enriched with genomic variations in ASD seldom detected WNT-related GOs. Therefore, we hypothesize that the co-occurrence of variations on genes in different GOs may be critically important for the etiology of ASD, including ASD with macrocephaly. Evaluation of co-occurrence in gene expression datasets, such as WGCNA (30), has proven to illuminate linked pathways dysregulated in a variety of disorders (22, 26).
In the current study, we exploit SNV/INDEL and SV (deletion) data from both the Simons Simplex Collection (SSC) dataset from the Simons Foundation Autism Research Initiative (SFARI) (SFARI-SSC dataset, 10) and sequencing data from previously published iPSC models for ASD with macrocephaly by our lab and collaborators (22) (validation dataset). Both datasets detected SNVs, INDELs and structural variations (SVs) enriched in ASD with macrocephaly and control individuals. Traditional GO analysis for genes with these variations identified a set of GOs that were highly consistent between the two datasets, including “cell-cell adhesion” and “neuron differentiation”, among others. Using a pipeline we developed, the correlated GO pairs specific to ASD with macrocephaly were identified, helping us to identify a set of variations that together contributed to the etiology of ASD with macrocephaly. With the same pipeline examining samples of ASD microcephaly patients and ASD with no brain size change, several GO pairs and a group of genomic variations were identified for each of these subtypes of ASD. These findings demonstrate the utility of using co-occurring variation to identify potential links among various genes participating in disparate GOs in the etiology of various subsets of ASD.

Materials And Methods

Selection of samples from SFARI-SSC database

Brain circumferences of ASD probands from 2760 Simons Simplex Collection (SSC) families (10) were normalized by the age of the measurement. Probands with larger than 2 standard deviations (SD) above average were labeled as ASD-macro (ASD probands with macrocephaly comorbidity) and those with 2 SD below average were labeled as ASD-micro (ASD with microcephaly). This standard resulted in the identification of 47 and 52 probands with macrocephaly and microcephaly, respectively. We next selected 52 probands from the same database with the smallest difference from average and labeled them as “ASD-other” (probands with no brain size phenotype).

Genomic variation data from SSC dataset

SNV/INDEL data from the 41 ASD-macro, 37 ASD-micro and 38 ASD-other probands and their fathers were downloaded from the Simons Foundation Autism Research Initiative (SFARI) SSC database (10). Fathers from ASD-other families (n=38) were used as controls for ASD macro and ASD micro probands. Fathers from the ASD micro group (n=37) were used as controls for ASD-other probands for subsequent genomic variation analysis (Table S1A).

In total 320,710 SNV/INDEL loci were downloaded. Since the minimum sample number for each group for specific loci was required to be 8 (n_sample >=8), a total of 60,617 loci were annotated using ANNOVAR (27). Variations in the intergenic, upstream, or downstream regions (> 10 Kb distance from TSS or TTS) were identified and excluded for further analysis. For each of the remaining loci, alternative allele (ALT) frequencies were compared between each ASD group and controls. Loci with an increased ALT frequency of 10% or more in ASD than in CTRL were considered “ASD-enriched loci.” There were 7,373, 5,233 and 2,458 ASD-enriched SNV/INDEL loci in ASD-macro, ASD-micro and ASD-other groups, respectively (Table
S1B-D). These loci affected 457, 411 and 562 genes in each of the groups (Table S1E). These gene lists were the input for GO analysis using Gene Set Enrichment Analysis (GSEA) (28).

Structural variation data was also obtained from the SFARI-SSC database. Only deletions were used for subsequent analysis. After filtering by $n_{\text{sample}} \geq 8$ in each group, 1,044, 1,330 and 1,223 deletion loci were included for ASD-macro, ASD-micro and ASD-other, respectively (Table S1 G-I). These loci were annotated by AnnotSV (29). The same standard identifying ASD-enriched loci as applied to SNV/INDEL data was used. GO analyses were performed for genes intersected with the ASD- or control-enriched deletions (Table S1F) using GSEA (see above).

**Mapping genomic variations to GOs**

The GO and gene lists were downloaded from a GSEA database ("msigdb.v7.1.symbols.gmt" under “All gene sets”). Genomic variation rate per GO was calculated based on these loci (in the formula listed below). Specifically, for each selected ASD or control individual from the SSC dataset, the number of ASD-enriched loci (increase of alternative allele frequency in probands over controls $\geq 0.1$) in each type of variation (SNV/INDELs and SVs) for each gene (genei) were summed for all genes in each GO (GOj) and normalized by the total number of genes in that GO to get $\text{variation\_rate}_{\text{GOj}}$. [FS1]

$$\text{variation\_rate}_{\text{GOj}} = \frac{\sum_{i=1}^{n_{\text{gene\_GOj}}} n_{\text{SNV\_genei}} + \sum_{i=1}^{n_{\text{gene\_GOj}}} n_{\text{INDEL\_genei}} + \sum_{i=1}^{n_{\text{gene\_GOj}}} n_{\text{SV\_genei}}}{n_{\text{gene\_GOj}}}$$

The variation-rate$_{\text{GO}}$ was calculated for all 10,131 GOs. GO with zero counts for all samples (n=1845) were excluded from subsequent analysis. The individuals from each of the 6 groups (ASD-macro, ASD-micro and ASD-other, CTRL-macro, CTRL-micro and CTRL-other) were randomly divided into group 1 and group 2, with 115 and 114 samples, respectively (Table S3A).

**gWGCNA analysis of genomic variation data**

We employed a weighted correlation network analysis (WGCNA), a data mining method used for studying biological networks based on pairwise correlations between variables (30), to analyze GO correlations. With WGCNA default settings selected (soft threshold power=6; minimum module size=30), the sample tree was first generated for group 1 based on inter-sample Spearman correlation using variation\_rate$_{\text{GO}}$. Next, the WGCNA network was constructed with dissimilarities among modules to be at least 15%. [FS2]

After the construction of the network, the eigengene for each module was associated with each of the two “phenotypes:” ASD\_vs\_CTRL (labeled as “ASD”) (all ASD samples coded as 1 and controls as 0); and brain size (labeled as “Brain”) (all SSC probands with macrocephaly coded as 2, SSC proband with microcephaly as 0 and probands from “ASD-other” group as 1, all control individuals (fathers) coded as...
1) using Pearson correlation. Modules with significant correlation (p<0.05) for either phenotype were detected for group 1 samples. Based on the correlation with ASD and brain phenotypes, GO modules were defined as several “module groups” such as ASD-macro, ASD-micro, etc. (Table S3B).

The same settings were then applied to network construction for group 2 samples to detect GO modules associated with ASD and brain size. Venn diagrams were generated to find results from group 1 that could be confirmed by those from group 2 for each module group.

In the “confirmed” GOs, if a few representative GO groups (e.g., GOs related to “WNT signaling”) were enriched in each module group, were tested using a proportions test (31).

**Correlation among GOs related to ASD**

The enriched representative GOs detected above were used as seeding GOs (sGOs). Pearson correlation among these seeding GOs and all other GOs for ASD and Control samples were calculated for each of 3 sets of samples: ASD-macro, ASD-micro and ASD-other. The correlation matrix was flattened using a R library “corrplot”, and the p value for each GO pair was calculated. Using p<0.05 as the cutoff, GO pairs that displayed positive and significant correlation in ASD samples and negative or insignificant correlation in control samples were considered “correlated GO” (cGO) for subsequent analysis.

**Detection of loci enriched in seeding and correlated GOs**

Total occurrence of each gene in each of the 8 GO groups was calculated (4 sGO groups and 4 cGO groups, Table S3F-I), with 4 seeding GO groups (GOs related to “neuron”, “WNT”, “organ morphogenesis” and “synapse”, Fig.5) and 4 groups of GOs correlated to each of the sGO. The occurrence value was standardized into a z score.

The ALT allele frequency difference between ASD and CTRL (>=0.1, see above) was also standardized into a z score. The “combined enrichment score” for each locus was defined as the product of the z score for gene frequency in a specific GO group and the z score for ALT allele enrichment. Loci with positive combined enrichment score were selected.

**Detection of SNV/INDEL loci enriched in ASD or controls in validation dataset**

Whole genome DNA was extracted from fibroblast cell lines of 8 ASD with macrocephaly and 5 gender/age matched control individuals previously published with the clinical phenotype (22). Genomic variations detected in these cell lines were called “validation dataset” in this paper. Exome libraries were produced in the CWRU Genome Core using the Illumina ultra-sensitive exome library protocol. The libraries were sequenced on Illumina HiSeq 2500 at 8 libraries/run, which yielded ~ 100-150 million reads per library. The reads were aligned to the hg19 reference genome using BWA (32) with default settings (Gap open penalty=6, Mismatch penalty=4, etc) and yielded about 20X coverage of the human genome in each library. VCFtools (33) were used to call variants. The single nucleotide variant (SNV) lists were generated after filtering out loci with low reading depth (<10), low number of reads in support of variant
(n<3), low alignment quality (q<20) and low base quality (Q<30). ANNOVAR (27) was used to annotate the SNVs. The ASD-enriched SNV/INDELs were those with significant p value for alternative allele frequency difference between ASD and control [5,000 times resampling using R, similar to previously reported method (34)]. All SNV/INDELs were compared with records in dbSNP. Filters used for calling INDELs were the same as those used for SNVs. The annotation of INDEL was performed using SeattleSeq Annotation 138 (35). GO analysis was performed on genes harboring selected SNV/INDELs using GSEA. SNVs under negative selection were found using Funseq (36).

**Matepair sequencing and calling of deletions in validation dataset**

“Jumping libraries” for matepair sequencing (37) from 8 ASD samples with macrocephaly and 2 control samples (described above) were built using Illumina Nextera Mate Pair Sample Prep Kit. The libraries were sequenced in the CWRU sequencing core using an Illumina HiSeq 2500 machine with 8 samples/run, which yield ~ 60 million reads per library. The reads were first trimmed off adapter sequences and reverse complemented using Nxtrim (38), and then aligned to hg19 using BWA, resulting in an average physical coverage of about 2X and actual coverage of human genome of ~ 50X.

The bam files were used as input for DELLY (39) and LUMPY (40) (version 0.2.13) to call deletions. The start and end intervals of each structural variant (SV) from both algorithms were intersected using BEDTools (41). The matched intervals were used for the final inference of deletion breakpoints (Fig. S4). The deletions from all samples were collapsed if they overlapped. The frequency of deletions among ASD samples was calculated for these collapsed deletion events. The deletions from 2 control samples were inferred following the same process (“control deletion list”). The deletions were classified as “ASD-Control-shared deletions” if the ASD deletions overlapped with the control list or “ASD-specific deletions” otherwise. Deletions were then annotated using AnnotSV (29). GO analysis for genes intersected with the deletions was performed using GSEA.

To test if the variations selected by our pipeline from SFARI-SSC samples were enriched by putative targets of any transcription factors, we applied CHEA analysis (42) for the genes that carried the 1,514 loci selected from SFARI-SSC ASD-macro samples and 410 loci selected from SFARI-SSC ASD-other samples. Using the CHEA online (https://maayanlab.cloud/chea3/#top) function “ChIP-Seq-> Literature” library. Similarly, genes affected by the 515 ASD-specific deletions in the validation dataset were tested with CHEA online tools. The results from the 3 input sets were compared using a Venn diagram (Fig. S5).

**RNASeq for neural progenitor cells (NPCs) from validation dataset**

Total RNA was extracted from NPCs from 8 ASD-macro and 5 control lines as mentioned above. One control line (COVE) was excluded from subsequent analysis as its karyotype was abnormal. The RNA was purified and quantified, and samples of high quality (RIN>= 7.0) were used. The Illumina TruSeq Stranded Total RNA kit with Ribo Zero Gold (for rRNA removal by hybridization/bead capture) was used for library preparation. Optimized libraries were then loaded onto the HiSeq 2500 flowcell (8 libraries/lane) for 50 bp single-end sequencing.
Adapter sequences were trimmed and filtered. Reads that passed quality filter were aligned to hg38 using HISAT2 (43) and converted to sorted BAM files with samtools (44). Identification of differentially expressed genes and statistical analyses was performed using DESeq2 (45).

ChIPSeq for NPCs from validation dataset

ChIPSeq libraries for BRN2 were prepared using NPCs from 3 ASD lines and 3 control lines. About 10 million NPCs for each library were collected between passages 6-9 at day 3 (about 80% confluency) and cross linked using 4% formaldehyde for 10 mins at room temperature. The cells were resuspended in lysis buffer, sonicated, and incubated with antibody (POU3F2 (D2C1L) from Cell Signaling) linked DynaBeads Protein G overnight at 4 °C. The DynaBeads were washed and then reverse crosslinked at 65 degrees for 12 hours. RNA and antibody were removed with RNase A (Ambion Cat # 2271) and proteinase K (Invitrogen, 25530-049). The pulled-down DNA was end-repaired, and a poly-A tail was added, linked with adapter and PCR amplified. The PCR product was gel purified and fragments in the 250–400 bp were excised and purified with Qiagen MinElute Gel Extraction Kit (Qiagen, 28606). The Bioanalyzer DNA 1000 assay (Agilent) was used to access the quality of the libraries. Seventy-five bp single-end reads were generated for high quality libraries using the HiSeq 2500 (8 libraries/lane on flowcell) sequencing pipeline. Reads were adaptor trimmed with fastx_clipper, aligned with hg19 using BWA (32), and further processed using SamTools (44). Peaks were called by MACS14 (46) with default settings using sorted bam files with redundant reads removed. Called peaks were overlapped with published BRN2 on human NPC (47) and ATAC-Seq data from sample NPC lines (48) (see below).

Further bioinformatic processing of RNASeq, ChIPSeq and HiC data

The genes with selected genomic variation loci and differentially expressed between ASD and control NPCs were illustrated using Venn diagrams (https://bioinformatics.psb.ugent.be/webtools/Venn/). The ATAC-Seq data for 3 ASD and 2 control NPC lines were previously published (48) and obtained from the Gage lab through collaboration. These lines were a subset of the NPC lines on which we performed RNASeq and ChIPSeq experiments. Overlap between ATAC-seq peaks and BRN2 ChIPSeq were found using BedTools (41). Published BRN2 ChIPSeq data at NPC stage (with 2 samples) (47) were downloaded. To ascertain that the BRN2 binding positions were in actively transcriptional sites, BRN2 peaks from at least 2 control lines or published BRN2 peaks from both NPC samples needed to overlap with ATAC-Seq peak from both control lines.

Similarly, published β-catenin ChIPSeq from hESCs (human embryonic stem cells) (49) were obtained from GEO and overlapped with control NPC ATAC-Seq data. The ATAC-Seq confirmed BRN2 and β-catenin ChIPSeq peaks were annotated using HOMER (50).

Published HiC data from human NPCs were obtained (51). Both sides of significantly associated HiC intervals were overlapped with selected genomic variations using BedTools.
Results

Genomic variations of genes that function in neurogenesis, neuron development, and cell adhesion were enriched in, but not unique to, ASD-macro probands

Based on $\geq 2$ SD larger/smaller than mean head circumference standard (3), 41 ASD macro and 37 ASD-micro probands were selected from the SFARI-SSC database (Table S1A). Another 38 probands were selected that displayed head circumferences close to average (ASD-other, Fig. 1A). The significant increase in head circumference of the ASD-macro probands over their siblings confirmed our method of sample selection (Fig.1B).

For genetic analysis, fathers of these 38 ASD-other probands were used as controls for ASD-macro and ASD-micro probands. Fathers from 37 ASD-micro families were used as controls for ASD-other probands. In total, 320,710 SNV/INDEL loci for these 191 individuals were retrieved from exome sequencing results of SFARI-SSC samples (Table S1A). After filtering for sample size, SNV/INDEL loci enriched in either ASDs or controls (enrich rate $\geq 0.1$, see method) from each of the 3 groups (ASD-macro, ASD-micro and ASD-other) were selected. A total of 7,373 loci for ASD-macro, 5233 loci for ASD-micro and 2458 loci for ASD-other were selected for subsequent analysis (Table S1B-D). Genes with these loci enriched in ASD probands were used as input for GO analysis (457 for ASD-macro, 411 for ASD-micro and 561 for ASD-other probands; 468, 480 and 491 for controls for these 3 groups, respectively) (Table S1E).

GO analysis demonstrated that, in ASD-macro samples, GO terms for neurogenesis, neuron development, organ morphogenesis, and biological adhesion were overrepresented by genes with ASD-enriched loci (Fig.1C, Table S2A). For ASD-micro samples, GO terms representing external encapsulating structure organization were most significant, together with GO terms related to neuron development and adhesion (Table S2B). The GO list for the ASD-other probands was similar to that of ASD-macro, with less significant levels for corresponding terms (Table S2C). These results suggested that the 3 types of ASD comorbidities could be characterized based on enriched genomic variations. Specifically, genes related to neurodevelopment and adhesion were more represented in ASD-macro individuals.

The GO results for structural variation data confirmed the conclusion from the SNV/INDEL data. GO terms related to adhesion remained significant in ASD-macro and ASD-micro probands, whereas GO terms for neuron development and neuron differentiation were not the top terms in ASD-macro individuals, but were still significant (Fig.1D, Table S1 F-I).

Finding GOs associated with ASD comorbidities by gWGCNA pipeline

We hypothesized that the co-occurrence of genomic variations affecting genes with different but related biological functions would be important for the etiology of ASD and that different co-occurrence patterns of genomic variations would be present in ASD probands with different comorbidities. To determine this co-occurrence, we used weighted correlation network analysis (WGCNA) with genomic variation data as
input (so called gWGCNA hereafter) and determined correlated “GO modules” associated with brain size and ASD phenotype.

The 229 SFARI-SSC ASD and control samples were randomly assigned into two groups (n\textsubscript{group1}=115 and n\textsubscript{group2}=114). Genes with SV and SNV/INDEL loci (enrichment rate >=0.1) were mapped to each GO for each sample (Table S3A). The gWGCNA pipeline was performed on each group with the same settings (n\textsubscript{GO} in module >=30, module similarity <= 0.85). Performance of gWGCNA algorithm was similar in these two groups: the “similarity score” was comparable for majority of group 1 and group 2 modules. Fig. 2A and 2B). The proportion of un-clustered GO module (grey module) was similar in the two groups. GO modules positively associated with ASD (ASD-only), with brain size (macrocephaly only) and with both (“ASD-macro”) were detected in both group 1 and group 2 (Fig. 2C and 2D). Modules associated with ASD and negatively associated with brain size were considered ASD-micro modules. Other modules (except un-clustered GOs (grey module)) were labeled as “non-significant” (Table S3C).

The GOs in corresponding module groups between group 1 and group 2 were conserved (Fig. 3A-E, Table S3D), with the highest proportion of overlap between group 1 and 2 being for “ASD-only” (73.9%, Fig. 3A), followed by “ASD-macro” (65.5%, Fig. 3B), then “Macro-only” (43.1%, Fig. 3C) and lowest for “ASD-micro” (37.1%, Fig. 3D). These results suggested that this pipeline reliably detected GO modules associated with ASD and ASD-macro.

**GO groups for WNT, neuron morphology/function and organ morphogenesis were enriched in GOs associated with ASD macrocephaly**

We next took the overlapped GOs for each module group as input to determine whether a few functionally similar GO groups (called “GO groups” hereafter) were enriched. We first tested GOs overrepresented by genes with ASD-enriched CNVs (11), which included cell proliferation, GTPase/Ras activity, and organ morphogenesis. In addition, WNT signaling was tested because WNT activity was decreased in ASD with macrocephaly (22). Furthermore, GOs presumably related to brain size change such as head/brain development and neuron morphology/function were examined.

A proportion test (31) demonstrated that WNT signaling, neuron function/morphology, and organ morphogenesis-related GOs were enriched in GO modules associated with ASD-macro (Fig. 4A, Table S3E). Synapse-related GOs were enriched in those associated with ASD-only. These results are supported by previous publications that found the importance of WNT signaling for ASD with macrocephaly (22). GOs for neuron and organ morphogenesis were also significant and plausible GO categories found for brain size change. On the other hand, synapse function, especially vesicle release related GOs, were enriched in ASD only associated GOs (Fig. 4A, Table S3E), suggesting that synapase function was dysregulated to affect neurological, behavioral and/or cognitive functions associated with ASD without brain size differences. That adhesion related GOs was not enriched in the modules associated with ASD-macro may be due to the big number of GOs related with adhesion in non-significant GO modules.
Therefore, even the number of adhesion related GOs were big in GO modules significantly associated with “ASD” or “Brain”, the proportion test result was not significant.

**Several GOs, including cell cycle, were enriched in correlated GOs for ASD with different comorbidities**

The four representative GO groups enriched in GOs associated with ASD phenotypes (WNT signaling activity, organ morphogenesis, neuron and synapse) were used as “seeding GOs” (sGOs) for further analysis (Table S3F). With these sGOs, we calculated the Pearson correlation for each seeding GO versus all other (n=8425) GOs in ASD macro, ASD micro and ASD-other (Table S3G-J). Consistent with our hypothesis that multiple genomic variations tend to affect genes with different but related functions and contribute to phenotype collaboratively in each individual, a set of “correlated GOs” (so called cGOs hereafter) was found. In ASD-macro probands, there were 630 cGOs positively correlated with WNT sGOs, 2,235 with neuron sGOs, and 3 011 with organ morphogenesis sGOs (Table S3 G-I). In the ASD-other probands, 4322 cGOs were significantly correlated with synapse sGOs ([MOU1] [CF2] [CF3] Table S3J). As these positive correlations were unique to ASD probands (control samples showed either negative or insignificant correlation), variations on genes belonging to these GOs may contribute collaboratively to ASD etiology.

Similar to the enrichment test we performed for sGOs (Fig. 4A), we tested if specific GO groups were enriched in these cGOs (Fig. 4B). Cell cycle was enriched in cGOs correlated with all three sets of sGOs (sGO WNT, sGOs neuron and sGO organ morphogenesis) for ASD-macro. On the other hand, the cell cycle was not enriched in cGOs correlated with sGO (synapse) for the ASD-other group. This finding suggested that cell cycle might be a very important biological process in brain size change in ASD. Interestingly, cell cycle-related GOs were significantly enriched in both cGOs found in ASD macro and ASD micro samples, except for the cGOs correlated with WNT sGOs, in which cell cycle-related GOs were only enriched in ASD macro samples (Fig. 4B). These observations may together suggest that variations in WNT signaling could trigger changes in cell cycle to cause ASD with macrocephaly, but some other factor(s) may be responsible for changes in cell cycle in ASD with microcephaly. Two example plots further elucidate the correlation between cell cycle related GOs and GOs related to neuron (Fig. 4C) or organ morphogenesis (Fig. 4D).

Adhesion-related GOs were never enriched in cGOs in either the ASD-macro or ASD-other group (Fig.4B). Only the ASD-micro group showed deprivation of adhesion-related GOs compared with either ASD-macro or ASD-other in cGOs correlated with sGO “neuron.” These observations do not exclude the contribution of adhesion to ASD etiology; the suggested adhesion was not specific to any of these 3 ASD subgroups. In other words, adhesion-related variations may not account specifically for brain size comorbidity in ASD.

“Neural process”-related GOs were all GOs related to the nervous system except for those with “neuron” in the name. This GO category is not enriched in cGOs correlated with most sGOs for ASD-macro except for sGO “organ morphogenesis.” Nonetheless, in cGOs correlated with sGO for ASD-other (“synapse”) “Neural process”-related cGOs were enriched in both the ASD macro and ASD microcephaly group, which may suggest that this GO group may be too broad to be specifically linked with any sGO.
“Development”-related GOs were enriched in cGOs correlated with sGOs for neuron and organ morphogenesis in the ASD-macro group (Fig. 4B). However, this enrichment was not observed in the ASD-micro group. Together, this finding could suggest that common variants on development-related genes may co-occur with variants affecting neuron/organ morphogenesis to cause ASD with macrocephaly.

“Synapse”-related GOs were enriched in cGOs for sGO neuron in the ASD-other group, suggesting that the biological process related to neuron and synapse were correlated closer in the ASD-other group. Again, this endorsed the effect on the behavioral rather than brain morphological side of the synapse-related variations.

Surprisingly, the “head/brain”-related GOs were not enriched in cGOs correlated with most sGOs except for sGO synapse in the ASD micro and ASD macro group. Our interpretation was that variations for brain or head morphology changes may not often correlate with genes function on WNT/neuron/organ morphogenesis. Instead, they could function independently or tend to co-occur with variants with other functions such as synapse.

“Transmembrane”-related GOs were enriched in cGOs for sGO neuron, sGO WNT and sGO organ morphogenesis in the ASD-other group, suggesting that the “transmembrane”-related GOs may “amplify” the effect of variants on WNT, neuron and organ morphogenesis-related genes in ASD without brain size changes. “Transmembrane”-related GOs were significantly enriched in cGOs correlated with sGO synapse in both the ASD microcephaly and ASD-other group. Considering that the enrichment of vesicle-related GO terms was enriched in sGO group synapse in the ASD-other group (11 of 43 sGOs for synapse, proportion test, p<2.2e-16, Fig. 4A, Table S3F), the correlation between transmembrane process and vesicle-related synapse function was specific for the ASD-other group. Some examples may further consolidate this interpretation. In Fig. 4E, the positive correlation between sGO Synaptic vesicle cycle and cGO sodium ion transmembrane was specific to the ASD-other group and the correlation in the ASD macro and microcephaly groups was negative.

The enrichment of WNT-related GOs in cGOs correlated with different type of sGOs was not systematically tested but there were examples that endorsed the correlation between WNT signaling and neuron-related GOs in ASD-macro. For example, the sGO “neuron apoptotic process” was positively associated with the cGO “cell cell signaling by WNT” (Fig. 4F), the correlation was positive in both the ASD-macro and ASD-other groups, and the correlation coefficient was much larger in the ASD-macro group, whereas the correlation was negative in the ASD-micro group.

**Variations identified on sGOs from ASD with macrocephaly**

With GOs directly (sGOs) and indirectly associated (cGOs) with ASD or brain size being identified, we next focused on identifying genomic variations important for phenotype in these GOs. Two factors were considered to prioritize the key variations: the number of the genes occurring in each of the sGOs or cGOs and the enrichment of specific gene loci in ASD over control samples. The product of both factors was defined as the “combined score” and given a standardized value (z score). Variation loci were then ranked
based on this “combined enrichment score” and only variations with a z score > 0 were selected.

In sGOs for ASD-macro, 126 variations (26 SV and 100 SNV/INDEL loci) were on genes from the 3 groups of sGOs (GOs related to WNT, GOs related to organ morphogenesis and GOs related to neuron development/function, inner circle in Fig. 5A, left panel). The length of each color bar in both the inner and outer circles was proportional to the number of variants selected in each corresponding sGO.

About 50% (70 of 126) of variant loci were on genes belonging to all 3 types of sGOs (WNT, neuron and organ morphogenesis) for the ASD-macro (Group III in Fig. 5A left panel). On the other hand, more than 30% of variants selected were unique to either neuron- (group VI) or organ morphogenesis-(group I) related sGOs (Fig. 5B). These variants may affect specific biological process in ASD etiology. The number of variants on genes belonging to both WNT- and morphogenesis-related sGOs (group IV) or on genes belonging to both neuron- and morphogenesis-related sGOs (group II) were low (Fig. 5B). The 126 variants on sGOs for ASD-macro affected in total 86 genes. These genes were enriched with genes known to relate to ASD (n=19, proportion test, p=7.146e-15), including GRIN2B, PTEN, SMG6, WNT2B, etc. PTEN was known to be important for macrocephaly (52); another gene on PI3K-AKT pathway (AKT3) further increased the confidence of the variants we found as new biomarkers for ASD macrocephaly.

Additional variations identified using sGO/cGO correlations from ASD macrocephaly

Next, we determined whether identifying cGOs with each of the sGOs would provide further insight into ASD with macrocephaly. In cGOs in ASD-macro group, 1,476 genomic loci (91 SV and 1,385 SNV/INDEL loci) were selected (outer circles in Fig. 5A, left panel).

Among these loci, more than 50% (45 SV and 852 SNV/INDEL) were on cGOs (group 3 in Fig 5C, n=375) correlated with all 3 types of sGOs (WNT, neuron and organ morphogenesis), suggesting that these variants could be “triggered” by variants on genes in all these 3 major biological processes to cause ASD macrocephaly. Of these variants, 136 were on known ASD genes (proportion test, p< 2.2e-16), including NRXN1, RELN, and SEMA5A, whereas 761 variations were on genes not yet related to ASD, including AKAP13, HTT, NAV2 and TET3. These identified variations are promising candidates for biomarkers of ASD with macrocephaly. On the other hand, there was a much smaller number of variants on cGOs correlated with single type of sGOs (Fig. 5C). For example, the number of cGOs correlated with sGO for morphogenesis was large (group 1 in Fig 5C, n=1279), and variations selected from these cGOs were few (n=154, 20 SVs, 134 SNV/INDELs). Similarly, from 528 cGOs correlated with sGO neuron (group 7), only 25 variants passed the filter standard of the combined score. These suggest the combined score, which take both frequency of gene in cGOs and the locus level enrichment in the population into consideration, would largely suppress potential false positive rate in candidate loci detection from cGOs.

One exception was cGOs specifically correlated with sGO for WNT signaling (group 5 in Fig 5C). In 68 cGOs, 105 variations passed the combined score filter. This higher ratio of variants selected per cGO
suggested that genes interacted with WNT signaling tend to be involved in multiple biological processes (GOs) and carry variants enriched in ASD.

Above average ratio of variants selected per cGO was also observed in the 106 cGOs correlated with both WNT and morphogenesis sGOs (group 4 in Fig 5C) and in 81 cGOs correlated with both neuron and WNT sGOs (group 6). 89 variants were selected in each group by combined score. These results together emphasized the variants, not only those on genes belonging to WNT signaling pathway but also those functionally interacting with WNT signaling, were important for ASD with macrocephaly, rendering them good biomarkers.

**High consistency between SFARI-SSC and validation dataset results**

We next determined whether any of our findings from the SFARI-SSC dataset could be replicated. Using the ASD and control fibroblast cell lines from a previous publication (22), we detected SNVs/INDELs using exome sequencing and detected structural variations using matepair sequencing (These variations were called validation dataset hereafter). In detail, there were 333,846 SNV loci (Fig. 6A), corresponding to 837,180 SNV events (Fig. 6B) in 8 ASD and 5 control samples. There were 33,908 INDEL loci (Fig. 6C), corresponding to 69,031 INDEL events (Fig. 6D). Importantly, 80.82% of the 1,924 selected genomic variation loci from SFARI-SSC ASD-macro samples had matches in the validation dataset (Fig. 6G, Table S4A, S6A, S6E).

In total, 761 deletions were detected in the validation dataset (Fig 6E). Exome sequencing data showed exonic regions overlapped with deletions we detected have reduced reads compared with nearby up/downstream regions (Fig. 6F). Of the deletion loci selected using combined score from SFARI-SSC dataset, 16.4% could find matches with these deletions (Fig. 6G). The relatively low matching rate could be accounted for by the lower frequency of deletions compared with SNVs/INDELs in population and would tend to be missed in the small sample size validation dataset (Fig. 6G, Table S8A).

GOs overrepresented by genes enriched with SNVs/INDELs/SVs in ASDs over controls in the validation dataset largely overlapped with GOs overrepresented by genes with ASD-enriched variations in the SFARI-SSC dataset in the ASD macrocephaly group (Fig. 6H, I). This GO pattern similarity confirmed that, although high individual heterogeneity exists at the gene level, the overall genetic machinery of ASD-macrocephaly was conserved at the level of GOs and could be distinguished from those of control individuals and other ASD subgroups. In this sense, the RNASeq, ATAC-Seq and ChIPSeq data collected from the validation dataset could be related to genomic variations selected from the SFARI-SSC dataset with high confidence.

**Interaction among genomic variations, gene expression and transcriptional regulation related to WNT/β-catenin in ASD macrocephaly**

WNT signaling activity was significantly reduced in both mouse and human NPC models of ASD with macrocephaly. For the 1514 loci we selected from the SFARI-SSC ASD-macro group, ChEA analysis
showed that transcription factors (TFs) related to the WNT pathway (e.g., TCF4, SOX2, SMAD4, etc.) were overrepresented (Table S8G).

These SFARI-SSC ASD-macro results overlapped largely with the results from the ChEA analysis on genes overlapped with SV (deletion) specific to ASD samples in validation dataset (Fig. S3, Table S8i). The overlap of ChEA results between SFARI-SSC ASD-other and the ASD-specific SV was much smaller (Fig. S3). Combined, these findings show that genomic variations affecting WNT signaling pathway-regulated transcriptional activity could contribute to the etiology of ASD with macrocephaly.

To further explore the effect of WNT-related transcription regulation on ASD with macrocephaly, we used publicly available ChIPSeq b-catenin binding data in human embryonic stem cells (hESCs; 49), since b-catenin is the transcriptional effector of the canonical WNT pathway. 11,621 binding peaks distributed among various genomic regions (Fig. 7A) that bound a total 4,036 genes that were active in NPCs by ATAC-seq (Fig. 7B). A total of 227 b-catenin-bound genes overlapped with selected genomic variation loci from SFARI-SSC ASD-macro samples (Fig. 7B), which was much more frequent than the overlap (n=74) with genes carrying loci selected from SFARI-SSC ASD-other samples (Fig. 7C). These results support the relationship between WNT/β-catenin signaling and ASD with macrocephaly.

We next examined whether any of the identified genes in the ASD loci displayed differences in gene expression in the NPCs from the iPSCs (define) from 8 ASD and 5 control lines used in our validation dataset. Consistent with previous studies with these cell lines (22, 53), the number of differentially expressed genes between ASD and control NPCs was small (n=191, Table S5A). Among them, 40 overlapped with genomic variations selected from the SFARI-SSC ASD-macro samples (Fig. 7B) and 8 overlapped with SFARI-SSC ASD-other samples (Fig. 7C). GO analysis of the 191 DE genes did not yield any significant results. When we compared expression levels in each of the NPC lines with the average expression level of all 4 control lines, an additional 664 “individualized DE genes” were detected (Fig 7C, D, Table S5A).

The single gene identified that was differentially expressed, bound by b-catenin and was a known ASD gene was SMG6 (Fig. 7B). An SNV in the second exon of SMG6 (SMG6, Chr17:2203025, T->G) was enriched in both ASD macro (by 17.5%) and ASD microcephaly probands (by 12.4%), but not in ASD-others from SFARI-SSC. This gene is a known ASD gene (54) and mainly functions in nonsense-mediated mRNA decay. This locus was also found in one ASD-macro proband (ARCH) from the validation dataset. This locus is close to the TSS of SMG6 (Fig. 7D), which increased the possibility that this variation would affect TF binding. The b-catenin ChIPSeq peaks overlapped with SMG6 but not with this locus, so the function of this locus is unclear. We speculate that this locus could affect the interaction between b-catenin binding to the intron region of SMG6 and some other TFs that bind to TSS of SMG6. The RNASeq results demonstrated a significant decrease in SMG6 in the ASD NPC line (ARCH) compared with control lines.

One of the five genes that were differentially expressed and bound by b-catenin but not identified as known ASD gene, GPR39, contains an ASD-macro enriched SNV (Chr2:133174999, A->G) (Fig. 7B, 7E).
This gene encodes a rhodopsin-type G-protein-coupled receptor (GPCR) and is related to the pathophysiology of depression (55). It contains a binding target of β-catenin, and this variation (Chr2:133174999, A->G) was in the ChIP binding area close to the TSS, which suggested that it could affect the binding of β-catenin to GPR39. The frequency of this allele in ASD in the SFARI-SSC ASD-macro samples was 18% higher than in control (Table S4A). This locus also displayed a 30% higher frequency in ASD than control in the validation dataset (Table S6A). RNASeq results demonstrated that GPR39 was downregulated in ASD (Fig. 7E). These findings combined suggest that this SNV in GPR39 may be a new biomarker for ASD with macrocephaly.

A second example within the five genes identified that was differentially expressed, bound by β-catenin and was not a known ASD gene, was AKAP13, which contains an SNV (chr15:86284342, C->T, Fig. 7B,7C, 7F). This gene encodes an A-kinase anchor protein, which is functionally related to both GPCR signaling and mTOR signaling (56). This SNV demonstrated a higher alternative allele frequency in ASD subjects than in controls in both the SFARI-SSC ASD-macro and SFARI-SSC ASD-micro samples (Table S6A). By contrast, the SFARI-SSC ASD-other samples did not demonstrate enrichment for this locus, suggesting that this locus might contribute to the brain size changes associated with ASD. The expression of AKAP13 was upregulated in one ASD NPC line (ABLE) but downregulated in another ASD sample (ARCH), which may be attributed to the stop gain mutation on CTNNB1 (coding for β-catenin) in the ARCH line. Considering the binding of β-catenin to both TSS and TTS region of AKAP13, this SNV loci could play a role linking WNT signaling and mTOR signaling to regulate brain development in the context of ASD.

Finally, we found TRIM2, which was bound by β-catenin and not a known ASD gene, overlapped with a deletion (chr4:154125517-154260572) that was enriched in ASD-macro samples (Fig.7B, Fig.S3B). This gene plays a neuroprotective role and functions as an E3-ubiquitin ligase in proteasome-mediated degradation of target proteins. This ASD-macro enriched deletion covered the middle part to the 3’ end of the TRIM2. This deletion overlapped with both the ATAC-Seq peak and β-catenin binding peak within a HiC interval, and this interval looped with another HiC interval covering the 5’ end of the same gene. These results suggested that TRIM2 may be the transcriptional target of β-catenin, and the deletion (chr4:154125517-154260572) may be an important biomarker for ASD with macrocephaly. However, we did not find this deletion in our small validation dataset, so the functional effect of this deletion (e.g., effect on TRIM2 gene expression) could not be directly tested.

[MOU1]THIS IS NOT CORRECT AND PROBABLY NEEDS TO BE DELETED.

[CF2]corrected

[CF3]

[MOU4]The combined score needs to be defined.

[CF5]May the previous sentence explain the term?

[MOU6]YES, thanks


### Discussion

#### Overall significance

ASD is characterized by phenotypic and genetic heterogeneity. The different comorbidities of ASD patients have been used to divide ASD patients into different phenotypic subgroups that may also reduce the genetic heterogeneity to facilitate the clarification of ASD genetic mechanism. However, one challenge of this approach is how to tease apart the genetic factors that mainly contribute to the comorbidities (macrocephaly, aggressive behavior, seizure, etc.) from those that mainly contribute to the defining behavioral abnormalities found in ASD. Here, we developed a new quantitative approach to detect the co-occurring genomic variations important for the development of ASD with and without macrocephaly, an important co-morbidity in approximately 20-25% of ASD individuals (3). In detail, 160 genomic variations were identified through GOs directly associated with ASD and brain size phenotype, 1,565 genomic variations were identified through GO-GO correlations, and 104 of these 1,715 variations were identified by both. These genes and variations, especially their combinatorial patterns, may provide novel biomarkers for different ASD subtypes. This pipeline could also be applied to analysis of other -omic data types, such as RNAseq, ChIPseq and ATACseq.

#### Dissection of ASD and brain size phenotype

In the current study, 922 GOs associated with ASD-only probands were identified. Synapse-related GOs were enriched in these GOs, which suggested that dysregulation at the synapse level might be specifically associated with behavioral abnormalities in ASD independent of brain size. For synapse-related sGOs, “transmembrane process” related GOs were significantly enriched. As transmembrane process was often functionally related to synapse development and function (e.g., synaptic vesicle recycling), this observation further supports the importance of synapse dysfunction in behavioral changes diagnostic for ASD (57).

On the other hand, WNT-, neuron- and organ morphogenesis-related GOs were enriched in 777 GOs associated with ASD-macro. This result is consistent with previous findings that increased neuron numbers in ASD with macrocephaly probands (3) and decreased WNT signaling pathway activity (22) are important for ASD with macrocephaly. For sGOs of WNT, neuron and organ morphogenesis, both ASD macro and micro probands showed enrichment of cell cycle-related cGOs, which is consistent with the conclusion based on gene expression change in blood from ASD with macrocephaly probands (21). We validated the importance of the Wnt pathway correlation by performing RNA-seq, ChIP-seq for b-catenin and ATAC-seq on the ASD and control lines used as the validation dataset, which led to the identification of SMG6 as a Wnt-regulated ASD gene (Fig. 7D).

#### New biomarkers for ASD with macrocephaly

We identified 6 genes (AKAP13, BSG, DNAH11, GPR39, SMG6, SUMF1) as potential biomarkers for ASD with macrocephaly, as they were supported by our gWGCNA pipeline computation results from the SFARI-
SSC dataset, β-catenin ChIPSeq and differential expression data from the validation dataset. Among these 6 genes, SMG6 had already been reported to be related to the ASD gene (Nguyen et al, 2013), which increased the confidence that this gene was an ASD-macro biomarker. The other 5 genes were not previously linked to ASD, and our results suggest they could be new biomarkers for ASD-macro and require independent replication. In ASD-other samples from SFARI-SSC, AKAP13 was selected by our pipeline, confirmed by ChIPSeq and differential expression data from validation dataset. Considering it is one of the 6 candidates found in ASD-macro samples, it may be a new biomarker for ASD, not necessarily limited to ASD with macrocephaly.

We noticed a sharp decrease in the number of candidate loci selected by our gWGCNA pipeline and the number of loci validated by RNASeq, ChIPSeq data. Two factors could account for this decrease: first, the purpose of the gWGCNA pipeline was to determine the co-occurrence of genomic variations, and the expression level changes for these co-occurred variations in ASD NPC lines would be more complex than simply differential gene expression from control individuals at the single gene level. Second, we only have 8 ASD-macro and 5 control cell lines in the validation dataset, which was more than 5-fold fewer than the SFARI-SSC dataset sample size. Some low frequency variations could not be found in the validation dataset and, more importantly, many co-occurrence patterns could not be found in the rather small dataset. In the future, when larger expression/ChIPSeq datasets with genomic/clinical information available can be used as validation datasets, we believe a much higher proportion of our co-expression results could be determined and validated.

Cell cycle-related genes may affect both macro- and microcephaly

We identified that cell cycle-related GOs were enriched in cGOs correlated with sGOs for WNT, neuron, organ morphogenesis in ASD-macro and ASD-micro probands compared with ASD-other probands (Fig. 4B, Table S3G,3I-J). This finding suggested that cell cycle is a biological process dysregulated in ASD micro/macro probands, potentially affecting neural precursor cell/neuron proliferation/differentiation and altered neuron number in the brain, which is not surprising in and of itself. However, the correlation between WNT-related sGOs and cGOs for cell cycle suggested that variations in WNT signaling genes may be upstream of variations in cell cycle-related genes in ASD with macrocephaly genetic architecture, a novel observation. Similarly, variations in genes from sGOs for neuron and organ morphogenesis may be downstream of cell cycle to affect specific neuronal/developmental functions in ASD macro/microcephaly. In the 88 GOs associated with ASD microcephaly, WNT-related GOs were not overrepresented (Fig.4A) but, using sGOs identified in ASD with macrocephaly, cell cycle-related cGOs were enriched in the ASD microcephaly samples (Fig.4B). This result suggests that a WNT-cell cycle correlation exists in ASD with microcephaly, but this correlation is not as strong as that in ASD with macrocephaly. Importantly, this finding suggests that WNT signaling may be one of the upstream factors for cell cycle change in ASD microcephaly but is probably not the major factor.

WNT signaling for ASD with macrocephaly
Previous publications showed \(\beta\)-catenin/BRN2 transcriptional activity was decreased in human (22) and mouse models (58) for ASD with macrocephaly. Also, the gene expression level for WNT-related genes were downregulated in ASD with macrocephaly postmortem brain samples (59). Our results provide genomic level evidence of the dysregulation of WNT signaling in the context of ASD with macrocephaly. More importantly, our results elucidated how genomic variations in the WNT signaling pathway interact with other genomic variations in the context of ASD with macrocephaly. In detail, 62 WNT-related variations were identified on genes from sGOs, including 14 known ASD genes, such as \textit{DLG1}, \textit{GRIN2B}, \textit{WNT2}, \textit{GPC6}, \textit{TNN}, etc. The other 48 genes may be potential new candidate loci for ASD with macrocephaly, including variations on \textit{HES1}, \textit{FZD3}, \textit{DKK1}, \textit{GLI2}, \textit{PRKN}, \textit{AKT3}, etc. Considering their potential effect on transcriptional regulation (Fig. 7B, Table S4A), \textit{TNN} and \textit{GLI2} (Fig.5B, Fig. 7B) are good candidate biomarkers for ASD with macrocephaly. These loci were correlated with 1,105 variations on cGOs (Table S4A), affecting genes including \textit{AKAP13}, \textit{ESR1}, \textit{CNTNAP2}, \textit{GPR39}, \textit{SEMA5A}, etc. These results provide an example of how a small number of variations within WNT signaling co-occurred with a large number of variations on genes with a higher variation rate (e.g., neuron development) to cause complex disease.

**Application of gWGCNA pipeline to analysis of RNASeq and ChIPSeq data**

We have shown that the gWGCNA pipeline could integrate multiple types of genomic variations (SNV, INDEL, SV). This pipeline could also be applied to analysis of RNASeq and ChIPSeq data. It would be a promising next step to analyze HiC data with the gWGCNA pipeline in the context of complex human disease. The correlated GO (or gene, variation) pairs identified by gWGCNA and looping patterns detected by HiC could show how physically interacting DNA regions and functionally related genes interplay to affect disease etiology. We believe the integration of different types of -omic level data would generate a more complete picture of ASD genetics.

**Declarations**

**Acknowledgments:**

This work was supported by NIH R01MH113106 to A.W.B. and the JPB Foundation to FHG.

**Conflict of Interest**

The authors have no conflicts of interest to declare.

**References**


Figures

A

B

C

D
Figure 1

SNV, INDEL and deletions in ASD with macrocephaly and microcephaly from the SFARI Simplex Simons Collection (SFARI-SSC) dataset.

(A) ASD with macrocephaly (red dots) and microcephaly (blue dots) individuals with head circumference 2 S.D greater than or 2 S.D. less than the mean were selected from 2,760 ASD probands in the SSC dataset, whereas ASD with no brain size phenotype (“ASD-other”, black dots) were selected from the individuals with head circumference closest to the mean. All other ASD probands were plotted as green dots. (B) Head circumferences of probands with macrocephaly were significantly larger than their siblings (p=5.8E-07). (C) The top 15 significant GOs for genes that displayed significant enrichment for probands over siblings based on exome sequencing (SNV and INDEL) data were selected. Adjusted p value for top 15 significant GOs from each of ASD-macro, ASD-micro and ASD-others list was plotted from most to least significant for ASD-macro from top to bottom. The color scale was proportional to the -log transformed adjust p value. (D) GOs for genes intersected with deletions found in probands but not control individuals. Adjusted p value for top 20 GOs from each of ASD-macro, ASD-micro and ASD-other list was plotted. The color scale was proportional to the -log transformed adjust p value.
Find GO modules associated with ASD phenotype using WGCNA

(A)–(B) The GO dendrograms from WGCNA based on SNV/INDEL/SV data mapped to GOs from SFARI-SSC families in (A) group 1 and (B) group 2. (C–D) GO modules associated with ASD or brain size phenotype in (C) group 1 and (D) group 2. Pearson correlation (r) and p value between phenotype (“ASD”
or “Brain”) and eigenvalue of each GO module was shown in each grid, with the density of the red color representing positive correlation and the density of the green color representing negative correlation).

![GO modules diagram](Image)

**Figure 3**

**GOs associated with ASD or brain size phenotype detected by the gWGCNA pipeline.** The GOs within each of the significant modules associated with ASD and/or brain size in Figure 2 were collected. The numbers
of GOs detected in group 1 (blue circle), group 2 yellow circle), and overlap between the two groups (in red color) are displayed in a Venn diagram. Below each Venn diagram, the number of GOs in the GO module and representative GOs were listed. (A) GO modules that displayed significant positive association with “ASD” and insignificant association with “brain” were considered as “ASD-only”. (B) GO modules that displayed significant positive association with “ASD” and significant positive association with “brain” were considered “ASD-macro”. (C) GO modules that displayed insignificant association with “ASD” and significant positive association with “brain” were considered “macro-only”. (D) GO modules that displayed significant association with “ASD” and significant negative association with “brain” were considered “ASD-micro”. (E) All GO modules that displayed significant association with neither “ASD” nor “brain” were considered as “non-significant”.
Figure 4

**GO groups enriched in seeding GOs and correlated GOs.** (A) The percentage of GOs related to specific functions (GO groups) were calculated for each of the 5 types of GO modules as described in Fig.3 A-E. GO significantly enriched GO groups were indicated with associated p values from the proportion test on the top of the bar. GO groups found to be significant from the analysis of CNVs from previously published results (11) were within the green shade. Other GO groups, such as “head/brain,” “cell cycle,” “neuron,”
“synapse,” and “wnt” were selected based on our previously published data (22). (B) The percentage of selected GO groups (cGOs) for each of the 4 seeding GOs significant in (A) (WNT, neuron, organ morphogenesis and synapse). Red bars represent the percentage of each of the GO groups from ASD-macro, blue bars represent the percentage of ASD microcephaly and green bars represent the percentage in ASD-other. Significance levels on top of the bars were suggested by asterisks (*: p<0.05; **: p<0.01; ***: p<0.001). (C-F) Example of GO-GO correlation. Genomic variation rate for each sGO was plotted on the X axis and for each cGO was on the Y axis. Linear regression equation (y) and variance explained (R2) was on top of the regression line. Each dot represents one proband (ASD-macro in red, ASD microcephaly in blue and ASD-other in green). (C) GO central nervous system neuron differentiation vs. GO cell cycle. (D) GO sensory organ morphogenesis vs. GO cell cycle phase transition. (E) GO synaptic vesicle cycle vs. GO sodium ion transmembrane transport. (F) GO neuron apoptotic process vs. GO cell-cell signaling by WNT.
Figure 5

Genomic variations enriched in seeding and correlated GOs. (A-C) Genomic variations detected in ASD-macro samples. (A) Summary of genomic variants on genes belonging to GOs directly associated with ASD phenotype (seeding GOs (sGOs), inner circle) and on genes belonging to correlated GOs (cGOs, outer circle). (B–C) Summary of genomic variants on sGOs (B) and cGOs (C). Number of SV, SNV/INDELs on
Summary of genomic variations detected in ASD-other samples. (D)

Figure 6

**SNVs INDELs and SVs in validation dataset. (A-D)** Venn diagrams of number of loci detected in the validation dataset for SV events (A), Blue: ASDs, Red: Controls. SNV events (B), INDEL loci (C), and INDEL
events (D). Blue: ASDs, Red: Controls. (E) SV (Deletions) detected in 8 ASD with macrocephaly and 2 control individuals. Horizontal bar beside each sample name represents the total number of deletions detected in that sample. Control samples were in red and ASD samples in blue. Each row of the dot matrix corresponds to one sample and each column corresponds to one set of deletion loci with the same distribution pattern among these samples. Each vertical histogram bar represents the recurrence of a specific distribution pattern. Also, the recurrence number for each combination was labeled on top of the histogram bar. (F) Z-score of reading depth of exonic regions overlapped with deletions in ASD samples (red) demonstrated a significant reduction compared with genomic regions 10 KB upstream (blue) and downstream (green) of the deletion event. (G) The proportion of selected genomic loci from the SFARI-SSC dataset (as in Fig.5B-D, nASD_macro =1514, n asd_other=410) that overlapped with the validation dataset. (H) GOs for genes with SNV/INDELs enriched in ASD from the validation dataset corresponding to top 15 GOs for each SFARI-SSC group (as in Fig. 1C). Color scale was proportional to -log transformed adjusted p value. The blue rectangle depicted SFARI-SSC results; the purple rectangle depicted validation results. (I) GOs for genes with ASD-specific SVs from validation dataset corresponding to top 20 GOs for each SFARI-SSC group (as in Fig. 1D).
Figure 7

Possible effect of selected loci on gene expression and β-catenin transcription regulation. (A) Functional annotation of β-catenin binding peaks. (B) Selected genomic loci from SFARI-SSC ASD-macro samples that overlapped with β-catenin targets, differential expressed (DE) genes and known ASD genes. Genes with loci selected by gWGCNA pipeline, showing differential expression and bound by β-catenin, were listed below the Venn diagram. SMG6, which meets all these criteria and is a known ASD gene, was
labeled in the Venn diagram. Names of loci selected to be plotted individually were bold. (C) Selected genomic loci from SFARI-SSC ASD-other samples overlapped with β-catenin targets, DE genes and known ASD genes. (D) Example plot for SNV (Chr17:2203025, T->G) on SMG6, black tracks: β-catenin ChIPSeq (upper: inactivated, lower: activated by WNT3A); orange tracks: ATACseq data; purple track: H3K27Ac ChIPSeq from UCSC. Right panel: RPKM for ASD line with this variation (ARCH) and CTRL NPC lines. (E) SNV (Chr2:133174999, A->G) on GPR39. Right panel: RPKM for ASD and CTRL NPC lines. (F) SNV (chr15:86284342, C->T) on AKAP13. Right panel: RPKM for ASD line with this variation (ABLE and ARCH) and CTRL NPC lines. Significance level of difference in gene expression between ASD and CTRL group was indicated using asterisks on top of the bar (*: p<0.05, **: p<0.01, ***: p<0.001).

Supplementary Files

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

- supplementarylegend09041.pdf
- supplementarylegend09042.pdf
- supplementarylegend09043.pdf
- supplementarylegend09044.pdf
- supplementarylegend09045.pdf
- supplementarylegend0904.docx
- supplementarytable1.xlsx
- supplementarytable2.xlsx
- Supplementarytable3.xlsx
- supplementarytable4.xlsx
- supplementarytable5.xlsx
- supplementarytable6.xlsx
- supplementarytable7.xlsx
- supplementarytable8.xlsx