Title: Calculating the Effects of Autism Risk Gene Variants on Dysfunction of Biological Processes Identifies Clinically-Useful Information

Authors: Olivia J. Veatch[1*], Diego R. Mazzotti[1], James S. Sutcliffe[2], Robert T. Schultz[3], Ted Abel[4], Birkan Tunc[3], Susan G. Assouline[5], Edward S. Brodkin[6], Jacob J. Michaelson[7], Thomas Nickl-Jockschat[7], Zachary E. Warren[8], Beth A. Malow[9], Allan I. Pack[1]

Affiliations:

[1]Center for Sleep and Circadian Neurobiology, Perelman School of Medicine, University of Pennsylvania.

[2]Vanderbilt Genetics Institute, Department of Molecular Physiology and Biophysics, Vanderbilt University Medical Center.

[3]Center for Autism Research, Children's Hospital of Philadelphia.

[4]Iowa Neuroscience Institute, University of Iowa.

[5]Belin-Blank Center for Gifted Education and Talent Development, University of Iowa.

[6]Department of Psychiatry, Perelman School of Medicine at the University of Pennsylvania.

[7]Department of Psychiatry, University of Iowa.

[8]Department of Pediatrics, Vanderbilt University Medical Center.

[9]Department of Neurology, Vanderbilt University Medical Center.

*To whom correspondence should be addressed: veatcho@upenn.edu.

One Sentence Summary: A novel approach we developed to interrogate previously reported risk genes for autism identified pharmacogenetics information that is clinically-relevant.

Abstract

Autism spectrum disorders (ASD) are neurodevelopmental conditions that are influenced by genetic factors and encompass a wide-range and severity of symptoms. The details of how genetic variation contributes to variable symptomatology are unclear, creating a major challenge for translating vast amounts of data into clinically-useful information. To determine if variation in ASD risk genes correlates with symptomatology differences among individuals with ASD, thus informing treatment, we developed an approach to calculate the likelihood of genetic dysfunction in Gene Ontology-defined biological processes that have significant overrepresentation of known risk genes. Using whole-exome sequence data from 2,381 individuals with ASD included in the Simons Simplex Collection, we identified likely damaging variants and conducted a clustering analysis to define subgroups based on scores reflecting genetic dysfunction in each process of interest to ASD etiology. Dysfunction in cognition-related genes distinguished a distinct subset of individuals with increased social deficits, lower IQs, and reduced adaptive behaviors when compared to individuals with no evidence of cognition-related gene dysfunction. In particular, a stop-gain variant in the pharmacogene encoding cycloxygenase-2 was associated with having an IQ<70 (i.e. intellectual disability), a key comorbidity in ASD. We expect that screening genes involved in cognition for deleterious variants in ASD cases may be useful for identifying clinically-informative factors that should be prioritized for functional follow-up. This has implications in designing more comprehensive genetic testing panels and may help provide the basis for more informed treatment in ASD.

Introduction

Autism spectrum disorders (ASD) are a group of neurodevelopmental conditions characterized by core symptoms that include impairments in social interactions, delays in language development and expression of repetitive interests and/or behaviors(*1*). ASDs manifest along a wide distribution of core symptom severity, and numerous different comorbidities are highly prevalent [e.g., intellectual disability(*2*), gastrointestinal issues(*3*)]. Evidence supports contributions from different types of common and rare genetic variation – including inherited and *de novo* single nucleotide variants (SNVs), small insertions or deletions (In/Dels), and large insertions or deletions (CNVs) – in hundreds of genes(*4*, *5*). The already large, and rapidly expanding, landscape of genetic factors involved in expression of ASD makes it difficult to determine how results from genetic studies can translate into clinically-useful information(*6*-*8*). A crucial step toward using genetics to inform more effective, personalized approaches for treatment of individuals with ASD is to better understand how variation in implicated genes influences expression of core symptoms and comorbidities.

While there are more than one hundred implicated genes, many function in the same biological process(*9*, *10*). Dysfunction in genetic mechanisms encoding different biological functions may contribute independently to increase risk for ASD. For example, one study observed that a subset of individuals with ASD had *de novo* and rare, inherited variants in synaptic genes but not chromatin modification genes, while another subset had these types of variants in chromatin modification genes but not synaptic genes(*11*). If some individuals with ASD have dysfunction in a particular biological process while others have dysfunction in a separate process, then it may be possible to use genetic data to inform a more personalized (i.e., precision medicine) approach to treatment of symptoms. However, the study mentioned above,

and others, have not observed a relationship between genetic and phenotypic differences(*11, 12*). As such, it is difficult to determine if distinguishing dysfunction across different underlying biological processes is clinically useful. Notably, previous studies have focused largely on evaluating contributions from specific *types* of genetic variants (e.g., solely *de novo* and rare variants, or common variants) to explain phenotypic differences(*11-16*). A more holistic approach that incorporates all relevant risk variation is better situated to ask how overall genetic risk is related to particular symptom profiles that are unique to the individual(*17, 18*).

Furthermore, to enable use of disparate genetic information in personalized medicine approaches for ASD, ability to predict functional effects of a given variant on the ASD risk gene and encoded protein is essential and may require functional analysis to test(*19*). While functional study of every suspected ASD risk variant is desirable in the long-term, reliance on such a strategy is not feasible if genetic findings are to be rapidly translated in the clinic. It may be more immediately useful to have computational approaches which incorporate evidence from multiple sources to allow for more thorough *in silico* predictions from patient data to help pinpoint specific genes and variants that should be prioritized for functional follow-up(*20-22*).

To determine if current genetic evidence could help explain variability in ASD symptoms, and ultimately inform treatment approaches, we developed an approach to calculate the likelihood that a biological process with overrepresentation of ASD candidate genes is dysfunctional. We evaluated the approach using whole-exome sequencing and phenotype data from the Simons Simplex Collection (SSC)(*23*). We hypothesized that incorporating evidence from all possible types of genetic variation to calculate cumulative risk of dysfunction overall in biological processes would identify underlying mechanisms contributing to differences in symptomatology among individuals with ASD. We also expected that careful evaluation of the

current genetic evidence would be useful to recognizing ASD-related variants that are already clinically-actionable as many individuals carry pharmacogenetics variants which influence how a patient responds to a drug(*24*).

Materials and Methods

Identification of Genetic Mechanisms Relevant to ASD

To assess the influences of predicted dysfunction in overall biological processes, we compiled a list of ASD candidate genes using the Autism Informatics Portal (AutDB, http://autism.mindspec.org/autdb/Welcome.do)(*25*), which is continuously updated with manual annotations as new scientific literature is published. As the goal was to determine if any genes evidenced to have a relationship with ASD were useful to understanding symptom variability and informing personalized treatment approaches, all genes were considered regardless of the strength of evidence supporting an association with ASD (December 2017 update). Official Gene Symbols were converted to Ensembl IDs using the Gene ID Conversion Tool available in the database for annotation, visualization and integrated discovery (DAVID)(*26*). Ensembl IDs for a subset of these genes could not be converted via DAVID and were manually identified by searching the Ensembl database. Gene set overrepresentation analyses were run on all candidate genes for ASD using the classic algorithm and Fisher's exact test from the TopGO package in R(*27*). Overrepresented processes were interrogated to identify terms representing processes useful to ASD etiology ('unique terms'; *Table S1*). Processes were considered biologically meaningful unique terms if they represented the initial process in each GO hierarchy that was system-, organ-, tissue-, or organelle-specific (e.g., 'GO:0007399=nervous system development'). GO term definitions were based on AmiGO version 1.8, GO version 2018-01-01.

Calculation of Overall Biological Process Dysfunction

Variants identified using whole-exome sequencing available for a total of 2,392 individuals with ASD whose data were included in the Simons Simplex Collection (SSC) dataset(*23*) are provided by the Simons Foundation Autism Research Initiative and WuXi NextCODE: A Contract Genomics Organization (https://www.wuxinextcode.com/). The SSC represents the largest collection of simplex autism families, with one affected child and at least one unaffected sibling, collected to date(*23*). Data are made available to approved researchers via the Sequence Miner Tool 5.24.7. Gender discrepancies were first identified using the 'Sex Check' report builder in Sequence Miner. This algorithm evaluates both the ratio of heterozygous SNPS on the X chromosome compared to autosomes and coverage of the Y chromosome gene, *SRY*. Seven individuals with unclear gender assignments, 2 individuals with 47,XYY and one individual with 47,XXX were excluded from analyses. Genome-wide genotyping and whole-exome sequence data for all but one individual in the evaluated dataset (n=2,381) was previously interrogated to identify *de novo* and rare, inherited copy number variants (CNVs)(*11*, *28*). The final analysis dataset included 2,381 individuals who were 4-18 years old at the time of data collection. The dataset was 86% male and 79% parent-reported white (*Table S2*).

Variation Annotation queries were performed in Sequence Miner (*29*) to identify single nucleotide variants (SNVs) and short insertions or deletions (<200bp; In/Dels) located in protein coding gene transcripts that had Variant Effect Predictor(*29*) consequences that were highly likely to be damaging to the encoded protein product (i.e., splice site alterations, gains or losses of stop codons, loss of start codons, or frameshifts). We considered variants were called by either the Genome Analysis Toolkit (GATK)(*30*) or FreeBayes(*31*) software across all 22 autosomes and both sex chromosomes. Quality Control thresholds included depth ≥8 reads and genotype

quality for variant calls of ≥20(*32*). Variants flagged as 'LowQuality' as indicated by GT Filter criteria were excluded.

The final list of variants passing QC, that were predicted by VEP to be very likely to be damaging, were interrogated to identify those located in transcripts for ASD risk genes (included the Autism Informatics Portal) that were protein coding (*Table S3*). Notably, the Sequence Miner platform reports Ensembl IDs for each gene in the query output. While these were used to ensure the appropriate VEP predictions and help search the current Ensembl database, some of the Ensembl IDs provided with this platform were outdated and are now represented by new IDs. As such, both Ensembl IDs and gene names were cross-referenced to compare those provided by Sequence Miner and those mapped using DAVID and manual searches. Discrepancies were further interrogated to verify that the VEP prediction was not based on a variant location in an alternate transcript that is not supported by evidence in Ensembl.

There is substantial variability in pathogenicity predictions depending on the algorithm employed (e.g., based on variant location, evolutionary conservation, protein structure/function)(*21, 33*). Therefore, to more completely assess the likelihood of a variant being damaging and ultimately resulting in a dysfunctional protein product, nine different variant prediction algorithms were run on all of the variants pulled from Sequence Miner using filter-based annotation from ANNOtate VARiation (ANNOVAR) software(*34*). *In silico* prediction algorithms included: 1) Sorts Intolerant From Tolerant (SIFT)(*35*), 2) Polymorphism Phenotyping v2 (Polyphen-2) HVAR(*36*), 3) Mutation Taster(*37*), 4) Mutation Assessor(*38*), 5) Likelihood Ratio Test (LRT)(*39*), 6) FATHMM-MKL(*40*), 7) PROVEAN, 8) MetaLR(*41*), and 9) Mendelian Clinically Applicable Pathogenicity (M-CAP)(*42*). Genomic locations of variants available from Sequence Miner are based on Human Genome Build GRCh37/hg19; all analyses

were conducted based on these genomic locations. Each prediction algorithm uses different nomenclature to denote variant predictions. To allow for cross-comparison of results from different predictors, scores were recoded as either benign (B), damaging (D), or unknown (U) as follows: SIFT: damaging (D)=D, tolerant (T)=B; Polyphen2 HVAR probably damaging (D)=D, possibly damaging (P)=U, benign (B)=B; LRT: deleterious(D)=D, unknown(U)=U, neutral(N)=B, Mutation Taster: disease causing automatic(A)=D, disease causing(D)=D, polymorphism(N)=B, polymorphism automatic(P)=B; Mutation Assessor: predicted functional (H, M)=D, predicted non-functional (L, N)=B; FATHMM-MKL: damaging(D)=D, neutral(N)=B; PROVEAN: deleterious(D)=D, neutral(N)=B; MetaLR: damaging(D)=D, tolerant(T)=B; M-CAP: pathogenic(D)=D, benign(T)=B.

We developed the following equation to calculate the likelihood that a variant was damaging to the function of the encoded protein product:

$$LDV = CR \times FD \times Z$$

Where $\boldsymbol{LDV}$ = the likelihood that the variant is damaging; $\boldsymbol{CR}$ = the number of variant callers that called the variant (based on GATK and FreeBayes software); $\boldsymbol{FD} = \left(\left(\boldsymbol{D} - \boldsymbol{B}\right) + 1\right)\right) / \left(\boldsymbol{N} + 1\right)$ where $\boldsymbol{D}$ = the number of *in silico* prediction algorithms that called the variant damaging, $\boldsymbol{B}$ = the number of algorithms that called the variant benign, $\boldsymbol{N}$ = the total number of algorithms that provided a prediction for the variant, and 1 = a constant to account for the fact that variants were preselected according to variant effect predictions indicating a high potential to be deleterious to at least one gene transcript based on the genetic location; and $\boldsymbol{Z}$ = zygosity where heterozygous calls=1 and homozygous calls=2. To reduce the likelihood of false positive calls overly influencing genetic risk scores, variants were weighted such that if only one variant caller recognized the base pair alteration compared to reference $\boldsymbol{CR} = 0.5$. If the variant was called by

both the GATK and FreeBayes callers $\boldsymbol{CR}=1$. $\boldsymbol{FD}$ scores ranged from -0.8-1.0; however, as the goal was to identify variants that were more likely to be deleterious, all negative scores were recoded to zero. Regarding zygosity for sex chromosomes, as it is difficult to determine which X chromosome is inactivated using the data available, female individuals with heterozygous variants on the X chromosome were weighted the same as autosomal variants. In addition, for X chromosome variants called as heterozygous in males, those located within Pseudoautosomal Regions (PAR) were weighted the same as autosomal variants. Male heterozygous X chromosome variants located outside of PAR1 and PAR2 were considered homozygous and were weighted as such in genetic risk scores.

Hg19 genomic locations of rare, inherited and validated, *de novo* CNVs previously reported in Sanders et al., 2015(*11*) and Krumm et al., 2015(*28*) that encompassed coding and regulatory regions of protein coding transcripts for ASD candidate genes were pulled from supplemental data included in these publications. Bedtools(*43*) was used to identify regions of overlap between CNVs reported across the previously published studies. Gene-based annotations in ANNOVAR were used to identify CNVs that encompassed portions of the coding (i.e., exonic, splice-site) and proximal promoter (i.e., 5'-UTR) regions (*Tables S4-S5*). CNVs were given weights equal to SNVs and In/Dels with the strongest likelihood of being damaging based on the distribution of $\boldsymbol{FD}$ scores described above and variant weights for all CNVs were set equal to 1. The currently published data for CNVs report only the presence of a deletion or duplication in a particular genomic region but not the predicted number of copies; however deletions were expected to occur on only one chromosome(*11, 28*). Deletions and amplifications were assumed to occur on only one chromosome. In addition, while some CNVs were not reported for the same evaluated individual, the analysis datasets across the two prior publications

did not completely overlap. Therefore, whether or not both studies reported the CNV was not included in variant weights.

Separate genetic risk scores were then calculated for each individual to assess the likelihood of dysfunction in overall genetic mechanisms that represented unique GO-defined biological processes with overrepresentation of ASD candidate genes. We developed the following equation to calculate the likelihood of genetic dysfunction in biological processes:

$$\boldsymbol{DBP_X} = \sum \left(\begin{array}{l} \left(\boldsymbol{LDV_{v1}^{GeneA}} \times \boldsymbol{EBP_X^{GeneA}} \right) + \left(\boldsymbol{LDV_{v2}^{GeneA}} \times \boldsymbol{EBP_X^{GeneA}} \right) + \left(\boldsymbol{LDV_{v1}^{GeneB}} \times \boldsymbol{EBP_X^{GeneB}} \right) + \\ \cdots \left(\boldsymbol{LDV_{v1}^{GeneZ}} \times \boldsymbol{EBP_X^{GeneZ}} \right) \end{array} \right) / \boldsymbol{nvBPx}$$

Where $\boldsymbol{DBP_X}$ = Dysfunction of Biological Process X and is the sum of the products of $\boldsymbol{LDV_{vn}^{GeneA}}$ = the likelihood that variant $\boldsymbol{n}$ is damaging to gene A, and $\boldsymbol{EBP_X^{GeneA}}$ = the sum of the frequencies of the GO evidence codes, across all genes assigned to biological process X, that were used to assign gene A to biological process X (*Fig. S1*) plus the number of assigned child terms for biological process X, divided by the total number of child terms available for biological process X. $\boldsymbol{nvBPx}$ = the number of variants assigned to biological process X. We expected that a gene having more than one likely damaging variant was increased evidence that the encoded protein product was dysfunctional. Furthermore, the size of the transcripts was not correlated with the number of variants identified in the gene (R^2=2.0x10^{-4}). Therefore, we did not correct for multiple variants per gene.

Clustering of Biological Process Dysfunction Scores

To cluster individuals based on overall genetic risk, we used an approach that we previously developed and showed was capable of identifying genetically-meaningful subgroups in ASD (*44*). Briefly, the correlation structure across the genetic risk scores was determined by

calculating pairwise Spearman's rank correlation coefficients. As score ranges varied by biological process, all scores were transformed into Hazen percentile ranks to be more comparable. To help ensure that correlated genetic risk scores did not overly influence results, Gower dissimilarity matrices were calculated using correlation-based weights with the 'FD' package v1.0-12 in R(*45*). The threshold for non-independence of genetic risk scores was $\rho \geq 0.50$, or moderate to strong correlation(*46*). Correlated scores were weighted to allow for only partial contributions to analyses. The 'clValid' package v0.6-6 in R was used to evaluate different methods for internal validity using connectivity, silhouette width, and the Dunn index while partitioning the dissimilarity matrix into anywhere from 2 to 5 clusters(*47*). Clustering methods that are available for evaluation in the clValid package include: 1) agglomerative hierarchical, 2) partitioning around medoids, 3) self-organizing tree algorithm, 4) model-based, 5) divisive hierarchical, and 6) fuzzy k-means. The final clustering solution was performed using the agglomerative hierarchical method via the 'cluster' package v2.0.7-1 in R(*48*). Final cluster solution validity was assessed by performing 1,000 data permutations and comparing clustering of real versus permuted genetic risk scores with the Adjusted Hubert-Arabie Rand index(*49*). Sensitivity and regression analyses were performed to determine if dysfunction in any particular biological process was important to definition of the final cluster solutions. Chi-square tests were used to determine if having variants with $\boldsymbol{LDV} > 0$ in any particular gene was associated with assignment of individuals to genetic clusters.

Differences in ASD-Related Phenotype Variables Based on Genetic Subgroup

Phenotype variables representing quantitative or ordinal severity measures for symptoms assessed in the SSC standard phenotype battery and medical history intakes were downloaded directly from SFARI Base (https://base.sfari.org/) and were available for the majority of the ASD

probands included in the genetic data analyses (99.66%, n=2,373). For more information on symptom severity measurements used for the SSC see Fischbach and Lord, 2010(*23*). When available, normalized z-scores or age-standardized scores were used. Head circumferences were transformed to *z*-scores by standardizing for age and sex using a typically developing population(*50*). Sleep duration was determined using current answers to the question "On average, how many hours/night [does your child sleep]?" obtained from the medical history intakes as described in our previous study(*51*). Student's t-tests were used to compare mean scores for symptom severity measures, that were available for at least half of the analysis dataset, between the individuals assigned to genetic clusters. Age was not associated with measures that were significantly different between clusters ($p \geq 0.43$). For measures with sex-specific differences, additional t-tests were conducted that were stratified by sex. Chi-square tests were used to determine if having variants with $\boldsymbol{LDV} > 0$ in any particular gene was associated with assignment to the genetic clusters. Logistic regression was used to test if having variants with $\boldsymbol{LDV} > 0$ in cluster-associated genes was associated with: 1) individuals with ASD compared to unaffected siblings in all races and only in white individuals, 2) increased risk for intellectual disability (IQ<70) or reports of irritable bowel syndrome while adjusting for gender and race. False discovery rate was controlled for using the Benjamini-Hochberg procedure(*52*).

Principal Component Analysis (PCA) was conducted while applying correlation-based weights to allow only partial contributions of moderately-strongly correlated phenotype variables ($\rho \geq 0.50$), similar to that described for clustering of genetic risk scores. Phenotype variables were transformed to Hazen percentile ranks prior to PCA. PCA was conducted without scaling as variables did not contribute equal weights. The number of dimensions of the PCA was estimated

via cross-validation. PCA was then performed on percentile ranked phenotype data using the 'FactoMineR' package v1.41 in R(*53*).

Results

Novel Approach Calculates Dysfunction in Biological Processes Underlying ASD

At the time of these analyses, there were 989 different protein coding ASD risk genes included in the Autism Informatics Portal (December 2017 update). 2,482 Gene Ontology (GO)(*54*, *55*) biological processes defined for humans were overrepresented for ASD risk genes based on a significance threshold of $p<0.05$; 16 terms had the lowest possible p-values ($p<1x10^{-30}$; *Fig. 1*, *Table S1*). Of the 16 top overrepresented terms, four GO terms – nervous system development (GO:0007399), synaptic signaling (GO:0099536), cognition (GO:0050890), and regulation of membrane potential (GO:0042391) – represented unique processes. There were 400 ASD candidate genes with evidence for involvement in at least one of these four biological processes. The genes that remained unassigned to any process were overrepresented in the chromosome organization process (GO:0051276, $p=7.10x10^{-12}$; *Fig. 1*, *Table S1*). An additional 82 genes were evidenced to be involved in chromosome organization. There were no unique biological processes with evidence of overrepresentation for the remaining 507 unassigned ASD candidate genes (*Table S1, Fig. S2*). The overlap in ASD risk genes assigned the five overrepresented biological processes representing unique terms is shown in Figure S3A.

There were 2,077 unique SNVs and In/Dels predicted by Variant Effect Predictor (VEP) to be damaging (*Table S3*). Predictions of variant effects based on nine other algorithms that use information in addition to genetic location indicated that 730 of the 2,077 variants were more often predicted damaging compared to benign (i.e., $\boldsymbol{LDV} > 0$). The majority of the individuals in the analysis dataset (n=2,295, 96.35%) had a variant with $\boldsymbol{LDV} > 0$ in an ASD risk gene. On

303 average, there were ~15 variants [μ=14.6(5.3)] observed per individual that was predicted to be
304 damaging more often than benign, and ~11 different [μ=11.3, (4.2)] ASD candidate genes per
305 person with possibly damaging variants. None of the variants that were *de novo* were predicted
306 to be benign and inherited variants were more often predicted to be damaging if the consequence
307 related to a frameshift, splice-site alteration, or incorporation of a premature stop codon (*Fig. 2*).
308 Screening data reported in previous studies(*11, 28*) for *de novo* and rare, inherited structural
309 variation in the SSC dataset identified 572 unique Copy Number Variants (CNVs) encompassing
310 coding regions or proximal promoter elements of 354 ASD candidate genes (*Tables S4-S5*).
311 There were 546 individuals in the analysis dataset with ≥1 CNV that was likely to cause
312 dysfunction in ≥1 ASD candidate gene; 292 CNVs encompassed more than one gene. In total,
313 there were 751 currently implicated genes with either a SNV, In/Del or CNV with $\boldsymbol{LDV} > 0$. Of
314 these, 355 were assigned to at least one unique process that was overrepresented for ASD risk
315 genes (*Fig. S3B*).

316 Most individuals in the dataset (98.1%) had evidence indicating genetic dysfunction in
317 more than one of the evaluated biological processes. There were five individuals with evidence
318 for dysfunction only in nervous system development, 35 with evidence for dysfunction only in
319 chromosome organization, and five with no evidence for dysfunction in any of the evaluated
320 processes. Scores for dysfunction in nervous system development, synaptic signaling, and
321 regulation of membrane potential were moderately to strongly correlated. Scores reflecting
322 dysfunction in cognition and chromosome organization were weakly correlated with each other
323 and other scores (*Fig. 3A*).

324 A clustering analysis was then performed on $\boldsymbol{DBP_X}$ scores reflecting the likelihood of
325 genetic dysfunction in each of the five unique biological processes. Agglomerative hierarchical

clustering identified two valid subgroups of individuals ($n_{Cluster\ 1}$=1,485, $n_{Cluster2}$=896) (*Fig. 3B*, *Fig. S4*). This solution was significantly different from clustering permuted datasets (HubertArabieRandIndex=-1.2x10^{-4}), further evidence supporting validity of the clustering analysis. Sensitivity analyses indicated that scores reflecting genetic dysfunction in the cognition biological process had the strongest influence on the stability of the clusters (*Fig. 3C*). Notably, all of the individuals assigned to the smaller cluster had evidence of dysfunction in genes involved in cognition ('cognition gene dysfunction cluster') while none of the individuals assigned to the larger cluster had evidence for dysfunction in these genes (*Fig. 3D*).

Three Cognition Genes are Associated with Distinct ASD Genetic Subgroup

Of the 61 cognition genes with likely damaging variants identified in the SSC dataset, there were three genes (*PTGS2*, *ABCA7*, and *SHANK3*) that were strongly associated with assignment to the cognition gene dysfunction cluster (*Table 1A, Table S6*). There were 196 individuals who were heterozygous for a stop-gain variant in exon 4 (rs200314986; transcript ENST00000367468.9:c.366C>A, ENSP00000356438.5:p.Tyr122Ter) of the *prostaglandin-endoperoxide synthase 2* (*PTGS2*) gene, which results in a shortened transcript that is missing the final 6 exons. This variant was more frequent in individuals with ASD compared to unaffected siblings (*Table 1B*). There were 17 different likely damaging variants observed in 280 individuals in the *ATP Binding Cassette Subfamily A Member 7* (*ABCA7*) gene. These included six frameshifts, four splice-sites, four stop-gains, one stop-loss, one inherited deletion of the first 11 exons (CNV size=18.7kb), and one inherited amplification encompassing exons 27-40 (CNV size=4.5kb). For the *SH3 and multiple ankyrin repeat domains 3* (*SHANK3*) gene, there were 294 individuals who were heterozygous for a splice-site variant (rs150909992) that changes the 5'

end of an intron in transcript variant ENST00000445220.2, and two individuals with *de novo* CNVs that deleted the entire coding region (CNV sizes>3Mb).

Individuals with Cognition Gene Variants Have More Severe Symptoms

Among the 27 ASD-related symptom measures that were available for at least half of the dataset (*Table S7*), the severity of social impairment based on teacher reports on Social Responsiveness Scales (SRS-TR), intelligence quotient (IQ) scores, personal and social skills measured using composite standard scores from the Vineland Adaptive Behavior Scales, receptive vocabulary measured via the Peabody Picture Vocabulary Test, and the severity of ASD-related abnormalities exhibited by 36 months of age (i.e., Developmental Abnormality scores) from the Autism Diagnostic Interview-Revised (ADI-R) were different between the genetic clusters (*Fig. 4A, Table S8A*). After false discovery rate corrections, the observations that individuals with dysfunction in cognition genes had increased severity of social impairment reported by teachers on the SRS-TR and reduced IQs remained significant (*Fig 4A, Table S8A*). Notably, both nonverbal and verbal IQ scores were lower in the genetic subgroup defined by dysfunction in cognition genes (*Fig 4A, Table S8B*). Sex-stratified mean comparisons indicated that differences between the genetic clusters for SRS-TR scores and verbal IQs were more significant in males compared to females (*Table S8C*).

To determine how much of the overall variability in ASD symptomatology was explained by symptoms that were different between the genetically distinct subgroups, principal components analysis (PCA) was conducted, while adjusting for correlated variables, on phenotype data from the subset of the dataset with all evaluated measures (n=543). Five principal components (PCs) were able to define the majority of the variability in symptoms (cumulative percentage of variance=46.97%). Of the 27 measures evaluated, teacher reports on SRS-TR were

the 5th strongest contributor to the cumulative variability defined by the first component of the data (*Fig. 4B, Fig. S5*). The strongest correlation for SRS-TRs (ρ=0.39) was with the variable which contributed the most to explaining the phenotypic heterogeneity defined by PC1, social and communication impairment observed via the Autism Diagnostic Observation Schedule (*Fig. 4C, Figs. S5-S6*). Full scale IQs were the 6th strongest contributor to the variability defined by PC1 (*Fig. 4B, Fig. S5*). Full scale IQ scores were moderately correlated with scores for dexterity (Purdue Pegboard Test, ρ=0.45; *Fig 4C*) and language acquisition (non-word repetition task, ρ=0.48; *Fig. 4C*) which were the 3rd and 4th largest contributors to PC1, respectively (*Fig. S5*).

Of the top three genes associated with assignment to the 'cognition dysfunction cluster', the stop-gain variant in the *PTGS2* gene was associated with increased risk for having an IQ score reflecting intellectual disability (*Table 2A*) and reports of comorbid irritable bowel syndrome, when adjusting for sex and race (*Table 2B*). The majority of the individuals with ASD, and all of the unaffected siblings with the variant inherited it from at least one parent. There were six individuals with ASD whose parents did not appear to have the variant.

Discussion

Novel Approach Identified Clinically-Relevant Genes to Prioritize for Functional Follow-up

Beginning with all 989 ASD candidate genes included in the December 2017 update of the AutDB Autism Informatics Portal, our approach identified a subset of 61 genes involved in cognition that were useful to defining a cluster of individuals with more severe teacher reported social impairment, lower IQ scores, and reduced daily living skills. We then identified three genes (i.e., *PTGS2*, *ABCA7*, and *SHANK3*) with likely damaging variants that were strongly associated with this ASD subgroup. This helped us to pinpoint the specific gene and variant that was associated with expression of important comorbidities in ASD, including intellectual

disability and irritable bowel syndrome. In particular, a stop-gain variant in the *PTGS2* gene (rs200314986) encoding Cycloxygenase-2 (COX2) – a target for non-steroidal anti-inflammatory drugs (NSAIDs) – was more frequent in individuals with ASD compared to unaffected siblings. The support for *PTGS2* as a candidate gene for ASD resides in the results of a small gene-centric association study(*56*). As such, it is considered to have weak evidence for an association with ASD based on the cumulative strength of evidence for individual variants in that gene as defined in the AutDB Autism Informatics Portal. Our work provides additional support not only for a relationship between the *PTGS2* gene and ASD risk but also for increased risk of intellectual disability in ASD. Notably, the encoded enzyme is involved in serotonergic synaptic transmission and oxytocin signaling, which are known to be impaired in some individuals with ASD(*57-60*). While not this specific variant, there are 13 other pathogenic variants reported in this gene (https://www.ncbi.nlm.nih.gov/clinvar/) relating to developmental abnormalities. *PTGS2* is also considered a very important pharmacogene by PharmGKB and has strong implications for functional follow-up studies and eventual translation to improve clinical care(*61, 62*). There are a number of variants reported in this gene that have been shown to influence individual response to NSAIDs in the typically developing population(*61*). Given the evidence that long-term use of NSAIDs has been linked to gastrointestinal issues(*63*), we also tested for and observed that individuals with the *PTGS2* stop-gain variant had increased risk for reports of irritable bowel syndrome. It is possible that drugs that selectively inhibit COX-2, as well as traditional NSAIDs that target COX-2 and COX-1 (e.g., ibuprofen) may be less effective in individuals with this loss-of-function variant. This indicates that it may be useful to test for the rs200314986 variant in individuals with ASD to help improve treatment for pain and avoid exacerbation of gastrointestinal issues.

Variants in *ABCA7* were not strongly associated with ASD. Notably, there were 17 different likely damaging variants identified in this gene. This suggests that *ABCA7* may be more tolerant to loss of function mutations. We looked at loss intolerance scores (pLI), available via DECIPHER (https://decipher.sanger.ac.uk/) which assess the probability that a gene is intolerant to a loss of function mutation(*64*). These scores indicate that *ABCA7* may tolerate deleterious variants (pLI=0.0). In comparison, there was only one stop-gain variant in *PTGS2* and three different variants (one splice-site and two CNVs) in *SHANK3*. *PTGS2* and *SHANK3* are predicted to be extremely intolerant (pLI for both genes=1.00) to loss of function mutations.

Unexpectedly, *SHANK3* variants were associated with decreased risk for ASD. *SHANK3* is considered a strong candidate gene for ASD and haploinsufficiency of *SHANK3* is implicated in Phelan-McDermid syndrome which is often comorbid with ASD and characterized by delayed speech and intellectual disability(*65*). Notably, as we conducted gene-based tests our results were likely driven by a splice-site variant (rs150909992) that was observed to be heterozygous in 294 individuals, and not by the two CNVs. The splice-site variant was identified based on the VEP consequence from a previous assembly of the reference human genome (GRCh37.p13). This variant was not predicted by any of the other algorithms tested. In the most recent update of Ensembl (GRCh38.p10), this variant is no longer predicted to be a splice-site variant for a protein coding transcript of *SHANK3*. The transcript it affects corresponds to *SHANK3-202* which is now evidenced to encode a non-coding RNA. It is possible that this variant has no effect on the SHANK3 protein, which may explain why we did not see significant effects of having a likely damaging variant in *SHANK3* on risk for ASD or intellectual disability. It may instead have regulatory effects on other genes, as there is evidence that some non-coding RNAs are functional(*66*), and it is located in a promoter flanking region which is active in neuronal

progenitor cells (ENSR00000147759). This is an excellent example of why it is important to consider that solely using the genetic location of the variant is potentially misleading in the ever-changing landscape of human genetics.

Multiple Prediction Algorithms are Necessary for Efficient Identification of Damaging Variants

By evaluating damaging variant predictions from multiple algorithms, we were able to identify the variants (whether *de novo* or inherited, rare or common) in ASD risk genes with more evidence to be damaging to the encoded protein function. It is unclear what the optimal approach is for *in silico* prediction of the likelihood a genetic variant is damaging to the encoded protein product(*21, 33, 67*). Predictions from available tools vary widely when applied to the same variant as they employ different algorithms and use different training data to determine the accuracy of predictions(*68*). As such, it is highly advisable to combine predictions from multiple tools to assess the overall likelihood a variant is damaging(*69*). We observed that ~13% of the SNVs and In/Dels that were expected to have a negative consequence on the encoded protein based on genetic location (i.e., the VEP prediction) were more often predicted to be benign by algorithms that incorporated additional information (e.g., the frequency of the variant in populations with no evidence of disease, the level of conservation of the genetic region across species). As such, if we had chosen to focus solely on VEP consequence predictions, we would have overestimated the likelihood of genetic dysfunction in the evaluated biological process. In addition, over half of the variants (52%) that were located in a genetic region that was likely to be damaging were not given predictions by any other algorithm. This is possibly because the variant being evaluated has not been observed in the populations that are used for training prediction algorithms. As such, it is currently difficult to determine the likelihood that an extremely rare variant is damaging without conducting functional follow-up studies. Only 0.14%

of the variants that VEP predicted to be highly likely of damaging the protein product based on the location in the coding region of the gene were predicted to be damaging by all of the other nine prediction algorithms evaluated. Fortunately, as the field of *in silico* variant prediction continues to develop novel methods, focused on advances like mapping variants to three-dimensional protein structures(*70*), predictions should become more accurate and variant prioritization more efficient.

Evidence of Intra-Individual Genetic Dysfunction in Multiple Biological Processes

The majority of the evaluated individuals had a variant in an ASD candidate gene that was predicted more often to be damaging compared to benign. By using these variants to calculate dysfunction in overall biological processes, we also observed that the majority of individuals had evidence of dysfunction in more than one process important to ASD etiology. The unique terms that were selected reflect validations of results from previous studies implicating genes involved in neural development, synaptic signaling, and chromosome packaging(*10, 11*). In addition, ASD risk genes were overrepresented in processes that encode the mental activities related to thinking, learning and memory (i.e., cognition) and regulate the difference in electric potential between the intra- and extra-cellular environments (i.e., regulation of membrane potential). While all of these processes had some degree of overlap in genes with likely damaging variants, there were also genes with variants that were uniquely assigned to only one process suggesting some genetic factors influencing these processes are distinct. Not surprisingly, individuals with more evidence for genetic dysfunction in development of the nervous system also had more evidence for dysfunction in the regulation of membrane potential and synaptic signaling. Dysfunction in genes influencing chromosome organization appeared independent from other processes. This provides additional support that mechanisms of

chromosome organization may contribute independently from genes influencing neurological function to increase risk for ASD(*10, 11*). Notably, predicted dysfunction in cognition genes robustly identified a genetically-distinct subgroup of individuals with ASD. Many of these genes are evidenced to be involved in human cognition because they are implicated in intellectual disability, dementia, executive function, long-term memory, and a number of aspects of learning (for details see http://amigo.geneontology.org/amigo/term/GO:0050890). These genes may be particularly relevant to developing more comprehensive genetic screening panels for ASD.

Individuals with More Cognition Gene Dysfunction Have More Severe ASD Symptoms

The genetic subgroup defined primarily based on evidence of cognition gene dysfunction had increased severity of social impairment as measures via teacher reports on the Social Responsiveness Scale (SRS-TR)(*71*). Previous studies of families with more than one child diagnosed with ASD (i.e., multiplex) have observed that SRS scores are heritable, and linked to loci on a number of different chromosomes(*72-74*). SRS scores are observed to have differential distributions when comparing male and female individuals with ASD(*72*), and simplex versus multiplex families(*75*). Our results indicate that genetic factors influence social impairment measured via the SRS in simplex families, primarily in males. It is not clear why the observations are limited to teacher reports and do not extend to parent reports on the SRS-PR. We observed weaker correlations between parent and teacher reports on the SRS than has been previously reported(*76*). It is possible that these results reflect the highly variable symptom severity of the subjects in the SSC as concordance between teacher and parent reports is influenced by severity of ASD with higher concordance as ASD severity increases(*77*). Moreover, many studies have observed that parents rate their children as being more impaired compared to teachers possibly due to the context of the social setting in which the child is being

observed(*78*). Notably, teacher reports on the SRS-TR were more correlated with social affect measured on the Autism Diagnostic Observation Schedule (ρ=0.39) when compared to SRS-PR parent reports (ρ=0.21) suggesting better agreement between teacher-reported and clinician-observed social impairment on average. Notably, PCA indicated that clinician-observed social/communication deficits measured via the Autism Diagnostic Observation Schedule (ADOS) were the largest contributors to the overall variability in quantitative ASD-related symptoms measured in the evaluated dataset and SRS-TR scores were among the top five.

Verbal and nonverbal IQ scores were also reduced in individuals with evidence of cognition gene dysfunction compared to those without. This was independent of social impairment measured via the SRS-TR, suggesting that individuals with more social impairment, and lower nonverbal and verbal IQ [as opposed to an 'IQ split'(*79*)] have genetic differences compared to individuals with less social and intellectual impairment or discordance between these two measures (e.g., higher IQs with more social impairment). Previous studies have also observed that social deficits ascertained by the SRS are generally unrelated to IQ(*71, 80*).

Limitations

Notably, many ASD candidate genes with likely damaging variants were assigned to more than one unique biological process. Therefore, to calculate scores for dysfunction in overall processes, genes with likely damaging variants were weighted to account for 1) the level of evidence supporting assignment of the gene to the biological process of interest, and 2) the proportion of the child terms of the unique biological process to which the gene was also assigned. For each gene assigned to a particular biological process, GO provides evidence codes that indicate the type of evidence supporting this assignment (http://www.geneontology.org/page/guide-go-evidence-codes). It is unclear what should be

considered the most reliable sources of evidence supporting assignment of genes to GO Terms. While experimental evidence would be preferred, it is potentially biased, as this code will likely be assigned more often to genes that are directly evaluated for a role in the process of interest and not genes that have yet to be experimentally assessed for a role in the process. The majority of genes are assigned to terms based on computational predictions that have been observed to be generally reliable in the absence of experimental data(*81*). To avoid bias in gene process assignment, weights were calculated for each gene to account for the level of evidence it was correctly assigned to the process. It is possible that this approach under- or over-estimated the level of biological process dysfunction.

It is also possible that by focusing on currently implicated ASD risk genes we did not take into account all evidence for genetic dysfunction in a process. Future work aimed at understanding genetic contributions to overall process dysfunction, regardless of the underlying evidence of genetic risk for ASD may help detect more robust differences in ASD-related symptoms. In lieu of these potential limitations, the approach we developed helped identify the variants in ASD risk genes with more evidence to be damaging to the encoded protein function. This approach was also able to identify subsets of candidate genes with common underlying biology that are dysfunctional in individuals with ASD and related to differences in symptomatology. Notably, an inherited stop-gain variant in *PTGS2* was prioritized which has strong implications for functional follow-up studies and may be a novel treatment target. This work constitutes a translational bioinformatics approach beneficial to gleaning clinically-useful information from whole-exome data and could be adapted and applied to identification of clinically-relevant genetic factors for a number of complex human disorders.

555 **Supplemental Data description:**

556 Supplemental Data include six figures and eight tables.

557 **Acknowledgments:** This work was supported by a National Library of Medicine grant K01-
558 LM012870 [OJV] and the Simons Foundation (SFARI) [JSS, ZEW]. We are grateful to all of the
559 families at the participating Simons Simplex Collection (SSC) sites, as well as the principal
560 investigators (A. Beaudet, R. Bernier, J. Constantino, E. Cook, E. Fombonne, D. Geschwind, R.
561 Goin-Kochel, E. Hanson, D. Grice, A. Klin, D. Ledbetter, C. Lord, C. Martin, D. Martin, R.
562 Maxim, J. Miles, O. Ousley, K. Pelphrey, B. Peterson, J. Piggot, C. Saulnier, M. State, W. Stone,
563 J. Sutcliffe, C. Walsh, Z. Warren, E. Wijsman). We appreciate obtaining access to phenotypic
564 and genetic data on SFARI Base. Approved researchers can obtain the SSC population dataset
565 described in this study (https://www.sfari.org/resource/simons-simplex-collection/) by applying
566 at https://base.sfari.org.

567 **Declaration of Interests**
568 The authors have no conflicts of interest to declare.

569 **Figure Legends**

570 **Figure 1. Selection of unique biological processes with overrepresentation of ASD**
571 **candidate genes for further study.** Shown is the distribution of significant terms in the GO
572 structure for biological processes (GO:0008150). Terms highlighted in yellow indicate unique
573 terms selected due to their place in the hierarchy and meaningfulness to ASD etiology. Terms
574 highlighted in blue indicate significant processes considered too broad to be meaningful and
575 green indicates significant child terms with complete genetic overlap to unique terms. Sig=the
576 number of ASD candidate genes assigned to the process, Exp=the expected number of genes
577 assigned by chance. *denotes terms that were significant at Fisher's exact FDR<1.0×10^{-30}
578 following the primary analysis of all 989 ASD risk genes, ** denotes terms that were significant
579 at Fisher's exact FDR ranging from 3.5×10^{-17} to 7.1×10^{-12} following the secondary analysis run
580 on genes unassigned to the top processes. Black lines connect terms that regulate each other, blue
581 lines connect terms that are part of each other.

582

583 **Figure 2. Proportion of VEP consequences predicted to be damaging based on nine**
584 **prediction algorithms.** Inherited variation resulting in frameshifts, splice-site and stop gains
585 were more often predicted damaging compared to benign, while variants predicted to cause the
586 loss of either stop or start sites were equally or more often predicted to be benign. *De novo*
587 variants, regardless of the consequence were more often damaging.

588

589 **Figure 3. Clustering individuals based on overall biological process dysfunction. A**)
590 Correlation across scores reflecting dysfunction in biological processes with overrepresentation
591 of ASD candidate genes indicates many individuals have dysfunction in >1 process. **B)**
592 Clustering identified two distinct subgroups of individuals with more similar scores for overall
593 biological process dysfunction (agglomerative coefficient=0.96). **C)** Sensitivity analyses indicate
594 removing the scores had the strongest effect on stability of the clustering solution. APN=average

595 proportion of non-overlap, AD=average distance, ADM=average distance between means,
596 FOM=figure of merit. **D)** Evidence of dysfunction in genes involved in cognition primarily
597 defined separation of individuals into either cluster 1 (no cognition gene dysfunction) or cluster 2
598 (cognition gene dysfunction).
599
600 **Figure 4. Relationship between genetic and phenotypic differences. A)** T-tests comparing
601 differences in the 27 quantitative ASD-related symptom measures between genetic clusters
602 identified that social impairment was more severe and IQs and daily living skills were reduced in
603 the cognition gene dysfunction cluster. **B)** Principal components analysis, while adjusting for
604 correlations, of all 27 symptom measures identified that symptoms that were different between
605 the genetic clusters majorly contributed to overall phenotype variability (as defined by
606 Dimension 1). Black indicates symptom differences that remained significant (FDR≤0.04)
607 following multiple testing correction, gray indicates symptom differences based on an unadjusted
608 significance threshold ($p \leq 0.03$), and unlabeled arrows indicate symptoms that were not different
609 but had strong contributions to phenotype variability. **C)** Significant ($p < 0.05$) correlations are
610 shown indicating that absolute values for many symptoms that were different between genetic
611 clusters were correlated with those contributing majorly to overall phenotype variability**.**
612

Tables:

Table 1A. Genes associated with the cognition gene dysfunction cluster

Genetic Cluster	*PTGS2*		*ABCA7*		*SHANK3*	
	No Variant	Variant	No Variants	Variants	No Variants	Variants
Cluster 1 Observed	1,485	0	1,485	0	1,485	0
Expected	1,363	122	1310	175	1,300	185
Cluster 2 Observed	700	196	616	280	600	296
Expected	822	74	791	105	785	111
Total	2185	196	2,101	280	2085	296
Pearson χ^2	353.98		525.91		560.23	
Fisher's exact	$<1.0x10^{-32}$		$<1.0x10^{-32}$		$<1.0x10^{-32}$	

Table 1B. Association of cluster-associated cognition genes with Autism Spectrum Disorder

ASD Diagnosis	*PTGS2*		*ABCA7*		*SHANK3*		
	No Variant	Variant	No Variants	Variants	No Variants	Variants	Total
All Reported Races							
ASD	2,185	196	2,101	280	2,085	296	2,381
Unaffected Siblings	1,700	100	1,602	198	1,431	396	1,800
Odds Ratio (95%C.I.)†	1.52 (1.18, 1.97)		1.08 (0.88, 1.31)		0.55 (0.46, 0.65)		
p-value	$5.0x10^{-4}$		$2.4x10^{-1}$		$<1.0x10^{-5}$		
FDR	$7.5x10^{-4}$		$2.4x10^{-1}$		$3.0x10^{-5}$		
Reported White Race							
ASD	1,644	184	1,634	194	1,594	234	1,828
Unaffected Siblings	1,280	85	1,240	125	1,095	270	1,365
Odds Ratio (95%C.I.)†	1.69 (1.28, 2.23)		1.18 (0.92, 1.50)		0.60 (0.49, 0.72)		
p-value	$1.0x10^{-4}$		$9.7x10^{-2}$		$<1.0x10^{-5}$		
FDR	$1.5x10^{-4}$		$9.7x10^{-2}$		$3.0x10^{-5}$		

Of the 61 genes used to calculate $\boldsymbol{DBP_{Cognition}}$ scores, **A)** three genes were significantly associated with assignment of individuals to the cluster with evidence for dysfunction in cognition genes (i.e., Genetic Cluster 2). **B)** In particular, a variant in *PTGS2* was more frequent in individuals with ASD compared to unaffected siblings. Tests were conducted in all individuals and only in individuals with white race reported to account for potential influences of population stratification. †Odds ratios denote the likelihood for an individual to have an ASD diagnosis given the presence of any likely damaging variant in *SHANK3*, or *ABCA7*, or the T-allele (i.e. a stop-gain variant) in *PTGS2*.

Table 2A. Association of cognition gene variants with intellectual disability in ASD

Cognition Gene	Chr	z (df=2262)	Odds Ratio (95%C.I.)	p-value
PTGS2	1q31.1	2.05	***1.40 (1.02, 1.92)*** ***2.08 (1.75, 2.40)^{Het}*** ***3.01 (2.65, 3.38)^{Hom}***	***0.040***
ABCA7	19p13.3	1.07	1.16 (0.88, 1.53)	0.287
SHANK3	22q13.33	1.56	1.24 (0.95, 1.62)	0.119

Table 2B. Association of *PTGS2* variant with irritable bowel syndrome in ASD

Cognition Gene	Chr	z (df=2264)	Odds Ratio (95%C.I.)	p-value
PTGS2	1q31.1	3.10	5.38 (1.85, 15.58) 2.01 (1.70, 2.33)^{Het} 2.94 (2.59, 3.30)^{Hom}	0.002

All tests were adjusted for sex and race. **A)** Odds ratios denote the risk for having a full scale IQ score <70 (n=690) compared to a full scale IQ ≥70 (n=1,638) given any likely damaging variant in the tested gene. Likely damaging variants were defined as those that were more often predicted damaging when comparing results from 10 different prediction algorithms. More than one of these variants was identified in *ABCA7* and *SHANK3.* For *PTGS2*, results of ordered logistic regression are shown testing effects of heterozygosity (het) or homozygosity (hom) for a stop-gain variant. df=degrees of freedom.; Chr=chromosomal location of gene. **B)** Odds ratios denote increased risk for an individual to have reports of irritable bowel syndrome (n=17) given the stop-gain variant in *PTGS2*.

References and Notes:

1. A. P. Association, *Diagnostic and Statistical Manual of Mental Disorders. 5th ed.* . (American Psychiatric Association, Washington DC, 2013).
2. J. L. Matson, M. Shoemaker, Intellectual disability and its relationship to autism spectrum disorders. *Research in developmental disabilities* **30**, 1107 (Nov-Dec, 2009).
3. E. Y. Hsiao, Gastrointestinal issues in autism spectrum disorder. *Harvard review of psychiatry* **22**, 104 (Mar-Apr, 2014).
4. G. Ramaswami, D. H. Geschwind, Genetics of autism spectrum disorder. *Handbook of clinical neurology* **147**, 321 (2018).
5. P. Chaste, K. Roeder, B. Devlin, The Yin and Yang of Autism Genetics: How Rare De Novo and Common Variations Affect Liability. *Annual review of genomics and human genetics* **18**, 167 (Aug 31, 2017).
6. E. Lacivita, R. Perrone, L. Margari, M. Leopoldo, Targets for Drug Therapy for Autism Spectrum Disorder: Challenges and Future Directions. *Journal of medicinal chemistry* **60**, 9114 (Nov 22, 2017).
7. L. de la Torre-Ubieta, H. Won, J. L. Stein, D. H. Geschwind, Advancing the understanding of autism disease mechanisms through genetics. *Nature medicine* **22**, 345 (Apr, 2016).
8. S. LeClerc, D. Easley, Pharmacological therapies for autism spectrum disorder: a review. *P & T : a peer-reviewed journal for formulary management* **40**, 389 (Jun, 2015).
9. N. N. Parikshak, M. J. Gandal, D. H. Geschwind, Systems biology and gene networks in neurodevelopmental and neurodegenerative disorders. *Nature reviews. Genetics* **16**, 441 (Aug, 2015).
10. D. Pinto, E. Delaby, D. Merico, M. Barbosa, A. Merikangas, L. Klei, B. Thiruvahindrapuram, X. Xu, R. Ziman, Z. Wang, J. A. Vorstman, A. Thompson, R. Regan, M. Pilorge, G. Pellecchia, A. T. Pagnamenta, B. Oliveira, C. R. Marshall, T. R. Magalhaes, J. K. Lowe, J. L. Howe, A. J. Griswold, J. Gilbert, E. Duketis, B. A. Dombroski, M. V. De Jonge, M. Cuccaro, E. L. Crawford, C. T. Correia, J. Conroy, I. C. Conceicao, A. G. Chiocchetti, J. P. Casey, G. Cai, C. Cabrol, N. Bolshakova, E. Bacchelli, R. Anney, S. Gallinger, M. Cotterchio, G. Casey, L. Zwaigenbaum, K. Wittemeyer, K. Wing, S. Wallace, H. van Engeland, A. Tryfon, S. Thomson, L. Soorya, B. Roge, W. Roberts, F. Poustka, S. Mouga, N. Minshew, L. A. McInnes, S. G. McGrew, C. Lord, M. Leboyer, A. S. Le Couteur, A. Kolevzon, P. Jimenez Gonzalez, S. Jacob, R. Holt, S. Guter, J. Green, A. Green, C. Gillberg, B. A. Fernandez, F. Duque, R. Delorme, G. Dawson, P. Chaste, C. Cafe, S. Brennan, T. Bourgeron, P. F. Bolton, S. Bolte, R. Bernier, G. Baird, A. J. Bailey, E. Anagnostou, J. Almeida, E. M. Wijsman, V. J. Vieland, A. M. Vicente, G. D. Schellenberg, M. Pericak-Vance, A. D. Paterson, J. R. Parr, G. Oliveira, J. I. Nurnberger, A. P. Monaco, E. Maestrini, S. M. Klauck, H. Hakonarson, J. L. Haines, D. H. Geschwind, C. M. Freitag, S. E. Folstein, S. Ennis, H. Coon, A. Battaglia, P. Szatmari, J. S. Sutcliffe, J. Hallmayer, M. Gill, E. H. Cook, J. D. Buxbaum, B. Devlin, L. Gallagher, C. Betancur, S. W. Scherer, Convergence of genes and cellular pathways dysregulated in autism spectrum disorders. *American journal of human genetics* **94**, 677 (May 1, 2014).
11. S. J. Sanders, X. He, A. J. Willsey, A. G. Ercan-Sencicek, K. E. Samocha, A. E. Cicek, M. T. Murtha, V. H. Bal, S. L. Bishop, S. Dong, A. P. Goldberg, C. Jinlu, J. F. Keaney,

3rd, L. Klei, J. D. Mandell, D. Moreno-De-Luca, C. S. Poultney, E. B. Robinson, L. Smith, T. Solli-Nowlan, M. Y. Su, N. A. Teran, M. F. Walker, D. M. Werling, A. L. Beaudet, R. M. Cantor, E. Fombonne, D. H. Geschwind, D. E. Grice, C. Lord, J. K. Lowe, S. M. Mane, D. M. Martin, E. M. Morrow, M. E. Talkowski, J. S. Sutcliffe, C. A. Walsh, T. W. Yu, D. H. Ledbetter, C. L. Martin, E. H. Cook, J. D. Buxbaum, M. J. Daly, B. Devlin, K. Roeder, M. W. State, Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. *Neuron* **87**, 1215 (Sep 23, 2015).

12. P. Chaste, L. Klei, S. J. Sanders, V. Hus, M. T. Murtha, J. K. Lowe, A. J. Willsey, D. Moreno-De-Luca, T. W. Yu, E. Fombonne, D. Geschwind, D. E. Grice, D. H. Ledbetter, S. M. Mane, D. M. Martin, E. M. Morrow, C. A. Walsh, J. S. Sutcliffe, C. Lese Martin, A. L. Beaudet, C. Lord, M. W. State, E. H. Cook, Jr., B. Devlin, A genome-wide association study of autism using the Simons Simplex Collection: Does reducing phenotypic heterogeneity in autism increase genetic homogeneity? *Biological psychiatry* **77**, 775 (May 1, 2015).

13. E. B. Robinson, K. E. Samocha, J. A. Kosmicki, L. McGrath, B. M. Neale, R. H. Perlis, M. J. Daly, Autism spectrum disorder severity reflects the average contribution of de novo and familial influences. *Proceedings of the National Academy of Sciences of the United States of America* **111**, 15161 (Oct 21, 2014).

14. K. E. Samocha, E. B. Robinson, S. J. Sanders, C. Stevens, A. Sabo, L. M. McGrath, J. A. Kosmicki, K. Rehnstrom, S. Mallick, A. Kirby, D. P. Wall, D. G. MacArthur, S. B. Gabriel, M. DePristo, S. M. Purcell, A. Palotie, E. Boerwinkle, J. D. Buxbaum, E. H. Cook, Jr., R. A. Gibbs, G. D. Schellenberg, J. S. Sutcliffe, B. Devlin, K. Roeder, B. M. Neale, M. J. Daly, A framework for the interpretation of de novo mutation in human disease. *Nature genetics* **46**, 944 (Sep, 2014).

15. M. Ronemus, I. Iossifov, D. Levy, M. Wigler, The role of de novo mutations in the genetics of autism spectrum disorders. *Nature reviews. Genetics* **15**, 133 (Feb, 2014).

16. S. J. Sanders, A. G. Ercan-Sencicek, V. Hus, R. Luo, M. T. Murtha, D. Moreno-De-Luca, S. H. Chu, M. P. Moreau, A. R. Gupta, S. A. Thomson, C. E. Mason, K. Bilguvar, P. B. Celestino-Soper, M. Choi, E. L. Crawford, L. Davis, N. R. Wright, R. M. Dhodapkar, M. DiCola, N. M. DiLullo, T. V. Fernandez, V. Fielding-Singh, D. O. Fishman, S. Frahm, R. Garagaloyan, G. S. Goh, S. Kammela, L. Klei, J. K. Lowe, S. C. Lund, A. D. McGrew, K. A. Meyer, W. J. Moffat, J. D. Murdoch, B. J. O'Roak, G. T. Ober, R. S. Pottenger, M. J. Raubeson, Y. Song, Q. Wang, B. L. Yaspan, T. W. Yu, I. R. Yurkiewicz, A. L. Beaudet, R. M. Cantor, M. Curland, D. E. Grice, M. Gunel, R. P. Lifton, S. M. Mane, D. M. Martin, C. A. Shaw, M. Sheldon, J. A. Tischfield, C. A. Walsh, E. M. Morrow, D. H. Ledbetter, E. Fombonne, C. Lord, C. L. Martin, A. I. Brooks, J. S. Sutcliffe, E. H. Cook, Jr., D. Geschwind, K. Roeder, B. Devlin, M. W. State, Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. *Neuron* **70**, 863 (Jun 9, 2011).

17. D. J. Weiner, E. M. Wigdor, S. Ripke, R. K. Walters, J. A. Kosmicki, J. Grove, K. E. Samocha, J. I. Goldstein, A. Okbay, J. Bybjerg-Grauholm, T. Werge, D. M. Hougaard, J. Taylor, D. Skuse, B. Devlin, R. Anney, S. J. Sanders, S. Bishop, P. B. Mortensen, A. D. Borglum, G. D. Smith, M. J. Daly, E. B. Robinson, Polygenic transmission disequilibrium confirms that common and rare variation act additively to create risk for autism spectrum disorders. *Nature genetics* **49**, 978 (Jul, 2017).

18. X. Ji, R. L. Kember, C. D. Brown, M. Bucan, Increased burden of deleterious variants in essential genes in autism spectrum disorder. *Proceedings of the National Academy of Sciences of the United States of America* **113**, 15054 (Dec 27, 2016).
19. R. Ben-Shalom, C. M. Keeshen, K. N. Berrios, J. Y. An, S. J. Sanders, K. J. Bender, Opposing Effects on NaV1.2 Function Underlie Differences Between SCN2A Variants Observed in Individuals With Autism Spectrum Disorder or Infantile Seizures. *Biological psychiatry* **82**, 224 (Aug 1, 2017).
20. M. Kircher, D. M. Witten, P. Jain, B. J. O'Roak, G. M. Cooper, J. Shendure, A general framework for estimating the relative pathogenicity of human genetic variants. *Nature genetics* **46**, 310 (Mar, 2014).
21. S. Richards, N. Aziz, S. Bale, D. Bick, S. Das, J. Gastier-Foster, W. W. Grody, M. Hegde, E. Lyon, E. Spector, K. Voelkerding, H. L. Rehm, Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in medicine : official journal of the American College of Medical Genetics* **17**, 405 (May, 2015).
22. D. Guala, E. L. L. Sonnhammer, A large-scale benchmark of gene prioritization methods. *Scientific reports* **7**, 46598 (Apr 21, 2017).
23. G. D. Fischbach, C. Lord, The Simons Simplex Collection: a resource for identification of autism genetic risk factors. *Neuron* **68**, 192 (Oct 21, 2010).
24. S. L. Van Driest, Y. Shi, E. A. Bowton, J. S. Schildcrout, J. F. Peterson, J. Pulley, J. C. Denny, D. M. Roden, Clinically actionable genotypes among 10,000 patients with preemptive pharmacogenomic testing. *Clinical pharmacology and therapeutics* **95**, 423 (Apr, 2014).
25. S. N. Basu, R. Kollu, S. Banerjee-Basu, AutDB: a gene reference resource for autism research. *Nucleic acids research* **37**, D832 (Jan, 2009).
26. X. Jiao, B. T. Sherman, W. Huang da, R. Stephens, M. W. Baseler, H. C. Lane, R. A. Lempicki, DAVID-WS: a stateful web service to facilitate gene/protein list analysis. *Bioinformatics* **28**, 1805 (Jul 1, 2012).
27. A. R. Alexa, J. (2016).
28. N. Krumm, T. N. Turner, C. Baker, L. Vives, K. Mohajeri, K. Witherspoon, A. Raja, B. P. Coe, H. A. Stessman, Z. X. He, S. M. Leal, R. Bernier, E. E. Eichler, Excess of rare, inherited truncating mutations in autism. *Nature genetics* **47**, 582 (Jun, 2015).
29. W. McLaren, L. Gil, S. E. Hunt, H. S. Riat, G. R. Ritchie, A. Thormann, P. Flicek, F. Cunningham, The Ensembl Variant Effect Predictor. *Genome biology* **17**, 122 (Jun 6, 2016).
30. A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, M. A. DePristo, The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20**, 1297 (Sep, 2010).
31. E. M. Garrison, G, Haplotype-based variant detection from short-read sequencing. *arXiv:1207.3907v2*, (2012).
32. A. R. Carson, E. N. Smith, H. Matsui, S. K. Braekkan, K. Jepsen, J. B. Hansen, K. A. Frazer, Effective filtering strategies to improve data quality from population-based whole exome sequencing studies. *BMC bioinformatics* **15**, 125 (May 2, 2014).

33. L. C. Walters-Sen, S. Hashimoto, D. L. Thrush, S. Reshmi, J. M. Gastier-Foster, C. Astbury, R. E. Pyatt, Variability in pathogenicity prediction programs: impact on clinical diagnostics. *Molecular genetics & genomic medicine* **3**, 99 (Mar, 2015).
34. H. Yang, K. Wang, Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. *Nature protocols* **10**, 1556 (Oct, 2015).
35. P. C. Ng, S. Henikoff, Predicting deleterious amino acid substitutions. *Genome research* **11**, 863 (May, 2001).
36. I. Adzhubei, D. M. Jordan, S. R. Sunyaev, Predicting functional effect of human missense mutations using PolyPhen-2. *Current protocols in human genetics* **Chapter 7**, Unit7 20 (Jan, 2013).
37. J. M. Schwarz, C. Rodelsperger, M. Schuelke, D. Seelow, MutationTaster evaluates disease-causing potential of sequence alterations. *Nature methods* **7**, 575 (Aug, 2010).
38. B. Reva, Y. Antipin, C. Sander, Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic acids research* **39**, e118 (Sep 1, 2011).
39. S. Chun, J. C. Fay, Identification of deleterious mutations within three human genomes. *Genome research* **19**, 1553 (Sep, 2009).
40. H. A. Shihab, M. F. Rogers, J. Gough, M. Mort, D. N. Cooper, I. N. Day, T. R. Gaunt, C. Campbell, An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics* **31**, 1536 (May 15, 2015).
41. C. Dong, P. Wei, X. Jian, R. Gibbs, E. Boerwinkle, K. Wang, X. Liu, Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Human molecular genetics* **24**, 2125 (Apr 15, 2015).
42. K. A. Jagadeesh, A. M. Wenger, M. J. Berger, H. Guturu, P. D. Stenson, D. N. Cooper, J. A. Bernstein, G. Bejerano, M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nature genetics* **48**, 1581 (Dec, 2016).
43. A. R. Quinlan, BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Current protocols in bioinformatics* **47**, 11 12 1 (Sep 8, 2014).
44. O. J. Veatch, J. Veenstra-Vanderweele, M. Potter, M. A. Pericak-Vance, J. L. Haines, Genetically meaningful phenotypic subgroups in autism spectrum disorders. *Genes, brain, and behavior* **13**, 276 (Mar, 2014).
45. E. L. Laliberté, P.; Shipley, B. (CRAN, 2014).
46. M. M. Mukaka, Statistics corner: A guide to appropriate use of correlation coefficient in medical research. *Malawi medical journal : the journal of Medical Association of Malawi* **24**, 69 (Sep, 2012).
47. G. P. Brock, V.; Datta, S.; Datta, S. (Journal of Statistical Software, 2008).
48. M. R. Maechler, P.; Struyf A.; Hubert, M.; Hornik, K.; Studer, M.; Roudier, P.; Gonzalez, J.; Kozlowski, K. (CRAN, 2018).
49. D. Steinley, Properties of the Hubert-Arabie adjusted Rand index. *Psychological methods* **9**, 386 (Sep, 2004).
50. A. F. Roche, D. Mukherjee, S. M. Guo, W. M. Moore, Head circumference reference data: birth to 18 years. *Pediatrics* **79**, 706 (May, 1987).
51. O. J. Veatch, J. S. Sutcliffe, Z. E. Warren, B. T. Keenan, M. H. Potter, B. A. Malow, Shorter sleep duration is associated with social impairment and comorbidities in ASD. *Autism research : official journal of the International Society for Autism Research*, (Mar 16, 2017).

52. Y. H. Benjamini, Y., Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57**, 289 (1995).
53. S. Lê, J. Josse, F. Husson, FactoMineR: An R Package for Multivariate Analysis. *2008* **25**, 18 (2008-03-18, 2008).
54. Expansion of the Gene Ontology knowledgebase and resources. *Nucleic acids research* **45**, D331 (Jan 4, 2017).
55. M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, G. Sherlock, Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics* **25**, 25 (May, 2000).
56. H. J. Yoo, I. H. Cho, M. Park, E. Cho, S. C. Cho, B. N. Kim, J. W. Kim, S. A. Kim, Association between PTGS2 polymorphism and autism spectrum disorders in Korean trios. *Neuroscience research* **62**, 66 (Sep, 2008).
57. I. Cataldo, A. Azhari, G. Esposito, A Review of Oxytocin and Arginine-Vasopressin Receptors and Their Modulation of Autism Spectrum Disorder. *Frontiers in molecular neuroscience* **11**, 27 (2018).
58. C. L. Muller, A. M. J. Anacker, J. Veenstra-VanderWeele, The serotonin system in autism spectrum disorder: From biomarker to animal models. *Neuroscience* **321**, 24 (May 3, 2016).
59. A. K. Singh, J. Zajdel, E. Mirrasekhian, N. Almoosawi, I. Frisch, A. M. Klawonn, M. Jaarola, M. Fritz, D. Engblom, Prostaglandin-mediated inhibition of serotonin signaling controls the affective component of inflammatory pain. *The Journal of clinical investigation* **127**, 1370 (Apr 3, 2017).
60. Y. Sugimoto, A. Yamasaki, E. Segi, K. Tsuboi, Y. Aze, T. Nishimura, H. Oida, N. Yoshida, T. Tanaka, M. Katsuyama, K. Hasumoto, T. Murata, M. Hirata, F. Ushikubi, M. Negishi, A. Ichikawa, S. Narumiya, Failure of parturition in mice lacking the prostaglandin F receptor. *Science* **277**, 681 (Aug 1, 1997).
61. C. F. Thorn, T. Grosser, T. E. Klein, R. B. Altman, PharmGKB summary: very important pharmacogene information for PTGS2. *Pharmacogenetics and genomics* **21**, 607 (Sep, 2011).
62. M. Whirl-Carrillo, E. M. McDonagh, J. M. Hebert, L. Gong, K. Sangkuhl, C. F. Thorn, R. B. Altman, T. E. Klein, Pharmacogenomics knowledge for personalized medicine. *Clinical pharmacology and therapeutics* **92**, 414 (Oct, 2012).
63. X. P. Miao, J. S. Li, Q. Ouyang, R. W. Hu, Y. Zhang, H. Y. Li, Tolerability of selective cyclooxygenase 2 inhibitors used for the treatment of rheumatological manifestations of inflammatory bowel disease. *The Cochrane database of systematic reviews*, CD007744 (Oct 23, 2014).
64. M. Lek, K. J. Karczewski, E. V. Minikel, K. E. Samocha, E. Banks, T. Fennell, A. H. O'Donnell-Luria, J. S. Ware, A. J. Hill, B. B. Cummings, T. Tukiainen, D. P. Birnbaum, J. A. Kosmicki, L. E. Duncan, K. Estrada, F. Zhao, J. Zou, E. Pierce-Hoffman, J. Berghout, D. N. Cooper, N. Deflaux, M. DePristo, R. Do, J. Flannick, M. Fromer, L. Gauthier, J. Goldstein, N. Gupta, D. Howrigan, A. Kiezun, M. I. Kurki, A. L. Moonshine, P. Natarajan, L. Orozco, G. M. Peloso, R. Poplin, M. A. Rivas, V. Ruano-Rubio, S. A. Rose, D. M. Ruderfer, K. Shakir, P. D. Stenson, C. Stevens, B. P. Thomas, G. Tiao, M. T.

Tusie-Luna, B. Weisburd, H. H. Won, D. Yu, D. M. Altshuler, D. Ardissino, M. Boehnke, J. Danesh, S. Donnelly, R. Elosua, J. C. Florez, S. B. Gabriel, G. Getz, S. J. Glatt, C. M. Hultman, S. Kathiresan, M. Laakso, S. McCarroll, M. I. McCarthy, D. McGovern, R. McPherson, B. M. Neale, A. Palotie, S. M. Purcell, D. Saleheen, J. M. Scharf, P. Sklar, P. F. Sullivan, J. Tuomilehto, M. T. Tsuang, H. C. Watkins, J. G. Wilson, M. J. Daly, D. G. MacArthur, Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285 (Aug 18, 2016).

65. M. Bidinosti, P. Botta, S. Kruttner, C. C. Proenca, N. Stoehr, M. Bernhard, I. Fruh, M. Mueller, D. Bonenfant, H. Voshol, W. Carbone, S. J. Neal, S. M. McTighe, G. Roma, R. E. Dolmetsch, J. A. Porter, P. Caroni, T. Bouwmeester, A. Luthi, I. Galimberti, CLK2 inhibition ameliorates autistic features associated with SHANK3 deficiency. *Science* **351**, 1199 (Mar 11, 2016).

66. A. F. Palazzo, E. S. Lee, Non-coding RNA: what is functional and what is junk? *Frontiers in genetics* **6**, 2 (2015).

67. V. Pejaver, S. D. Mooney, P. Radivojac, Missense variant pathogenicity predictors generalize well across a range of function-specific prediction challenges. *Human mutation* **38**, 1092 (Sep, 2017).

68. M. A. Care, C. J. Needham, A. J. Bulpitt, D. R. Westhead, Deleterious SNP prediction: be mindful of your training data! *Bioinformatics* **23**, 664 (Mar 15, 2007).

69. C. Knecht, M. Mort, O. Junge, D. N. Cooper, M. Krawczak, A. Caliebe, IMHOTEP-a composite score integrating popular tools for predicting the functional consequences of non-synonymous sequence variants. *Nucleic acids research* **45**, e13 (Feb 17, 2017).

70. G. Glusman, P. W. Rose, A. Prlic, J. Dougherty, J. M. Duarte, A. S. Hoffman, G. J. Barton, E. Bendixen, T. Bergquist, C. Bock, E. Brunk, M. Buljan, S. K. Burley, B. Cai, H. Carter, J. Gao, A. Godzik, M. Heuer, M. Hicks, T. Hrabe, R. Karchin, J. K. Leman, L. Lane, D. L. Masica, S. D. Mooney, J. Moult, G. S. Omenn, F. Pearl, V. Pejaver, S. M. Reynolds, A. Rokem, T. Schwede, S. Song, H. Tilgner, Y. Valasatava, Y. Zhang, E. W. Deutsch, Mapping genetic variations to three-dimensional protein structures to enhance variant interpretation: a proposed framework. *Genome medicine* **9**, 113 (Dec 18, 2017).

71. J. N. Constantino, S. A. Davis, R. D. Todd, M. K. Schindler, M. M. Gross, S. L. Brophy, L. M. Metzger, C. S. Shoushtari, R. Splinter, W. Reich, Validation of a brief quantitative measure of autistic traits: comparison of the social responsiveness scale with the autism diagnostic interview-revised. *Journal of autism and developmental disorders* **33**, 427 (Aug, 2003).

72. J. K. Lowe, D. M. Werling, J. N. Constantino, R. M. Cantor, D. H. Geschwind, Social responsiveness, an autism endophenotype: genomewide significant linkage to two regions on chromosome 8. *The American journal of psychiatry* **172**, 266 (Mar 1, 2015).

73. H. Coon, M. E. Villalobos, R. J. Robison, N. J. Camp, D. S. Cannon, K. Allen-Brady, J. S. Miller, W. M. McMahon, Genome-wide linkage using the Social Responsiveness Scale in Utah autism pedigrees. *Molecular autism* **1**, 8 (Apr 8, 2010).

74. J. A. Duvall, A. Lu, R. M. Cantor, R. D. Todd, J. N. Constantino, D. H. Geschwind, A quantitative trait locus analysis of social responsiveness in multiplex autism families. *The American journal of psychiatry* **164**, 656 (Apr, 2007).

75. Y. V. Virkud, R. D. Todd, A. M. Abbacchi, Y. Zhang, J. N. Constantino, Familial aggregation of quantitative autistic traits in multiplex versus simplex autism. *American*

journal of medical genetics. Part B, Neuropsychiatric genetics : the official publication of the International Society of Psychiatric Genetics **150B**, 328 (Apr 5, 2009).

76. S. Bolte, F. Poustka, J. N. Constantino, Assessing autistic traits: cross-cultural validation of the social responsiveness scale (SRS). *Autism research : official journal of the International Society for Autism Research* **1**, 354 (Dec, 2008).
77. G. Azad, E. Reisinger, M. Xie, D. S. Mandell, Parent and Teacher Concordance on the Social Responsiveness Scale for Children with Autism. *School mental health* **8**, 368 (Sep, 2016).
78. D. S. Murray, L. A. Ruble, H. Willis, C. A. Molloy, Parent and teacher report of social skills in children with autism spectrum disorders. *Language, speech, and hearing services in schools* **40**, 109 (Apr, 2009).
79. D. O. Black, G. L. Wallace, J. L. Sokoloff, L. Kenworthy, Brief report: IQ split predicts social symptoms and communication abilities in high-functioning children with autism spectrum disorders. *Journal of autism and developmental disorders* **39**, 1613 (Nov, 2009).
80. J. N. Constantino, R. D. Todd, Autistic traits in the general population: a twin study. *Archives of general psychiatry* **60**, 524 (May, 2003).
81. N. Skunca, A. Altenhoff, C. Dessimoz, Quality of computationally inferred gene ontology annotations. *PLoS computational biology* **8**, e1002533 (May, 2012).

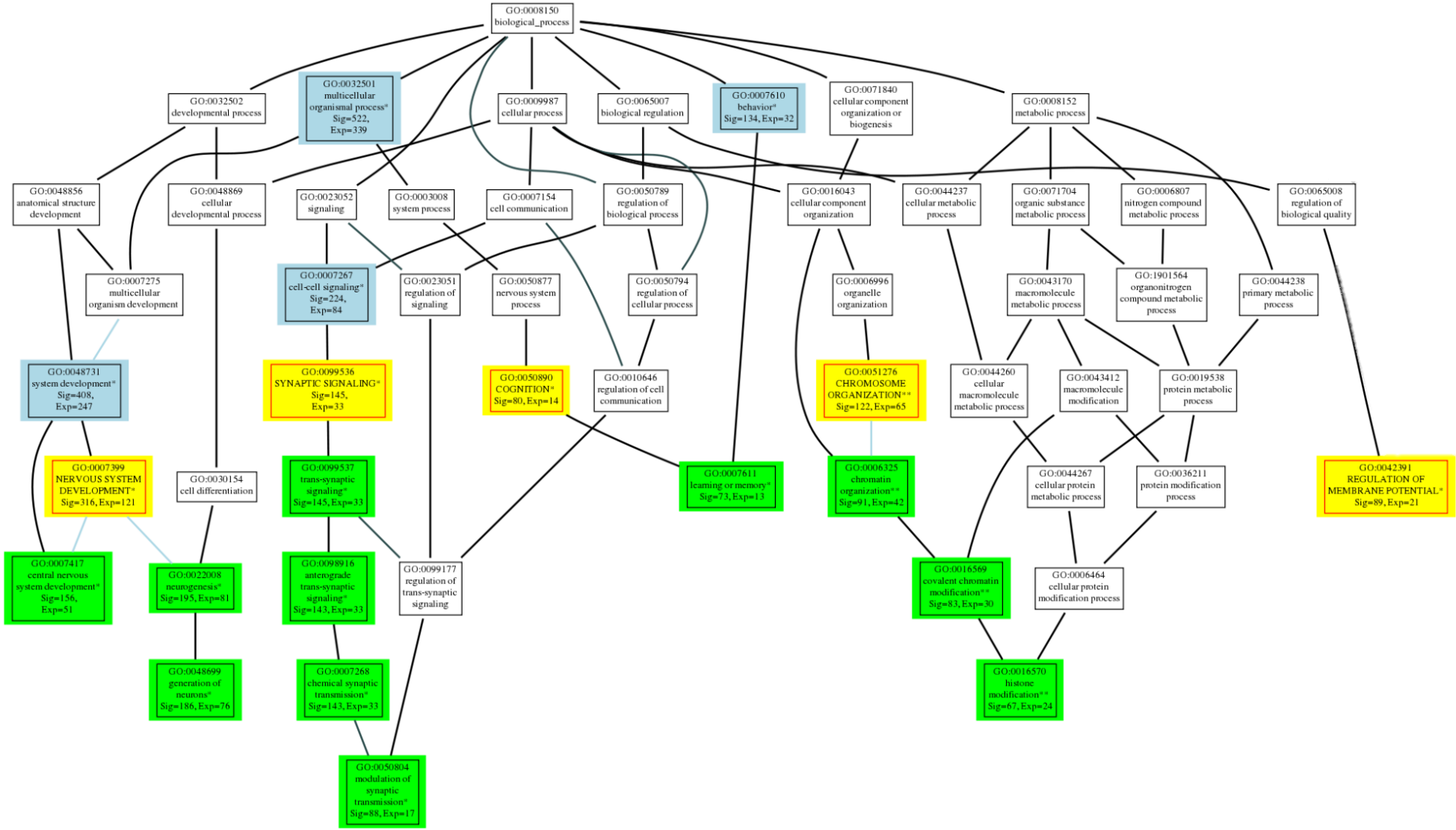


Figure 1. Selection of unique biological processes with overrepresentation of ASD candidate genes for further study. Shown is the distribution of significant terms in the GO structure for biological processes (GO:0008150). Terms highlighted in yellow indicate unique terms selected due to their place in the hierarchy and meaningfulness to ASD etiology. Terms highlighted in blue indicate significant processes considered too broad to be meaningful and green indicates significant child terms with complete genetic overlap to unique terms. Sig=the number of ASD candidate genes assigned to the process, Exp=the expected number of genes assigned by chance. *denotes terms that were significant at Fisher's exact FDR<1.0×10^{-30} following the primary analysis of all 989 ASD risk genes, ** denotes terms that were significant at Fisher's exact FDR ranging from 3.5×10^{-17} to 7.1×10^{-12} following the secondary analysis run on genes unassigned to the top processes. Black lines connect terms that regulate each other, blue lines connect terms that are part of each other.

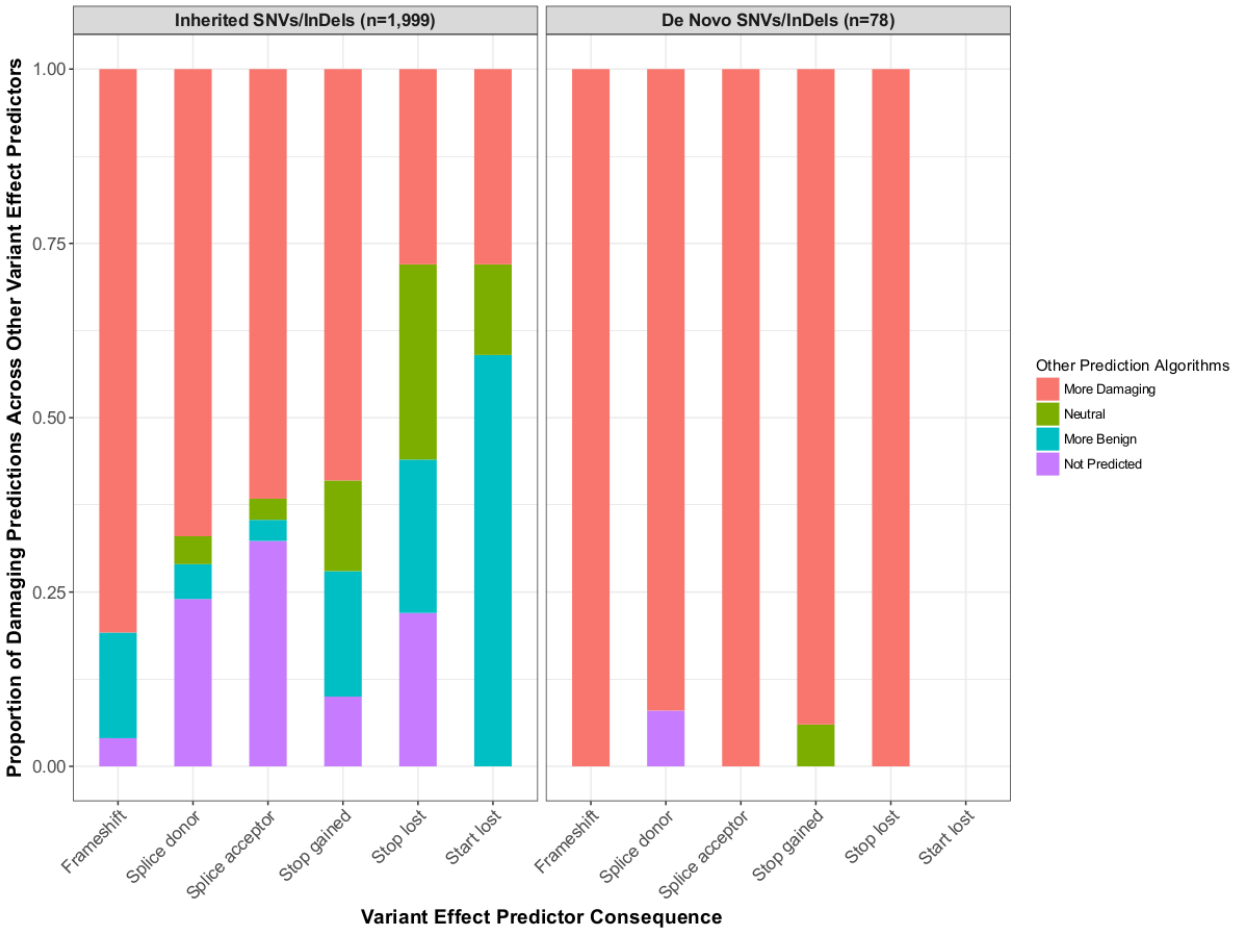


Figure 2. Proportion of VEP consequences predicted to be damaging based on nine prediction algorithms. Inherited variation resulting in frameshifts, splice-site and stop gains were more often predicted damaging compared to benign, while variants predicted to cause the loss of either stop or start sites were equally or more often predicted to be benign. *De novo* variants, regardless of the consequence were more often damaging.

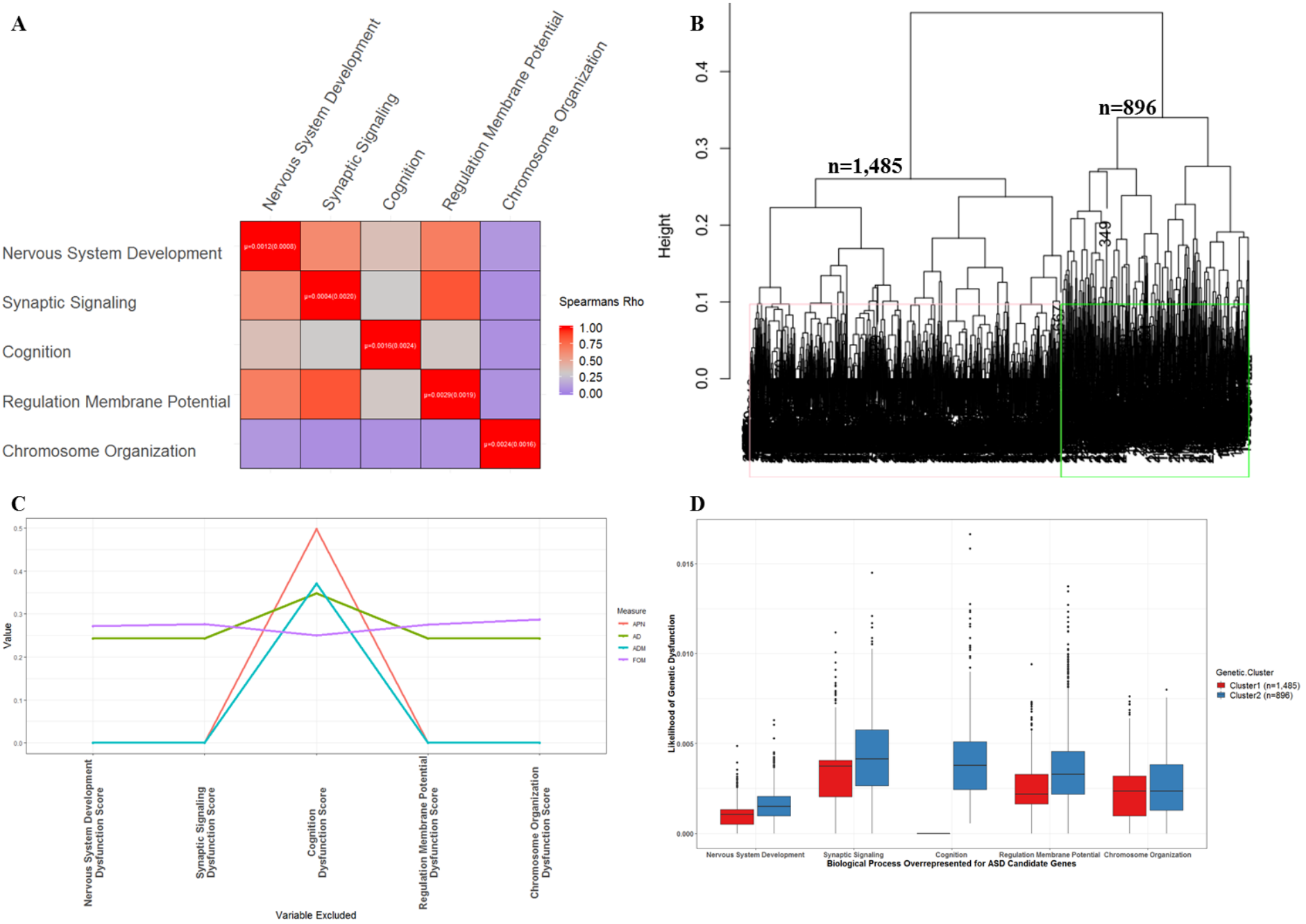
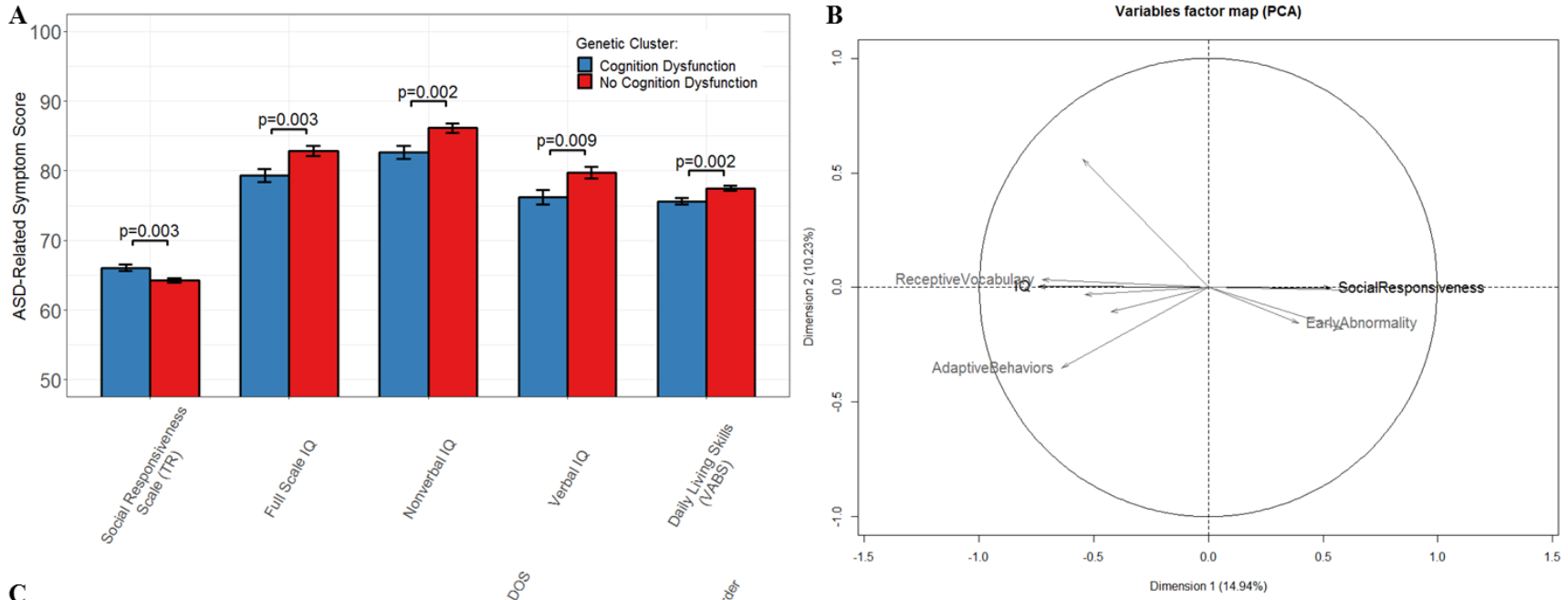


Figure 3. Clustering individuals based on overall biological process dysfunction. A) Correlation across scores reflecting dysfunction in biological processes with overrepresentation of ASD candidate genes indicates many individuals have dysfunction in >1 process. **B)** Clustering identified two distinct subgroups of individuals with more similar scores for overall biological process dysfunction (agglomerative coefficient=0.96). **C)** Sensitivity analyses indicate removing the scores had the strongest effect on stability of the clustering solution. APN=average proportion of non-overlap, AD=average distance, ADM=average distance between means, FOM=figure of merit. **D)** Evidence of dysfunction in genes involved in cognition primarily defined separation of individuals into either cluster 1 (no cognition gene dysfunction) or cluster 2 (cognition gene dysfunction).



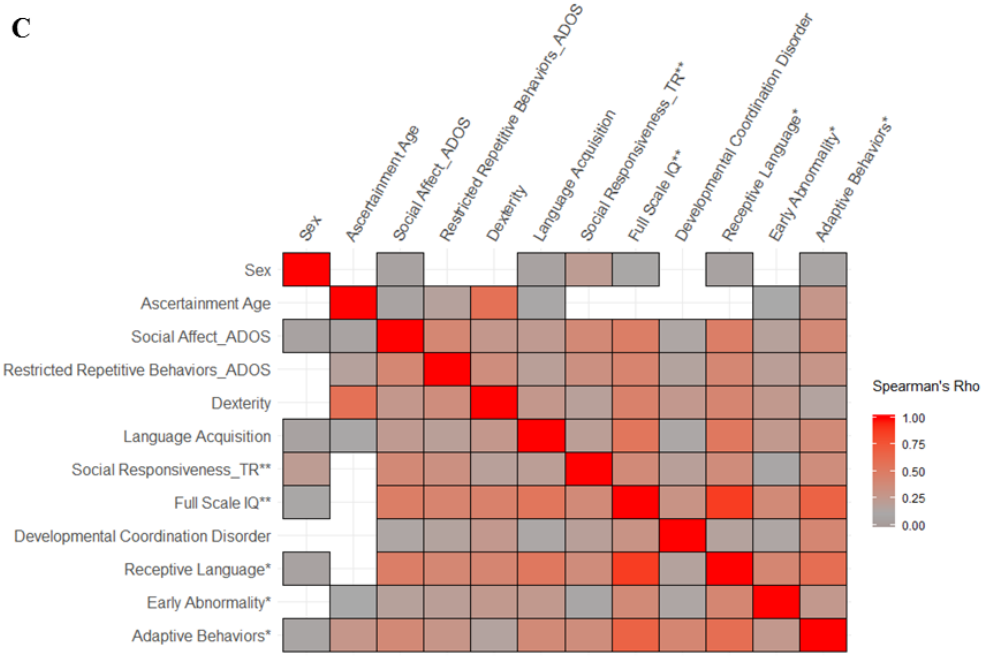


Figure 4. Relationship between genetic and phenotypic differences. A) T-tests comparing differences in the 27 quantitative ASD-related symptom measures between genetic clusters identified that social impairment was more severe and IQs and daily living skills were reduced in the cognition gene dysfunction cluster. **B)** Principal components analysis, while adjusting for correlations, of all 27 symptom measures identified that symptoms that were different between the genetic clusters majorly contributed to overall phenotype variability (as defined by Dimension 1). Black indicates symptom differences that remained significant (FDR≤0.04) following multiple testing correction, gray indicates symptom differences based on an unadjusted significance threshold (p≤0.03), and unlabeled arrows indicate symptoms that were not different but had strong contributions to phenotype variability. **C)** Significant (p<0.05) correlations are shown indicating that absolute values for many symptoms that were different between genetic clusters were correlated with those contributing majorly to overall phenotype variability**.**