


Coalitional Game Theory Facilitates Identification of Non-Coding Variants Associated With Autism

Min Woo Sun, Anika Gupta*, Maya Varma, Kelley M Paskov, Jae-Yoon Jung, Nate T Stockham and Dennis P Wall 

Departments of Pediatrics (Division of Systems Medicine), Psychiatry (by courtesy), and Biomedical Data Science, Stanford University, Stanford, CA, USA.

Biomedical Informatics Insights
Volume 11: 1–6
© The Author(s) 2019
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1178222619832859



ABSTRACT: Studies on autism spectrum disorder (ASD) have amassed substantial evidence for the role of genetics in the disease's phenotypic manifestation. A large number of coding and non-coding variants with low penetrance likely act in a combinatorial manner to explain the variable forms of ASD. However, many of these combined interactions, both additive and epistatic, remain undefined. Coalitional game theory (CGT) is an approach that seeks to identify players (individual genetic variants or genes) who tend to improve the performance—association to a disease phenotype of interest—of any coalition (subset of co-occurring genetic variants) they join. This method has been previously applied to boost biologically informative signal from gene expression data and exome sequencing data but remains to be explored in the context of cooperativity among non-coding genomic regions. We describe our extension of previous work, highlighting non-coding chromosomal regions relevant to ASD using CGT on alteration data of 4595 fully sequenced genomes from 756 multiplex families. Genomes were encoded into binary matrices for three types of non-coding regions previously implicated in ASD and separated into ASD (case) and unaffected (control) samples. A player metric, the Shapley value, enabled determination of individual variant contributions in both sets of cohorts. A total of 30 non-coding positions were found to have significantly elevated player scores and likely represent significant contributors to the genetic coordination underlying ASD. Cross-study analyses revealed that a subset of mutated non-coding regions (all of which are in human accelerated regions (HARs)) and related genes are involved in biological pathways or behavioral outcomes known to be affected in autism, suggesting the importance of single nucleotide polymorphisms (SNPs) within HARs in ASD. These findings support the use of CGT in identifying hidden yet influential non-coding players from large-scale genomic data, to better understand the precise underpinnings of complex neurodevelopmental disorders such as autism.

KEYWORDS: coalitional game theory, autism spectrum disorder, non-coding genome

RECEIVED: December 4, 2018. **ACCEPTED:** December 17, 2018.

TYPE: Precision and Individualized Medicine - Methodology

FUNDING: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported in awards to DPW by the Hartwell Foundation's Autism Research and Technology Initiative (IHART) the Stanford's BioX Program, Beckman Center, and Precision Health and Integrated Diagnostics Center (PHIND).

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Dennis P Wall, Division of Systems Medicine, Departments of Pediatrics and Biomedical Data Science, Stanford University, Stanford, CA 94305, USA. Email: dpwall@stanford.edu

Introduction

Autism spectrum disorder (ASD) is a complex neurodevelopmental disease that has frustrated traditional analysis methods due to its etiological heterogeneity and diverse and subtle phenotypic presentation. Recent large twin and sibling studies^{1–3} support the hypothesis of an extremely strong genetic effect with an estimated 64% to 91% broad-sense heritability. This study focuses on the role of inherited variation acting in a combinatorial manner for autism risk.

Epistatic interactions have been associated with autism risk^{4,33} using extremely strong priors on likely molecular pathways, but combinatorial approaches in general have been hindered by the prohibitively large search space and lack of statistical power.^{5,6} Recent studies have assessed individual epigenomic contributions but have not yet tackled cases involving multiple regulators.⁷ In general, the non-coding regions of the genome are largely unexplored from a combinatorial perspective.

Variations in non-coding chromosomal regions such as microRNAs (miRNAs) and human accelerated regions (HARs) are particularly relevant to ASD.³⁴ MicroRNAs play a major role in the regulation of gene expression and have been implicated in ASD and related disorders due to their aberrant

expression in affected individuals.^{8,9} HARs are regions of DNA that are conserved in vertebrates but differ in humans. Copy number variations (CNVs) in these regions have been implicated in autism.¹⁰

This article seeks to explore combinatorial interactions between non-coding variants using coalitional game theory (CGT). Coalitional game theory is a technique that enhances signal detection by modeling synergistic interactions between variants. Variants with the greatest average marginal contribution across all coalitions are selected as significant. Coalitional game theory has been a useful tool in the analysis of gene expression data^{11,12} and has helped to elucidate the genetic underpinnings of Alzheimer disease.¹³

We have previously shown that CGT can be effectively applied to genome-wide alteration data to highlight candidate ASD genes.¹⁴ Here, we expand on these results to include an analysis of non-coding variants in miRNA regions and HARs in genomes of 4595 individuals (2182 cases and 2413 controls). This work explores the contribution of combinatorial interactions to autism susceptibility, focusing on the role of non-coding variants. Pinpointing inherited alterations linked to ASD, such as those presented, could improve accuracy of diagnoses



and enable precise therapeutic development against the combinations of underlying genomic causes.

Methods

Data source and preprocessing

We analyzed 30×-coverage whole genome sequencing data from the Hartwell Foundation's Autism Research and Technology Initiative (iHART), as used by Gupta et al.¹⁴ Specifically, we assessed the genomes and phenotypic measurements from 756 multiplex families containing at least two children affected by ASD. The following is the distribution of the 756 families grouped by the phenotype of the children (count of children with ASD, count of neurotypical children): 380 (2, 0), 243 (2, 1), 62 (3, 0), 29 (3, 1), 23 (2, 2), 5 (3, 2), 4 (4, 0), 2 (3, 3), 2 (3, 4), 2 (4, 1), 2 (5, 0), 1 (4, 4), and 1 (5, 2) families. Removal of non-Mendelian variants for quality control eliminated sequencing error and de novo mutations.

We restricted our attention to inherited, non-coding mutations to test the hypothesis of their effect on the ASD phenotype. We examined the following three non-coding regions: dysregulated miRNA segments associated with psychiatric disorders, schizophrenia-associated miRNA regions, and HARs.¹⁵⁻¹⁷ We used MirBase and the UCSC Genome Browser to identify the chromosomal coordinates of all segments in each of the regions. Filtering the variant call format files to keep only variants located within these coordinates resulted in 261 dysregulated variants, 292 schizophrenia-associated variants, and 1962 HAR variants.^{18,19}

Replicating the representation presented in the study of Gupta et al.¹⁴ for coding variants, we encoded variant data for each of the three non-coding regions into binary matrices. Specifically, 1 indicates the presence of a variant (homozygous alternate or heterozygous) and 0 indicates the absence of the variant in a particular sample.

CGT method

We applied CGT as introduced in the work of Moretti et al.²⁰ to the whole genome sequencing data. The goal of CGT is to study the interaction among groups of players in a game. In our case, players represent variants. Let N be a finite number of players $\{1, 2, \dots, n\}$ and $T \subseteq N$ be a coalition. A coalitional game (N, v) is given by N players and a characteristic function $v(T)$, with $v(\emptyset) = 0$. Let $R = \{r_1, r_2, \dots, r_k\}$ set of samples and $M_j \subseteq N$ be the set of altered genes for a given individual j . Coalitional games can be represented by a linear combination of unanimity games as shown in the study of Moretti et al.²⁰ Unanimity game (N, u) is defined such that $u(T) = 1$ if $M_j \subseteq T$ and $u(T) = 0$ otherwise. We can write the game v as a linear combination of unanimity games

$$v = \sum_{S \subseteq N, S \neq \emptyset} \lambda_S(v) u_S$$

where

$$\lambda_S(v) = \frac{|\{j \in R | M_j = S\}|}{|R|}$$

We are interested in players who tend to increase the score of any team they join. This property is quantified by the Shapley value, which measures the average marginal contribution of each player across all possible coalitions. The following is an alternative way of representing the Shapley value using unanimity coefficients that makes the Shapley value calculation computationally more tractable as shown by Moretti et al.²⁰

$$\phi_i(v) = \sum_{S \subseteq N, i \in S} \frac{\lambda_S(v)}{|S|}$$

For a more detailed explanation of the Shapley value calculation, see the electronic supplementary material provided by Moretti et al.¹²

CGT analysis

Coalitional game theory was performed using R version 3.4.0 and Bioconductor version 3.5. To compute the Shapley value differences, the Boolean matrices corresponding to each of the three aforementioned genomic regions were split into two Boolean matrices by case and control (2182 cases and 2413 controls): B^{case} and B^{control} . We adapted the script provided by Moretti et al.¹² to compute the Shapley value differences.

We performed Comparative Analysis of Shapley (CASH), a resampling-based multiple hypothesis testing procedure introduced by Moretti et al.¹² on the matrices to filter out non-coding variants with Shapley value differences that could be high by chance. We used the MTP function from the Bioconductor package "multtest" to generate a CASH P -value for each variant. A total of 1000 non-parametric bootstrap resamples were ran with replacement on the matrices. The MTP produces unadjusted P -values calculated via simulations for each variant. We selected variants that were significant at the .05 and .01 significance levels. Application of the CGT pipeline on the non-coding region is presented in Figure 1.

Functional analyses

To elucidate potential associations with known ASD variant candidates, we cross-referenced the CGT variants list with previously published findings. We checked the chromosomal location of each of the non-coding variants using dbSNP, the National Center for Biotechnology Information database for variants known across the human genome (<https://www.ncbi.nlm.nih.gov/projects/SNP>). To search for variants according to their genomic context, we converted the chromosomal coordinates to RS IDs, using the GRCh37 build of the genome as the reference, with Kaviar (<http://db.systemsbiology.net/kaviar/cgi-pub/Kaviar.pl>). We filtered to only look for single

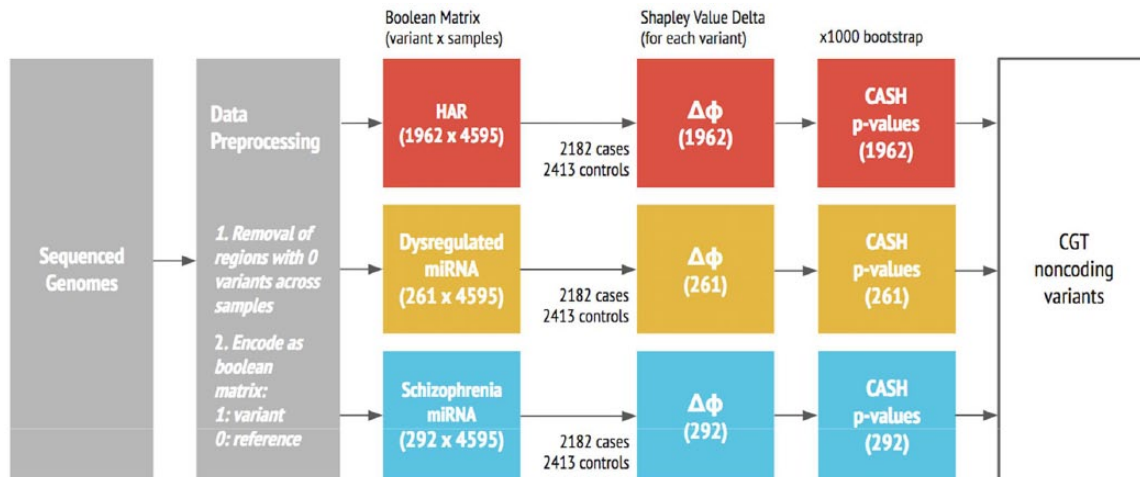


Figure 1. Data analysis flow diagram, starting from the sequenced genomes to identification of statistically significant non-coding variants through coalitional game theory. Data from human accelerated regions (red), psychiatric disorder-associated dysregulated miRNA (yellow), and schizophrenia-associated miRNA (blue) were analyzed independently.

nucleotide polymorphisms (SNPs) and recorded whether a given CGT variant fell in a functionally annotated region of biological significance.

Results

The genome pre-processing steps yielded 2182 cases and 2413 controls for 2515 non-coding variants. The differences in Shapley value between cases and controls highlighted certain non-coding variants as key contributors in the genetic coordination of ASD (Figure 1).

A total of 30 non-coding variants showed statistical significance at the .05 significance level ($P < .05$), with four of those positions significant at the .01 level ($P < .01$). Of these variants, 25 are in HARs and 5 are in schizophrenia-associated miRNA regions, 1 of which is in a dysregulated segment associated with psychiatric disorders (Table 1). Within the functionally annotated regions, the proportion of SNPs was 0.0337 for HARs (1962 variants/58 171 base pairs), 0.0169 for dysregulated segments (261 variants/15 477 bp), and 0.0216 for miRNA regions (292 variants/13 503 bp). To account for the relatedness of multiplex structure, we tried randomly sampling one sample from each family for case and

control which decreased the sample size for case from 2182 to 996 samples and control from 996 to 468 samples. Running the CGT algorithm on the simplex matrices did not identify the same variants as the 30 non-coding variants. The stark decrease in sample size precludes us from being able to definitively conclude that the family structure confounded the results. Increasing the sample size may help mitigate this challenge.

Cross-referencing CGT variants with high confidence variants previously implicated in ASD or related neurological disorders extracted the known biological functions represented by these candidate ASD variants. None of the studies included for validation were used for the selection of regions analyzed in this study. We deem non-coding SNPs as functionally relevant if the dbSNP database ensures they have been validated by multiple, independent submissions, genotyped by the HapMap project, have been sequenced as part of the 1000Genomes project, and the alleles have been observed in at least two chromosomes.

At the highest confidence level ($P < .01$), the X chromosome variant rs724600 (X:147783665) is located in the intron region of the transcriptional activator AFF2 (Figure 2, adapted

Table 1. Non-coding variants highlighted through coalitional game theory at two levels of significance.

SIGNIFICANCE	CHROMOSOMAL POSITION
$P < .05$	chr2:176990625, chr2:208297661, chr3:70655368, chr 4:182669844, chr4:35519558, chr6:31237991, chr6:31238010, chr6:31238053, chr6:31238029, chr6:31238135, chr6:31238138, chr6:64877038, <u>chr7:130496132</u> , chr12:16350701, chr12:16350704, chr14:34130323, chr14:92789362, chr14:92789365, chr14:92789363, chr16:79199908, chr16:79451667, chr17:48037183, chr20:30753270, chr20:58817633, chr21:16093111, chrX:139924185
$P < .01$	chr7:25357732, chr8:116717451, <i>chrX:139924153</i> , chrX:147783665

The 30 variants in non-coding chromosomal positions fall into one of three categories: dysregulated segments associated with psychiatric disorders (underlined), schizophrenia-associated miRNA regions (italicized), and HARs. The 26 variants listed in $P < .05$ are the subset of the 30 variants not in $P < .01$. All coordinates are relative to build GRCh37.

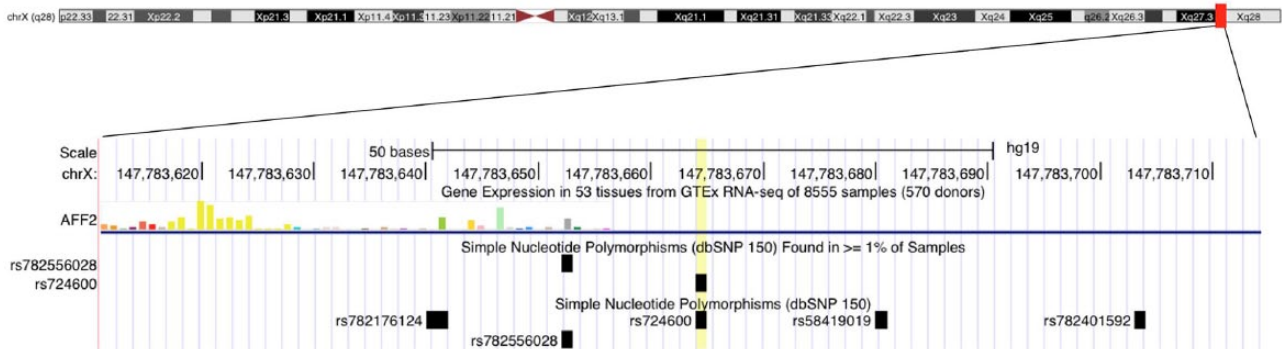


Figure 2. A non-coding variant (rs724600) identified by CGT as significantly ($P < .01$) associated with ASD. The position of the variant, X:147783665, is highlighted in yellow. This variant falls within an intron of the gene *AFF2* (indicated by the red bar). Alternative splicing and regulatory interactions could implicate the variant with different phenotypic outcomes.

from the UCSC Genome Browser), which has been implicated in both Fragile X Syndrome and ASD.^{21,22} Some individuals with Fragile X Syndrome display cognitive disability, as do a significant proportion of autistic individuals.

At the .05 significance level for non-coding variants, rs6552520 (4:182669844) falls in the intron region of *TENM3*, a promoter involved in neural development, and in regulating the establishment of proper connectivity within the nervous system.²³ Two variants in chromosome 16 (rs8051054 and rs4641754) are located in the intron region of *WWOX*—disruption of this gene is also associated with autosomal recessive spinocerebellar ataxia 12, a disease characterized by degenerative changes in the part of the brain related to movement control.²⁴ Furthermore, a group of closely located variants in chromosome 6 (rs2308628, rs1131015, rs2308622, rs1050317, and rs1050320) all fall within the intron of the gene that encodes for human leukocyte antigen-C (*HLA-C*), a major histocompatibility complex class I (MHC I) protein (Figure 3). These SNPs are in linkage disequilibrium (LD) with each other and thus likely act together (LD r^2 values are as follows: rs2308628 and rs1131015=0.937, rs2308628 and rs2308622=0.998, rs2308628 and rs1050317=0.996, rs2308628 and rs105032=0.996, rs1131015 and rs2308622=0.937, rs1131015 and rs1050317=0.934, rs1131015 and rs1050320=0.935, rs2308622 and rs1050317=0.996, rs2308622 and rs1050320=0.997, rs1050320 and rs1050317=0.999). MHC proteins play a central role in cytokine signaling in the immune system, which has been previously associated with ASD.^{25,26}

Discussion

Further characterizing the heritability of ASD across the genome remains a challenging task of widespread significance.^{27–32} In this study, we applied CGT to a large collection of whole genomes from multiplex autism families in an effort to find signal among coalitions that relate specifically to the autism phenotype. We focused our analysis on inherited, non-coding variants to pursue the hypothesis of ASD being a largely inherited disorder with multiple underpinnings extending beyond coding regions. By analyzing relative

cooperative contributions, we found 30 non-coding variants that were significantly associated with the autism phenotype.

All of the 30 non-coding variants identified through CGT are in regions that have been previously implicated in either ASD or related neurodevelopmental disorders.^{15,16} Of note are the high confidence variants that fall within introns of genes involved in processes related to neuronal development and cognitive ability, biological processes that are often impaired in ASD. The overlap between CGT's findings and prior, orthogonal studies addressing non-coding importance in the disease indicates that variants in such regions may play an important role in its progression.

A prominent goal of this work was to identify specific loci within previously implicated regions, hence the rationale for starting off with the regions that have prior hypothesized associations. A finding of particular biological relevance is that 25 of the 30 significant variants found through CGT are in HARs, which are areas that are conserved throughout vertebrates, but that are different in humans. This finding implies that from a CGT perspective, HARs are more important to autism than the other two functionally annotated regions included in this work. While HARs had more selected variants than the others, this difference is not statistically significant after accounting for region size. Furthermore, only CNVs in HARs have been previously implicated in ASD,¹⁰ so our findings support the importance of HARs while adding new variants (SNPs, rather than CNVs) to the list of candidate variants involved in the phenotypic manifestation of ASD.

Potential limitations of this work include the sparsity of functionally annotated non-coding regions as they pertain to disease, which bias and limit our ability to find links between the disease phenotype and more subtle non-coding variation. Exploring such nuanced alterations could shed light on additional high-impact molecular mechanisms in ASD. Orthogonally validating only the new HAR SNPs identified with known ASD associations ensures no functional bias in our findings. Replicating this work on additional datasets of larger cohorts and functional characterization of candidate variants

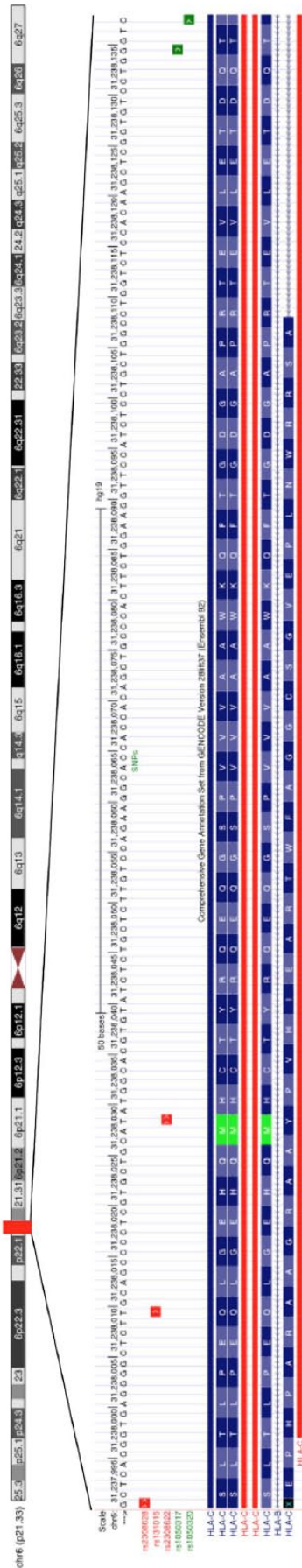


Figure 3. Example of five non-coding variants identified by CGT as significantly ($P < .05$) associated with ASD. These variants all fall within the HLA-C gene on chromosome 6. Red variants are missense, and green are synonymous (adapted from the UCSC Genome Browser).

will be necessary for evaluating the broader biological implications of our findings.

Probing into groups of co-altered non-coding variants in case subgroups may provide further insights into the mechanisms underlying ASD. Stratifying patients according to their landscape of co-alterations could improve the precision of diagnoses, and knocking out groups of genes or non-coding variants identified in functional assays could reveal potent combinations in therapeutically targeting the molecular underpinnings of ASD.

Coalitional game theory thus serves as a powerful approach to understand interactions that may only emerge in a multi-variant model. Capitalizing on the unparalleled rate of genomes being sequenced in increasingly diverse demographic and disease populations, unconventional yet rigorous tools such as CGT may accelerate the search for biomarkers, particularly in complex conditions of neurodevelopmental disorders.

Author Contributions

Min Woo Sun and Anika Gupta are co-primary contributors.

ORCID iD

Dennis P Wall  <https://orcid.org/0000-0002-7889-9146>

REFERENCES

- Colvert E, Tick B, McEwen F, et al. Heritability of autism spectrum disorder in a UK population-based twin sample. *JAMA Psychiatry*. 2015;72:415–423.
- Tick B, Bolton P, Happe F, Rutter M, Rijdsdijk F. Heritability of autism spectrum disorders: a meta-analysis of twin studies. *J Child Psychol Psychiatr Allied Discip*. 2016;57:585–595.
- Sandin S, Lichtenstein P, Kuja-Halkola R, Hultman C, Larsson H, Reichenberg A. The heritability of autism spectrum disorder. *JAMA*. 2017;318:1182–1184.
- Mitra I, Lavillaureix A, Yeh E, et al. Reverse pathway genetic approach identifies epistasis in autism spectrum disorders. *PLoS Genet*. 2017;13:e1006516.
- Phillips PC. Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems. *Nat Rev Genet*. 2008;9:855–867.
- de la Torre-Ubieto L, Won H, Stein JL, Geschwind DH. Advancing the understanding of autism disease mechanisms through genetics. *Nat Med*. 2016;22:345–361.
- Koberstein JN, Poplawski SG, Wimmer ME, et al. Learning-dependent chromatin remodeling highlights non-coding regulatory regions linked to autism. *Sci Signal*. 2018;11:eaan6500.
- Mellios N, Sur M. The emerging role of microRNAs in schizophrenia and autism spectrum disorders. *Front Psychiatry*. 2012;3:39.
- Turner TN, Hormozdiari F, Duyzend MH, et al. Genome sequencing of autism-affected families reveals disruption of putative noncoding regulatory DNA. *Am J Human Genet*. 2016;98:58–74.
- Doan RN, Bae B-I, Cubelos B, et al. Mutations in human accelerated regions disrupt cognition and social behavior. *Cell*. 2016;167:341.e12–354.e12.
- Esteban EJ, Wall DP. Using game theory to detect genes involved in autism spectrum disorder. *TOP*. 2011;19:121–129.
- Moretti S, van Leeuwen D, Gmuender H, et al. Combining Shapley value and statistics to the analysis of gene expression data in children exposed to air pollution. *BMC Bioinform*. 2008;9:361.
- Vardarajan BN. Identification of gene-gene interactions for Alzheimer’s disease using co-operative game theory. *ProQuest Dissert Theses Global*. 2013;7:S197.
- Gupta A, Sun MW, Paskov KM, Stockham NT, Jung J-Y, Wall DP. Coalitional game theory as a promising approach to identify candidate autism genes. *Pac Symp Biocomput*. 2018;23:436–447.
- Chao Y, Chen C. An introduction to microRNAs and their dysregulation in psychiatric disorders. *Tzu Chi Med J*. 2013;25:1–7.
- Beveridge NJ, Cairns MJ. MicroRNA dysregulation in schizophrenia. *Neurobiol Dis*. 2011;46:263–271.

17. Lindblad-Toh K, Garber M, Zuk O, et al. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature*. 2011;478:476–482.
18. Kozomara A, Griffiths-Jones S. MiRBase: annotating high confidence MicroRNAs using deep sequencing data. *Nucleic Acids Res*. 2013;42:68–73.
19. Kent WJ, Sugnet CW, Furey TS, et al. The human genome browser at UCSC. *Genome Res*. 2002;12:996–1006.
20. Moretti S, Patrone F, Bonassi S. The class of microarray games and the relevance index for genes. *TOP*. 2007;15:256–280.
21. Mondal K, Ramachandran D, Patel VC, et al. Excess variants in *AFF2* detected by massively parallel sequencing of males with autism spectrum disorder. *Human Molec Genet*. 2012;21:4356–4364.
22. Sahoo T, Theisen A, Marble M, et al. Microdeletion of Xq28 involving the *AFF2* (*FMR2*) gene in two unrelated males with developmental delay. *Am J Med Genet*. 2011;155A:3110–3115.
23. Young TR, Leamey CA. Teneurins: important regulators of neural circuitry. *Int J Biochem Cell Biol*. 2009;41:990–993.
24. Mallaret M, Synofzik M, Lee J, et al. The tumour suppressor gene *WWOX* is mutated in autosomal recessive cerebellar ataxia with epilepsy and mental retardation. *Brain*. 2014;137:411–419.
25. Davidson WF, Kress M, Khoury G, Jay G. Comparison of HLA class I gene sequences. Derivation of locus-specific oligonucleotide probes specific for HLA-A, HLA-B and HLA-C genes. *J Biol Chem*. 1985;260:13414–13423.
26. Goines P, Van de Water J. The immune system's role in the biology of autism. *Curr Opin Neurol*. 2010;23:111–117.
27. Abrahams BS, Geschwind DH. Advances in autism genetics: on the threshold of a new neurobiology. *Nat Rev Genet*. 2008;9:341–355.
28. Eichler EE, Flint J, Gibson G, et al. Missing heritability and strategies for finding the underlying causes of complex disease. *Nat Rev Genet*. 2010;11:446–450.
29. Gratten J, Wray NR, Keller MC, Visscher PM. Large-scale genomics unveils the genetic architecture of psychiatric disorders. *Nat Neurosci*. 2014;17:782–790.
30. Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases. *Nature*. 2009;461:747–753.
31. Robinson EB, Neale BM, Hyman SE. Genetic research in autism spectrum disorders. *Curr Opin Pediatrics*. 2015;27:685–691.
32. Sanders SJ, He X, Willsey AJ, et al. Insights into autism spectrum disorder genomic architecture and biology from 71 risk loci. *Neuron*. 2015;87:1215–1233.
33. Weiner DJ, Wigdor EM, Ripke S, et al. Polygenic transmission disequilibrium confirms that common and rare variation act additively to create risk for autism spectrum disorders. *Nat Genet*. 2017;49:978–985.
34. Varma M, Paskov KM, Jung JY, et al. Outgroup machine learning approach identifies single nucleotide variants in noncoding DNA associated with autism spectrum disorder. *Pac Symp Biocomput*. 2019;260–271.