ARTICLE | Genetics inMedicine

Genome sequencing identifies multiple deleterious variants in autism patients with more severe phenotypes

Hui Guo, PhD[1,2], Michael H. Duyzend, PhD[1], Bradley P. Coe, PhD[1], Carl Baker, BSc[1], Kendra Hoekzema, MSc[1], Jennifer Gerdts, PhD[3], Tychele N. Turner, PhD[1], Michael C. Zody, PhD[4], Jennifer S. Beighley, PhD[3], Shwetha C. Murali, MSc[1], Bradley J. Nelson, MSc[1] University of Washington Center for Mendelian Genomics[1], Michael J. Bamshad, MD[5], Deborah A. Nickerson, PhD[1], Raphael A. Bernier, PhD[3] and Evan E. Eichler, PhD[1,6]

Purpose: To maximize the discovery of potentially pathogenic variants to better understand the diagnostic utility of genome sequencing (GS) and to assess how the presence of multiple risk events might affect the phenotypic severity in autism spectrum disorders (ASD).

Methods: GS was applied to 180 simplex and multiplex ASD families (578 individuals, 213 patients) with exome sequencing and array comparative genomic hybridization further applied to a subset for validation and cross-platform comparisons.

Results: We found that 40.8% of patients carried variants with evidence of disease risk, including a de novo frameshift variant in *NR4A2* and two de novo missense variants in *SYNCRIP*, while 21.1% carried clinically relevant pathogenic or likely pathogenic variants. Patients with more than one risk variant (9.9%) were more severely affected with respect to cognitive ability compared with patients with a single or no-risk variant. We observed no instance among the 27 multiplex families where a pathogenic or likely pathogenic variant was transmitted to all affected members in the family.

Conclusion: The study demonstrates the diagnostic utility of GS, especially for multiple risk variants that contribute to the phenotypic severity, shows the genetic heterogeneity in multiplex families, and provides evidence for new genes for follow up.

Genetics in Medicine (2018) https://doi.org/10.1038/s41436-018-0380-2

Keywords: autism spectrum disorders; genome sequencing; exome sequencing; multiple-hit events; diagnostic utility

INTRODUCTION

Detailed phenotyping coupled with sequencing of patient cohorts with autism spectrum disorder (ASD) and related neurodevelopmental disorders (NDDs) have allowed the identification of hundreds of risk genes and related variants.[1–5] While useful, most of these studies have largely focused on a particular subset of patients or have imposed strict enrollment criteria that have led to phenotypic ascertainment biases. One of the most useful and deeply phenotyped cohorts, the Simons Simplex Collection (SSC),[6] for example, was restricted to simplex cases and carried a relatively lower proportion of intellectual disability (ID) cases (<25%). Such biases have likely skewed our understanding of the relative contribution of de novo and private variants as well as the potential diagnostic or predictive utility of genome sequencing (GS) in a clinical setting.

The genetic architecture of ASD has become clearer in the last decades and hundreds of risk genes and related variants have now been identified for both syndromic and idiopathic autism, based on genome-wide microarrays,[7] exome sequencing (ES),[3,4] and more recently, GS.[8,9] Nevertheless, rare genetic variants, including de novo single-nucleotide variants (SNVs) and insertions/deletions (indels) and copy-number variants (CNVs), still account for only a limited fraction of simplex cases (10–30%) (refs.[3,10]). The high heritability of ASD (50–80%) (ref. [11]) suggests that the monogenic model is likely too simplistic and that other risk variants await discovery. Although hundreds of risk variants have been

[1]Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA, USA; [2]Center for Medical Genetics, School of Life Sciences, Central South University, Changsha, Hunan, China; [3]Department of Psychiatry, University of Washington, Seattle, WA, USA; [4]New York Genome Center (NYGC), New York, NY, USA; [5]Department of Pediatrics, University of Washington School of Medicine, Seattle, WA, USA; [6]Howard Hughes Medical Institute, University of Washington, Seattle, WA, USA. Correspondence: Evan E. Eichler (eee@gs.washington.edu)
These authors contributed equally: Hui Guo and Michael H. Duyzend

Submitted 6 September 2018; accepted: 13 November 2018
Published online: 03 December 2018

identified, many of them demonstrate reduced penetrance and/or variable expressivity, including the transmission of potentially pathogenic variants from an unaffected parent to offspring.

One possibility may be that the penetrance of such risk alleles depends upon the genetic background on which these variants occur. Multiple gene-disruptive events, for example, may co-occur in probands and act synergistically or additively to lead to a more severe phenotype as suggested by several recent studies.[12–14] Among multiplex families where more than one sibling is affected, differential transmission of such variants in conjunction with additional de novo variants may lead to phenotypic variability, even when Mendelian inheritance seems likely.[15] These situations make genetic diagnosis or risk prediction of individuals with ASD and related NDDs particularly challenging.

Comprehensive variant discovery is key to disease association and gene discovery. GS is now regarded as the preferred approach to identify the full spectrum of risk variants and explore the individual-level genetic architecture. In this study, we applied three platforms—GS, ES, and array comparative genomic hybridization (aCGH)—to study a local cohort of families presenting to the clinic with at least one child with ASD features. Our goals were to maximize the discovery of potentially pathogenic variants to better understand the diagnostic utility of GS compared with a multiplatform approach, identify/validate novel disorder-related variants, and assess how the presence of multiple pathogenic variants might affect the phenotypic severity in individuals with ASD.

MATERIALS AND METHODS

Patients

We selected autism families for genetic investigation where at least one proband had been diagnosed with ASD and had been clinically evaluated at the Seattle Children's Autism Center over the past five years from the Study of Autism Genetics Exploration (SAGE) collection (Supplementary Methods). SAGE included individuals with ASD and ID as well as individuals with intact cognitive abilities and included children from both multiplex and simplex families. We only selected samples where DNA from both parents was available and probands were diagnosed with ASD by meeting cutoff criteria on the Autism Diagnostic Observation Schedule and/or Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5). In total, we investigated 180 families (578 individuals, 213 patients and 5 unaffected siblings), including 149 trios, 23 multiplex quads, 3 simplex quads, 3 multiplex five-member families, 1 simplex five-member family, and 1 seven-member family (Supplementary Figure S1, Supplementary Table S1). Clinical information was extracted from medical record review, standardized psychological evaluation, and/or parent report. For 64 affected individuals from 55 families, quantitative intelligence quotient metrics (full-scale IQ [FSIQ] and/or nonverbal IQ [NVIQ]) were available (Supplementary Table S1). All participants provided informed consent prior to participation in the study (institutional review board [IRB] protocol #44219).

GS and analysis

Sequencing and quality control (QC)

All GS samples were analyzed at the New York Genome Center (NYGC) using 1 microgram of DNA, an Illumina PCR-free library protocol, and sequencing on the Illumina HiSeq X Ten platform. Sequence analysis was performed using a Centers for Common Disease Genomics (CCDG)-compliant pipeline as described elsewhere.[16] Generated reads were aligned to the genome (GRCh38) using BWA-MEM[17] (v0.7.15), duplicate reads were marked using Picard (v2.4.1), and base scores were recalibrated using Genome Analysis Toolkit (GATK)[18] (v3.5). Quality control (QC) analysis included GS metrics estimation (Picard v2.4.1), flagstat estimation (SAMtools v1.3.1), and insert-size estimation (WHAM-Graphening v1.7.0) (Supplementary Figure S2a–d). Genomes were sequenced to a mean coverage of 35.4× (37.3× for the CCDS region). Full QC statistics are available in Supplementary Table S2. Kinship coefficients (ϕ) by KING[19] were used to assess family relationships. All family relatedness was estimated as reported (Supplementary Figure S2e, Supplementary Table S2). Mitochondrial haplogroup analysis (Supplementary Figure S2f) indicates that most families are of European descent (consistent with self-reporting).

SNV/indel calls

We used the same pipeline to call single-nucleotide variants (SNVs) and small indels as described previously.[8] In summary, SNVs and indels were called using the GATK HaplotypeCaller (v3.5) on a multiple-samples joint-calling basis and FreeBayes (v1.0.2) on a per-family basis. De novo SNVs and indels were called using a custom pipeline with family-level VCFs for both FreeBayes and GATK. First, a BCFtools (v.1.8) norm was used to left-align and normalize indels. Then, candidate sites were chosen where the father's genotype was 0/0, the mother's genotype was 0/0, and the child's genotype was either 0/1 or 1/1. Finally, we applied allele count, read-depth, and allele balance filters: the father alternate allele count = 0, mother alternate allele count = 0, child allele balance >0.25, father depth >9, mother depth >9, child depth >9, and either child genotype quality >20 (GATK) or sum of quality of the alternate observations >20 (FreeBayes). Any sites in low-complexity regions were removed from further analysis.

CNV calls

We use the same pipeline to call copy-number variants (CNVs) as described previously[8] with several changes to the callers applied. In our original pipeline, CNV detection was performed by five SV-calling programs (dCGH,[20] Genome STRiP,[21] LUMPY,[22] WHAMG,[23] and VariationHunter). In this study, we excluded VariationHunter and added CNVnator[24] and DELLY[25] for six total algorithms. Calls generated from those six CNV callers were then merged on a per-sample

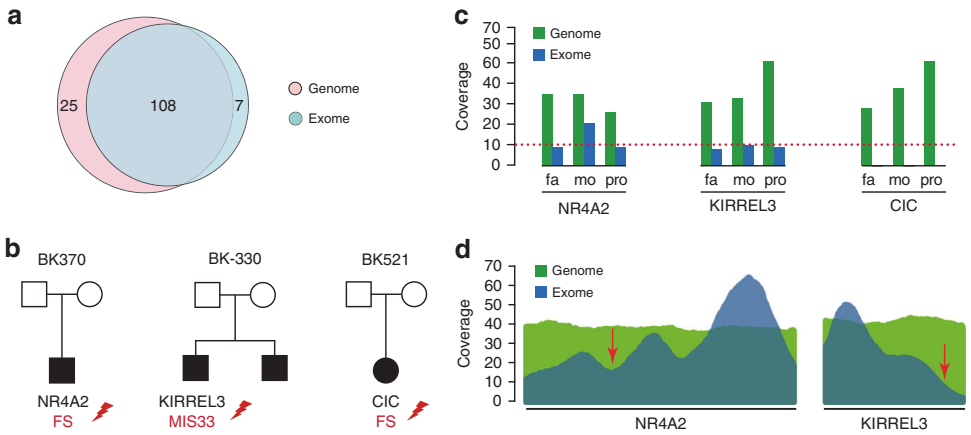

Fig. 1 Genome sequencing (GS) versus exome sequencing (ES) detection. (**a**) Comparison of the number of de novo coding single-nucleotide variants (SNVs) and indels detected in exome-targeted regions by GS and ES. (**b**) Families with de novo variants in neurodevelopmental disorder (NDD) risk genes missed by ES. (**c**) Comparison of ES and GS coverage of variants in *NR4A2*, *KIRREL3*, and *CIC* in mother, father, and proband. (**d**) Mean coverage across all samples sequenced by both GS and ES across *NR4A2* and *KIRREL3*. Arrows point to location of the two variants discovered by GS but missed in ES.

basis with calls being reported with the breakpoints from one algorithm and supporting algorithms annotated. Breakpoint selection was accomplished by our previously described[8] algorithm, which utilizes a combination of relative known breakpoint accuracy (Genome STRiP, LUMPY, WHAMG, DELLY, CNVnator, and finally dCGH), read depth, and SVtyper support. In addition, we manually visualized all high-quality, private de novo CNVs using samplot (https://github.com/ryanlayer/samplot) and WSSD read-depth line plot, and only considered the ones that passed our visualization for further analysis.

ES and microarray-based CNV analysis

A subset of families were also subjected to ES and CNV analyses using standard procedures (Supplementary Methods).

RESULTS

GS and variant discovery

We performed GS on all 180 families (578 DNA samples) and applied GATK and FreeBayes to detect SNVs/indels (Supplementary Figure S3). After filtering, 35,384 putative de novo SNVs/indels were detected by both GATK and FreeBayes. We randomly selected 150 putative de novo SNVs for validation distributed from unique ($n = 50$), ancient repetitive ($n = 50$), and recent repetitive ($n = 50$) regions as described previously.[8] After excluding variants that cannot be amplified or reliably Sanger sequenced, we estimated a validation rate of 95.5% (42/44) in unique regions, 65.7% (23/35) in ancient repeat regions, and 16.2% (6/37) in recent repeat regions. Correcting for differential validation, we estimated a genome-wide rate of 79 de novo SNVs per child.

We selected 70 families (90 affected offspring, 230 individuals) for ES (Supplementary Figure S1, Supplementary Methods). To compare the ES and GS results, we analyzed both data sets using the same analysis software and filtering pipeline for de novo variants, while also applying GATK hard filtering to remove high-frequency variants (minor allele frequency >1% in ExAC). Of the putative de novo coding events detected by either GS or ES, 108 variants were supported by both, 25 by GS only, and 8 by ES only. We attempted to validate all GS-only and ES-only variants. Six GS-only variants could not be validated by Sanger sequencing, one ES-only variant was maternal, and all the others were validated as de novo (Fig. **1a**, Supplementary Table S3). Of the GS-only de novo events, we discovered a frameshift variant in *NR4A2* and a missense variant in *KIRREL3* (Combined Annotation Dependent Depletion [CADD] score = 33), both NDD-associated genes missed by ES data due to reduced sequence coverage across these exons (Fig. **1**b–d). GS also detected eight rare nonsynonymous variants not present in the exome target, including a frameshift variant in *CIC*, a known autism-associated gene[26] (Fig. **1b**). No variants within NDD-associated genes were identified by only ES.

In addition to SNVs/indels, we detected 3498 private CNVs (specific to a SAGE family) in the offspring of which 623 intersected at least one RefSeq gene (Supplementary Table S4). After visualizing the de novo CNVs, we predicted 41 private CNVs to be de novo, including 19 deletions ranging in size from 302 bp to 6.6 Mbp, and 22 duplications ranging in size from 1 kbp to 9.2 Mbp.

Classification of disorder-related SNVs/indels

We set criteria to define and comprehensively characterize disorder-related SNVs/indels in this study (Supplementary Methods, Supplementary Table S5). We identified and validated 56 de novo disorder-related SNVs/indels (15 likely gene-disruptive [LGD], 40 missense, and 1 in-frame) from 52 genes in 49 affected individuals (Table **1**, Supplementary Table S6). We identified and validated seven inherited disorder-related SNVs/indels (Table **1**, Supplementary Table S7). We estimate that 25.4% of the affected offspring carry disorder-related SNVs/indels in one or more candidate genes. To evaluate how many patients carried clinically

Table 1 Disorder-related single-nucleotide variants (SNVs)/indels

Gene	Func[a]	NT_AAChange[b]	Inheritance[c]	Clinical significance[d]	Sample ID	Gene sets[e]
ARID1B[f]	SG	NM_020732:c.C2692T:p.R898*	DN	P	BK-303-03	SFARI(1)|DD93|BC253|ID526|MG237
ARID1B[f]	SG	NM_020732:c.C2536T:p.Q846*	DN	P	BK-418-03	SFARI(1)|DD93|BC253|ID526|MG237
MED13L	SG	NM_015335:c.G4076A:p.W1359*	DN	P	BK-242-03	SFARI(2)|DD93|BC253|ID_C628
WDFY3	SG	NM_014991:c.A2932T:p.R978*	DN	P	BK-283-03	SFARI(2)|BC253|ID_C628
RERE	SG	NM_001042681:c.C2278T:p.Q760*	DN	P	BK-186-03	SFARI(4)
ADNP	FS	NM_001282532:c.2250_2274del:p.V751Mfs*13	DN	P	BK-246-05	SFARI(1)|BC253|ID_C628
CIC	FS	NM_001304815:c.884_893del:p.A295Pfs*26	DN	P	BK-521-03	SFARI(2)
OPHN1	FS	NM_002547:c.932_933insCA:p.Q311Hfs*7	DN	P	BK-359-03	SFARI(3)|ID526
CHD8	SP	NM_020920:c.2682-2 A > G	DN	P	BK-186-03	SFARI(1)|DD93|BC253|MG237
FOXP1	SP	NM_001244815:c.1728 + 1- > TGCAGCTTTACAG	DN	P	BK-248-03	SFARI(2)|DD93|BC253|ID526|MG237
ASXL1	SG	NM_015338:c.C1045T:p.Q349*	MI	P	BK-483-03	DD93|ID526
GNB1	MIS	NM_001282539:c.G229A:p.G77S	DN	P	BK-328-03	-
MEF2C	MIS	NM_001193348:c.C43T:p.R15C	DN	P	BK-192-03	SFARI(4)|DD93|BC253|ID526|MG237
KMT5B	MIS	NM_016028:c.G791C:p.W264S	DN	P	BK-255-03	SFARI(1)|DD93|BC253|ID526
NR4A2	FS	NM_006186:c.601_602insGTCC:p.P201Rfs*82	DN	LP	BK-370-03	BC253|ID526|MG237
SMC3	MIS	NM_005445:c.C2413T:p.R805C	DN	LP	BK-255-03	SFARI(4)|BC253|MG237|ID_C628
STXBP1	MIS	NM_001032221:c.C560T:p.P187L	DN	LP	BK-277-03	SFARI(3)|DD93|BC253|ID526|MG237
GRIA1	MIS	NM_001258021:c.G2264A:p.G755D	DN	LP	BK-401-03	SFARI(2)|MG237 |ID_C628
KAT6A	MIS	NM_001305878:c.C1582T:p.P528S	DN	LP	BK-523-03	SFARI(3)|DD93|BC253
SATB2	MIS	NM_001172509:c.A1861T:p.I621F	DN	LP	BK-550-03	SFARI(4)|DD93|BC253|ID526|MG237
POGZ	MIS	NM_015100:c.G3048T:p.E1016D	DN	LP	BK-219-03	SFARI(1)|DD93|BC253|MG237|ID_C628
DYNC1H1	MIS	NM_001376:c.A13088C:p.K4363T	DN	LP	BK-283-03	SFARI(3)|DD93|BC253|ID526
SYNCRIP[f]	MIS	NM_001159676:c.T629C:p.F210S	DN	PDR	BK-611-01	BC253|ID_C628
SYNCRIP[f]	MIS	NM_001159676:c.1573_1574CA_TT:p.Q525L	DN	PDR	BK-252-03	BC253|ID_C628
JMJD1C	FS	NM_032776:c.667_668insA:p.M223Nfs*3	DN	PDR	BK-418-03	SFARI(4)
THBS1	SG	NM_003246:c.C2875T:p.R959*	DN	PDR	BK-396-04	SFARI|ID_C628
LARP4B	FS	NM_015155:c.801_802del:p.C267*	DN	PDR	BK-205-03	BC253
MCM3AP	FS	NM_003906:c.276delC:p.F93Lfs*42	DN	PDR	BK-302-03	ID_C628
KIRREL3	MIS	NM_032531:c.G1985A:p.R662H	DN	PDR	BK-330-03	SFARI(3)|ID526
RPS6KA2	MIS	NM_001006932:c.G1720A:p.G574R	DN	PDR	BK-222-03	SFARI(4)
TUB	MIS	NM_003320:c.G139A:p.G47S	DN	PDR	BK-141-03	MG237
DMXL2	MIS	NM_001174116:c.C6137T:p.A2046V	DN	PDR	BK-175-03	SFARI(4)
TOP1	MIS	NM_003286:c.A1217T:p.H406L	DN	PDR	BK-254-03	SFARI(5)|ID_C628
SLC9A3	MIS	NM_001284351:c.C914T:p.S305L	DN	PDR	BK-523-03	MG237
KCNK9	MIS	NM_001282534:c.C907T:p.R303C	DN	PDR	BK-227-03	ID_C628
SPG11	MIS	NM_001160227:c.C4955G:p.T1652R	DN	PDR	BK-599-07	MG237
SLITRK5[f]	MIS	NM_015567:c.G175T:p.G59C	DN	PDR	BK-354-03	SFARI|MG237
SLITRK5[f]	MIS	NM_015567:c.C976T:p.P326S	DN	PDR	BK-372-03	SFARI|MG237
CLSTN3	MIS	NM_014718:c.T599C:p.I200T	DN	PDR	BK-187-04	SFARI(5)

Table 1 continued

Gene	Func[a]	NT_AAChange[b]	Inheritance[c]	Clinical significance[d]	Sample ID	Gene sets[e]
SMG9	MIS	NM_019108:c.A947G:p.H316R	DN	PDR	BK-198-03	ID_C628
MYT1	MIS	NM_004535:c.C2138T:p.S713F	DN	PDR	BK-516-03	ID_C628
ACACB	MIS	NM_001093:c.A1963G:p.S655G	DN	PDR	BK-162-03	ID_C628
GRM5	MIS	NM_001143831:c.A523G:p.T175A	DN	PDR	BK-307-03	SFARI|ID_C628
CSMD1	MIS	NM_033225:c.A2381C:p.H794P	DN	PDR	BK-146-03	SFARI
SDK1	MIS	NM_152744:c.G6016A:p.E2006K	DN	PDR	BK-401-03	SFARI
TRANK1[f]	MIS	NM_014831:c.G2701A:p.V901I	DN	PDR	BK-358-03	-
TRANK1[f]	MIS	NM_014831:c.C6326A:p.T2109K	DN	PDR	BK-413-03	-
RRP8	MIS	NM_015324:c.G803A:p.R268H	DN	PDR	BK-590-01	BC253
UPF2	MIS	NM_080599:c.G91T:p.V31L	DN	PDR	BK-328-03	SFARI(5)|ID_C628
PTPRT	MIS	NM_007050:c.G548A:p.R183Q	DN	PDR	BK-261-04	SFARI|ID_C628
KCNS3	MIS	NM_001282428:c.G601A:p.A201T	DN	PDR	BK-144-03	BC253
PAFAH1B3	MIS	NM_001145939:c.T571C:p.Y191H	DN	PDR	BK-460-03	ID_C628
NALCN	MIS	NM_052867:c.C682T:p.H228Y	DN	PDR	BK-428-03	ID_C628
BIRC6	MIS	NM_016252:c.A3931G:p.I1311V	DN	PDR	BK-280-03	SFARI|ID_C628
STARD9	MIS	NM_020759:c.G4802A:p.R1601Q	DN	PDR	BK-473-03	BC253
FASN	MIS	NM_004104:c.G2719A:p.V907I	DN	PDR	BK-135-03	ID_C628
DST	NFS	NM_001144769:c.97_98insCCACCATCG:p.V33delinsATIV	DN	PDR	BK-522-03	SFARI(4)|ID_C628
TNRC6B	SP	NM_001024843:c.46-2A>G	PI	PDR	BK-182-03	SFARI(2)
DLG4	SP	NM_001365:c.20-1G>C	MI	PDR	BK-201-03	SFARI|Coe124|ID_C628
LAMB1	FS	NM_002291:c.144delG:p.K49Sfs*4	PI	PDR	BK-445-03	SFARI(3)
ATP10A	SG	NM_024490:c.C2397A:p.Y799*	MI	PDR	BK-254-03	SFARI(3)
ELP4	FS	NM_001288725:c.284delC:p.S95Yfs*64	PI	PDR	BK-219-03	SFARI(3)
LZTR1	FS	NM_006767:c.772delT:p.F258Lfs*93	MI	PDR	BK-384-03	SFARI(3)|ID_C628

[a]*FS* frameshift, *NFS* nonframeshift, *SG* stop-gain, *SP* splicing site.
[b]Canonical isoform presented.
[c]*DN* de novo, *MI* maternal inheritance, *PI* paternal inheritance.
[d]*LP* likely pathogenic, *P* pathogenic, *PDR* potentially disorder-related variants beyond the clinically relevant P and LP variants.
[e]List of neurodevelopmental disorder (NDD) gene sets that the genes belong to. SFARI, 970 ASD-associated genes from SFARI gene database; *DD93*, 93 developmental delay genes identified from DDD study 2017; BC253, 253 significant NDD genes from Coe et al.;[5] MG237, 237 NDD genes with nominal significance for enrichment or clustering of missense de novo variants from Geisheker et al.;[41] ID526, ID_C628, 526 intellectual disability (ID) genes and 628 candidate ID genes curated by Gilissen et al.[42] (also see Supplementary Methods for details and corresponding references).
[f]Recurrent variant identified.

relevant variants, we further curated the disorder-related SNVs/indels following the standards and guidelines for the interpretation of sequence variants from the American College of Medical Genetics and Genomics (ACMG).[27] In total, we classified 14 as pathogenic and 7 as likely pathogenic variants. Clinically relevant pathogenic or likely pathogenic SNVs/indels account for 8.9% of patients.

Multiple occurrences of de novo variants were identified in three NDD genes. We identified and validated two LGD variants in *ARID1B*, two missense variants in *SYNCRIP*, and two missense variants in *SLITRK5* (Table 1). *ARID1B* de novo variants have been strongly implicated in ASD and ID. Recurrent *SYNCRIP* LGD variants were identified in ID[1] and the probability of finding two de novo missense variants within this gene in a cohort of this size is significantly low ($P = 8.7 \times 10^{-7}$, $P_{adj} = 0.02$, one-tailed binomial test; Supplementary Methods). For *SLITRK5*, after integrating two de novo missense variants from denovo-db v.1.5, a compendium of primarily human de novo NDD variants,[28] we identified a potential cluster of them for future investigation (Supplementary Figure S4).

We also identified and validated variants in two other NDD genes, *NR4A2* and *MYT1*, with putative missense clusters. We discovered a de novo frameshift variant in *NR4A2*, of which a cluster of four missense variants from denovo-db associated with the DNA-binding domain was observed (Supplementary Figure S4). We similarly discovered a de novo missense variant in *MYT1* (CADD score = 25), a paralog of the autism-associated *MYT1L*. Sporadic case reports of *MYT1* de novo missense variants were identified in patients with oculo-auriculo-vertebral spectrum (OAVS), which presents with autism-like features.[29] Interestingly, a de novo missense variant, which is in close proximity to the one identified in this study, was recently detected in a patient with developmental delay from denovo-db (Supplementary Figure S4).

Classification of disorder-related CNVs

We also set criteria to define the disorder-related CNVs in this study (Supplementary Methods, Supplementary Table S8) and identified 46 disorder-related CNVs and two abnormal karyotypes (48, XXXY and 47, XXY). We attempted to validate these CNVs by two approaches: aCGH validation for relatively large CNVs (>50 kbp) and Sanger sequencing of small deletions that could not be assessed by the aCGH platform (Supplementary Methods). We successfully validated 45/46 disorder-related CNVs or abnormal karyotypes (2/2) accounting for 20.2% of participants (Supplementary Table S9). Once again, we further triaged these 47 CNVs or abnormal karyotypes into those that are clinically relevant following the ACMG standards and guidelines.[30] In total, we classified 19 as pathogenic and 11 as likely pathogenic variants accounting for 13.1% of the patients.

Clinically relevant pathogenic CNVs included 13 de novo (10 del, 3 dup) and 4 inherited (1 del, 3 dup) from 11 genomic regions among 12 affected offspring (Supplementary Table S9). One pathogenic CNV region was recurrent and observed in multiple families: the chromosome 16p11.2 CNV (5 del, 2 dup). The remaining ten pathogenic CNVs were observed in eight de novo instances: 8p12-11.1 duplication, 5p15.33 deletion, 6p25.3-25.2 deletion, 17p12 deletion, 16p11.2 distal deletion, 15q11-13 duplication, 1q42.11-42.12 deletion, 22q13.32-13.33 deletion; one paternal: 17q12 duplication; and one maternal: 1q21.1-21.2 duplication.

Likely pathogenic CNVs included two de novo (1 del, 1 dup) and nine inherited CNVs (6 del, 3 dup) from seven genomic regions among ten affected offspring from six families (Supplementary Table S9). A chromosome 22 duplication (1.6 Mbp) was identified in the five-member family. This segmental duplication-mediated duplication was transmitted from the affected father (high-functioning autism formerly classified as Asperger syndrome) to two affected children and was recently identified in an ASD family.[31] A deletion involving almost the entire 3' untranslated region (UTR) of *FOXP2* was detected in the seven-member family and transmitted from the affected father to 4/5 affected siblings. The other likely pathogenic CNVs include large de novo deletions or duplications, or encompass known neurodevelopmental or neuropsychiatric genes. The set includes a 18p11 duplication (4.7 Mbp, de novo), 18p11 deletion (3.2 Mbp, de novo), 2q32.1 duplication (2.8 Mbp, maternal), 13q21.1 deletion (2.9 Mbp, maternal), and *TCF4* deletion (4.5 kbp, paternal).

Patients with multiple variants and phenotypic severity

We classified pathogenic or likely pathogenic variants according to ACMG guidelines as stated above. Other disorder-related SNVs or CNVs were classified as potentially disorder-related (PDR) variants (Supplementary Methods). We estimate that 40.8% (87/213) of the affected offspring carry de novo or rare events in pathogenic, likely pathogenic, or PDR variants, while 21.1% (45/213) carry one or more events that would be classified as pathogenic or likely pathogenic (Fig. 2). One goal of this study is to understand the individual-level genetic architecture of ASD and determine if patients with multiple events are more clinically impaired. Considering only validated events among NDD candidate genes, we identified 21 affected offspring from 21 families with more than one event, accounting for 9.9% (21/213) of the affected offspring (Fig. 3). A subset of these (seven affected individuals from seven families) carried multiple events in different genes that would be classified as pathogenic or likely pathogenic. We observed families with all combinations of variant event types (e.g., SNV + SNV, CNV + CNV, SNV + CNV), which are only accessible in a single experiment by GS. Neither ES nor aCGH could detect 52.4% multiple-hit events in this study. Those include the combination of both SNVs and CNVs, and small CNVs (<50 kbp) that could not be detected by the aCGH platform used in this study (Fig. 2a).

To assess the relationship between the number of disorder-related variants and phenotypic severity, we performed two analyses. First, we compared the median distribution of the IQ

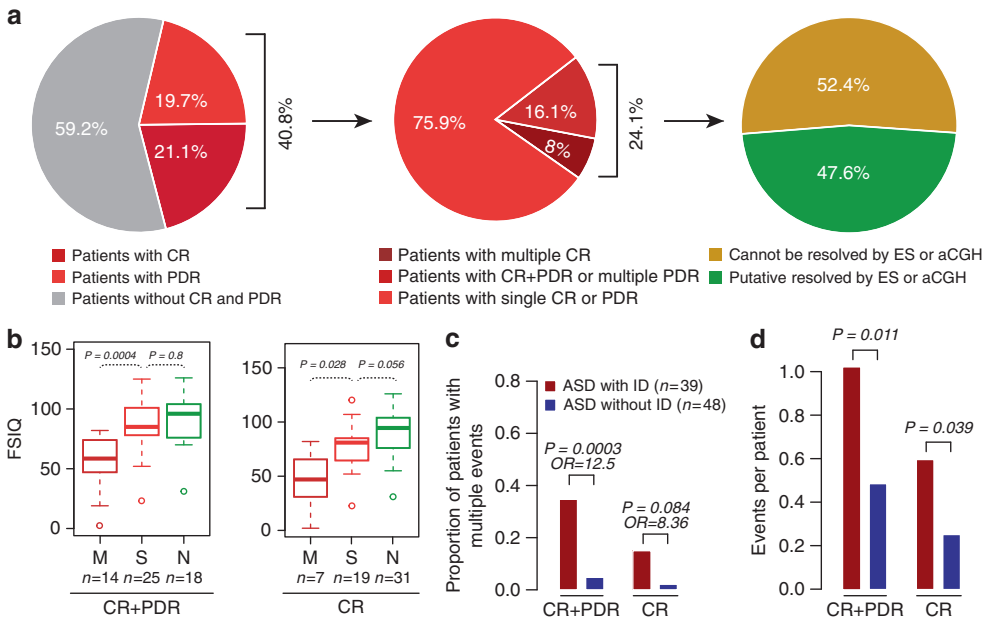

Fig. 2 Diagnostic yield of genome sequencing (GS) and phenotypic severity in multiple-hit patients. (**a**) Diagnostic yield of GS. Pie chart on left compares the proportion of patients with clinically relevant (CR) (pathogenic or likely pathogenic), potentially disorder-related (PDR) (e.g., candidate neurodevelopmental disorder [NDD] risk genes), and no risk variant identified. Middle pie chart compares the number of patients with multiple variants versus those with a single event. Right pie chart compares the number of such multiple events that can be resolved by exome sequencing (ES) or array comparative genomic hybridization (aCGH) (green) and those that cannot (yellow). (**b**) Comparison of full-scale IQ (FSIQ) for patients with multiple events (M), single events (S), and no event (N). Left panel considers both CR and PDR events (CR + PDR); right panel considers CR events only. (**c**) Burden analysis comparing the proportion of autism spectrum disorder (ASD) patients with and without intellectual disability (ID) for all CR + PDR or CR only events. (**d**) Overall burden analysis for patients with and without ID considering all CR + PDR or CR only events. *P* values were adjusted for multiple comparisons.

data across affected offspring with multiple genetic events, single events, and no event. We observed significantly lower FSIQ ($P_{adj} = 0.0004$, Mann–Whitney U test) in affected offspring with multiple variants in different NDD risk genes when compared with the affected offspring with a single event (Fig. 2**b**). Although there is a weak trend toward lower IQ between affected offspring with a single event and no identified genetic lesion, this difference does not reach significance ($P_{adj} = 0.8$, Mann–Whitney U test). When we restrict the analysis to variants deemed to be pathogenic or likely pathogenic, the trend still holds ($P_{adj} = 0.028$) (Fig. 2**b**).

Second, we performed a burden analysis comparing the proportion of individuals with multiple hits with (ASD+ID) and without (ASD–ID) ID. As expected, we observed more individuals with multiple NDD risk gene variants in the ASD +ID group compared with the ASD–ID group ($P_{adj} = 0.0003$, odds ratio [OR] = 12.5, Fisher's exact test) (Fig. 2**c**). The trend still holds (OR = 8.36) when we restrict to pathogenic or likely pathogenic events, although not yet significant ($P_{adj} = 0.084$). The same trend was also observed in the overall burden analysis considering all disorder-related events ($P_{adj} = 0.011$, one-way analysis of variance [ANOVA]) or clinically relevant events only ($P_{adj} = 0.039$) (Fig. 2**d**).

Genetic and clinical heterogeneity in multiple affected siblings

In this cohort, there are 27 multiplex families, including a total of 60 affected offspring (Supplementary Figure S1). We identified five clinically pathogenic or likely pathogenic variants in nine affected offspring from five of these families (Supplementary Figure S5). This suggests a reduced diagnostic yield (15%) when compared with simplex and trio families (23%) ($P = 0.35$, OR = 0.66, Fisher's exact test). Interestingly, there is no case among these multiplex families where a pathogenic or likely pathogenic variant was transmitted to all affected members in the family, thus implying considerable locus heterogeneity (Supplementary Figure S5) as previously observed.

It is also noteworthy that some of the clinical variability within these families correlates with the number and overall impact of such gene-disruptive variants. In family BK246 (Supplementary Figure S5), for example, the proband with two variants (the de novo frameshift variant in *ADNP* and the paternally inherited 1.6-Mbp duplication) is the only individual in the family with severe ID (FSIQ = 19, NVIQ = 20). This contrasts with the sibling with only the inherited duplication who is diagnosed with anxiety and attention-deficit/hyperactivity disorder (ADHD) without ID (FSIQ = 94, NVIQ = 100). The third sibling with no detected pathogenic variants has the highest IQ (FSIQ = 104, NVIQ = 107), although the difference is still within the realm of test–retest noise. Similarly in family BK599 (Supplementary Figure S5), the child (BK599.07) with both variants (4-kbp deletion of the 3' UTR of *FOXP2* and a de novo missense variant within *SPG11*) is more impaired than the other siblings (FSIQ = 60, NVIQ = 64). Once again, the sibling

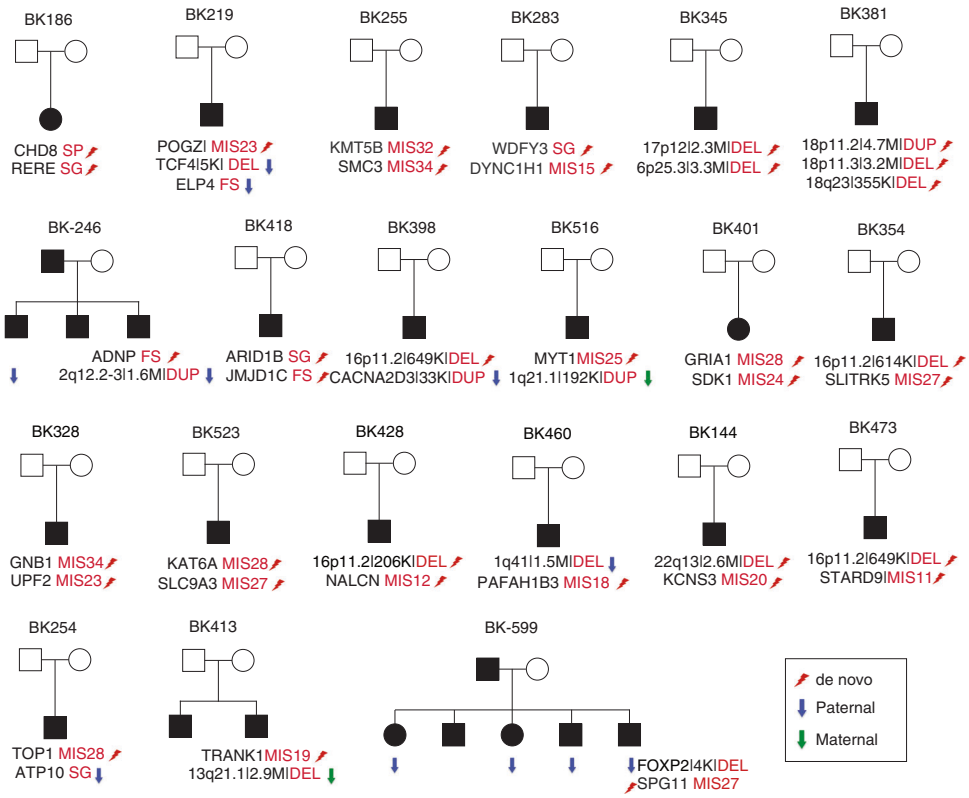

Fig. 3 Patients with multiple variants. Family structures shown for autism spectrum disorder (ASD) patients with multiple disorder-related variants. The first seven patients from seven families carried multiple pathogenic or likely pathogenic variants. De novo (lightning bolt), paternally (blue arrow), and maternally (green arrow) inherited single-nucleotide variant (SNV) and copy-number variant (CNV) (duplication or deletion) events are indicated as well as the severity of the missense variants as determined by Combined Annotation Dependent Depletion (CADD) score (i.e., MIS27 denotes a missense variant with a CADD score of 27).

without any disorder-related event has the highest IQ (NVIQ = 129) among all affected siblings.

DISCUSSION

In this study, we set out to determine the phenotypic and genotypic heterogeneity of a minimally ascertained clinical cohort of families with ASD. We demonstrated the diagnostic utility of GS for the discovery of disease-related variants, especially for multiple rare risk variants that contribute to the phenotypic severity of ASD; the genetic heterogeneity within multiplex families with ASD; and the identification of new ASD risk genes for future investigation. Given the narrow clinical definition of pathogenic and likely pathogenic variants, we used available neurodevelopmental gene lists and the literature to define the PDR variants. The expanded definition was necessary to explore the full heterogeneity and correlation with phenotypic severity.

Approximately 10% of the ASD-affected offspring in this study carried multiple risk variants, and multiple hits correlated with increased phenotypic severity. We observed a significant difference in FSIQ and NVIQ scores when comparing affected individuals with multiple hits with those with one or zero hits. The finding is crucial from a clinical perspective as the genetic workup of children with autism and developmental delay often ends if a likely pathogenic SNV or CNV is found by microarray or ES. Because such cases are unlikely to proceed to GS, variants contributing more significantly to the phenotype may remain undiscovered unless such families are subject to full GS. Furthermore, this finding provides support to the oligogenic model of ASD, specifically where multiple rare disruptive variants lead to more severe phenotypes.

We observed considerable genetic heterogeneity within families consistent with earlier observations.[15,32] Although such multiplex families are thought to share the same genetic risk event(s), 92% (12/13) of the families failed to segregate phenotype and genotype faithfully when a disorder-related event was discovered (i.e., affected individuals that did not carry the disorder-related variant were present in most families albeit such members tended to be less severely affected). This genetic heterogeneity was not only restricted to de novo variant events (e.g., the de novo LGD within *ADNP* in family BK246 or de novo 16p11.2 duplication in family BK187) but also observed for transmitted variants (e.g., paternally inherited 16p11.2 duplication in family BK313 and a maternally inherited 13q21.3 deletion in family BK413). These complicated combinations of disorder-related events and phenotypic diversity within families highlight the importance of GS for affected and unaffected members prior to genetic counseling of families.

Variants discovered in this study add substantial evidence confirming ASD candidate genes described in the literature, including but not limited to *NR4A2*, *SYNCRIP*, *MYT1*, and *TRANK1*. *NR4A2* encodes a transcription factor essential for the differentiation of dopaminergic neurons.[33] Recently, several de novo deletions covering *NR4A2* only or *NR4A2* and the adjacent *GPD2* were reported in patients with ASD, ID, and/or language impairment.[34] In this study, we identified a de novo frameshift variant and a cluster of de novo missense variants within the DNA-binding domain of the predicted protein. Similarly, an excess of de novo truncating variants within *SYNCRIP* was previously identified in patients with ID.[1] In this study, we identified two individuals with ASD with *SYNCRIP* de novo missense variants. SYNCRIP is a component of messenger RNA (mRNA) granules bound for the dendrites where it contributes to synaptic plasticity[35] and, in *Drosophila*, is thought to play a role in the decommissioning of neural stem cells.[36] We also identified two individuals with de novo missense variants in *TRANK1*. Patients with ID and de novo variants in *TRANK1* were previously reported and the locus has been associated with bipolar disorder in different genome-wide association studies;[37,38] however, the function of this gene is largely unknown. *MYT1* is a paralog of the autism-associated gene *MYT1L*. Sporadic case reports of *MYT1* de novo missense variants were reported in patients with OAVS, which often presents with autism-like features.[29]

As a diagnostic test, consistent with other recent reports,[8,12] GS provides a slight advantage over ES for the detection of protein-encoding risk variants. In this study, GS enabled the discovery of potential de novo ASD-associated variants missed by ES, including a frameshift variant in *NR4A2*, a frameshift variant in *CIC*, and a missense variant in *KIRREL3*. Improvements in capture design and increases in ES coverage have continued to minimize such differences with false negative rates now estimated at less than 2.5% (ref. [8]). Most GS advantages lie in the greater uniformity of sequence coverage and improved detection of gene-disruptive CNVs. This is especially relevant with respect to the detection of multiple variant events where ES and aCGH could detect and confirm fewer than half of such cases independently in this study.

The complete genetic architecture of ASD remains to be elucidated. Analysis of the SAGE cohort demonstrated the utility of GS in a clinical setting. The ability to capture most genetic variants enables the discovery of multiple hits that are clinically relevant in determining the severity of presentation of ASD. Our analysis showed that in multiplex families, it is crucial to not assume Mendelian inheritance and suggests that a combination of factors, including genetic background, play a role in phenotypic severity. Moving forward, it will be important to elucidate the full spectrum of genetic variation in clinically relevant cohorts. This will include not only the characterization of unrelated ASD patients with different variants in the same gene but also the comparison of affected siblings with one or more risk alleles within the same family. In addition, GS will provide a platform for assessing the contribution of noncoding regulatory variants and the interplay between rare and common variants in contributing to the risk of ASD and other NDDs.[39,40] Such studies will require much larger sample sizes but will provide an unprecedented opportunity to develop an integrated model for the genetic architecture of autism that will be valuable for future clinical diagnosis.

ELECTRONIC SUPPLEMENTARY MATERIAL

The online version of this article (https://doi.org/10.1038/s41436-018-0380-2) contains supplementary material, which is available to authorized users.

ACKNOWLEDGEMENTS

We thank Sunday Stray, Mary Eng, James Moore, Hannah Kortbawi and Anne Thornton from the laboratory of Mary-Claire King for isolation of DNA from whole blood and Tonia Brown for manuscript editing. We are especially grateful to the families who participated in the SAGE study. This work was supported by the following grants: the Simons Foundation Autism Research Initiative (SFARI 303241) and National Institutes of Health (NIH R01MH101221) to E.E.E., NIH (R01MH100047) to R.A.B., and NIH (1K99MH117165) to T.N.T. This work was also supported by the NYGC CCDG (UM1HG008901) and the Genome Sequencing Program (GSP) Coordinating Center (U24HG008956). The CCDG is funded by the National Human Genome Research Institute and the National Heart, Lung, and Blood Institute. The GSP Coordinating Center contributed to cross-program scientific initiatives and provided logistical and general study coordination. Exome sequencing was provided by the University of Washington Center for Mendelian Genomics (UW-CMG) and was funded by NHGRI and NHLBI grants UM1 HG006493 and U24 HG008956. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. E.E.E. is an investigator of the Howard Hughes Medical Institute.

DISCLOSURE

E.E.E. is on the scientific advisory board (SAB) of DNAnexus, Inc. The other authors declare no conflicts of interest.

DATA AVAILABILITY

The SAGE genome sequencing data is available at the database of Genotypes and Phenotypes (dbGaP) under accession: phs001740.v1.p1.

REFERENCES

1. Lelieveld SH, Reijnders MR, Pfundt R, et al. Meta-analysis of 2,104 trios provides support for 10 new genes for intellectual disability. Nat Neurosci. 2016;19:1194–1196.
2. Deciphering Developmental Disorders Study. Prevalence and architecture of de novo mutations in developmental disorders. Nature. 2017;542:433–438.
3. Iossifov I, O'Roak BJ, Sanders SJ, et al. The contribution of de novo coding mutations to autism spectrum disorder. Nature. 2014;515:216–221.
4. De Rubeis S, He X, Goldberg AP, et al. Synaptic, transcriptional and chromatin genes disrupted in autism. Nature. 2014;515:209–215.

5. Coe BP, Stessman HAF, Sulovari A, et al. Neurodevelopmental disease genes implicated by de novo mutation and CNV morbidity. Nat Genet (in press).
6. Fischbach GD, Lord C. The Simons Simplex Collection: a resource for identification of autism genetic risk factors. Neuron. 2010;68: 192–195.
7. 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.
8. Turner TN, Coe BP, Dickel DE, et al. Genomic patterns of de novo mutation in simplex autism. Cell. 2017;171:710–22 e712.
9. RK CY, Merico D, Bookman M, et al. Whole genome sequencing resource identifies 18 new candidate genes for autism spectrum disorder. Nat Neurosci. 2017;20:602–611.
10. Krumm N, Turner TN, Baker C, et al. Excess of rare, inherited truncating mutations in autism. Nat Genet. 2015;47:582–588.
11. Hallmayer J, Cleveland S, Torres A, et al. Genetic heritability and shared environmental factors among twin pairs with autism. JAMA Psychiatry. 2011;68:1095–1102.
12. Turner TN, Hormozdiari F, Duyzend MH, et al. Genome sequencing of autism-affected families reveals disruption of putative noncoding regulatory DNA. Am J Hum Genet. 2016;98:58–74.
13. Schaaf CP, Sabo A, Sakai Y, et al. Oligogenic heterozygosity in individuals with high-functioning autism spectrum disorders. Hum Mol Genet. 2011;20:3366–3375.
14. Girirajan S, Rosenfeld JA, Cooper GM, et al. A recurrent 16p12.1 microdeletion supports a two-hit model for severe developmental delay. Nat Genet. 2010;42:203–209.
15. Yuen RK, Thiruvahindrapuram B, Merico D, et al. Whole-genome sequencing of quartet families with autism spectrum disorder. Nat Med. 2015;21:185–191.
16. Regier AA, Farjoun Y, Larson D, et al. Functional equivalence of genome sequencing analysis pipelines enables harmonized variant calling across human genetics projects. Nat Commun. 2018;9:4038.
17. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics. 2010;26:589–595.
18. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20:1297–1303.
19. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. Robust relationship inference in genome-wide association studies. Bioinformatics. 2010;26:2867–2873.
20. Sudmant PH, Kitzman JO, Antonacci F, et al. Diversity of human copy number variation and multicopy genes. Science. 2010;330: 641–646.
21. Handsaker RE, Korn JM, Nemesh J, McCarroll SA. Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. Nat Genet. 2011;43:269–276.
22. Layer RM, Chiang C, Quinlan AR, Hall IM. LUMPY: a probabilistic framework for structural variant discovery. Genome Biol. 2014;15:R84.
23. Kronenberg ZN, Osborne EJ, Cone KR, et al. Wham: identifying structural variants of biological consequence. PLoS Comput Biol. 2015;11: e1004572.
24. Abyzov A, Urban AE, Snyder M, Gerstein M. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res. 2011;21:974–984.
25. Rausch T, Zichner T, Schlattl A, Stutz AM, Benes V, Korbel JO. DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics. 2012;28:i333–i339.
26. Lu HC, Tan Q, Rousseaux MW, et al. Disruption of the ATXN1-CIC complex causes a spectrum of neurobehavioral phenotypes in mice and humans. Nat Genet. 2017;49:527–536.
27. Richards S, Aziz N, Bale S, et al. 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. Genet Med. 2015;17:405–424.
28. Turner TN, Yi Q, Krumm N, et al. denovo-db: a compendium of human de novo variants. Nucleic Acids Res. 2017;45:D804–D811.
29. Lopez E, Berenguer M, Tingaud-Sequeira A, et al. Mutations in MYT1, encoding the myelin transcription factor 1, are a rare cause of OAVS. J Med Genet. 2016;53:752–760.
30. Kearney HM, Thorland EC, Brown KK, Quintero-Rivera F, South ST, Working Group of the American College of Medical Genetics Laboratory Quality Assurance Committee. American College of Medical Genetics standards and guidelines for interpretation and reporting of postnatal constitutional copy number variants. Genet Med. 2011;13:680–685.
31. Guo H, Peng Y, Hu Z, et al. Genome-wide copy number variation analysis in a Chinese autism spectrum disorder cohort. Sci Rep. 2017;7:44155.
32. Leppa VM, Kravitz SN, Martin CL, et al. Rare inherited and de novo CNVs reveal complex contributions to ASD risk in multiplex families. Am J Hum Genet. 2016;99:540–554.
33. Joseph B, Wallen-Mackenzie A, Benoit G, et al. p57(Kip2) cooperates with Nurr1 in developing dopamine cells. Proc Natl Acad Sci U S A. 2003;100:15619–15624.
34. Levy J, Grotto S, Mignot C, et al. NR4A2 haploinsufficiency is associated with intellectual disability and autism spectrum disorder. Clin Genet. 2018;94:264–268.
35. Bannai H, Fukatsu K, Mizutani A, et al. An RNA-interacting protein, SYNCRIP (heterogeneous nuclear ribonuclear protein Q1/NSAP1) is a component of mRNA granule transported with inositol 1,4,5-trisphosphate receptor type 1 mRNA in neuronal dendrites. J Biol Chem. 2004;279:53427–53434.
36. Yang CP, Samuels TJ, Huang Y, et al. Imp and Syp RNA-binding proteins govern decommissioning of Drosophila neural stem cells. Development. 2017;144:3454–3464.
37. Chen DT, Jiang X, Akula N, et al. Genome-wide association study meta-analysis of European and Asian-ancestry samples identifies three novel loci associated with bipolar disorder. Mol Psychiatry. 2013;18:195–205.
38. Ikeda M, Takahashi A, Kamatani Y, et al. A genome-wide association study identifies two novel susceptibility loci and trans population polygenicity associated with bipolar disorder. Mol Psychiatry. 2017;23:639–647.
39. 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.
40. Mek N, Martin HC, Rice DL, et al. Common genetic variants contribute to risk of rare severe neurodevelopmental disorders. Nature. 2018;562:268–271.
41. Geisheker MR, Heymann G, Wang T, et al. Hotspots of missense mutation identify neurodevelopmental disorder genes and functional domains. Nat Neurosci. 2017;20:1043–1051.
42. Gilissen C, Hehir-Kwa JY, Thung DT, et al. Genome sequencing identifies major causes of severe intellectual disability. Nature. 2014;511:344–347.