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

Supplementary Methods, Figures and Tables

Supplementary Methods

The Study of Autism Genetics Exploration (SAGE) collection. The SAGE collection is a local repository of samples from simplex and multiplex ASD families with autism-like features and deep phenotypic information collected from clinicians with the ability to recontact patients, providing the potential to correlate genetic findings with particular phenotypes. Study participants were ascertained based on the presence of a diagnosis of ASD, ID or developmental delay (DD) and recruited using three approaches: (1) the Seattle Children's Hospital Autism Center Clinic Registry, which consists of families with a diagnosis of ASD, ID or DD who expressed interest in participating in research; (2) Seattle-area listservs for families with ASD, ID or DD, including listservs pertaining to Autism Speaks, Parent to Parent (P2P), The Arc, FEAT, and others; and (3) providers who work with individuals with ASD, ID or DD. DNA was extracted from blood samples. ASD diagnoses were confirmed by meeting cutoff criteria on the Autism Diagnostic Observation Schedule[1] and DSM-5[2] criteria using all available information. Cognitive abilities were assessed using age-appropriate cognitive batteries, including DAS-2[3]; Wechsler tests: WPPSI-IV[4], WISC-V[5], WASI-2[6]; and the Mullen[7]. This study was approved for sequencing by the local institutional review board (IRB) at the New York Genome Center (Biomedical Research Alliance of New York [BRANY] IRB File #17-08-26-385) and for local SAGE enrollment and recontact at the University of Washington (IRB protocol #44219).

ES and analysis. We selected 42 families from the SAGE cohort on which to perform ES. These families were: (1) multiplex or had an individual with high-functioning autism (HFA) and (2) did not contain a >100 kbp rare CNV (present in <0.1% of controls).

Genomic DNA was derived from whole blood, and exomes for the 42 families (148 individuals) were captured with NimbleGen EZ Exome V2.0. Reads were mapped to the human genome reference assembly GRCh37 with decoy sequence using BWA-MEM. The positions of the variants were lifted over to GRCh38 for comparison with GS data. Further analysis was performed as for the GS. All ES was performed at the Northwest Genome Sequencing Center (Seattle, WA).

aCGH. Sample DNA and control DNA were labeled differentially with either cy3 or cy5 dye according to a NimbleGen oligo array labeling protocol. Sample and control DNA were mixed and hybridized on a NimbleGen 12-plex array with 144K probes and scanned on an Axon or Agilent scanner, then uploaded to UCSC Genome Browser. The microarray was designed for targeting regions of genomic rearrangement (density of probes) with genomic backbone coverage in the range of one probe every 14 kbp. Event calls (duplications or deletions compared to control hyb) were made by a Hidden Markov model (HMM), with criteria for calls with a Z score >1.5, size >50 kbp and probes >10. Manual curation was done to confirm calls and also to identify events not found by HMM. Manual curation involved viewing hybridization results on UCSC Genome Browser in genomic disorder regions and also chromosome by chromosome. Parental DNA was hybridized to establish inheritance where possible in probands where CNVs of significance were discovered. Rare CNVs are defined as <0.1% of 19,584 controls[8].

Classification of disorder-related SNVs/indels. In order to comprehensively characterize disorder-related SNVs/indels, we considered six overlapping NDD gene sets: 970 ASD-associated genes from the SFARI gene database[9] (SFARI970; https://gene.sfari.org), 93 DD genes identified from the DDD study[10] (DD93), 253 genes enriched for *de novo* variants in autism and DD cases from Coe et al.[11] (BC253), 237 genes with nominal significance for enrichment or clustering of missense *de novo* variants in NDDs from Geisheker et al.[12] (MG237), and 526 ID genes (ID526) and 628 candidate ID genes (ID_C628) compiled by Gilissen et al.[13] (Supplementary Table S5). The combined set includes 2,049 candidate genes with various degrees of confidence and support. We also defined a stricter, high-confidence NDD gene list containing only SFARI genes with a confidence score rank 1-3 or S (SFARI294 genes), 124 genes predicted as genome-wide significant in NDDs from Coe et al. (BC124)[11], and 93 genes from the DDD study (DD93). The combined set includes 361 high-confidence NDD genes (Supplementary Table S5).

We only considered rare variants (MAF < 0.1% in ExAC) in NDD genes that are intolerant to variation as defined by a Residual Variation Intolerance Score (RVIS) below the 50th percentile (ExAC v2 release 2.0) (http://genic-intolerance.org/index.jsp) or the probability of being loss of function (pLI score) above 50%. We defined *de novo* amino-acid-changing variants as disorder-related SNVs/indels if they occurred in the intolerant NDD gene list. In addition to the *de novo* events, we considered private (specific to a SAGE family) and rare (MAF < 0.1% in ExAC) inherited LGD variants as disorder-related SNVs/indels if they occurred in the intolerant high-confidence NDD gene list, and we exclude the heterozygous variants if there is strong evidence of recessive inheritance of the corresponding genes. For the *de novo* nonsynonymous variants and rare inherited LGD variants that are not in NDD gene list, we also examined the literature for evidence of potential pathogenicity.

Classification of disorder-related CNVs. We defined disorder-related CNVs as those having a 50% reciprocal overlap with a curated list of CNVs enriched in NDDs, including genome-wide significant CNVs by leveraging large ASD cohorts from Sanders et al. *Neuron* 2016[14] and genome-wide significant CNVs of the comorbid DD map from Coe et al. *Nat Genet* 2015[8] (Supplementary Table S8), and other CNVs if they met the following criteria: (i) all *de novo* genic CNVs with size greater than 1 Mbp, and CNVs less than 1 Mbp if they overlapped intolerant NDD genes as defined above, and (ii) all private inherited gene-disruptive CNVs mapping to intolerant high-confident NDD genes. Besides the above criteria, we also manually curated all private (specific to a SAGE family) genic CNVs and searched the literature for evidence of potential pathogenicity.

Variant validation. To evaluate the *de novo* variant-calling efficiency, we randomly selected 50 *de novo* SNVs from three categories of genomic regions: unique regions, ancient repetitive regions, and recent repetitive regions, respectively, for validation. Similar to our previous study, we concluded 95.5% validation rate (VR) in unique regions, 65.7% VR in ancient repeats regions, and 16.2% VR in recent repeats regions.

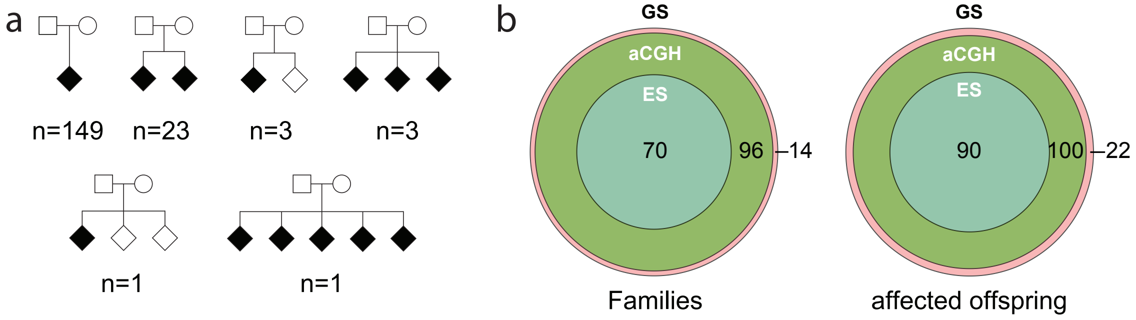
We attempted to validate three SNVs/indels sets of interest by Sanger sequencing: (i) all disorder-related variants defined above; (ii) all *de novo* LGD and *de novo* severe missense variants (CADD1.3 > 30) within genes that are not included in the NDD genes; and (iii) GS-only and ES-only *de novo* coding variants from samples applied to both GS and ES. Primers for Sanger sequencing were designed using Batch Primer3 and followed by manual *in silico* PCR. All disorder-related CNVs were validated by the following approaches: (i) all CNVs identified by both GS and aCGH were considered as validated and (ii) smaller CNVs that failed detection by aCGH were validated by either a higher density microarray (Agilent SurePrint G3 2X400K) or Sanger sequencing (for small deletions).

Statistical analysis. To identify significant enrichments for missense variants within *SYNCRIP*, we applied a probabilistic model that incorporates sequence context and human–chimpanzee fixed differences to generate a null model for the distribution of missense variation across the genome and applied a one-tailed binomial test to test for enrichment as we described before[15]. P-values were corrected for 18,945 genes by using Bonferroni correction. Comparison of the IQ data was performed by Mann–Whitney U test.

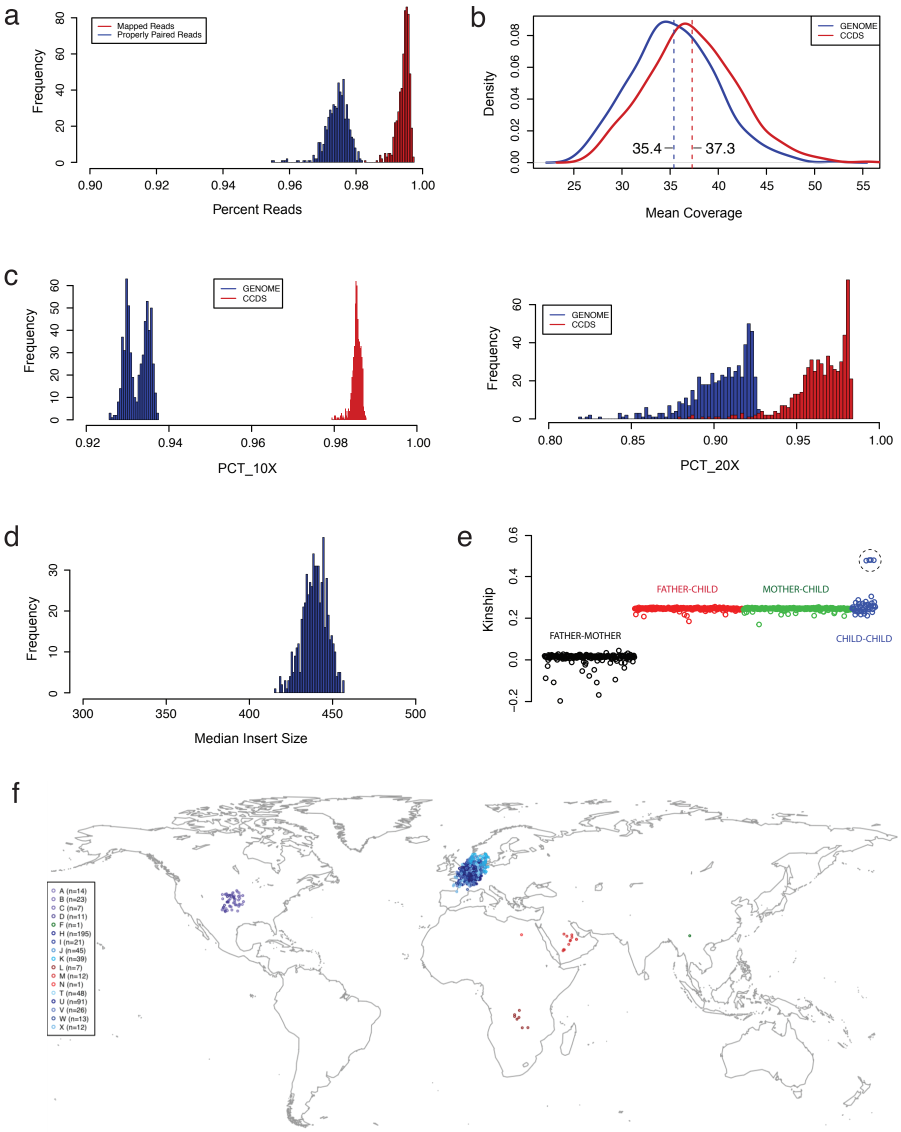
References

1. Lord CR, M.; DiLavore, P. C.; Risi, S.; Gotham, K.; Bishop, S. *Autism diagnostic observation schedule–Second edition (ADOS-2).* Los Angeles: Western Psychological Services; 2012.
2. Association AP. *Diagnostic and statistical manual of mental disorders: DSM-5.* 2013.
3. Elliott C. *Differential Ability Scales®-II (DAS-II®).* Bloomington: Pearson; 2007.
4. Wechsler D. *Wechsler Preschool and Primary Scale of Intelligence - Fourth Edition.* Vol Pearson: Bloomington; 2012.
5. Wechsler D. *Wechsler intelligence scale for children-fifth edition.* Bloomington: Pearson; 2014.
6. Wechsler D. *Wechsler Abbreviated Scale of Intelligence - Second Edition.* Bloomington: Pearson; 2011.
7. Mullen EM. *Mullen scales of early learning.* Circle Pines: AGS; 1995.
8. Coe BP, Witherspoon K, Rosenfeld JA, et al. Refining analyses of copy number variation identifies specific genes associated with developmental delay. *Nature genetics.* 2014;46(10):1063-1071.
9. Basu SN, Kollu R, Banerjee-Basu S. AutDB: a gene reference resource for autism research. *Nucleic acids research.* 2009;37(Database issue):D832-836.
10. Deciphering Developmental Disorders S. Prevalence and architecture of de novo mutations in developmental disorders. *Nature.* 2017;542(7642):433-438.
11. Coe BP, Stessman HAF, Sulovari A, et al. Neurodevelopmental disease genes implicated by de novo mutation and CNV morbidity. *Nature genetics.* In press.
12. Geisheker MR, Heymann G, Wang T, et al. Hotspots of missense mutation identify neurodevelopmental disorder genes and functional domains. *Nat Neurosci.* 2017;20(8):1043-1051.
13. Gilissen C, Hehir-Kwa JY, Thung DT, et al. Genome sequencing identifies major causes of severe intellectual disability. *Nature.* 2014;511(7509):344-347.
14. Sanders SJ, He X, Willsey AJ, et al. Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. *Neuron.* 2015;87(6):1215-1233.
15. O'Roak BJ, Vives L, Fu W, et al. Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders. *Science.* 2012;338(6114):1619-1622.

Supplementary Figures

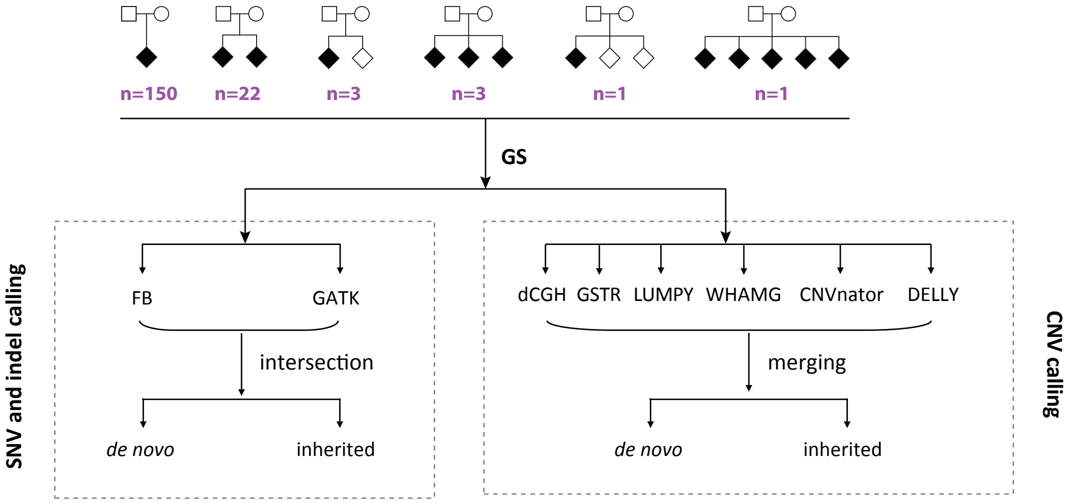


Supplementary Figure S1 Summary of ASD family structures and genetic analyses. (**a**) Simplex and multiplex family structures included in this study. Black indicates affected individuals with autism. (**b**) Number of families investigated with each type of genomic technology (GS=genome sequencing, ES=exome sequencing, aCGH=array comparative genomic hybridization).

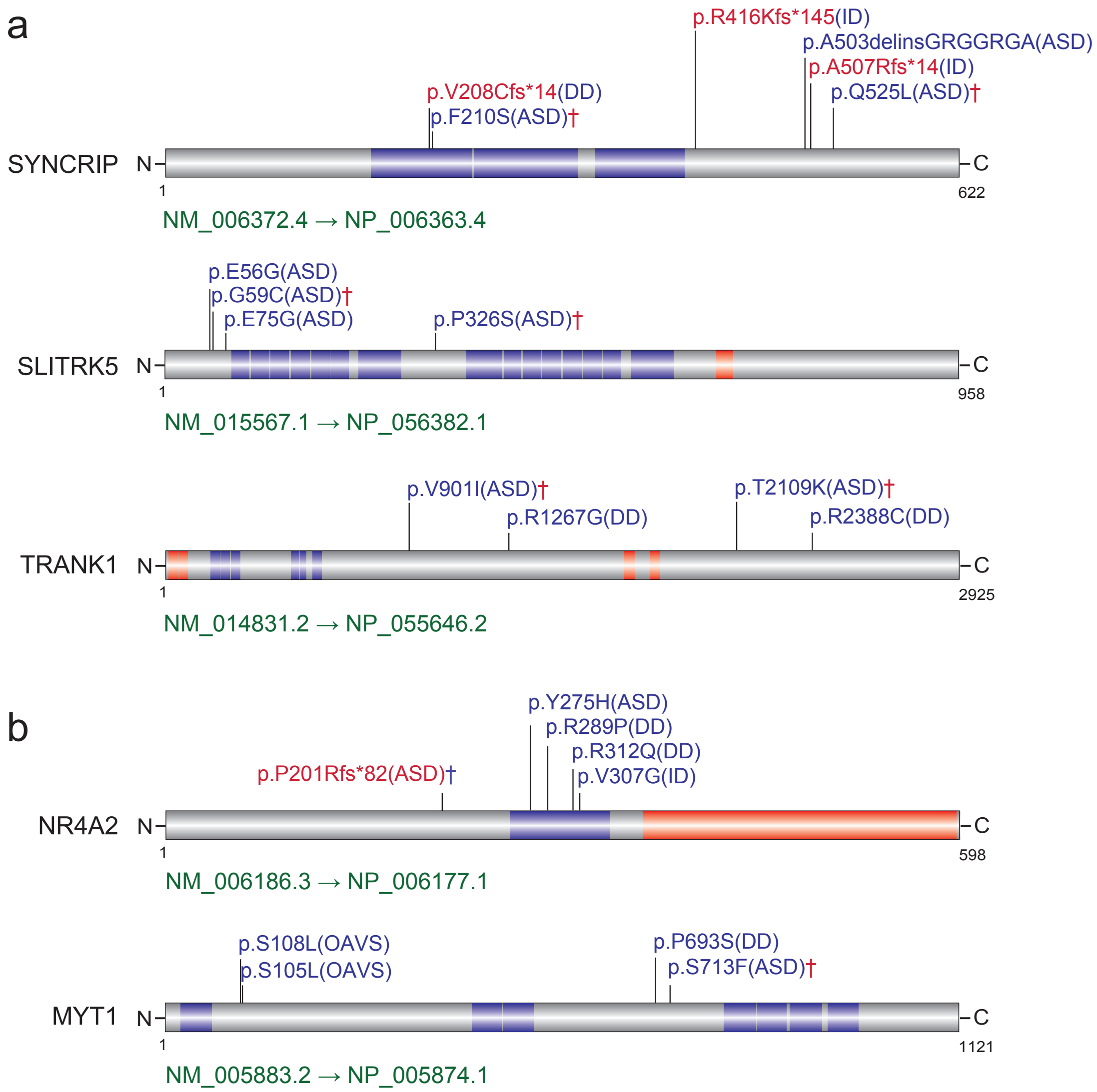


Supplementary Figure S2 QC metrics for GS and mitochondria haplogroup distribution. (**a**) Histogram showing the percentage of reads mapped and properly mapped to the reference genome for all 578 samples. (**b**) Mean coverage distribution for all 578 samples. (**c**) Histogram showing the fraction of each genome sequenced over 10X coverage and 20X coverage, respectively. (**d**) Histogram showing the distribution of median insert size for all 578 samples. (**e**) Kinship coefficients showing that expected familial relationship is supported by the genome data. Four identical twins (circled) were identified. (**f**) The world map shows the location where the individuals' mitochondria haplogroups are thought to have been derived in the world in the context of the past 7,000-170,000 years (personal communication with Marie Lott at The Children's Hospital of Philadelphia). Coordinates were approximated from looking at the following map:

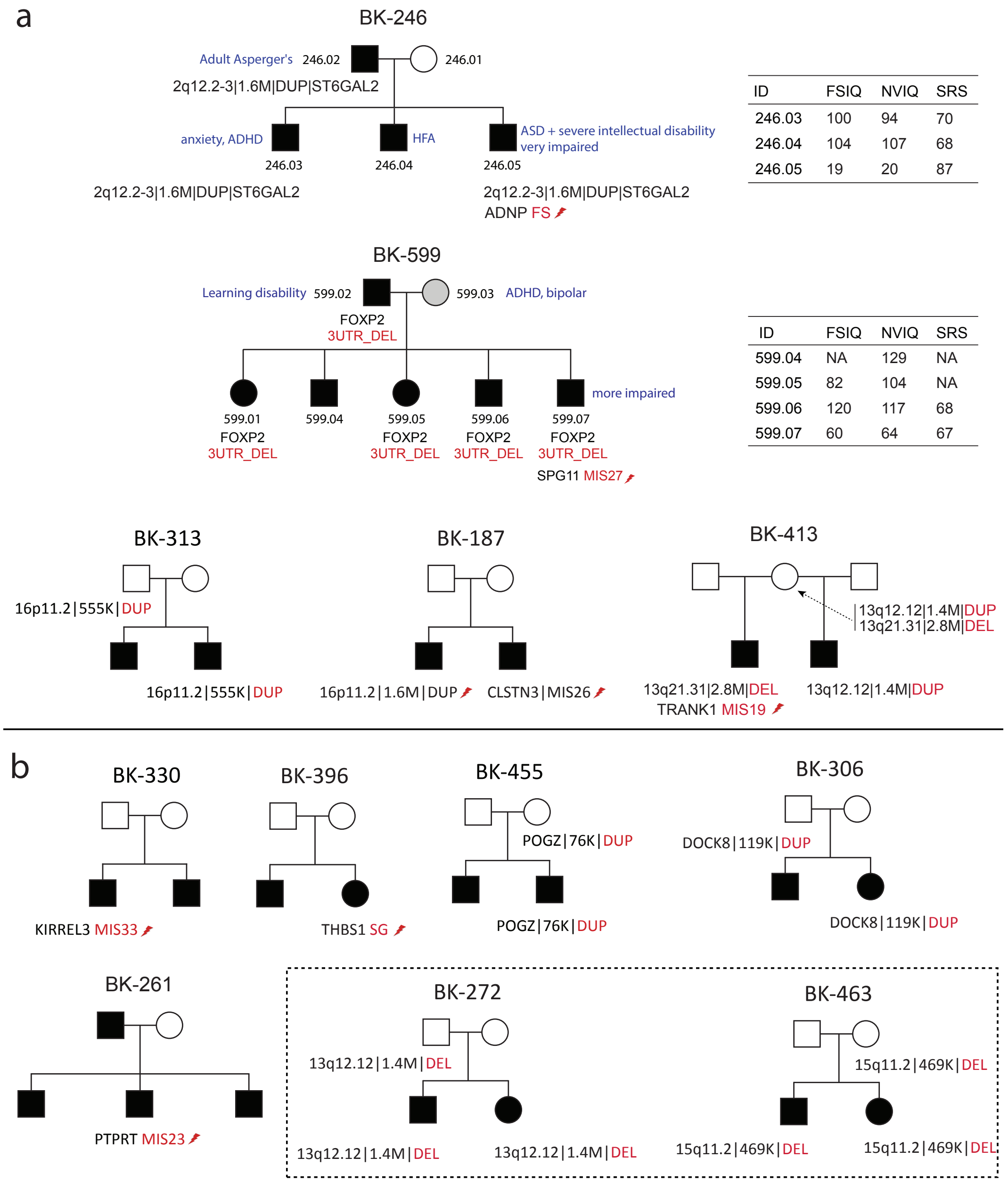
https://www.mitomap.org/foswiki/pub/MITOMAP/MitomapFigures/WorldMigrations2012.pdf.



Supplementary Figure S3 GS-calling methods for SNVs/indels and CNVs. We called SNVs/indels based on the intersection of GATK and FreeBayes (FB). We called CNVs using six different CNV callers that applied different combinations of sequence signatures (i.e., read-depth and paired-end). All calls identified using digital comparative genomic hybridization (dCGH) were considered, as well as calls supported by at least two of the methods. Putative *de novo* and inherited events were filtered differently. GSTR=Genome STRiP.



Supplementary Figure S4 Recurrent *de novo* variants and clustered missense variants. (**a**) NDD candidate genes with recurrent missense *de novo* variants identified in this study. (**b**) NDD candidate genes with potentially clustered missense variants. Validated *de novo* variants identified in this study are indicated with dagger (†) alongside other variants from denovo-db (10,927 NDD cases). LGD variants are in red while missense variants are indicated in blue.



ID	FSIQ	NVIQ	SRS
246.03	100	94	70
246.04	104	107	68
246.05	19	20	87

ID	FSIQ	NVIQ	SRS
599.04	NA	129	NA
599.05	82	104	NA
599.06	120	117	68
599.07	60	64	67

Supplementary Figure S5 Multiplex families with disorder-related events. (**a**) Multiplex families with pathogenic or likely pathogenic events. HFA=high-functioning autism; SRS=Social Responsiveness Scale score. (**b**) Families with disorder-related events not classified as pathogenic or likely pathogenic. All disorder-related events are not transmitted to all affected members within the same multiplex family, except 13q12.12 deletion in BK272 and 15q11.2 deletion in BK463.

Supplementary Tables

*Supplementary tables provided as separate Excel documents.

Supplementary Table S1. Sample information and annotation for disorder-related events for each offspring.

Supplementary Table S2. GS QC metrics and family relationship data (kinship) generated by KING program.

Supplementary Table S3. *De novo* SNVs and indels detected from families applied to both GS and ES.

Supplementary Table S4. All private CNVs detected by GS.

Supplementary Table S5. NDD gene sets.

Supplementary Table S6. All rare amino-acid changing *de novo* SNVs and indels detected by GS.

Supplementary Table S7. Inherited disorder-related SNVs and indels.

Supplementary Table S8. NDD CNV set.

Supplementary Table S9. Disorder-related CNVs.