The American Journal of Human Genetics, Volume *110*

Supplemental information

Systematic evaluation of genome sequencing for the diagnostic assessment of autism spectrum disorder and fetal structural anomalies

Chelsea Lowther, Elise Valkanas, Jessica L. Giordano, Harold Z. Wang, Benjamin B. Currall, Kathryn O'Keefe, Emma Pierce-Hoffman, Nehir E. Kurtas, Christopher W. Whelan, Stephanie P. Hao, Ben Weisburd, Vahid Jalili, Jack Fu, Isaac Wong, Ryan L. Collins, Xuefang Zhao, Christina A. Austin-Tse, Emily Evangelista, Gabrielle Lemire, Vimla S. Aggarwal, Diane Lucente, Laura D. Gauthier, Charlotte Tolonen, Nareh Sahakian, Christine Stevens, Joon-Yong An, Shan Dong, Mary E. Norton, Tippi C. MacKenzie, Bernie Devlin, Kelly Gilmore, Bradford C. Powell, Alicia Brandt, Francesco Vetrini, Michelle DiVito, Stephan J. Sanders, Daniel G. MacArthur, Jennelle C. Hodge, Anne O'Donnell-Luria, Heidi L. Rehm, Neeta L. Vora, Brynn Levy, Harrison Brand, Ronald J. Wapner, and Michael E. Talkowski

SUPPLEMENTAL INFORMATION

Table of Contents

SUPPLEMENTAL FIGURES

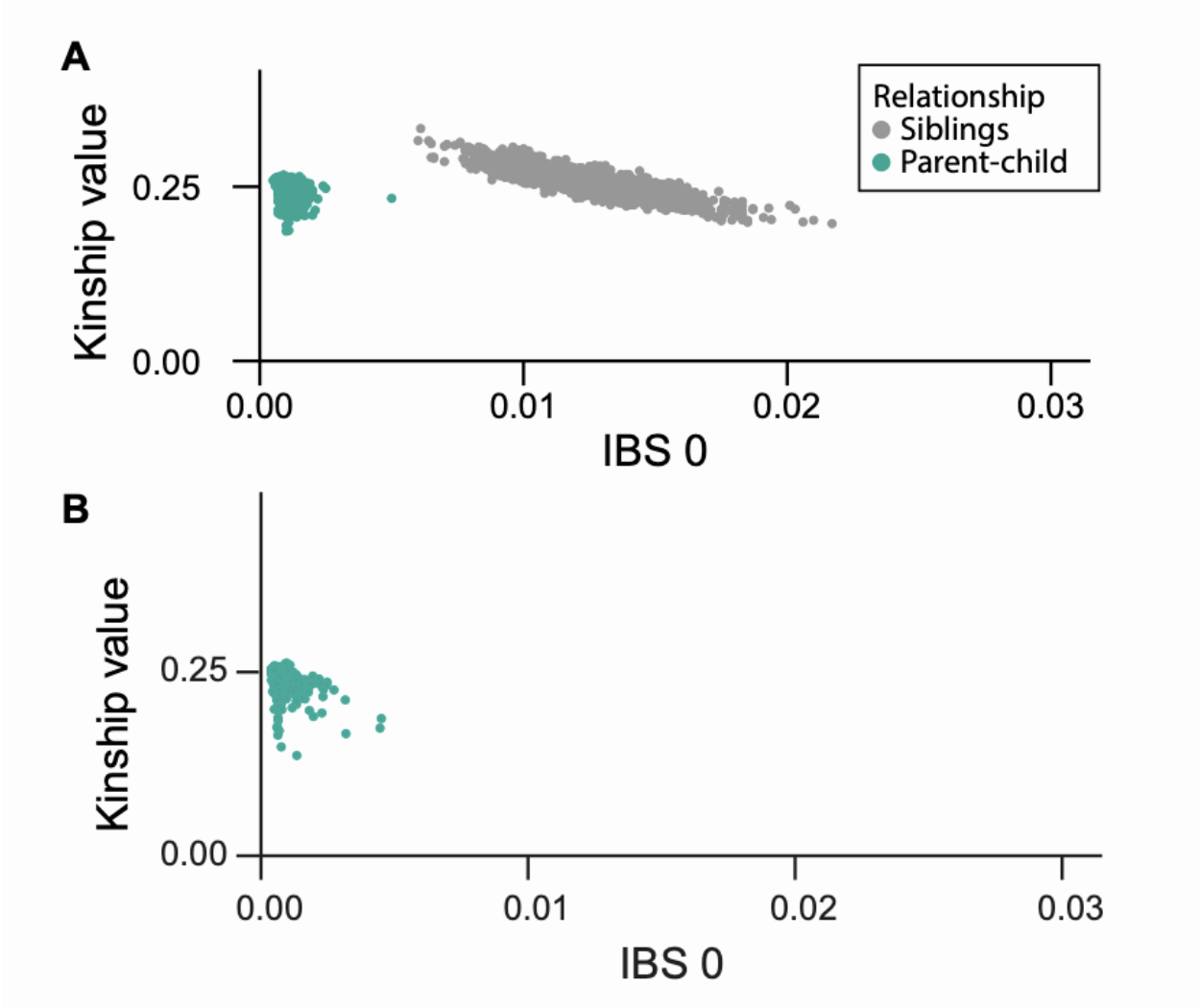

Figure S1. Confirmation of sample relatedness using kinship values
Kinship values for GS were calculated using KING[1] after restricting to SNVs with an alternate allele frequency greater than 5% in gnomAD genomes.[2] Each point on the plot represents a related pair of individuals, colored by relationship status. **(A)** Relatedness metrics for 6,448 individuals from the 1,612 ASD quartet families. **(B)** Relatedness metrics for 747 individuals from the 249 FSA trios.

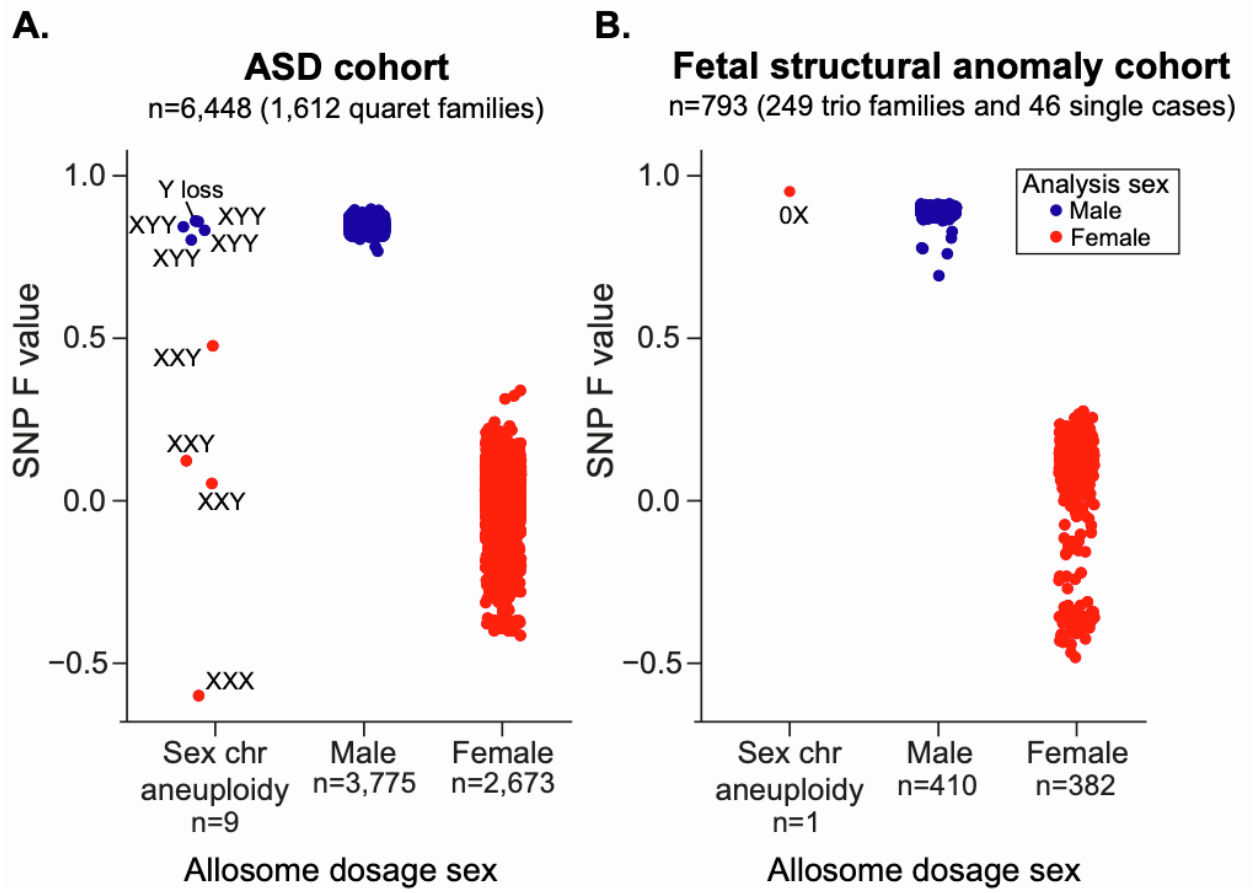

Figure S2. Sample sex QC

Confirmation of sample sex using single nucleotide polymorphism (SNP) and chromosomal read-depth information from GS data. Sex was inferred two ways from GS data: 1) using the F value generated with PLINK[3] based on sex chromosome SNP genotypes, and 2) using read depth (dosage) scores[4] derived from chrX and chrY. Each point represents a sample, colored by final sex used for analysis. **(A)** Sex metrics for the 6,448 individuals from the 1,612 ASD quartet families. Cases with sex chromosomal abnormalities (n=9) have been labeled. **(B)** Sex metrics for the 249 trios individuals from the FSA trios (n=747) that were pre-screened with standard-of-care diagnostic tests and the 46 singleton benchmarking cases.

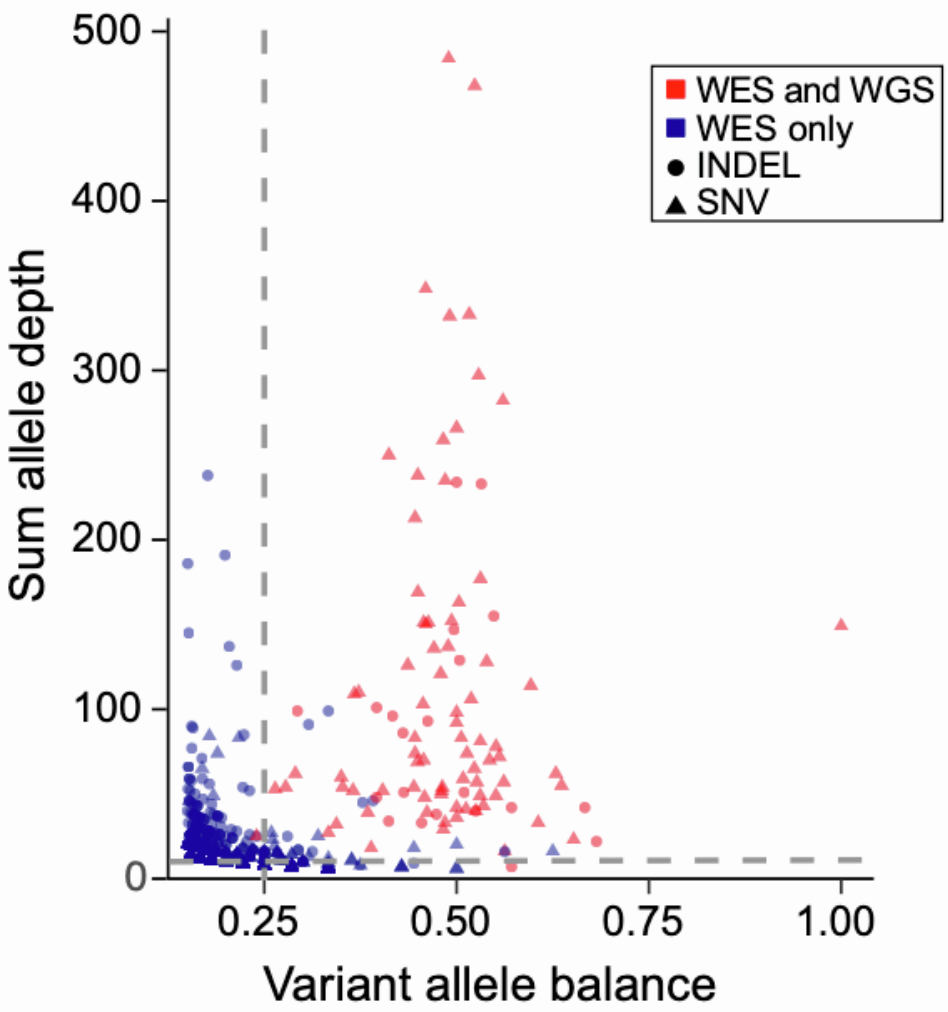

Figure S3. Modified exome sequencing depth and allele balance thresholds
The plot displays the allele balance (AB) and sum allele depth (AD) for all 1,453 *de novo* SNVs and indels detected from ES using the standard GS filters. Color indicates if the variant is unique to ES (blue; n=1,353) or was found in both the ES and GS data (red; n=100). Shape indicates variant type (SNV or indel). The dotted line represents the modified thresholds ultimately used for filtering the ES data before manual review: AB > 0.25 and sum(AD) ≥ 1.

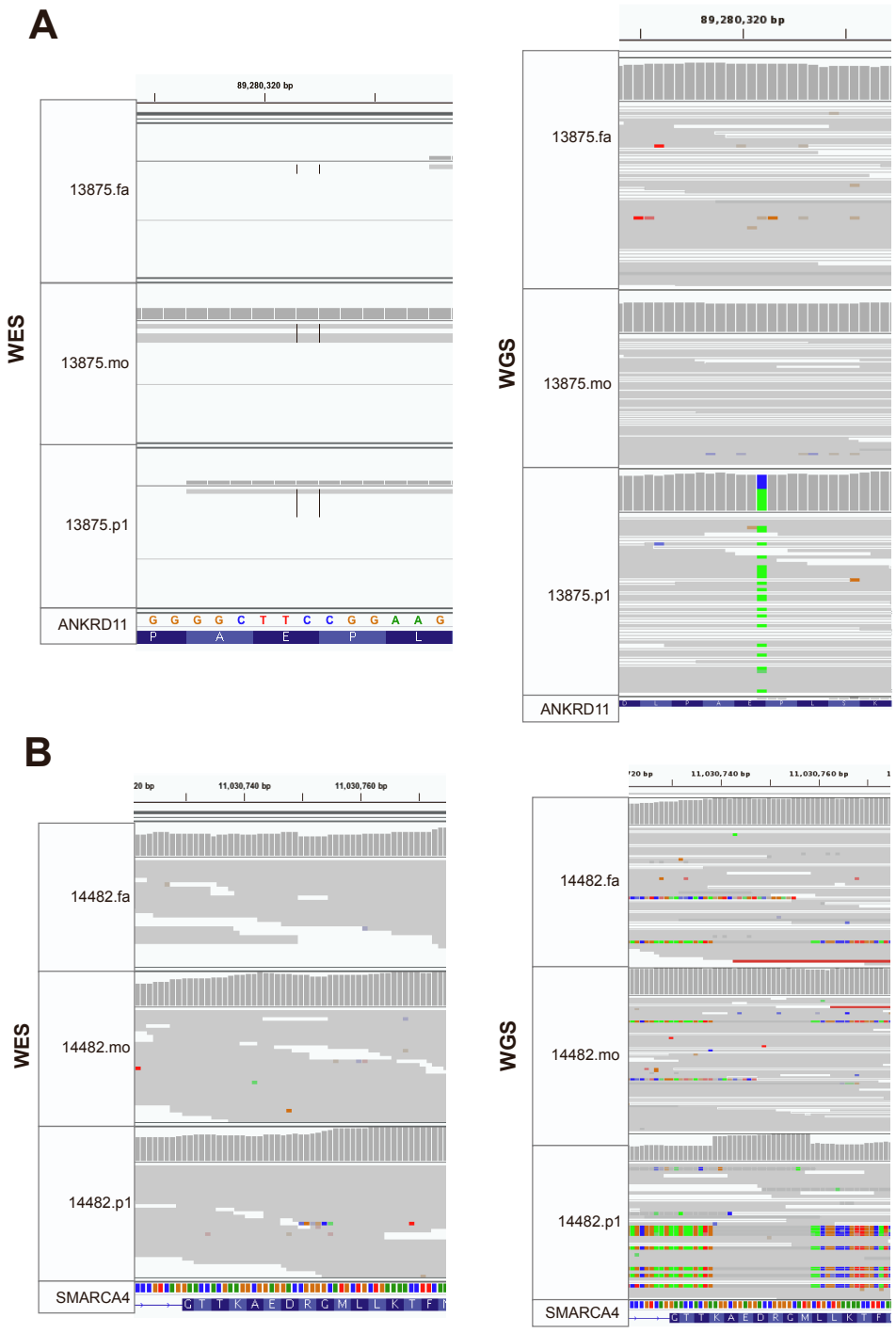

Figure S4. Two pathogenic sequence variants unique to GS in ASD probands
Alignment visualization for two *de novo* variants uniquely identified in GS alongside their raw read evidence from ES. Images were generated using IGV.[5] For each site, the ES and GS screenshots are shown side-by-side. **(A)** A stopgain SNV in *ANKRD11* in proband 13875.p1 that was absent from the ES VCF and has low coverage in the ES CRAM files **(B)** A 44 bp insertion in *SMARCA4* in 14482.p1 that was absent from the ES VCF and shows no supporting evidence in ES CRAM files.

SUPPLEMENTAL METHODS

Participant ascertainment and genome sequencing

We included 1,612 deeply phenotyped quartet families ascertained as part of the Simons Simplex Collection in this study.[6–8] As previously described,[9] each family included two unaffected parents, one unaffected sibling, and an affected proband with autism spectrum disorder (ASD). All affected probands underwent a battery of diagnostic tests, including the Autism Diagnostic Observation Schedule (ADOS) and the Autism Diagnostic Interview-Revised (ADI-R) to confirm the ASD diagnosis, as well as detailed evaluations of intellectual/cognitive functioning, adaptive behavior, physical/dysmorphic features, developmental milestones, medical comorbidities, and family history. We also included 295 fetuses that met criteria for diagnostic testing due to the presence of a structural anomaly (n=281) or advanced maternal age (n=14) (Figure 1). This included 46 singleton fetuses that were pre-selected for having a clinically reportable variant from karyotype, chromosomal microarray (CMA), and/or exome sequencing (ES), as well as 249 fetal structural anomaly (FSA) trios that had been pre-screened by karyotype, CMA, and/or ES. Recruitment and phenotyping protocols for the fetuses have been previously described.[10–12] All 7,241 individuals underwent paired-end genome sequencing (GS) to a mean target coverage of 30X (see Tables S1-3 for specific sequencing metrics).

Sample-level QC

To confirm sample relatedness, we performed a kinship inference analysis with KING[1] (http://people.virginia.edu/~wc9c/KING) using the GS data after restricting to single nucleotide polymorphisms (SNPs) with an alternate allele frequency (AF) >5% in gnomAD (Figure S1).[2] In parallel, we also predicted genetic sex using two independent approaches: first, we used PLINK to infer sex based on sex chromosome genotypes.[3] Second, we used GATK-SV to calculate copy number estimates for each chromosome per sample, which permitted inference of genetic sex as well as the identification of chromosomal aneuploidies. We compared the predicted sex for each individual between both methods and observed high concordance (Figure S2). Using the relatedness and sex results we resolved sample swaps based on discrepancies in family structures deviating from the expected relatedness metrics for parent-child (IBS0 ≤ 0.005 and kinship coefficient > 0.2) and sibling relationships (IBS0 > 0.005 and kinship coefficient > 0.2; Figure S1).

We also confirmed cross-technology sample relatedness to assure comparisons were performed on the same 6,448 individuals from the 1,612 ASD quartet families. This was accomplished by restricting the ES and GS VCFs to high-quality common SNPs from Purcell et al. 2014[13] that were lifted over to GRCh38/hg38 and limited to 5,862 SNPs common to both ES and GS. Samples were renamed based on their technology of origin and were merged into a single cross-technology ASD VCF for relatedness analysis with KING.[1] ES and GS samples with a kinship coefficient > 0.45 were considered to be identical samples. Finally, we used the confirmed sample metadata from the GS and ES comparisons to identify matching CMA data.[14]

Genome sequencing analysis framework

We developed a GS analytic framework to discover, filter, and interpret nine different classes of variation that are described in detail below and also summarized in Table S4. The aim of this pipeline was to retain as many pathogenic or likely pathogenic (P/LP) variants as possible while reducing the total number of variants requiring manual review. The same GS analysis pipeline was applied to both the ASD and prenatal cohorts with only modifications to the phenotype-specific gene list used.

1.0. Variant discovery

1.1. Sequence variants (GATK)

As previously described,[7,8] the ASD GS data was generated from PCR-free libraries and processed using the Center for Common Disease Genomics functional equivalence pipelines (https://github.com/CCDG/Pipeline-Standardization) and following the Genome Analysis Toolkit (GATK) Best Practices Workflows for sequence variant, single nucleotide variant (SNV) and small insertion/deletion (indel), discovery (https://gatk.broadinstitute.org/hc/en-us/sections/360007226651-Best-Practices-Workflows).[15] Briefly, this included aligning the raw FASTQ reads to the hg38/GRCh38 human reference genome using BWA-mem 0.7.15,[16] sorting and removing duplicate reads with Picard 2.4.1. (http://broadinstitute.github.io/picard/), performing base quality score recalibration, indel realignment, generating single sample gVCFs with GATK HaplotypeCaller 3.5-0,[17] merging single sample gVCFs into batch specific VCFs (ranging in size from 40 to 588 quartets),[8] joint-calling the merged VCFs, and performing Variant Quality Score Recalibration (VQSR). The aligned CRAM and gVCF files were transferred to the Amazon Web Services (AWS) S3 storage system and can be accessed with permission from the Simons Foundation Autism Research Initiative (https://www.sfari.org/resource/sfari-base/).

The GS data from the prenatal samples were generated at the Broad Institute Genomics Platform. After sequencing, individual FASTQ files were transferred to a Google Cloud bucket for storage. All GS data pre-processing and sequence variant discovery was performed using the GATK Best Practices Workflows on the cloud-enabled and freely available Terra platform (https://terra.bio/). Sequence variant calling followed the same steps described above for ASD.

1.2. Structural Variants (GATK-SV)

Structural variant (SV) discovery and genotyping was performed with GATK-SV, which was deployed on the Terra platform (https://terra.bio/). The code for GATK-SV is publicly available at https://github.com/broadinstitute/gatk-sv. All individuals were grouped into batches based on: their dosage bias score (a metric that quantifies the non-uniformity of coverage for a given GS sample),[4] sex, family status, PCR status, and cohort assignment. The ASD cohort included batches comprising 200-400 samples each, and the FSA cohort included one batch of PCR plus (n=186) and two batches of PCR free samples (n=345 and n=346, respectively). All families were kept intact during batching. All 7,241 individuals were analyzed with six SV discovery algorithms, including three paired-end/split-read algorithms (Manta v.1.4.0, Smoove v.0.2.3

[https://github.com/brentp/smoove], and WHAM-GRAPHENING v.1.7.0),[18–20] two read-depth algorithms (GATK-gCNV and cnMOPS v.1.12.0),[21,22] and one mobile element insertion algorithm, MELT v.2.0.5.[23] SV discovery generated six algorithm-specific VCFs per individual that were used as input for GATK-SV, which was run in cohort mode. The GATK-SV pipeline is organized into modules that harmonize predicted SVs across all input algorithms, reduce false positives, resolve overlapping SVs with disparate copy number, identifies complex variants (*e.g.*, inversions flanked by one or more copy number variants [CNVs]),[4,24] and provides cohort-wide SV genotypes and quality metrics available for *post hoc* filtering. We generated a cohort-wide SV VCF for the ASD and FSA cohorts, respectively, that was used as input for all downstream analyses. Further details on the GATK-SV methods can be found in Collins et al. 2020.[4]

1.3. Short tandem repeats (Expansion Hunter)

We identified short tandem repeat (STR) expansions across 18 loci that were selected from the gnomAD disease-associated STR catalog (https://github.com/broadinstitute/str-analysis/tree/main/str_analysis/variant_catalogs) based on conferring an early-onset developmental disorder phenotype (Table S8). STR expansions were genotyped using Expansion Hunter[25] v5.0.0 across 6,435/6,448 individuals from the ASD quartet families (n=9 ASD probands with a sex chromosomal abnormality were removed as well as one quartet family that revoked consent after all other analyses were completed). We also applied Expansion Hunter to all 295 prenatal samples (n=793 individuals in total).

2.0. Variant annotation

Details describing variant annotation are described in the methods of the main text.

3.0 Variant filtering

3.1. Variant QC

After variant discovery, we applied quality control (QC) filters intended to maximize sensitivity for candidate P/LP variants while removing false variant calls. For SVs, this included removing variants with a GATK-SV QUAL score ≤ 1 and multiallelic copy number variants (CNVs). For sequence variants, we removed multiallelic variants, variants with an allele balance (AB) < 0.15 in the case of interest, indels > 50bp, and variants where the sum of the reference and alternate allele depth (AD) was ≤ 5 in any family member. We also removed SNVs that did not pass GATK VQSR. To reduce false positives, we applied additional quality control filters to samples with outlier variant counts, defined as any sample with a variant count (based on raw GATK haplotype caller or individual SV algorithm output) above Q3 + 6*IQR. This definition resulted in relatively few SV outlier samples (n=12 SV in the FSA cohort and none in the ASD cohort) and sequence variant outliers (n=5 in the FSA cohort and n=1 in the ASD cohort). To control the false positive rate in these outlier samples, we removed SVs present in >2 SV outlier individuals and sequence variants with GQ <75.

3.2. Variant functional consequence

All variants were filtered for functional impact. SVs predicted to be loss-of-function (LoF) or full gene copy gain were retained for further filtering.[4] Partial gene duplications, defined as duplications with one breakpoint located outside the gene boundary and one within, were excluded given their unknown functional impact.[26] Any sequence variants predicted to be stop-gain, stop-loss, frameshift insertion, frameshift deletion, splicing (within 2 bp of a splice junction), or missense according to RefSeq or Gencode annotations were retained for additional filtering. We further filtered missense variants based on three tiers (described below) to identify those that are increasingly likely to be functionally damaging and thus classified as P/LP (Tier 1 = most likely to be P/LP, and Tier 3 = least likely to be P/LP). We removed missense variants classified as benign, likely benign, risk factor, association, drug response, or protective in ClinVar from all tiers.

Tier 1 missense:

- Classified as P/LP in ClinVar

Tier 2 missense:

- Classified as P/LP in ClinVar or
- Any missense variant with a CADD score > 30[27]

Tier 3 missense:

- Classified as P/LP in ClinVar or
- Any missense variant with a CADD score > 30 or
- Missense variants with a CADD score between 15 and 30 also located in a missense constrained region[28]

3.3. Disease genes and genomic regions

To facilitate variant filtering, we computationally built a candidate disease gene list for the ASD and FSA cohorts, respectively. The ASD gene list comprised 901 genes (Table S5) broadly associated with neurodevelopmental disorders (NDDs) from the DDG2P database[29] classified as having a 'confirmed' or 'probable' association with developmental disorders that conferred a brain/cognitive phenotype. To account for the variable phenotypes observed in the FSA cohort (Tables S2-3), we compiled 2,535 developmental disorder genes (Table S6) based on the union of eight gene lists, described below:

1) 374 dominant developmental disorder genes from the DDG2P database (accessed July 29, 2019)[29] with a "confirmed" disease association and monoallelic, imprinted, mosaic, x-linked dominant, and x-linked over-dominance modes of inheritance.

2) 800 recessive developmental disorder genes from the DDG2P database[29] with a "confirmed" disease association and biallelic or hemizygous modes of inheritance.

3) 93 genes that were significantly enriched for rare *de novo* variants in the Deciphering Developmental Disorders study.[30]

4) 26 dominant genes significantly enriched for rare *de novo* protein-truncating variants in ASD.[6]

5) 358 genes from the Clinical Genome (ClinGen) Resource Dosage Sensitivity Map with "some evidence for dosage pathogenicity" (haploinsufficiency/triplosensitivity score = 2) or "sufficient evidence for dosage pathogenicity" (haploinsufficiency/triplosensitivity score = 3) (downloaded July 29, 2019; https://www.clinicalgenome.org/curation-activities/dosage-sensitivity/).

6) 708 autosomal dominant and 1,182 recessive disease genes curated from the Online Mendelian Inheritance in Man (OMIM) database.[31,32]

7) 217 recessive and dominant X-linked genes from OMIM (tables were accessed June 12, 2017).

8) 117 genes that have been robustly associated with fetal structural anomalies detectable by ultrasound that were curated by the Prenatal Assessment of Genomes and Exomes study.[33]

Each gene was classified as being associated with a disorder that had a dominant and/or recessive pattern of inheritance based on existing annotations from DDG2P and OMIM. We categorized the inheritance labels provided by DDG2P as recessive: biallelic, and hemizygous or dominant: imprinted, monoallelic, mosaic, x-linked dominant, and x-linked over dominant. When disease inheritance was not available for a gene (n=4 missing from DDG2P), variants in that gene were retained under both dominant and recessive modes of inheritance.

We also compiled a list of 64 known genomic disorder (GD) loci to assess overlap with SVs in both our cohorts. We took all of the known CNV syndromes located on the autosomes and chromosome X from DECIPHER[34] and the haploinsufficient (HI) and triplosensitive (TS) regions from the Clinical Genome (ClinGen) Resource Dosage Sensitivity Map if they had a HI or TS score ≥ 2 ("sufficient evidence for dosage pathogenicity"). We removed any regions that were only associated with late-onset conditions, resulting in 64 candidate regions (Table S7). All SVs that overlapped ≥ 50% of a GD locus were retained for manual review. Following the most recent guidelines for CNV interpretation,[26] we also manually reviewed any rare (<1% frequency in gnomAD-SV)[4] deletion or duplication that overlapped ≥ 25 or ≥ 35 protein-coding genes, respectively, even if it did not overlap a disease gene or GD region from our lists. Finally, we also retained all SVs that overlapped one of 17 non-coding loci known to confer pathogenic long-range position effects (LRPEs; Table S8). To define the non-coding search space, we used topologically-associated domain (TAD) boundaries from the IMR90 fetal fibroblast cell line,[35] which have been previously shown to be associated with pathogenic LRPEs if disrupted,[36,37] that contained each LRPE target gene.

3.4. Inheritance

We filtered variants under the five inheritance modes described below. For the ASD quartets, the unaffected sibling and both parents were treated as independent trios during inheritance filtering. We applied more stringent missense variant filters (tiers described in the variant functional consequence section) to rare inherited and compound heterozygous variants as these two categories resulted in a large number of variants requiring manual review despite there being little evidence supporting their contribution to the etiology of ASD or FSAs.[6,12,33,38,39] The specific functional consequence considered for each inheritance type are as follows:

Dominant disease genes:

- *De novo*
  - All LoF
  - Missense Tier 3

- Rare inherited
  - All LoF
  - Missense Tier 1

Recessive disease genes:

- Homozygous
  - All LoF
  - Missense Tier 3

- X-linked recessive
  - All LoF
  - Missense Tier 3

- Compound heterozygous
  - At least one variant in the pair had to be LoF or Tier 2 missense

The identification of compound heterozygous variants comprised three steps, including: 1) compiling heterozygous SNVs, indels, and LoF SVs located in the same recessive disease gene, 2) annotating each variant with inheritance status, and 3) retaining only the instances where individuals had more than one variant in a recessive disease gene with disparate inheritance patterns (*e.g.*, one maternally inherited, one *de novo*). We required that at least one variant per compound heterozygous grouping be inherited from a parent due to the lack of phasing information from short-read GS.

3.5. Allele frequency

All variants (SNVs, indels, and SVs) meeting the above thresholds were retained if they had an alternate allele frequency (AF) <1% for variants in dominant disease genes or regions and <5% for recessive disease genes. Given that some GDs can occur at an appreciable frequency in disease cohorts,[40] we did not apply any AF cut-off when considering SV that overlapped ≥50% of a known GD locus.

4.0. Variant interpretation

Details describing manual variant curation are described in the methods of the main text.

Benchmarking the performance of GS against conventional tests

Filtering CMA data

As previously described,[7,14] SNP genotyping data was generated for the ASD cases using three microarray platforms, the Illumina 1Mv1, 1Mv3, or Omni2.5. CNV calls for each individual were identified using PennCNV,[41] QuantiSNPv2.3,[42] and GNOSIS/CNVision.[14] CNVs were filtered for rarity based on overlap with CNVs from the Database of Genomic Variants (in GRCh36/hg18) and overlap with CNVs from the ASD parents.[14,43] All CNV coordinates were lifted over from GRCh36/hg18 to GRCh38/hg38 and those classified as high-quality (CNV p-value [pCNV] ≤ 1.0×10^{-9})[14] were filtered following the same steps outlined in the GS SV pipeline (Table S4). There were 14 variants detected by GS that were also detected by CMA but failed filtering because they were not lifted over from hg18 to hg38 (n=6), failed the pCNV high-quality filter (n=5), or were removed due to incorrect CNV coordinates that suggested the variant did not overlap coding sequence (n=3; GS coordinates were used as truth). These variants were recovered and counted towards the overall yield of CMA. We also removed one deletion from CMA manual review that was identified to be rare by CMA but was found in 65 (2.0%) of our 3,224 ASD parents based on GS, which was above our allele frequency threshold.

Filtering exome sequencing data

The ES data for the ASD cases was generated as part of a larger sequencing initiative and has been extensively described.[6,38] We realigned sequencing data from GRCh37/hg19 to GRCh38/hg38[38] and applied the same filtering steps as those outlined in the GS filtering pipeline (Table S4) with minor modifications to account for differences in depth between ES and GS (Figure S3). These included increasing our AB and total allele depth filters for *de novo* variants (both in *de novo* dominant inheritance and as part of a compound heterozygous pair) to account for the higher ES coverage, increased rate of false positives, and potential for somatic variant detection. The new thresholds (AB >0.25 and total AD ≥10) were chosen based on retaining >95% of the variants that were also detected by GS (Figure S3).

To identify CNVs from ES data, we applied GATK-gCNV,[21] a publicly available Bayesian model for germline detection of CNVs. Briefly, this is a read-depth based tool that uses a negative-binomial factor analysis to adjust for known and unknown biases of exome sequencing, while modeling sample and genomic region copy number through a hierarchical hidden Markov model. In this analysis, we jointly processed the 6,448 individuals from 1,612 ASD quartet families described in this study with an additional 66,000 samples.[38] Samples were assigned to batches based on 3D clustering of the first three principal components of coverage depth after normalizing for average depth. The 72,448 samples were processed across 126 batches, with a median batch size of 449 samples (min 136 and maximum 2,259). After raw calling with GATK-gCNV, we applied

our calibrated sample-level quality filters, resulting in 5.49% of the total samples being removed. The GATK-gCNV quality score statistic (QS>400 for homozygous deletions, QS>100 for heterozygous deletions, and QS>50 for duplications) was applied to individual calls to extract rare CNVs with predicted sensitivity and positive predictive value of >90%, which resulted in a resolution of three exons or more. With these filtering metrics, the average ES sample harbored 1-2 rare high-quality CNVs.

REFERENCES

1. Manichaikul, A., Mychaleckyj, J.C., Rich, S.S., Daly, K., Sale, M., and Chen, W.M. (2010). Robust relationship inference in genome-wide association studies. Bioinformatics *26*, 2867–2873.

2. Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alfoldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature *581*, 434–443.

3. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.

4. Collins, R.L., Brand, H., Karczewski, K.J., Zhao, X., Alfoldi, J., Francioli, L.C., Khera, A.V., Lowther, C., Gauthier, L.D., Wang, H., et al. (2020). A structural variation reference for medical and population genetics. Nature *581*, 444–451.

5. Sudmant, P.H., Rausch, T., Gardner, E.J., Handsaker, R.E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Fritz, M.H.-Y., et al. (2015). An integrated map of structural variation in 2,504 human genomes. Nature *526*, 75–81.

6. Satterstrom, F.K., Kosmicki, J.A., Wang, J., Breen, M.S., De Rubeis, S., An, J.Y., Peng, M., Collins, R., Grove, J., Klei, L., et al. (2020). Large-scale exome sequencing study implicates both developmental and functional changes in the neurobiology of autism. Cell *180*, 568-584 e23.

7. Werling, D.M., Brand, H., An, J.Y., Stone, M.R., Zhu, L., Glessner, J.T., Collins, R.L., Dong, S., Layer, R.M., Markenscoff-Papadimitriou, E., et al. (2018). An analytical framework for whole-genome sequence association studies and its implications for autism spectrum disorder. Nat. Genet. *50*, 727–736.

8. An, J.Y., Lin, K., Zhu, L., Werling, D.M., Dong, S., Brand, H., Wang, H.Z., Zhao, X., Schwartz, G.B., Collins, R.L., et al. (2018). Genome-wide de novo risk score implicates promoter variation in autism spectrum disorder. Science *362*, eaat6576.

9. Fischbach, G.D., and Lord, C. (2010). The Simons Simplex Collection: a resource for identification of autism genetic risk factors. Neuron *68*, 192–195.

10. Wapner, R.J., Martin, C.L., Levy, B., Ballif, B.C., Eng, C.M., Zachary, J.M., Savage, M., Platt, L.D., Saltzman, D., Grobman, W.A., et al. (2012). Chromosomal microarray versus karyotyping for prenatal diagnosis. N. Engl. J. Med. *367*, 2175–2184.

11. Vora, N.L., Gilmore, K., Brandt, A., Gustafson, C., Strande, N., Ramkissoon, L., Hardisty, E., Foreman, A.K.M., Wilhelmsen, K., Owen, P., et al. (2020). An approach to integrating exome sequencing for fetal structural anomalies into clinical practice. Genet. Med. *22*, 954–961.

12. Petrovski, S., Aggarwal, V., Giordano, J.L., Stosic, M., Wou, K., Bier, L., Spiegel, E., Brennan, K., Stong, N., Jobanputra, V., et al. (2019). Whole-exome sequencing in the evaluation of fetal structural anomalies: a prospective cohort study. Lancet *393*, 758–767.

13. Purcell, S.M., Moran, J.L., Fromer, M., Ruderfer, D., Solovieff, N., Roussos, P., O'Dushlaine, C., Chambert, K., Bergen, S.E., Kahler, A., et al. (2014). A polygenic burden of rare disruptive

mutations in schizophrenia. Nature *506*, 185–190.

14. Sanders, S.J., He, X., Willsey, A.J., Ercan-Sencicek, A.G., Samocha, K.E., Cicek, A.E., Murtha, M.T., Bal, V.H., Bishop, S.L., Dong, S., et al. (2015). Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. Neuron *87*, 1215–1233.

15. van der Auwera, G., and O'Connor, B.D. (2020). Genomics in the Cloud: Using Docker, GATK, and WDL in Terra (O'Reilly Media, Incorporated).

16. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760.

17. Poplin, R., Ruano-Rubio, V., DePristo, M.A., Fennell, T.J., Carneiro, M.O., Van der Auwera, G.A., Kling, D.E., Gauthier, L.D., Levy-Moonshine, A., Roazen, D., et al. (2018). Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv, doi: https://doi.org/10.1101/201178.

18. Chen, X., Schulz-Trieglaff, O., Shaw, R., Barnes, B., Schlesinger, F., Kallberg, M., Cox, A.J., Kruglyak, S., and Saunders, C.T. (2016). Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics *32*, 1220–1222.

19. Rausch, T., Zichner, T., Schlattl, A., Stutz, A.M., Benes, V., and Korbel, J.O. (2012). DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics *28*, i333–i339.

20. Layer, R.M., Chiang, C., Quinlan, A.R., and Hall, I.M. (2014). LUMPY: a probabilistic framework for structural variant discovery. Genome Biol. *15*, R84.

21. Babadi, M., Fu, J.M., Lee, S.K., Smirnov, A.N., Gauthier, L.D., Walker, M., Benjamin, D.I., Karczewski, K.J., Wong, I., Collins, R.L., et al. (2022). GATK-gCNV: A Rare Copy Number Variant Discovery Algorithm and Its Application to Exome Sequencing in the UK Biobank. bioRxiv. doi: https://doi.org/10.1101/2022.08.25.504851.

22. Klambauer, G., Schwarzbauer, K., Mayr, A., Clevert, D.A., Mitterecker, A., Bodenhofer, U., and Hochreiter, S. (2012). cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. Nucleic Acids Res. *40*, e69.

23. Gardner, E.J., Lam, V.K., Harris, D.N., Chuang, N.T., Scott, E.C., Pittard, W.S., Mills, R.E., Genomes Project, Consortium, and Devine, S.E. (2017). The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology. Genome Res. *27*, 1916–1929.

24. Collins, R.L., Brand, H., Redin, C.E., Hanscom, C., Antolik, C., Stone, M.R., Glessner, J.T., Mason, T., Pregno, G., Dorrani, N., et al. (2017). Defining the diverse spectrum of inversions, complex structural variation, and chromothripsis in the morbid human genome. Genome Biol. *18*, 36.

25. Dolzhenko, E., Deshpande, V., Schlesinger, F., Krusche, P., Petrovski, R., Chen, S., Emig-Agius, D., Gross, A., Narzisi, G., Bowman, B., et al. (2019). ExpansionHunter: a sequence-graph-based tool to analyze variation in short tandem repeat regions. Bioinformatics *35*, 4754–4756.

26. Riggs, E.R., Andersen, E.F., Cherry, A.M., Kantarci, S., Kearney, H., Patel, A., Raca, G., Ritter, D.I., South, S.T., Thorland, E.C., et al. (2020). Technical standards for the interpretation

and reporting of constitutional copy-number variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics (ACMG) and the Clinical Genome Resource (ClinGen). Genet. Med. *22*, 245–257.

27. Rentzsch, P., Witten, D., Cooper, G.M., Shendure, J., and Kircher, M. (2019). CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. *47*, D886–D894.

28. Samocha, K.E., Kosmicki, J.A., Karczewski, K.J., O'Donnell-Luria, A.H., Pierce-Hoffman, E., MacArthur, D.G., Neale, B.M., and Daly, M.J. (2017). Regional missense constraint improves variant deleteriousness prediction. BioRxiv. doi: https://doi.org/10.1101/148353.

29. Wright, C.F., Fitzgerald, T.W., Jones, W.D., Clayton, S., McRae, J.F., van Kogelenberg, M., King, D.A., Ambridge, K., Barrett, D.M., Bayzetinova, T., et al. (2015). Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. Lancet *385*, 1305–1314.

30. Deciphering Developmental Disorders, Study (2017). Prevalence and architecture of de novo mutations in developmental disorders. Nature *542*, 433–438.

31. Berg, J.S., Adams, M., Nassar, N., Bizon, C., Lee, K., Schmitt, C.P., Wilhelmsen, K.C., and Evans, J.P. (2013). An informatics approach to analyzing the incidentalome. Genet. Med. *15*, 36–44.

32. Blekhman, R., Man, O., Herrmann, L., Boyko, A.R., Indap, A., Kosiol, C., Bustamante, C.D., Teshima, K.M., and Przeworski, M. (2008). Natural selection on genes that underlie human disease susceptibility. Curr. Biol. *18*, 883–889.

33. Lord, J., McMullan, D.J., Eberhardt, R.Y., Rinck, G., Hamilton, S.J., Quinlan-Jones, E., Prigmore, E., Keelagher, R., Best, S.K., Carey, G.K., et al. (2019). Prenatal exome sequencing analysis in fetal structural anomalies detected by ultrasonography (PAGE): a cohort study. Lancet *393*, 747–757.

34. Firth, H.V., Richards, S.M., Bevan, A.P., Clayton, S., Corpas, M., Rajan, D., Van Vooren, S., Moreau, Y., Pettett, R.M., and Carter, N.P. (2009). DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. Am. J. Hum. Genet. *84*, 524–533.

35. Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature *485*, 376–380.

36. Lowther, C., Mehrjouy, M.M., Collins, R.L., Bak, M.C., Dudchenko, O., Brand, H., Dong, Z., Rasmussen, M.B., Gu, H., Weisz, D., et al. (2022). Balanced chromosomal rearrangements offer insights into coding and noncoding genomic features associated with developmental disorders. medRxiv. doi: https://doi.org/10.1101/2022.02.15.22270795.

37. Redin, C., Brand, H., Collins, R.L., Kammin, T., Mitchell, E., Hodge, J.C., Hanscom, C., Pillalamarri, V., Seabra, C.M., Abbott, M.A., et al. (2017). The genomic landscape of balanced cytogenetic abnormalities associated with human congenital anomalies. Nat. Genet. *49*, 36–45.

38. Fu, J.M., Satterstrom, F.K., Peng, M., Brand, H., Collins, R.L., Dong, S., Wamsley, B., Klei,

L., Wang, L., Hao, S.P., et al. (2022). Rare coding variation provides insight into the genetic architecture and phenotypic context of autism. Nat. Genet. *54*, 1320–1331.

39. Doan, R.N., Lim, E.T., De Rubeis, S., Betancur, C., Cutler, D.J., Chiocchetti, A.G., Overman, L.M., Soucy, A., Goetze, S., Autism Sequencing, Consortium, et al. (2019). Recessive gene disruptions in autism spectrum disorder. Nat. Genet. *51*, 1092–1098.

40. Coe, B.P., Witherspoon, K., Rosenfeld, J.A., van Bon, B.W., Vulto-van Silfhout, A.T., Bosco, P., Friend, K.L., Baker, C., Buono, S., Vissers, L.E., et al. (2014). Refining analyses of copy number variation identifies specific genes associated with developmental delay. Nat. Genet. *46*, 1063–1071.

41. Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S.F., Hakonarson, H., and Bucan, M. (2007). PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res. *17*, 1665–1674.

42. Colella, S., Yau, C., Taylor, J.M., Mirza, G., Butler, H., Clouston, P., Bassett, A.S., Seller, A., Holmes, C.C., and Ragoussis, J. (2007). QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. Nucleic Acids Res. *35*, 2013–2025.

43. MacDonald, J.R., Ziman, R., Yuen, R.K., Feuk, L., and Scherer, S.W. (2014). The Database of Genomic Variants: a curated collection of structural variation in the human genome. Nucleic Acids Res. *42*, D986-92.