Supplemental methods

DNA extraction

For all blood samples, DNA was extracted using the Autopure system (QIAGEN). For all buccal samples, DNA was extracted manually using the Puregene 300 Kit (QIAGEN). For saliva samples from PGPC-0002, PGPC-0006, and PGPC-0050, DNA was extracted using the QIAsymphony instrument with the QIAsymphony DSP DNA Midi Kit (QIAGEN), whereas for PGPC-0005, DNA was extracted using the Autopure system (QIAGEN). All extractions were performed following the manufacturer's instructions.

Bacterial DNA quantification

The relative proportion of human and bacterial DNA in each sample was quantified using the QX200 Droplet Digital PCR system (Bio-Rad Laboratories) following the manufacturer's instructions. Bacterial DNA content was measured using the 16S rRNA gene with TaqMan probe All1NUW (Life Technologies), and human DNA measured using *RPPH1* with the Copy Number Reference Assay RNase P probe (Life Technologies).

DNA library preparation and sequencing

From each individual, the blood sample plus one saliva sample and one buccal sample (representing a range of bacterial DNA concentrations) were selected for further analysis. Five DNA libraries were prepared per individual, from the following sources of DNA: 1) blood; 2) saliva without prior enrichment for eukaryotic DNA; 3) saliva with enrichment; 4) buccal swab without enrichment; and 5) buccal swab with enrichment. Eukaryotic DNA enrichment was performed using the NEBNext® Microbiome DNA Enrichment Kit (New England Biolabs) following the manufacturer's instructions to elute and collect the eukaryotic DNA from the beads. DNA library preparation was performed using the NxSeq AmpFREE Low DNA kit (Lucigen). Sequencing was performed using the HiSeq X (Illumina) as previously described [1], generating 151 bp paired-end reads.

Data preprocessing and variant detection

Base calling was performed using HiSeq Analysis Software (HAS) v2-2.5.55.1311. Reads were aligned to the GRCh37/hg19 reference genome using BWA-MEM v0.7.12 [2]. BAM files were subsampled prior to variant detection using SAMtools v1.3.1 [3] to give each sample approximately the same mean depth as the sample with the lowest original mean depth (specifically, the proportion of aligned reads to retain from each sample was calculated as D_M/D_X, where D_M is the minimum original mean read depth among all the samples and D_X is the original mean read depth of sample X). Picard v2.5.0 was used to mark duplicate reads. SNVs and indels were detected using Genome Analysis Toolkit (GATK) v3.7 following best practices recommendations [4–6]. SNVs and indels for which the FILTER column of the VCF file was not equal to PASS after performing variant quality score recalibration were discarded. Due to their higher error rate and propensity to be expansions or contractions of homopolymers or simple repeats, a given indel was retained only if it 1) was marked as PASS, 2) had a genotype quality (GQ) score ≥ 99 (for heterozygous indels only), 3) had an allele fraction between 0.3 and 0.7 (for heterozygous indels) or greater than 0.9 (for homozygous indels), 4) was not adjacent to a homopolymer of size 8 or greater, and 5) did not overlap with low-complexity regions marked by dustmasker [7]. An SNV or indel was defined as novel if it was absent from genomes and exomes in the Genome Aggregation Database (gnomAD) [8], the data for which was downloaded from ANNOVAR [9] on March 10, 2017. CNVs were detected using a workflow involving ERDS v1.1 [10] and CNVnator v0.3.2 [11] as previously described [1]. Specifically, a given CNV was retained if 1) it was larger than 1 kb, 2) less than 70% of it overlapped with gaps and segmental duplications [1], and 3) it was either detected by both ERDS and CNVnator with 50% reciprocal overlap or the CNV was a duplication smaller than 50 kb detected by ERDS alone. A CNV was defined as rare if it was present in fewer than 1% of parents in the Autism Speaks MSSNG WGS dataset (2,241 unrelated individuals) according to both ERDS and CNVnator [12]. As an alternative CNV-detection workflow, CNVs were identified using Canvas v1.39.0.1598 [13] and were retained if they

satisfied criteria 1 and 2 above. Structural variants were detected using Manta v1.5.0 [14] and were not subjected to any filtering.

Statistical analysis

The Friedman repeated-measures test was used to determine whether there were statistically-significant differences in the distributions of sequencing metrics or variant counts among blood, non-enriched saliva, and non-enriched buccal samples, followed by the post-hoc Conover-Iman test to assess pairwise differences. The Wilcoxon signed-rank test was used to compare the distributions of enriched saliva samples and non-enriched saliva samples, as well as enriched buccal samples and non-enriched buccal samples. Exact P-value tables were used for the Friedman test and the Wilcoxon signed-rank test [15, 16], while approximate P-values were used for the Conover-Iman test. P-values <0.05 were considered significant, except for the Wilcoxon test, for which a P-value of 0.062 (the lowest possible P-value with $n = 4$ [16]) was considered significant.

BLAST searches of unmapped reads

To determine the sources of unmapped reads, the Basic Local Alignment Search Tool (BLAST) program blastn v2.7.1 was used to search unmapped reads against the National Center for Biotechnology Information (NCBI) nucleotide (nt) database. Specifically, we selected 10,000 read pairs from each sample for which both reads in the pair were unmapped, and used the first read in each pair as a BLAST query. Default parameters were used to blastn except "-task blastn" and "-evalue 1e-5". Only the best BLAST match (smallest E-value) for each read was considered. For each match, the esummary program from NCBI was used to determine the taxonomic ID of the organism from which the database sequence was derived. The domain (e.g., bacteria or eukaryota) associated with that taxonomic ID was determined using the fullnamelineage.dmp file, which can be downloaded from https://ftp.ncbi.nlm.nih.gov/pub/taxonomy/new_taxdump/new_taxdump.tar.gz.

Concordance and accuracy of single nucleotide variants and indels

Using hap.py (https://github.com/Illumina/hap.py), SNV and indel concordance was calculated between blood samples and non-enriched saliva samples, blood samples and non-enriched buccal samples, enriched saliva samples and non-enriched saliva samples, and enriched buccal samples and non-enriched buccal samples. In addition to whole-genome concordance, concordance was computed for variants found in coding exons, all exons, introns, and intergenic regions, as defined by the RefSeq gene annotations obtained from the University of California Santa Cruz genome browser.

To determine whether there were differences in the accuracy of SNV and indel detection among the five sample types, we performed visual inspection of alignments [1] using IGV [17] to categorize as true or false 400 randomly-selected variants that were detected in the blood sample but not the non-enriched saliva or buccal sample from the same individual, or in an enriched sample but not in the corresponding non-enriched sample (or vice versa). These variants were distributed roughly evenly among the four study participants (PGPC-0002, PGPC-0005, PGPC-0006, or PGPC-0050), presence in gnomAD (known or novel), and variant type (SNV or indel). Evidence displayed by IGV that was considered to support the correctness of a variant include the variant being present in approximately equal proportions on forward and reverse reads; the allele fraction being between 0.3 and 0.7 (for heterozygous variants); consistent length (for indels); and absence of reads surrounding the variant that were poorly mapped or soft-clipped (where only part of the read aligns).

Concordance and accuracy of copy number variants

Two CNVs were deemed the same if the first CNV overlapped with at least 50% of the second CNV and vice versa. Concordance was calculated by enumerating the CNVs that were detected in blood samples but not non-enriched saliva or buccal samples (or vice versa) or in enriched but not non-enriched saliva

or buccal samples (or vice versa). To compare the accuracy of different sample types, read alignments corresponding to discordant CNVs were visually inspected using IGV as previously described [1]. All rare CNVs were visually inspected, along with three common CNVs per category, where a category was defined as a combination of individual (PGPC-0002, PGPC-0005, PGPC-0006, or PGPC-0050), CNV type (deletion or duplication), size bin (between 1 and 5 kb, between 5 and 10 kb, or greater than 10 kb), comparison type (blood versus non-enriched saliva, blood versus non-enriched buccal, enriched versus non-enriched saliva, or enriched versus non-enriched buccal) and comparison direction (CNV detected in the first sample type but not the second sample type or vice versa). When there were fewer than three discordant CNVs in a given category, all CNVs in that category were inspected. For the two HuRef blood replicates, all rare discordant CNVs were inspected, along with up to three common CNVs per combination of CNV type, size bin, and comparison direction (CNV detected in the first blood sample but not the second or vice versa).

References

1 Trost B, Walker S, Wang Z, Thiruvahindrapuram B, MacDonald JR, Sung WWL, Pereira SL, Whitney J, Chan AJS, Pellecchia G, Reuter MS, Lok S, Yuen RKC, Marshall CR, Merico D, Scherer SW. A comprehensive workflow for read depth-based identification of copy-number variation from whole-genome sequence data. *Am J Hum Genet* 2018;**102**:142–55.

2 Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009;**25**:1754–60.

3 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009;**25**:2078–9.

4 DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 2011;**43**:491–8.

5 McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010;**20**:1297–303.

6 Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, Banks E, Garimella KV, Altshuler D, Gabriel S, DePristo MA. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinforma* 2013;**43**:11.10.1-33.

7 Morgulis A, Gertz EM, Schäffer AA, Agarwala R. A fast and symmetric DUST implementation to mask low-complexity DNA sequences. *J Comput Biol* 2006;**13**:1028–40.

8 Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, Tukiainen T, Birnbaum DP, Kosmicki JA, Duncan LE, Estrada K, Zhao F, Zou J, Pierce-Hoffman E, Berghout J, Cooper DN, Deflaux N, DePristo M, Do R, Flannick J, Fromer M, Gauthier L, Goldstein J, Gupta N, Howrigan D, Kiezun A, Kurki MI, Moonshine AL, Natarajan P, Orozco L, Peloso GM, Poplin R, Rivas MA, Ruano-Rubio V, Rose SA, Ruderfer DM, Shakir K, Stenson PD, Stevens C, Thomas BP, Tiao G, Tusie-Luna MT, Weisburd B, Won H-H, Yu D, Altshuler DM, Ardissino D, Boehnke M, Danesh J, Donnelly S, Elosua R, Florez JC, Gabriel SB, Getz G, Glatt SJ, Hultman CM, Kathiresan S, Laakso M, McCarroll S, McCarthy MI, McGovern D, McPherson R, Neale BM, Palotie A, Purcell SM, Saleheen D, Scharf JM, Sklar P, Sullivan PF, Tuomilehto J, Tsuang MT, Watkins HC, Wilson JG, Daly MJ, MacArthur DG, Exome Aggregation Consortium. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 2016;**536**:285–91.

9 Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 2010;**38**:e164.

10 Zhu M, Need AC, Han Y, Ge D, Maia JM, Zhu Q, Heinzen EL, Cirulli ET, Pelak K, He M, Ruzzo EK, Gumbs C, Singh A, Feng S, Shianna KV, Goldstein DB. Using ERDS to infer copy-number variants in high-coverage genomes. *Am J Hum Genet* 2012;**91**:408–21.

11 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**:974–84.

12 Yuen RKC, Merico D, Bookman M, L Howe J, Thiruvahindrapuram B, Patel RV, Whitney J, Deflaux N, Bingham J, Wang Z, Pellecchia G, Buchanan JA, Walker S, Marshall CR, Uddin M, Zarrei M, Deneault E, D'Abate L, Chan AJS, Koyanagi S, Paton T, Pereira SL, Hoang N, Engchuan W, Higginbotham EJ, Ho K, Lamoureux S, Li W, MacDonald JR, Nalpathamkalam T, Sung WWL, Tsoi FJ, Wei J, Xu L, Tasse A-M, Kirby E, Van Etten W, Twigger S, Roberts W, Drmic I, Jilderda S, Modi BM, Kellam B, Szego M, Cytrynbaum C, Weksberg R, Zwaigenbaum L, Woodbury-Smith M, Brian J, Senman L, Iaboni A, Doyle-Thomas K, Thompson A, Chrysler C, Leef J, Savion-Lemieux T, Smith IM, Liu X, Nicolson R, Seifer V, Fedele A, Cook EH, Dager S, Estes A, Gallagher L, Malow BA, Parr JR, Spence SJ, Vorstman J, Frey BJ, Robinson JT, Strug LJ, Fernandez BA, Elsabbagh M, Carter MT, Hallmayer J, Knoppers BM, Anagnostou E, Szatmari P, Ring RH, Glazer D, Pletcher MT, Scherer SW. Whole genome sequencing resource identifies 18 new candidate genes for autism spectrum disorder. *Nat Neurosci* 2017;**20**:602–11.

13 Roller E, Ivakhno S, Lee S, Royce T, Tanner S. Canvas: versatile and scalable detection of copy number variants. *Bioinformatics* 2016;**32**:2375–7.

14 Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Källberg M, Cox AJ, Kruglyak S, Saunders CT. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* 2015;**32**:1220–2.

15 Martin L, Leblanc R, Toan NK. Tables for the Friedman rank test. *Can J Stat* 1993;**21**:39–43.

16 Dixon WJ, Massey FJ. *Introduction to statistical analysis*. 2nd ed. McGraw-Hill 1957.

17 Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. Integrative genomics viewer. *Nat Biotechnol* 2011;**29**:24–6.