

**Genomic architecture of Autism Spectrum
Disorder from comprehensive whole-genome
sequence annotation**

Supplementary information

Supplementary Results

Sample quality control

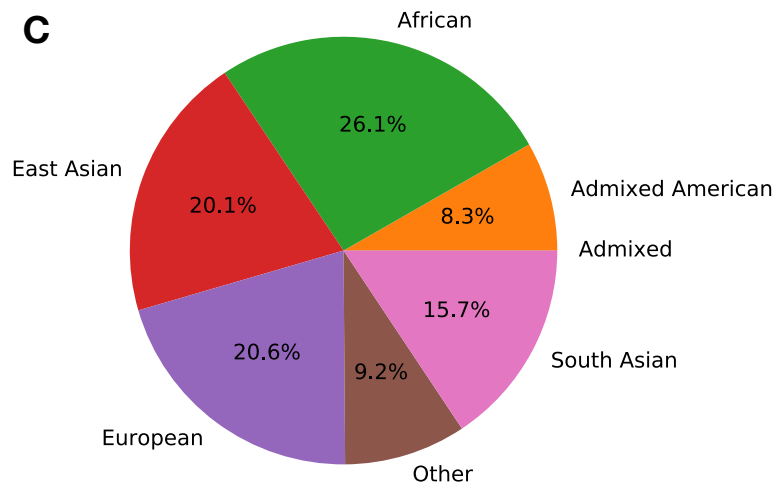
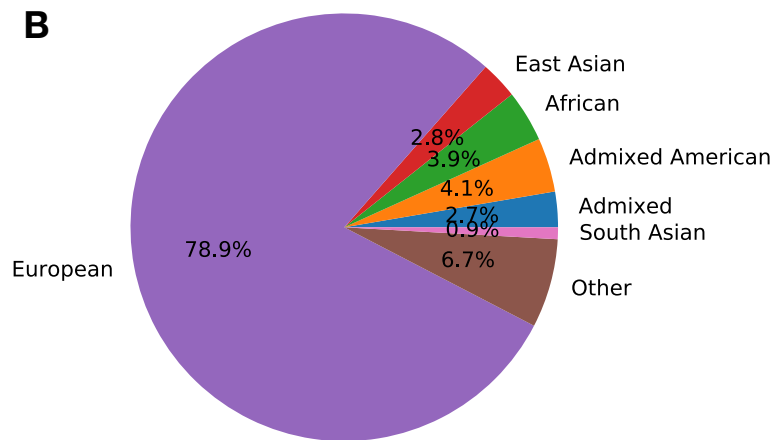
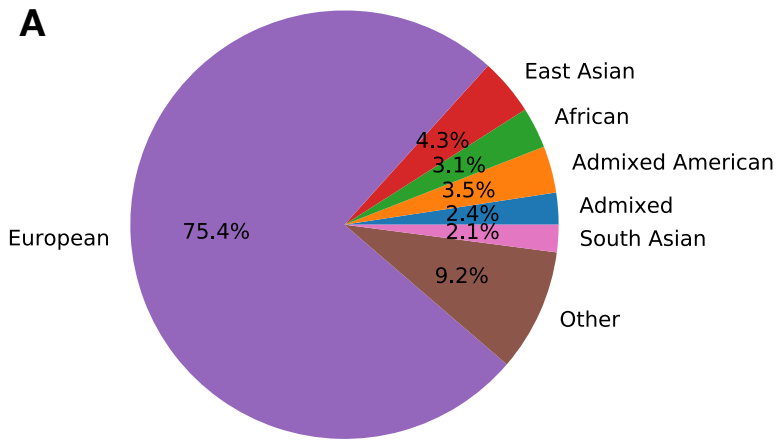
To assess the quality and comparability of the WGS samples used in this study, we stratified the samples into categories based on dataset (MSSNG, SSC, or 1000G), sequencing platform (Complete Genomics or Illumina HiSeq 2000/2500/X), and DNA library preparation method (PCR-based or PCR-free) and then examined the distributions of SNV ([Supplementary Figure 19](#)) and indel ([Supplementary Figure 20](#)) counts for various classes of variants (all, rare, rare exonic, rare damaging missense (DMis), rare protein-truncating variants (PTVs), or *de novo*). Only high-quality variants (as defined in Methods) were included. All categories had similar distributions, except that Complete Genomics samples differed from the others in terms of indel counts, in particular with fewer indels overall but more *de novo* indels ([Supplementary Figure 20](#)).

Tandem repeat data quality control

Because samples were added to MSSNG after our study on tandem repeat expansions in ASD (Trost et al., 2020), we ran ExpansionHunter Denovo v0.7.0 (EHdn) (Dolzhenko et al., 2020) on the additional samples, and then repeated QC measures using the full dataset. Each cohort comprising MSSNG had similar distributions of the number of tandem repeat loci detected by EHdn ([Supplementary Figure 21 \(A\)](#)). As we observed previously (Trost et al., 2020), samples sequenced on the Illumina HiSeq 2000 platform had substantially higher call counts than those sequenced on the HiSeq X or HiSeq 2500 platforms ([Supplementary Figure 21 \(A\)](#)). Thus, HiSeq 2000 samples were excluded. Samples sequenced using PCR-based DNA library preparation generally had higher call counts than those using PCR-free library preparation ([Supplementary Figure 21 \(B\)](#)); however, there was substantial overlap in the PCR-based versus PCR-free distributions, so PCR-based samples were retained.

Subsequently, we tagged samples that were outliers in terms of call counts or principal components (PCs). After removing HiSeq 2000 samples, samples having call counts greater than three standard deviations higher than the mean were tagged as failing QC. For QC based on PCs, the EHdn output was converted to a matrix, where the value of a given cell was the size reported by EHdn for a given sample-locus combination (loci not called by EHdn in a given sample were given a value of 0). Principal component analysis was then performed with that matrix as input using the R function *prcomp*. A handful of samples were extreme outliers in terms of some PCs ([Supplementary Figure 22](#)), hindering outlier detection for the remaining samples. Thus, two rounds of outlier detection were performed: first removing samples with PC outside mean \pm 3SD, recomputing the mean and standard deviation, and then again tagging samples outside mean \pm 3SD. Any sample that was an outlier in either round of outlier detection for any of the first five PCs was deemed to have failed QC.

Supplementary figures



Supplementary Figure 1: Genotype-computed ancestry of individuals from (A) MSSNG, (B) Simons Simplex Collection, and (C) 1000 Genomes Project.

CNV Query

One Box CNV Query

All coordinates are 0-based and use the GRCh38/hg38 reference.

- [CNV Query History](#)
- [Saved CNV Queries](#)
- [Advanced CNV Query](#)

What Can You Do?

With the one box CNV query, you can quickly make certain queries as shown in the table below.

Copy number variations for a specific gene symbol
LAMB2P1
Copy number variations for a genomic interval
chr2:50497000-50498000, chr2:50497000-50498000
Copy number variations for a specific subject
2-1116-003
Copy number variations for a specific family
FAM_1-0007-003

The Default Criteria

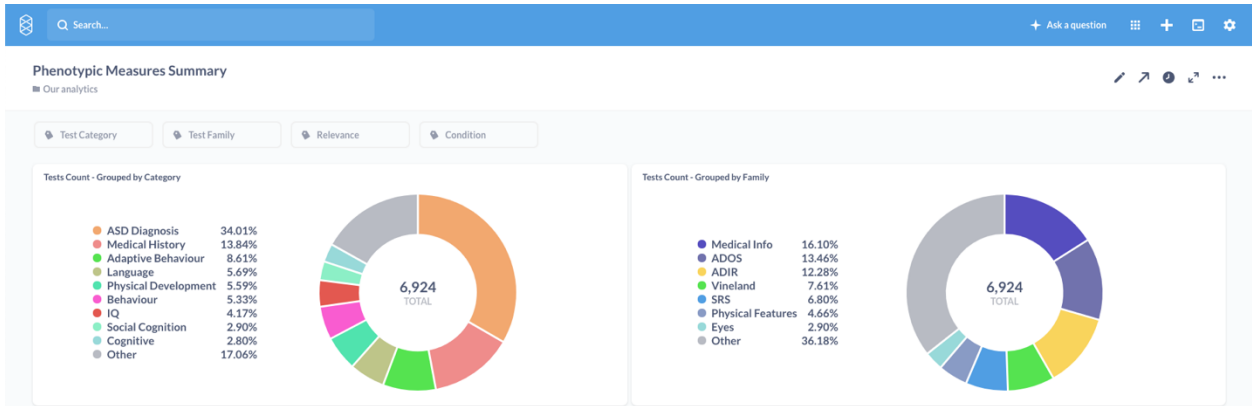
When you use the One Box Query, here are the default criteria.

Variant quality	Passing
Minimum CNV Length	>= 1000

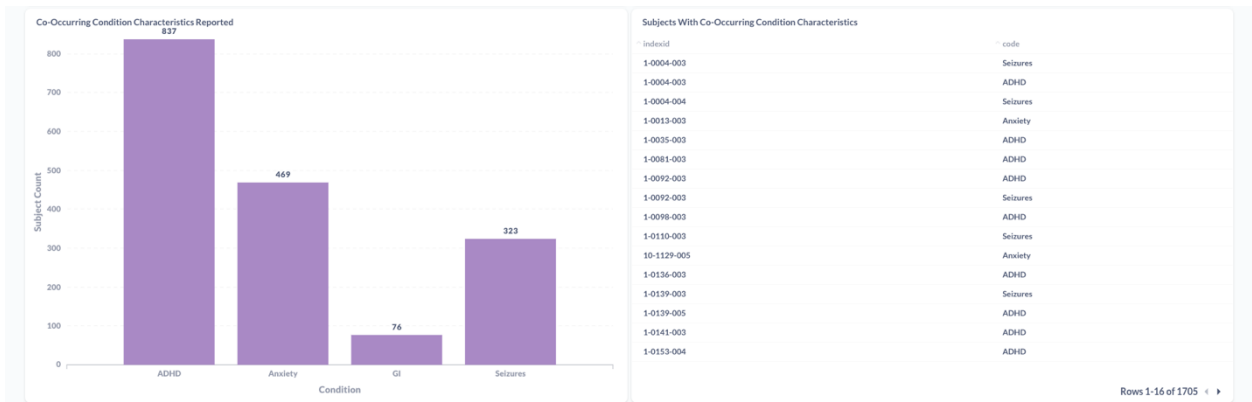
Use the Advanced Query for alternative query parameters.

Supplementary Figure 2: Screenshot of the interface for querying copy number variants in the MSSNG portal.

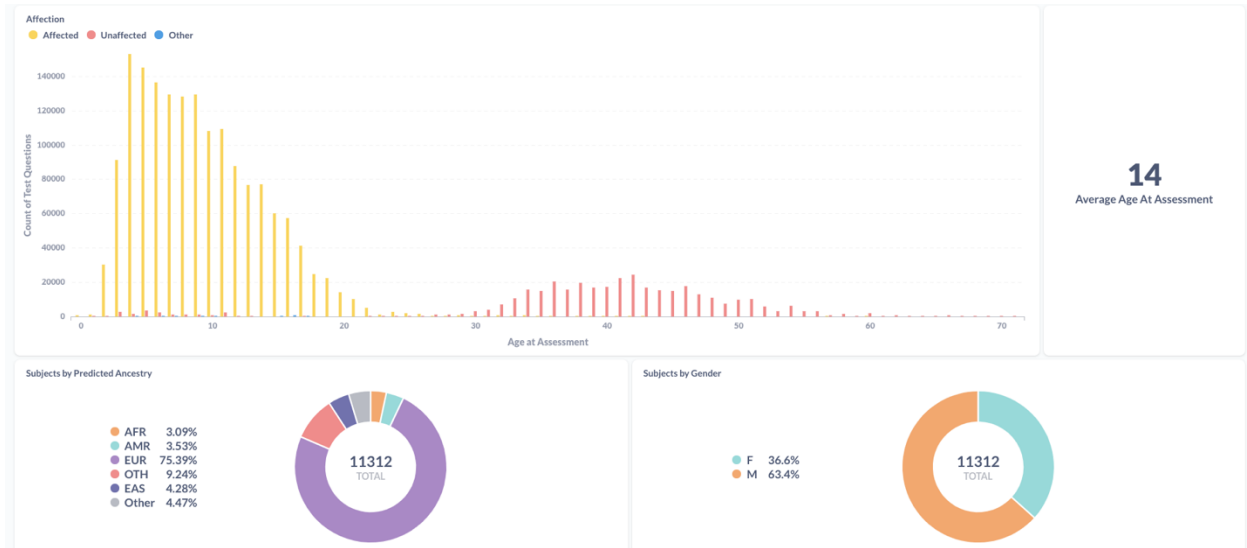
A



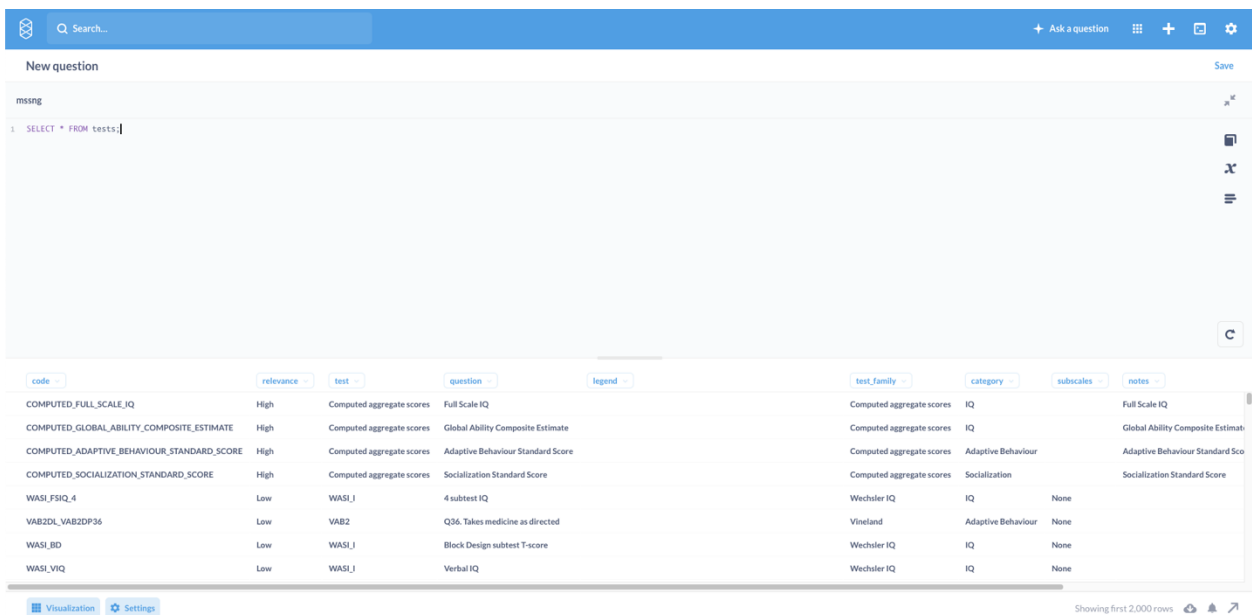
B



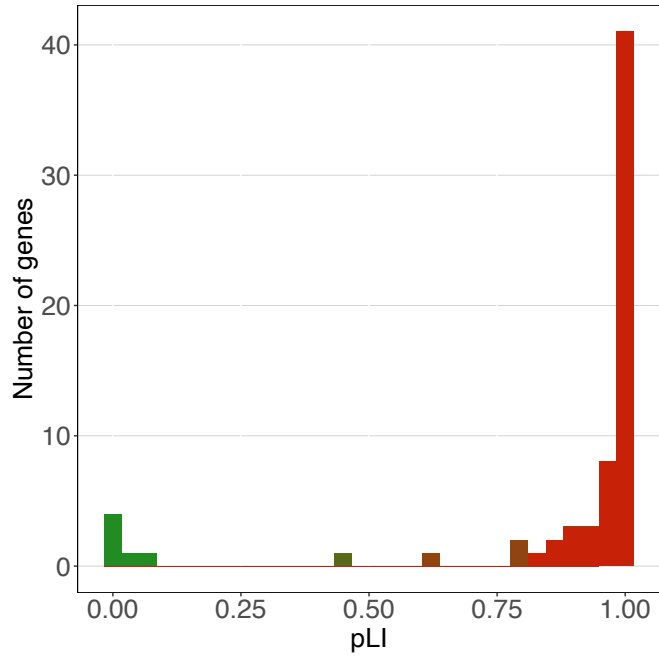
C



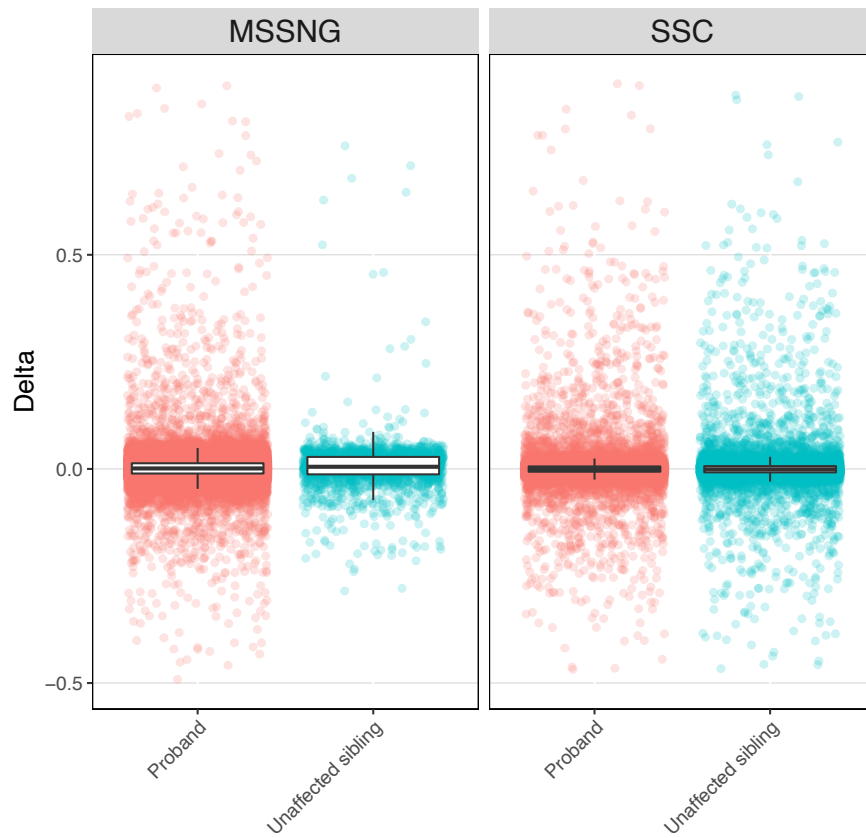
D



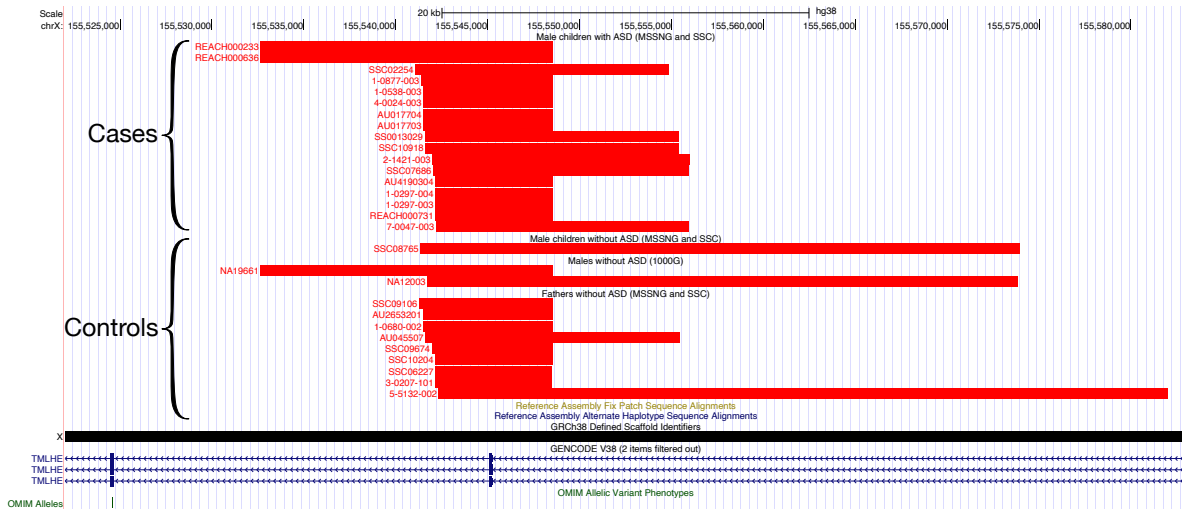
Supplementary Figure 3: Screenshots of the Phenotype Data Explorer, showing (A) categorization of tests, (B) co-occurring conditions, (C) summary information involving ancestry, sex, and age at assessment; and (D) results when querying individual test scores.



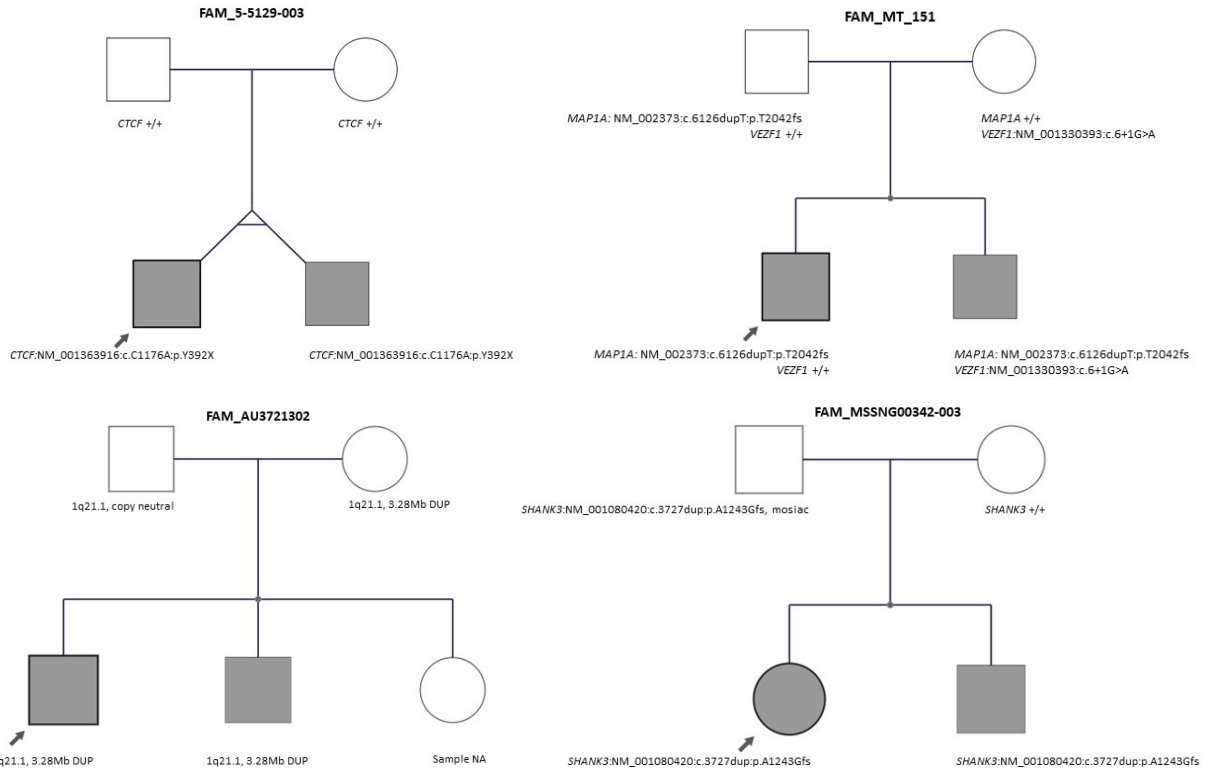
Supplementary Figure 4: Distribution of pLI values for the 68 ASD-associated genes not reported in the previous TADA+ analysis (Satterstrom et al., 2020).



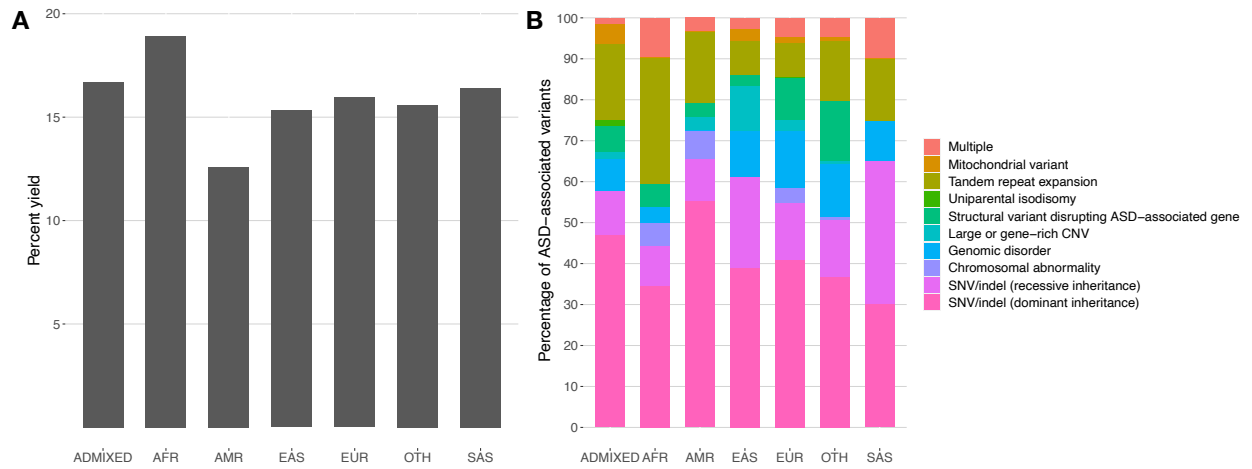
Supplementary Figure 5: Distributions of heteroplasmy deltas (child heteroplasmy minus mother heteroplasmy for a given position in the mitochondrial genome sequence) in individuals with ASD versus unaffected siblings in MSSNG (variants tested: ASD n=35,279; unaffected sibling n=2,171) and SSC (ASD n=29,005; unaffected sibling n=18,328).



Supplementary Figure 6: Deletions in males overlapping exon 2 of *TMLHE*. Deletions are categorized as being in children with ASD in MSSNG or SSC, children without ASD in MSSNG or SSC (i.e., unaffected siblings), population controls from the 1000 Genomes Project, and unaffected fathers in MSSNG or SSC.

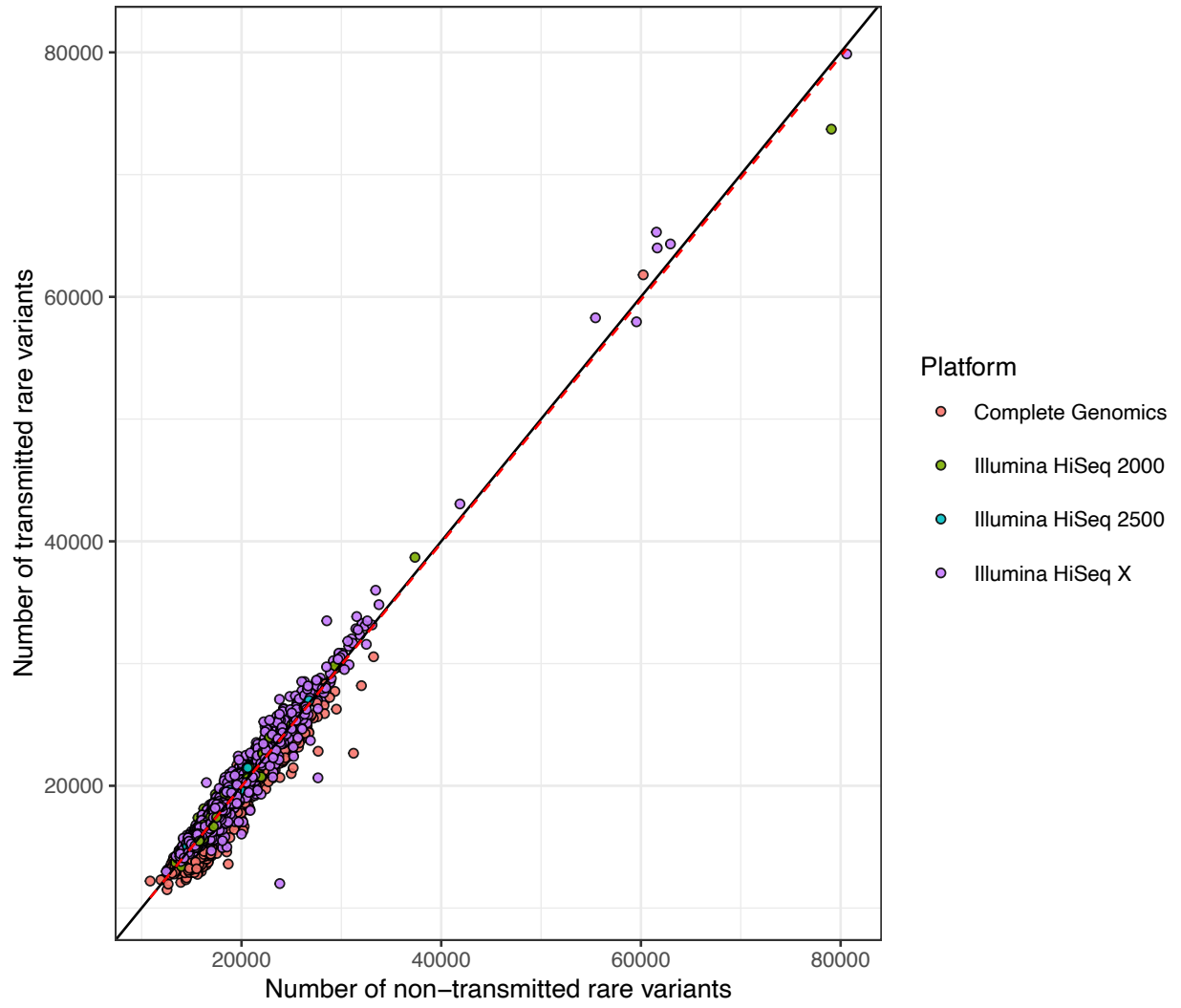


Supplementary Figure 7: Examples of multiplex families from MSSNG with ASD-associated rare variants.

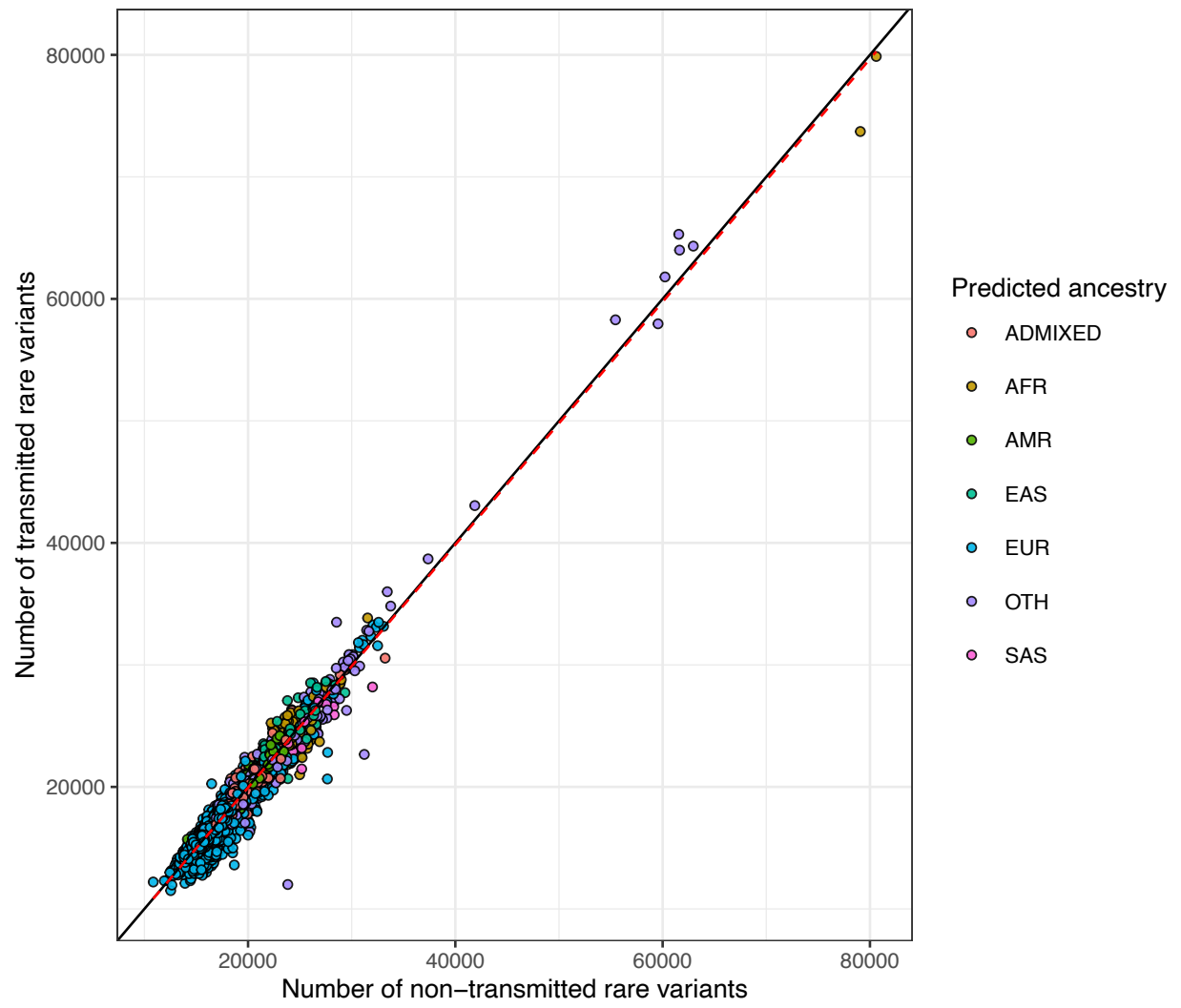


Supplementary Figure 8: ASD-associated rare variant yield and composition, stratified by ancestry group. (A) Percentage of individuals with ASD in each ancestry group having an ASD-associated rare variant. (B) Percentage of individuals with ASD with an ASD-associated rare variant having a rare variant in each category.

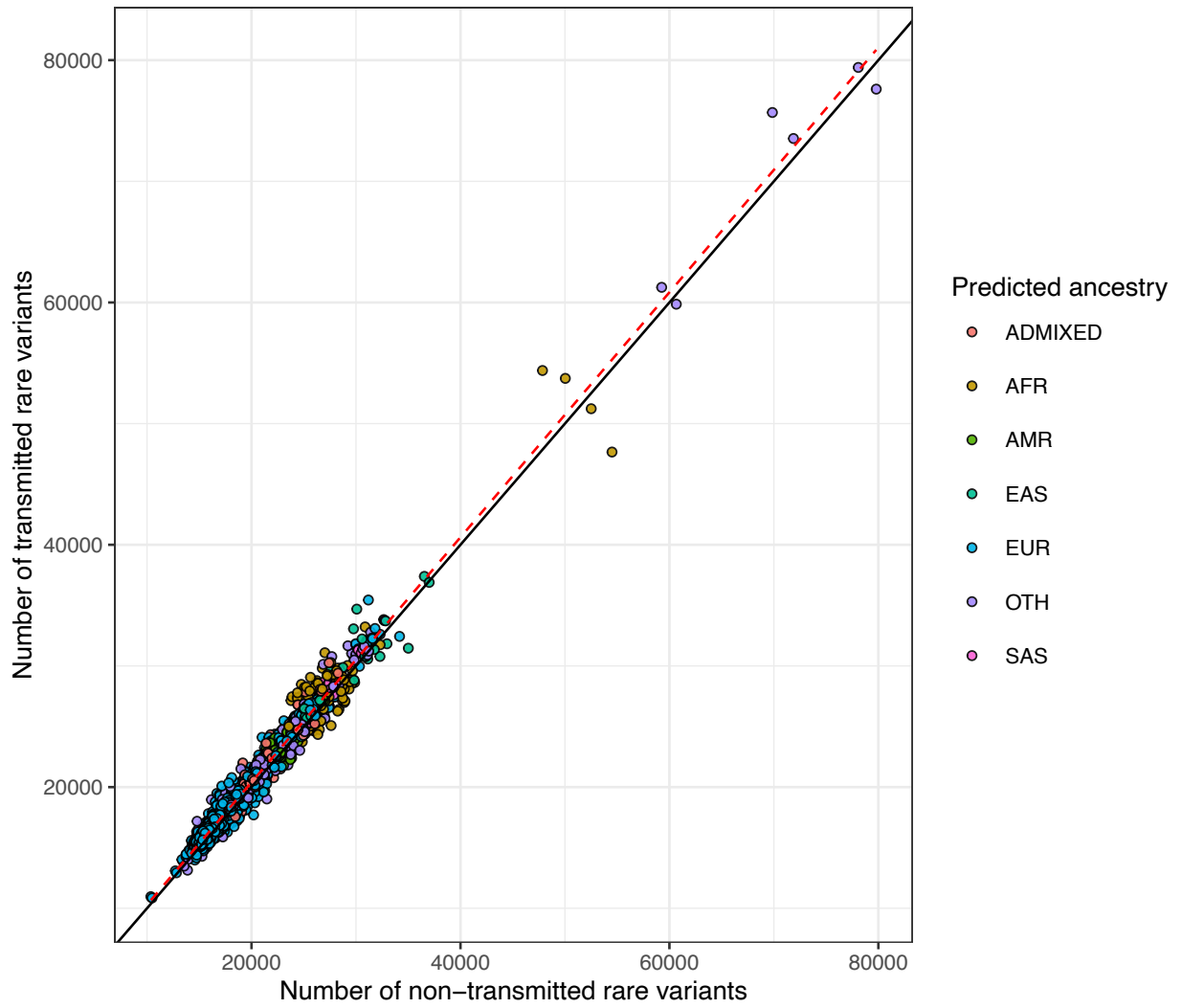
A



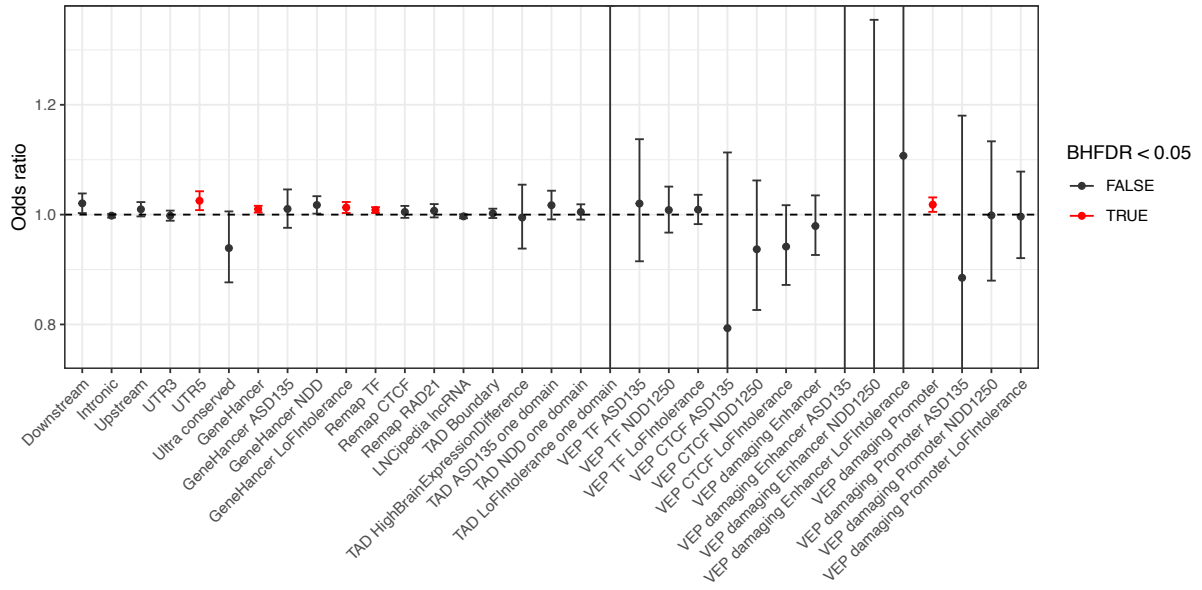
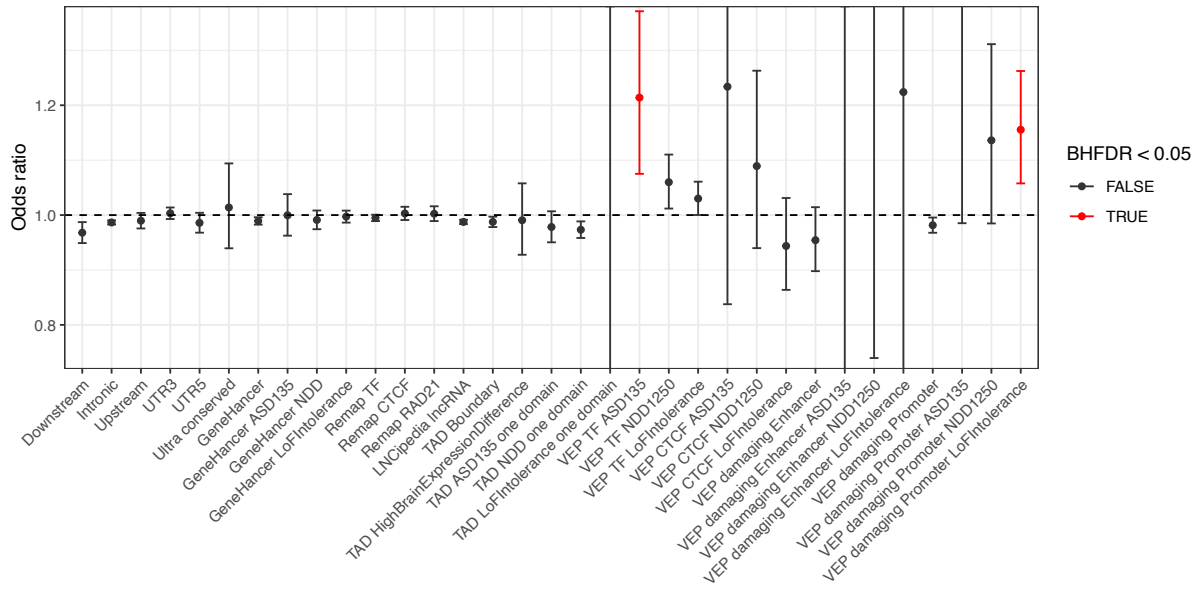
B

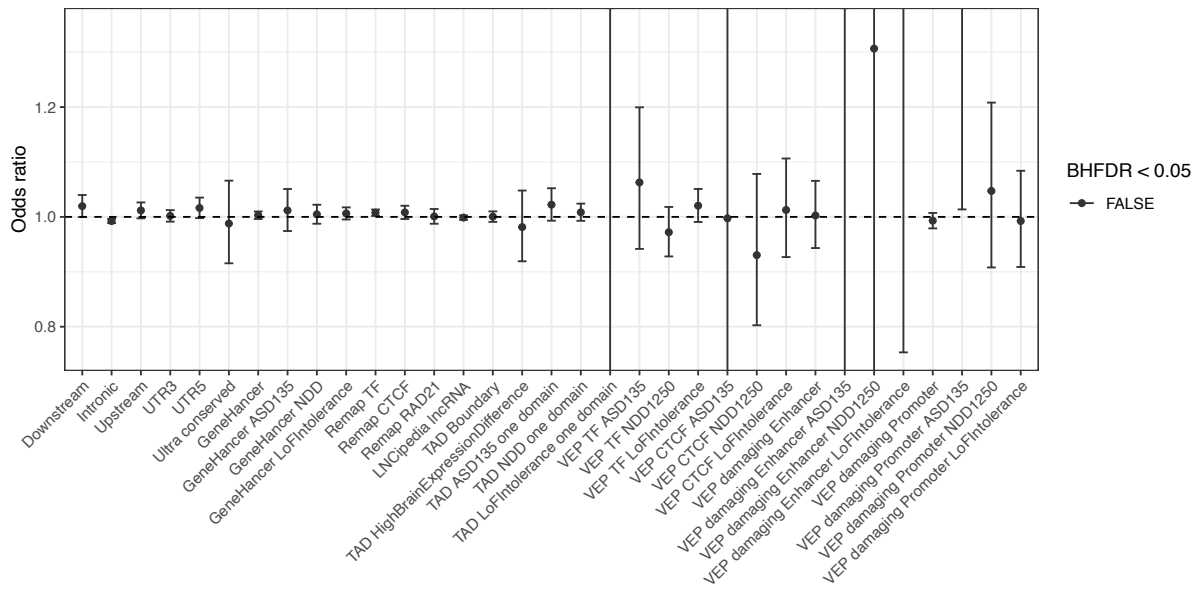


C

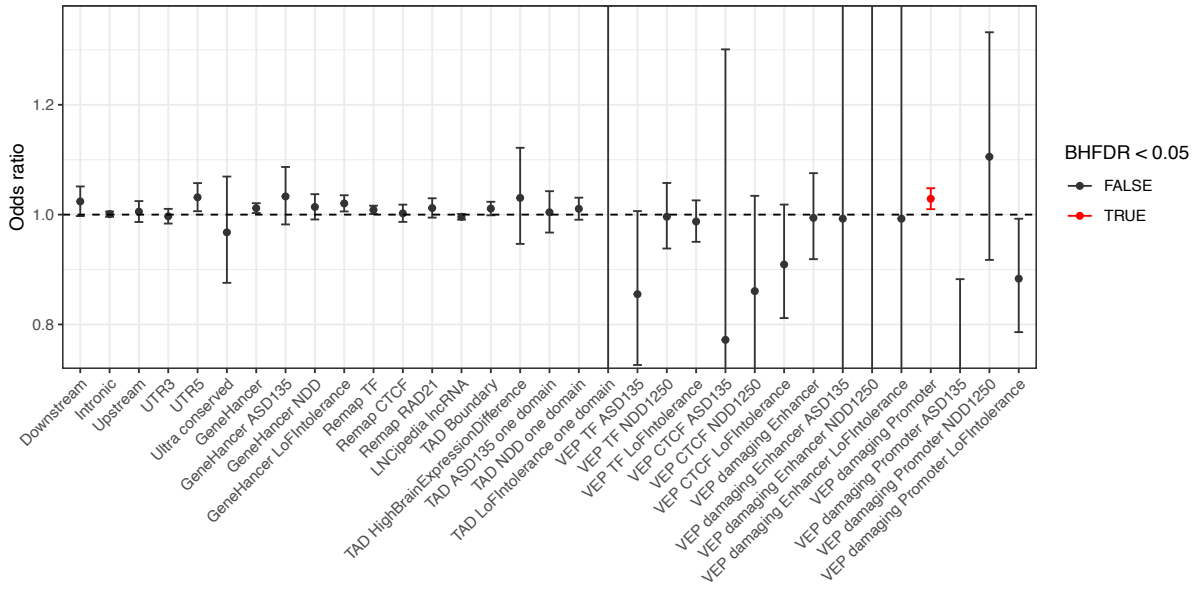
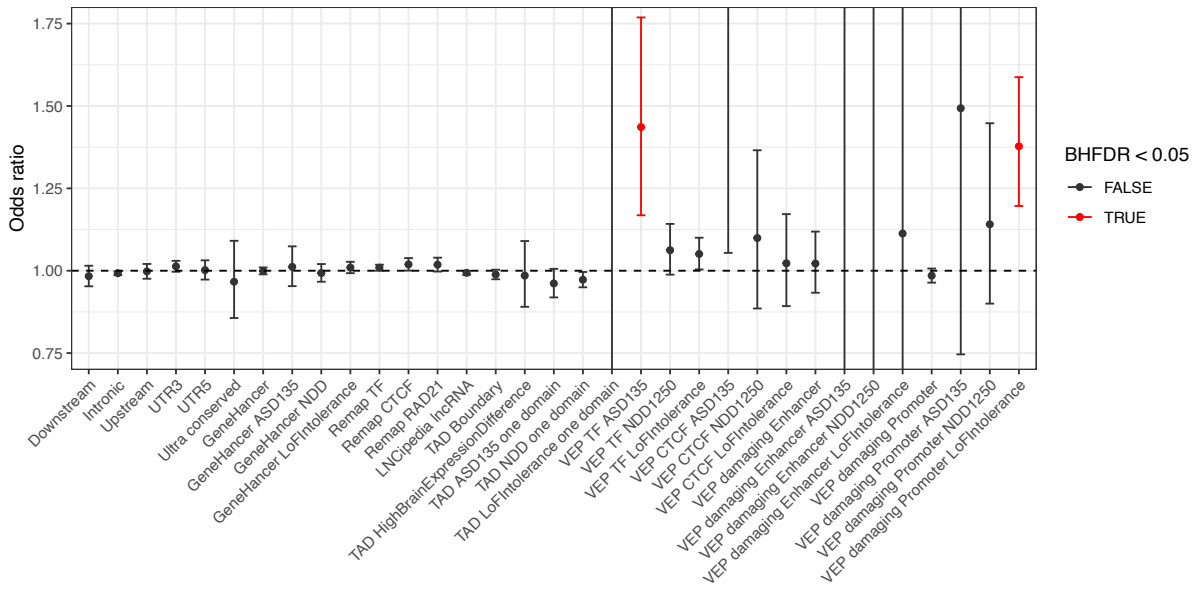


Supplementary Figure 9: Number of transmitted and non-transmitted rare variants in parents from MSSNG and SSC. (A) MSSNG parents, stratified by sequencing platform. (B) MSSNG parents, stratified by predicted ancestry. (C) SSC parents, stratified by predicted ancestry.

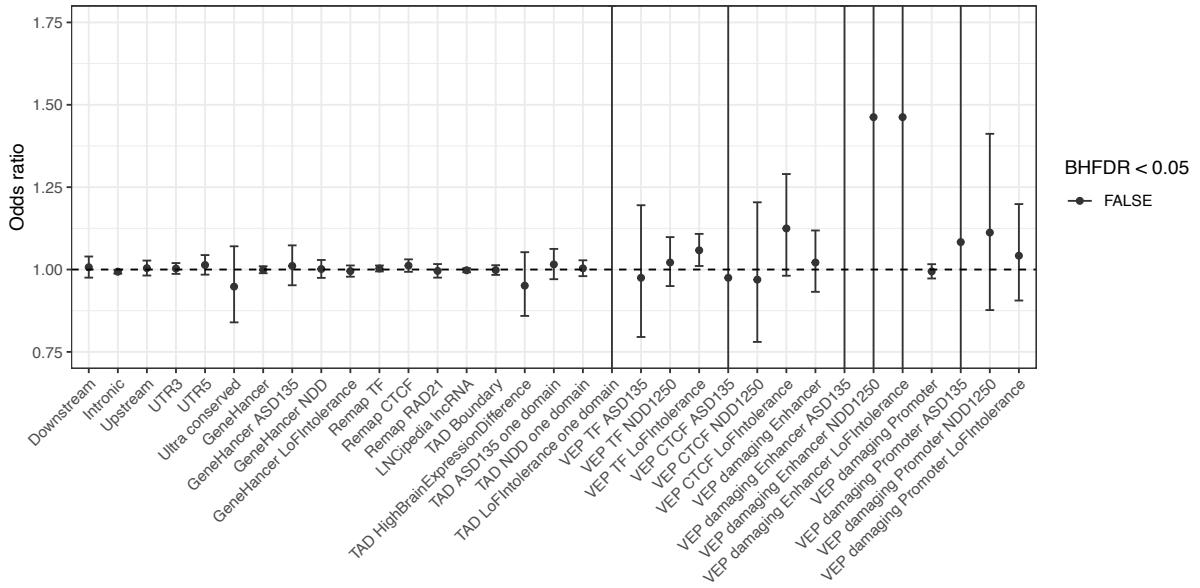
A**B**

C

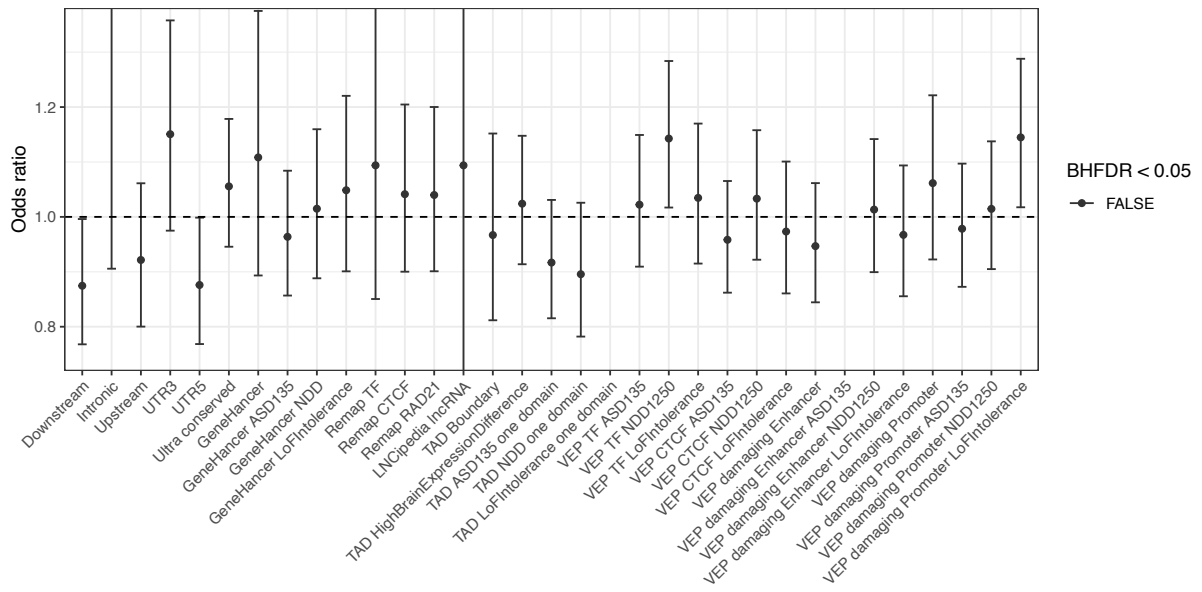
Supplementary Figure 10: Odds ratios for the number of rare variants (<0.1% frequency) in a given category that were transmitted versus non-transmitted in (A) Individuals with ASD in MSSNG, (B) individuals with ASD in SSC, and (C) unaffected siblings in SSC.

A**B**

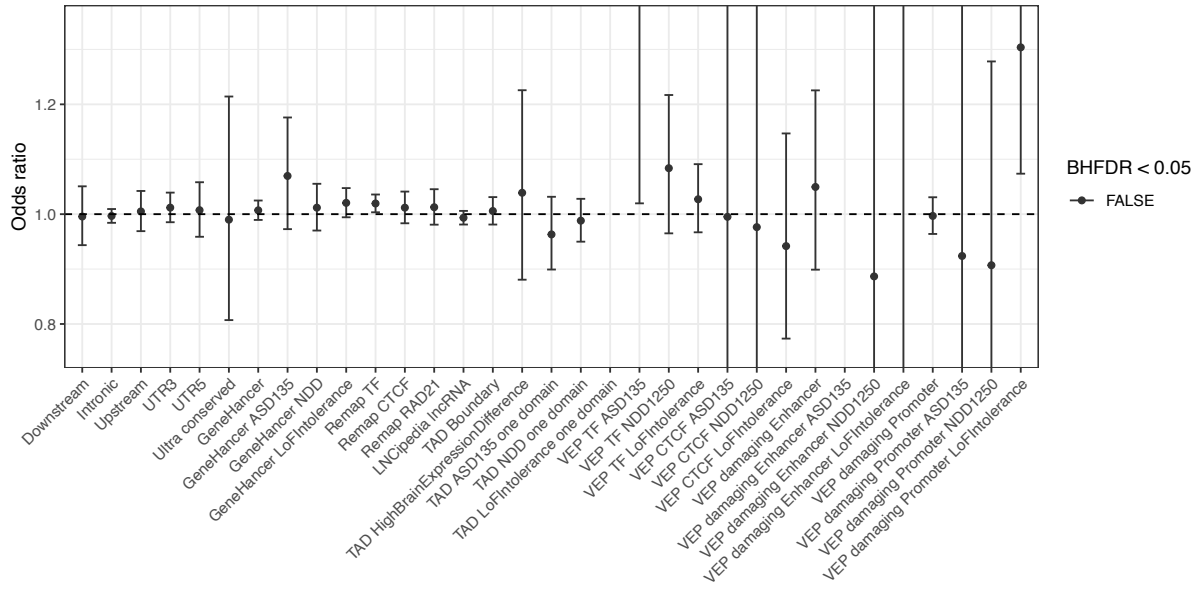
C



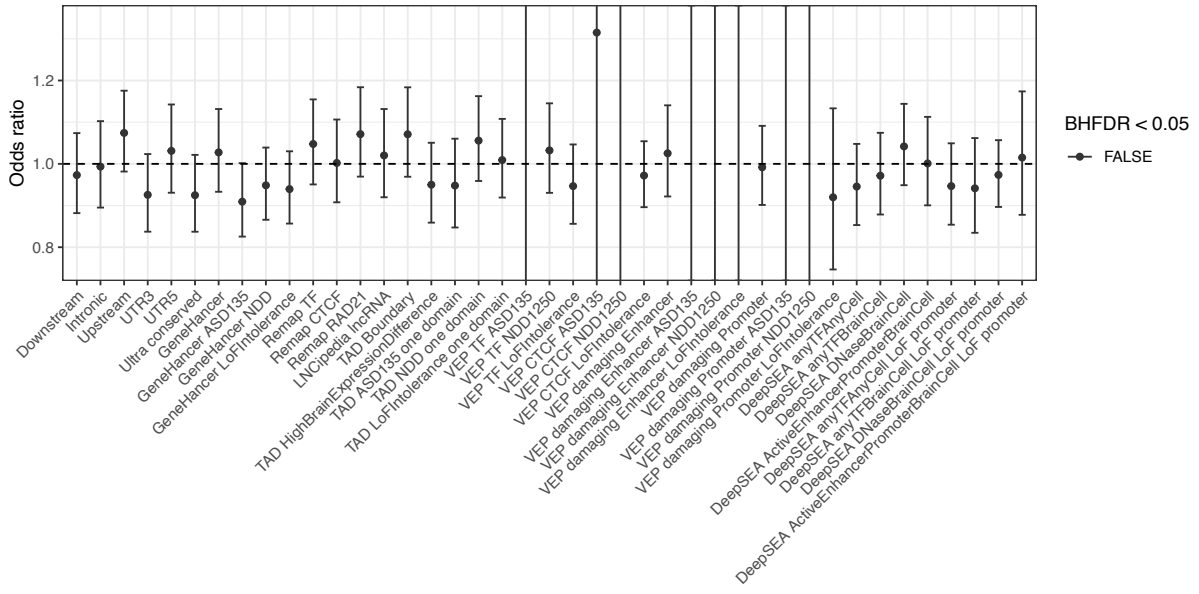
Supplementary Figure 11: Odds ratios for the number of singleton variants (private to a family) in a given category that were transmitted versus non-transmitted in (A) individuals with ASD in MSSNG, (B) individuals with ASD in SSC, and (C) unaffected siblings in SSC.



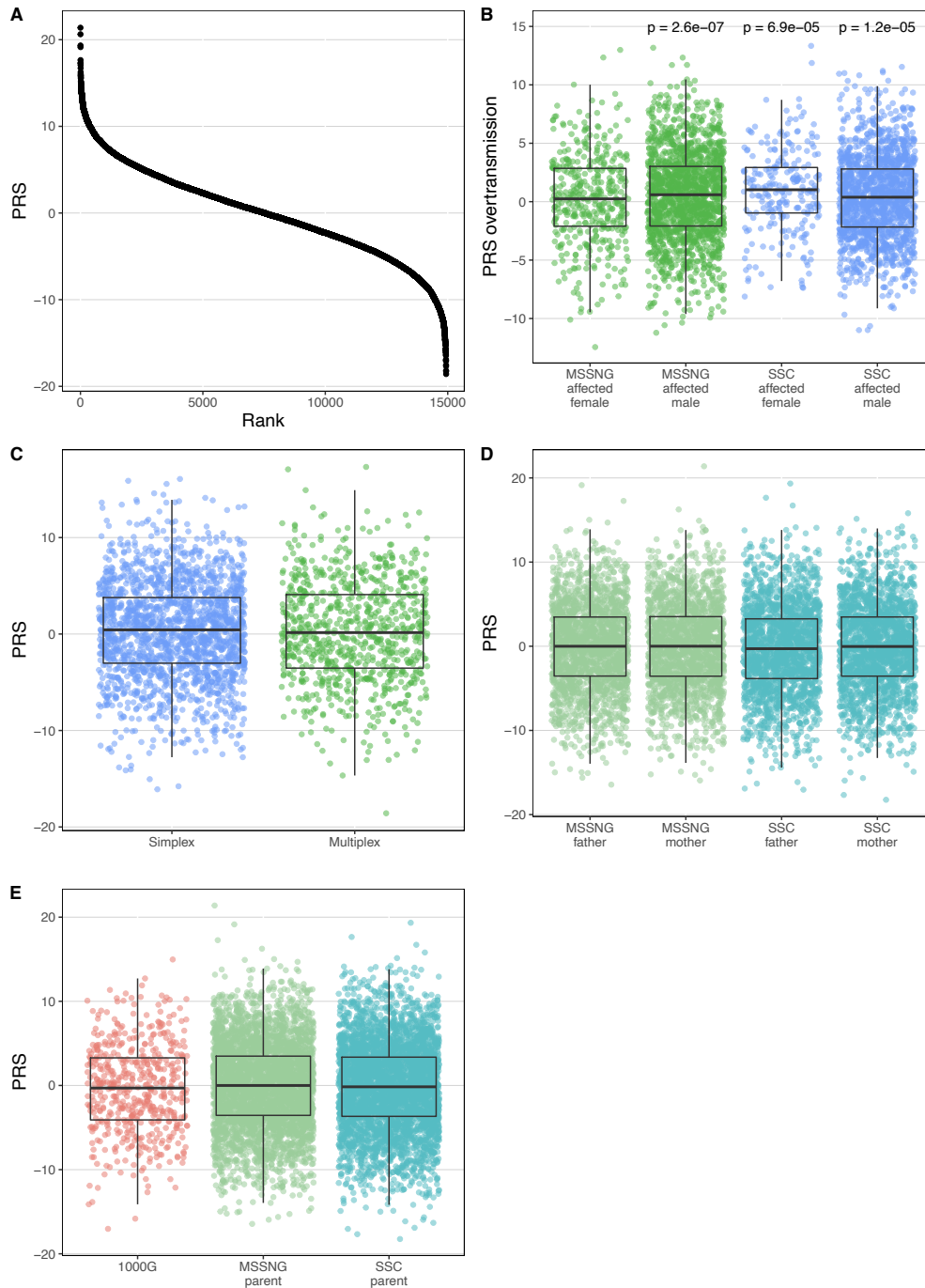
Supplementary Figure 12: Odds ratios for the number of rare variants in a given category in individuals with ASD from SSC versus their unaffected siblings.



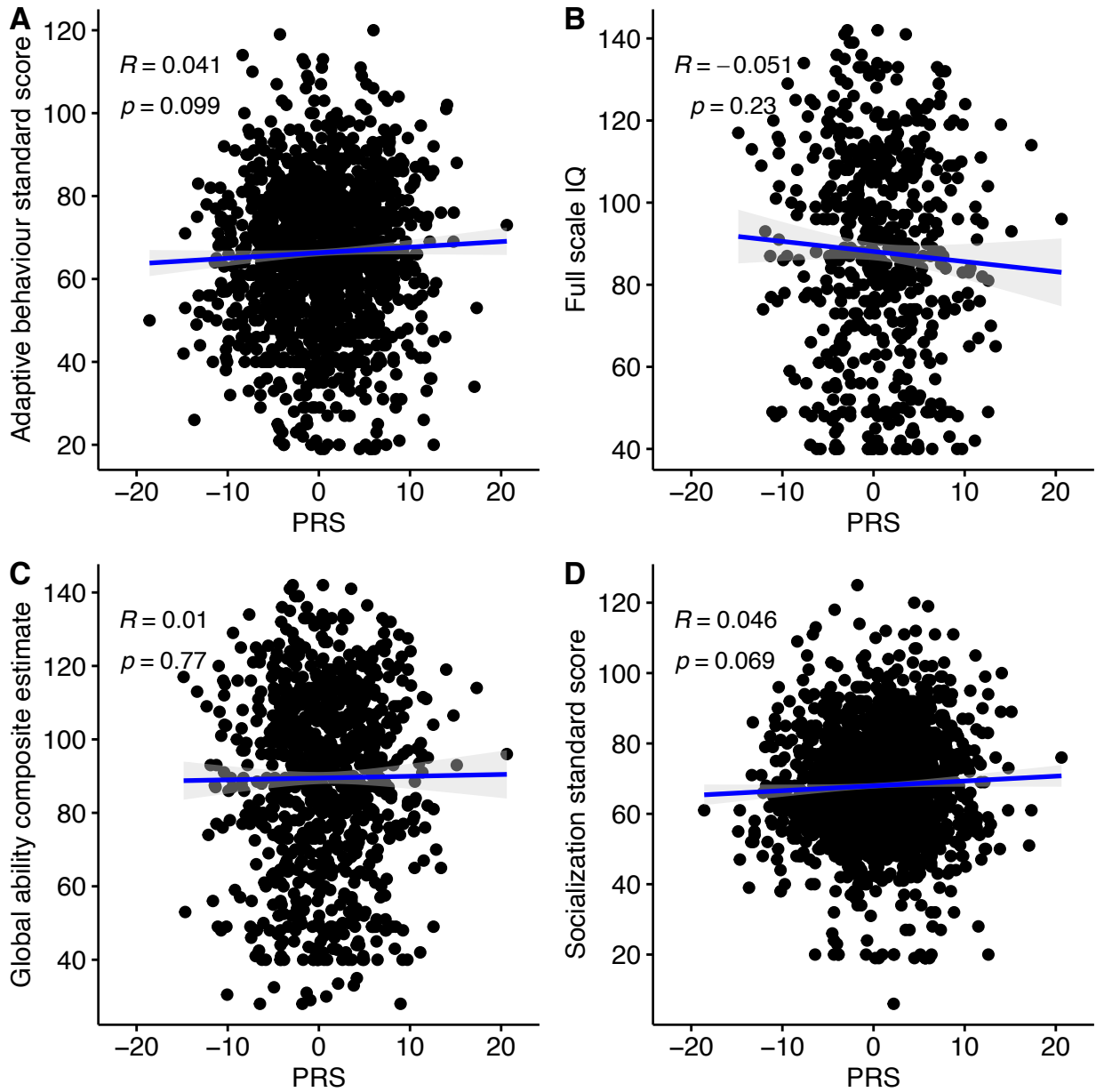
Supplementary Figure 13: Odds ratios for the number of singleton variants (private to a family) in a given category in individuals with ASD from SSC versus their unaffected siblings.



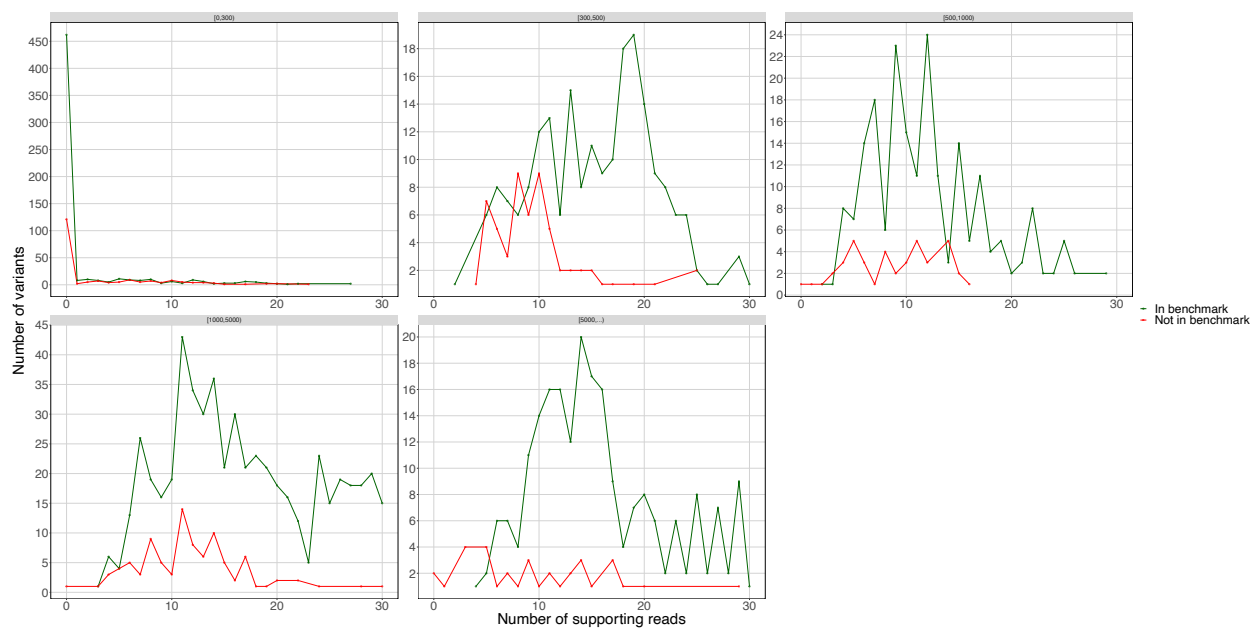
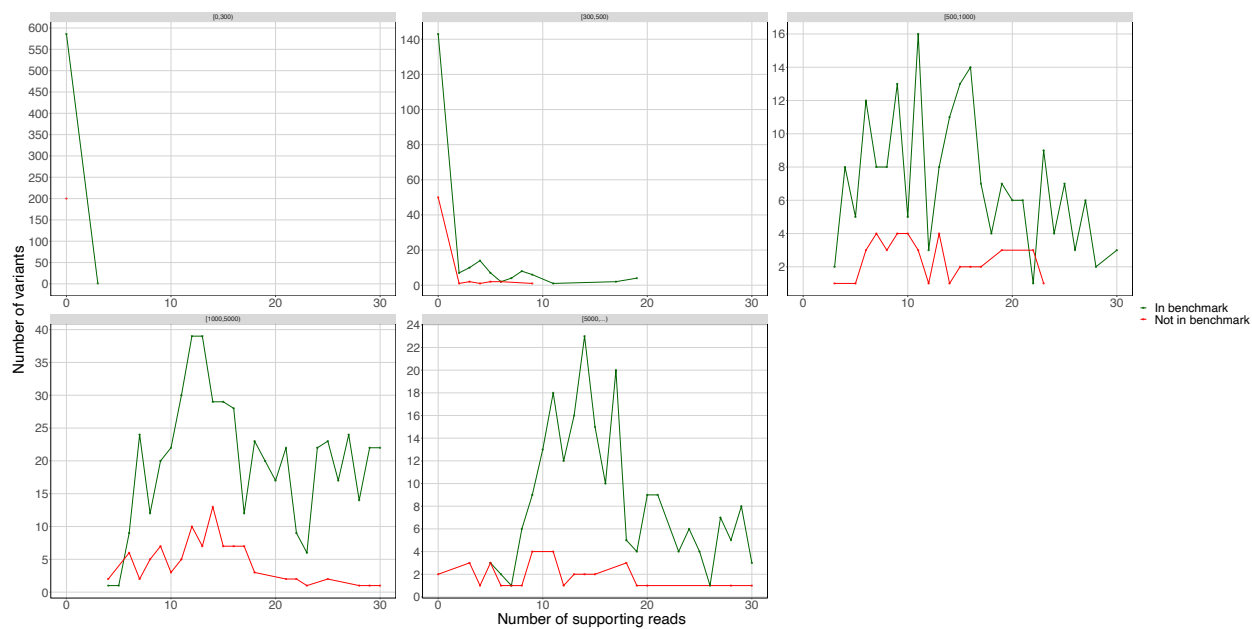
Supplementary Figure 14: Odds ratios for the number of *de novo* variants in a given category in individuals with ASD in SSC versus their unaffected siblings.



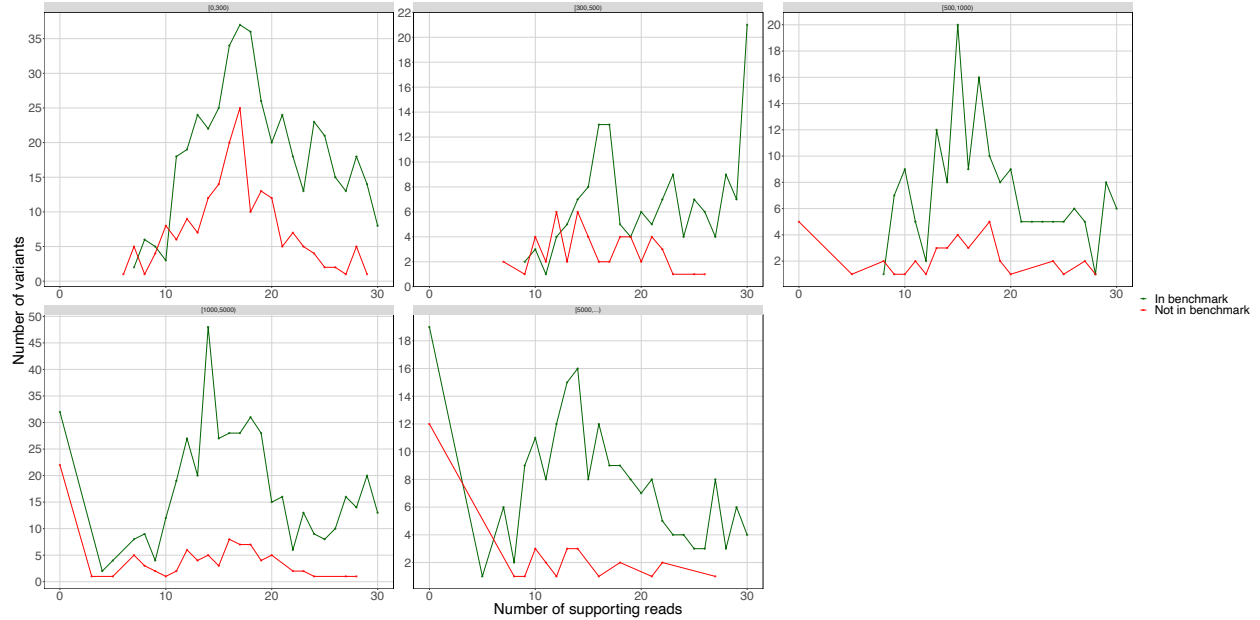
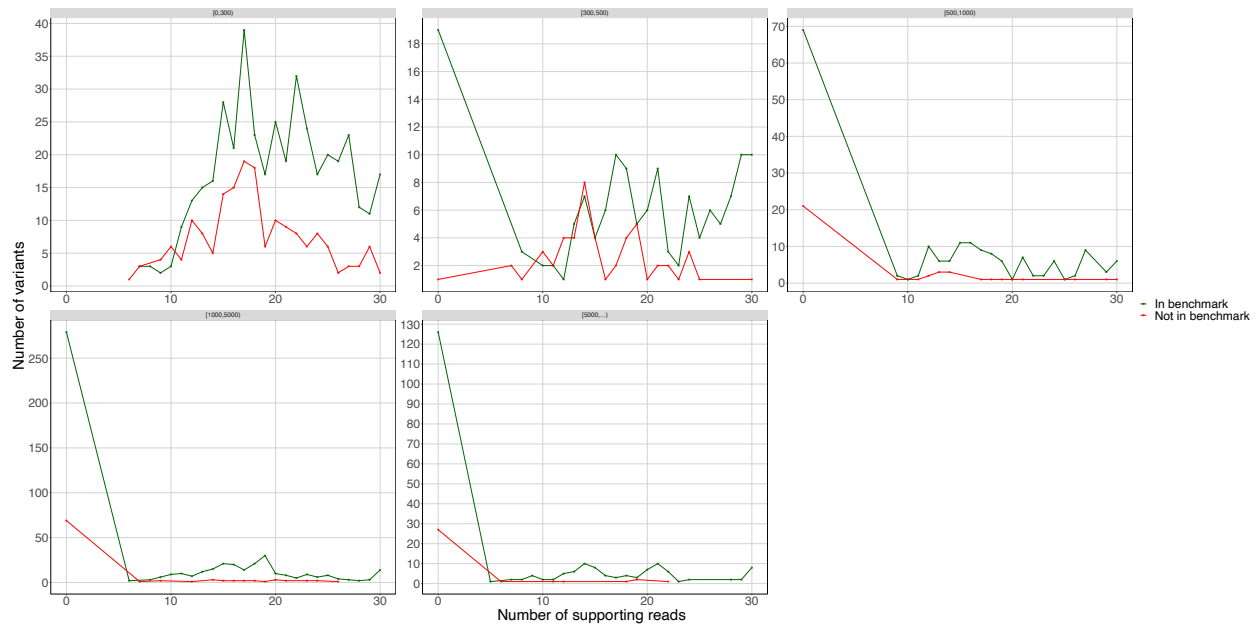
Supplementary Figure 15: Additional polygenic risk score (PRS) analysis. (A) PRS distribution when samples are ranked from highest to lowest score. (B) Sex-stratified polygenic transmission disequilibrium tests. (C) Comparison between individuals with ASD from multiplex families in MSSNG and individuals with ASD in SSC (which are all simplex families). (D) Comparison between unaffected fathers and unaffected mothers, stratified by dataset (MSSNG or SSC). (E) Comparison between unaffected parents and 1000 Genomes Project population controls.



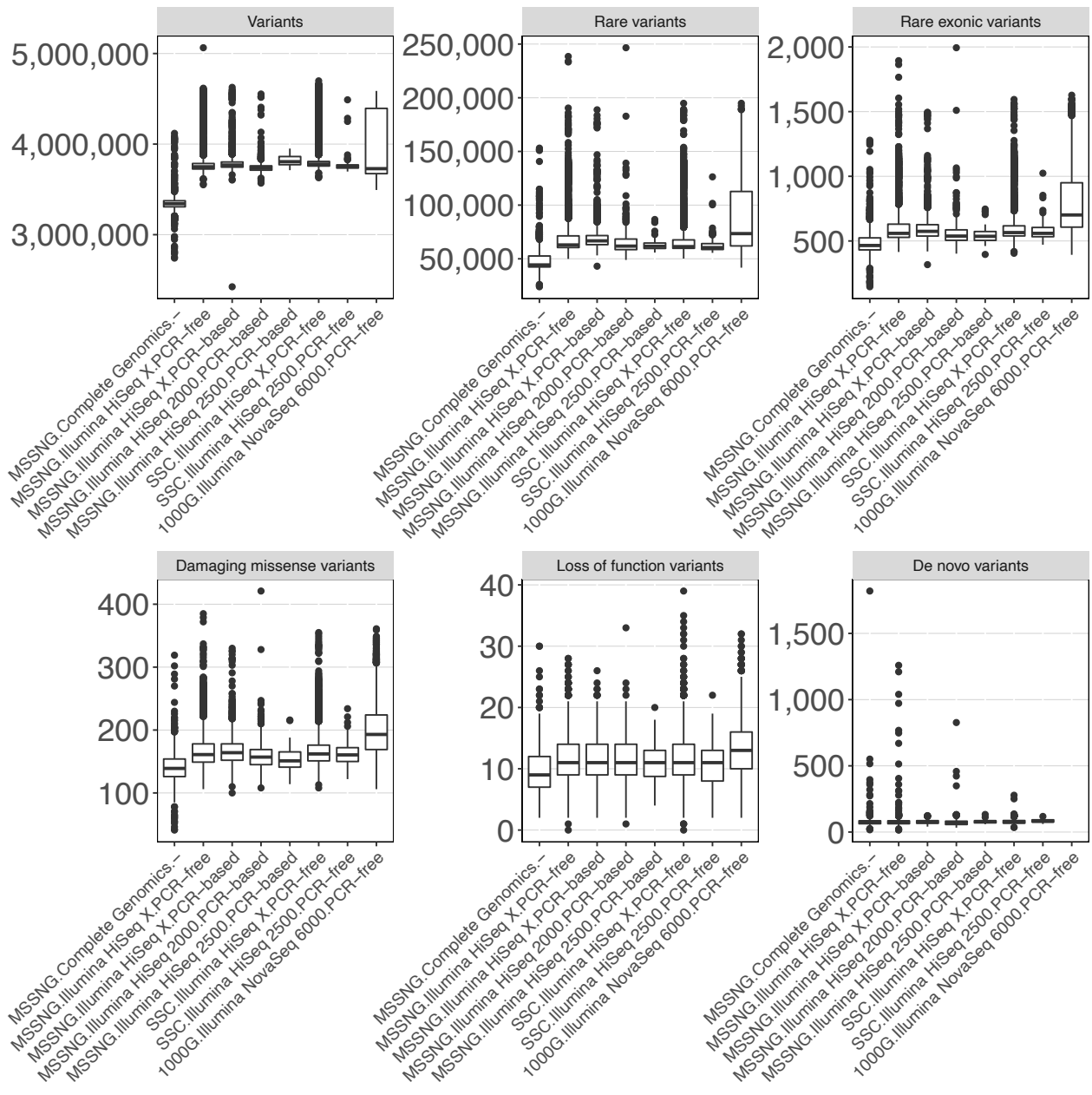
Supplementary Figure 16: Relationship between polygenic risk score (PRS) and composite phenotype measures. (A) Adaptive behavior standard score (n=2,787); (B) full-scale IQ (n=1,279); (C) global ability composite estimate (n=1,782); (D) socialization standard score (n=2,765).

A**B**

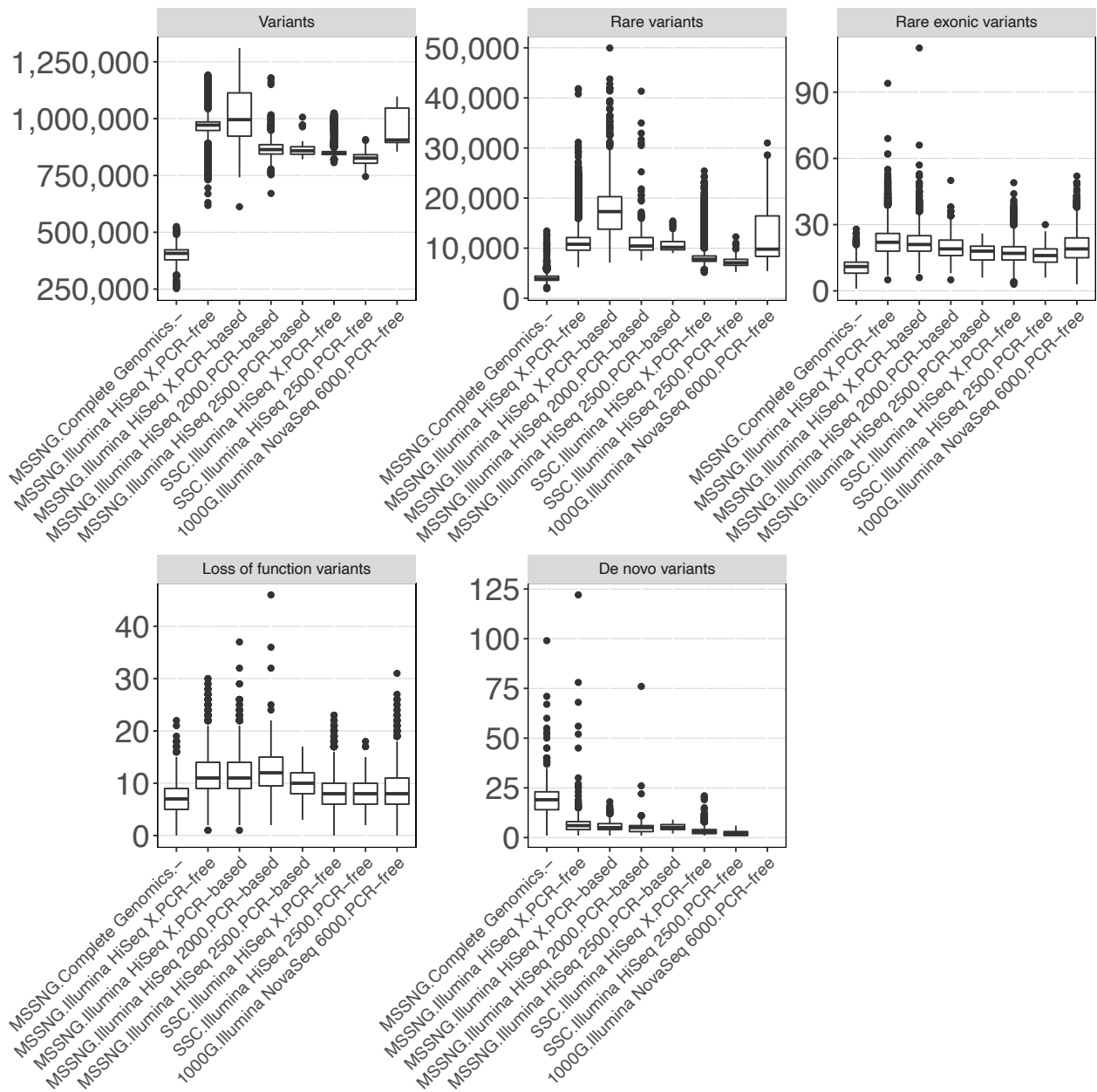
Supplementary Figure 17: Distributions of the number of deletions detected in the HuRef genome by (A) Manta or (B) DELLY that were or were not confirmed by the HuRef structural variant benchmark as a function of the number of supporting anomalously mapped paired-end reads. Similar trends were observed for the other two genomes/benchmarks tested (NA12878 and HG002) and for other variant types (duplications, insertions, and inversions). The facets indicate different deletion size bins; for example, the first facet represents deletions between 0 and 300 bp in size.

A**B**

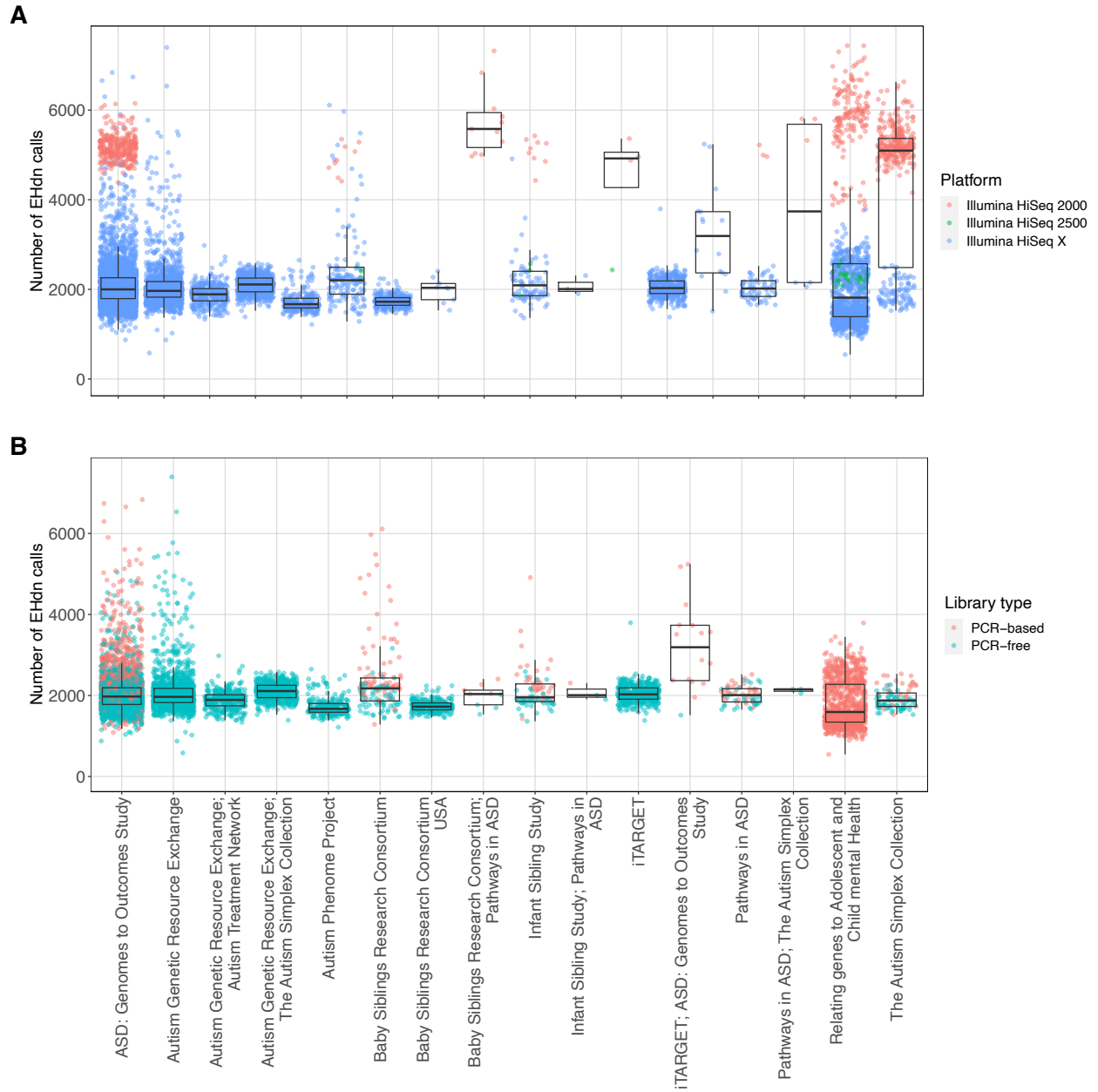
Supplementary Figure 18: Distributions of the number of deletions detected in the HuRef genome by (A) Manta or (B) DELLY that were or were not confirmed by the HuRef structural variant benchmark as a function of the number of supporting split reads. Similar trends were observed for the other two genomes/benchmarks tested (NA12878 and HG002) and for other variant types (duplications, insertions, and inversions). The facets indicate different deletion size bins; for example, the first facet represents deletions between 0 and 300 bp in size.



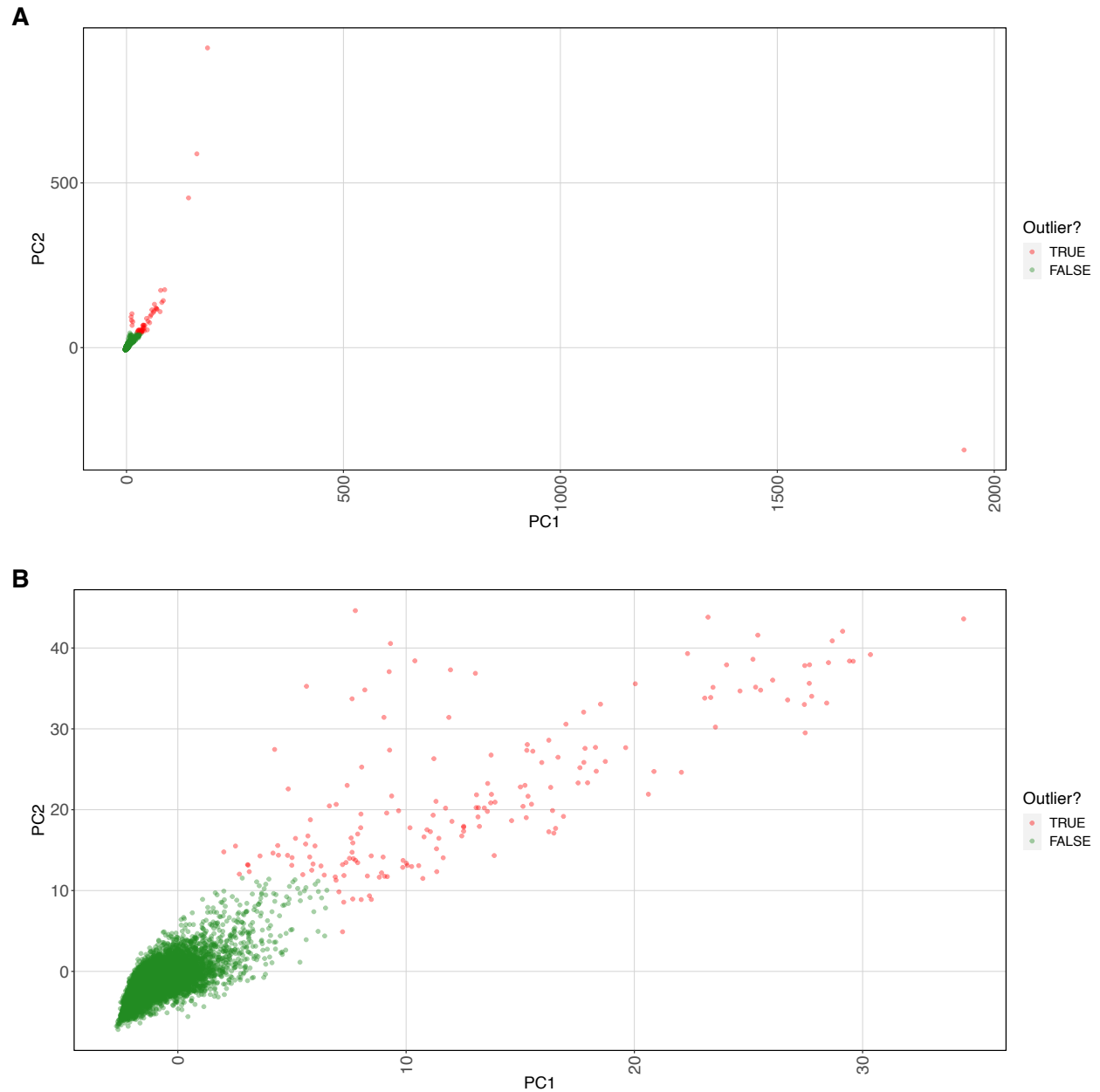
Supplementary Figure 19: Distributions of different categories of single nucleotide variants (SNVs), stratified by sequencing platform and DNA library preparation method.



Supplementary Figure 20: Distributions of different categories of indels, stratified by sequencing platform and DNA library preparation method.



Supplementary Figure 21: Number of tandem repeats detected by ExpansionHunter Denovo in each MSSNG sample, stratified by cohort and colored by either (A) sequencing platform or (B) DNA library preparation method. In part (B), only samples sequences on the Illumina HiSeq X platform are included.



Supplementary Figure 22: Principal Component Analysis of tandem repeats detected by ExpansionHunter Denovo in MSSNG samples sequenced on the Illumina HiSeq 2500 or HiSeq X platforms. Outliers are indicated after (A) one round of outlier detection, and (B) two rounds of outlier detection (i.e., after removing the outliers in part A and recomputing the mean and standard deviation of each principal component).