science.sciencemag.org/content/367/6477/569/suppl/DC1

Supplementary Materials for

Genetics of schizophrenia in the South African Xhosa

S. Gulsuner, D. J. Stein, E. S. Susser, G. Sibeko, A. Pretorius, T. Walsh, L. Majara, M. M. Mndini, S. G. Mqulwana, O. A. Ntola, S. Casadei, L. L. Ngqengelele, T. Korchina, C. van der Merwe, M. Malan, K. M. Fader, M. Feng, E. Willoughby, D. Muzny, A. Baldinger, H. F. Andrews, R. C. Gur, R. A. Gibbs, Z. Zingela, M. Nagdee, R. S. Ramesar, M.-C. King*, J. M. McClellan

*Corresponding author. Email: mcking@u.washington.edu

Published 31 January 2020, *Science* **367**, 569 (2020)
DOI: 10.1126/science.aay8833

This PDF file includes:

Genetics of Schizophrenia in the South African Xhosa: Supplementary materials

Genetics of Schizophrenia in the South African Xhosa

Supplementary Materials and Methods

Recruitment and sample collection

Subjects were individuals with schizophrenia and controls, self-identifying as Xhosa, and matched for sex, age, education level and region of residence (Table S1). Cases were recruited from provincial psychiatric hospitals and clinics in the Eastern Cape Province and Western Cape Province of South Africa, and included both in-patients and out-patients (Table S2). Participants met DSM-IV diagnostic criteria for schizophrenia or schizoaffective disorder (26) over at least a two-year period. Controls were recruited from university-affiliated general medical hospitals and community health centers that draw from catchment areas similar to the psychiatric hospitals.

A Community Advisory Board was created to protect the rights of research subjects, to aid in the development of consent processes and recruitment strategies, and to promote the respect and dignity of research participants and their communities (27). Cases and controls were assessed using the Structured Diagnostic Interview for DSM-IV Axis I Disorders (SCID-I) (28). Medical records information was reviewed and incorporated into a consensus diagnosis. All recruitment and clinical interview materials were administered in isiXhosa, following the World Health Organization (WHO) translation guidelines (29). In addition to the psychiatric assessment, participants provided blood samples for DNA analysis and HIV screening.

The study inclusion and exclusion criteria were:

Cases: patients were potentially eligible for the study based on the following: Xhosa speaking; clinical diagnosis of schizophrenia or schizoaffective disorder; duration of illness at least two

years; age 21-54 years; lived in the appropriate metro region; capable of informed consent when approached in pre-discharge open wards.

Controls: participants were recruited from community health clinics, and excluded if they reported an injury requiring hospitalization, a history of brain injury or seizures, a history of self-injury, or presented with injuries clearly related to alcohol or drug use and previous history of psychosis. Controls were generally recruited from clinics that provided basic medical care, including treatments for common conditions such as hypertension and diabetes. Given the nature of the clinics, we cannot rule out that some controls had subtle cognitive or neurodevelopmental conditions.

The sample was predominately male (88%), which was unexpected since gender ratios reported in epidemiological studies of schizophrenia range from approximately equal to a slight male predominance (30). The sampling strategy may have led to a higher proportion of males, in that recruitment sites included psychiatric hospitals and institutions. Males with schizophrenia generally have greater impairment, and are more likely to be chronically institutionalized than are females (31).

Exome sequencing and genotyping

Whole exome sequencing was carried out on genomic DNA of 1024 cases and 1004 controls at the Human Genome Sequencing Center (HGSC) at Baylor College of Medicine in 4 waves. Illumina paired-end pre-capture libraries were constructed using 0.5ug of DNA according to the manufacturer's protocol (*Illumina Multiplexing_SamplePrep_Guide_1005361_D*) with modifications as described in the HGSC protocol (https://www.hgsc.bcm.edu/content/protocols-sequencing-library-construction). For each capture hybridization, six pre-capture libraries were

pooled and then hybridized in solution to the HGSC VCRome 2.1 design (32) (42Mb, NimbleGen) essentially according to the manufacturer's protocol *NimbleGen SeqCap EZ Exome Library SR User's Guide (Version 2.2)*. Sequencing was performed using the Illumina HiSeq 2000 platform in paired-end mode, with sequencing-by-synthesis reactions extended for 101 cycles and an additional 7 cycles for the index read. With an average sequencing yield of 7.6 Gb, the samples achieved 95% of targeted exome bases covered to a depth of 20X or greater.

Primary sequence data analysis was performed using the HGSC Mercury analysis pipeline (33) (https://www.hgsc.bcm.edu/software/mercury) which moves data through various analysis tools from the initial sequence generation on the instrument to annotated variant calls (SNPs and intra-read in/dels). The whole exome sequence data from 1024 cases and 1004 controls was mapped to GRCh37/hg19 human reference genome with BWA-aln v.0.6.235. Aligned reads were further processed by PICARD v1.93 (broadinstitute.github.io/picard) and GATK v2.5-2 pipeline (34). Variants were called using Atlas-SNP2 and Atlas-Indel2 v1.4.337. Analyses included all sequence targeted by VCRome capture library probes +20bp.

Raw and processed data were transferred to the University of Washington by secure FTP, for further quality assessment and interpretation of variants and statistical analysis.

Variant calling and quality control

Genotypes from exome data at sites present in Omni v2.5 were used to determine variant-level quality thresholds. Based on the comparison of genotypes at Omni sites to genotypes at other variant sites, quality thresholds were defined as >4 median variant reads, >0.30 median variant read proportion, >40 median variant quality and Atlas filters PASS in >70% of all variant carriers (Figure S1). All variants not meeting these thresholds were omitted. We also removed variant sites that are flagged as problematic by ExAC and chromosome X variants located in pseudo-autosomal regions or with heterozygous genotype calls in males.

Samples were removed if their sequences were >3 standard deviations below median average coverage, percent coverage at 20X, number of variant calls, and number of high quality variant calls; 54 samples were removed by these criteria. Using PLINK v1.07 software (35) to identify sample relatedness, we identified and removed any first-degree relatives and included only one individual from each family.

In parallel to the whole exome sequencing workflow, all samples were independently genotyped with Human Exome-12v1.2, Human Core Exome-24 v1.0 or Infinium Core Exome-24v1.1 exome arrays to detect possible contamination or sample mislabeling. SNP array and exome data were analyzed through a locally developed software pipeline to calculate self-concordance ratios and contamination calculations (www.hgsc.bcm.edu/software/eris). Genotypes from SNP arrays were compared to the corresponding sites from whole exome data and a concordance score calculated from each site. A concordance score <0.90 indicated a possible sample swap and a contamination score >5% a potential sample contamination. No samples passing previous quality filters had suspicious contamination or concordance values (Figure S2).

In order to avoid biased results due to outliers, we removed samples that harbored a disproportionately high number of rare variants. For this step, rare variants were defined as a variant found in only 1 Xhosa participant (case or control), without population filtering. Outliers were defined as samples carrying >3 standard deviations above the median number of rare variants for the entire cohort, which corresponded to >135 rare variants per person. The number of rare variants harbored by samples included in the analyses ranged from 10 to 134, with median 35.5. Final analyses were then further restricted to Xhosa-specific private variants (Figure S3). There were no significant differences in the number of private variants within sequencing waves (Figure S4).

Finally, individuals were removed from the analyses based on diagnostic misclassification discovered during the data reconciliation process. Overall, 909 cases and 917 controls passed all quality filters (Figure S2). Distributions of different classifications of variants were similar across cases and controls (Figure S5).

Population structure and variation

Average heterozygosity for all exome sequence (31.7MB) was calculated for the Xhosa, and for all other populations in phase 3 of the 1000 Genomes project (36). Heterozygosity was defined as number of heterozygous sites per 10kb coding basepairs (Figure 1B, main text).

Principal component analysis was carried out to evaluate ancestral relationships between the Xhosa and other populations. Variant sites in the analysis were those found on the Omni V2.5 SNP platform and located in whole-exome-targeted regions. Genotypes were extracted from whole exome sequences of the Xhosa cohort, and from Omni V2.5 SNP array data of 1000 Genomes Project (36) and the EGAS00001000959 dataset of the African Genome Variation Project (AGVP) (37). Only autosomal SNPs were included because the AGVP dataset does not include genotypes from X or Y chromosomes. Populations from the 1000 Genomes Project were Afro-Caribbeans from Barbados (ACB), African-Americans from the U.S. southwest (ASW), Luhya from Kenya (LWK), Maasai from Kenya (MKK), Yoruba from Nigeria (YRI), Colombians (CLM), Mexican-Americans from Los Angeles (MXL), Peruvians (PEL), Puerto Ricans (PUR), Chinese Dai (CDX), Han Chinese (CHB) from Beijing, Southern Han Chinese (CHS), Japanese (JPT), Kinh from Vietnam (KHV), Utah residents of northern and western European ancestry (CEU), Finns (FIN), British (GBR), Spanish (IBS), Tuscans from Italy (TSI), Gujarati Indians living in Houston (GIH). Populations included from the AGVP were Baganda,

Banyarwanda, Barundi, Ethiopia, Kalenjin, and Kikuyu from East Africa; Fula, Ga-Adangbe, Igbo, Jola, and Wolof from West Africa; and Sotho and Zulu from South Africa.

Analyses included all autosomal SNPs with MAF ≥ 0.01 after removing samples with ≥5% missing genotypes and SNPs with ≥5% genotyping failures. SNPs were then pruned for linkage disequilibrium using PLINK (35) command 'indep 50 5 2'. A total of 26,146 common, exonic SNPs passed these filters and were included in population structure analyses. Principal components were estimated by multidimensional scaling with Euclidean distance using KING 2.1.2 software (38) (Figures S6, S7).

Ancestral population structures were evaluated using ADMIXTURE v1.3.0 (39) for the Xhosa, other African populations, and populations from other continents. For African populations, we constructed an African dataset including 19 African subpopulations from 1000 Genomes, the AGVP populations, and our data for the Xhosa. Global analyses included all 1000 Genome populations in addition to the African dataset. The 26,146 common exonic SNPs identified as above were included, with K values between 2 to 20. The ADMIXTURE software uses an n-fold cross-validation (CV) procedure to find the best K value that fits the data with minimal CV error. For our population data, a 5-fold CV procedure generated the best estimate of K, yielding K=6 for African populations and K=11 for global analysis (Figures S8, S9, S10).

Variant annotation

Variants were annotated using Annovar 2018-04-16 (40), with RefSeq as the reference for transcript definition. Population allele frequencies were drawn from 1000 Genomes, ExAC 3.0, gnomAD 2.0, and AGVP whole genome sequence (wgs) data. AGVP wgs data consists of autosomal variants of 100 Baganda individuals from Uganda, 100 Zulu individuals from South Africa, and 120 Gumuz, Wolayta, Oromo, Somali and Amhara persons from Ethiopia.

The potential impact of missense variants were predicted using PolyPhen-2 v2.2.2 and SIFT prediction tools (41 ,42), multiple sequence alignment data of 45 vertebrates to human sequence (multiz46way) and InterPro protein domains. For splice site variants located ±2bp of exon-intron junctions, MaxEnt scores (43) were calculated for splice motifs for reference and variant sequences. MaxEnt calculates a large range of scores (0-15) for reference splice junctions. We used a conservative region specific normalization to generate a normalized score difference for splice variants. For each site we calculated MaxEnt scores for the reference motif and for all possible mutations within a window defined as splice motif; 3 bases in exon and 6 bases in intron for donor sites and 20 bases in intron and 3 bases in exon for acceptor sites. For each site we selected the lowest score as the worst possible outcome and used it in the normalization. The normalized MaxEnt splice score was ***1 -*** *(Variant score – lowest score) / (Reference score – lowest score).* This score reflects the strength of a particular mutation in comparison to reference sequence and other possible mutations at the particular splice junction. We selected 0.80 as the threshold to classify splice mutations as damaging. We used gene constraint scores generated from ExAC excluding psychiatric cohorts.

Criteria for variant classification were as follows:

Functional variants: All frame-shifts, nonsenses, splice variants predicted to alter splicing, and missense variants.

Damaging variants: Nonsense and frame-shift mutations in the first 80% of the protein; splices at ±2bp of junctions and with a normalized MaxEnt score ≥0.80; missenses predicted damaging by polyphen-2 (humdiv ≥0.9, humvar ≥0.8) and SIFT (≤0.05) or by polyphen-2 alone (humdiv ≥0.9) if the variant is located in an InterPro protein domain; and for which the variant amino acid did not appear as the reference residue in any vertebrate.

Damaging mutations in mutation-intolerant genes: Predicted damaging splice, nonsense and

frame-shift variants in genes intolerant to loss of function mutations (pLI ≥0.5); predicted damaging missense variants in gene with missense Z score ≥2.5 (44). For genes with missense Z scores ≥4.0, missense variants predicted damaging with any of the algorithms (PPH2_humdiv ≥0.9 or PPH2_humvar ≥0.8 or SIFT ≤0.05).

Xhosa-specific private mutations

A "private variant" was defined as a variant appearing in only one case or only one control among our participants and not in other population databases (with the exception of minor allele frequency <0.00005 on gnomAD (45) to allow for re-occurrence of the same mutation and for database errors). Associations between Xhosa-specific private mutations and schizophrenia were tested using Firth's penalized-likelihood logistic regression model (46), corrected for the number of Xhosa-specific private variants, gender, sequencing wave, coverage and first ten principal components. Statistics were generated for each case-control test independently. Odds ratios (OR) and P values for each gene were obtained using the *logistf* function implemented in R.

The distribution of various classes of variants in cases and controls for all Xhosa-specific private mutations, for private damaging mutations in all genes and for private damaging mutations in case-only or control-only genes, is presented in Table S3. Rates of mutations in cases and in controls are compared by logistic regression with the covariates as defined above.

Numbers of cases and numbers of controls were compared for presence of private predicted-damaging mutations in genes intolerant to such mutations (47). From this set, a subset of genes was generated with private predicted-damaging mutations only in cases or only in controls. We identified 2370 private predicted-damaging variants in 1388 mutation-intolerant genes, of which 547 were "case only" genes and 493 were "control only" genes. Thresholds for "case only" and

"control only" genes were set at 1, 2, or 3 variants per gene to create text Figure 2. For this analysis, we tested the effect of having at least one qualifying variant to case-control status. We used the presence of a qualifying variant (yes or no) as the dependent variable. We repeated the analyses using a more conservative pLI threshold of 0.9. The results are virtually identical to those with pLI threshold of 0.5 (Table S4).

Using the same methods, we also compared numbers of cases versus controls harboring at least one private benign mutation in mutation-intolerant genes (text Figure 2). Benign variants were defined as private variants in genes with missense Z-score $\geq$2.5 that were not predicted to be damaging by the criteria above.

To test for an oligogenic model, we assessed whether harboring an increasing number of qualifying variants predicted case-control status; the dependent variable was the number of qualifying variants observed in each sample (Table S5).

Tissue expression profiles

Tissue expression profiles were examined for genes harboring private damaging mutations in cases and controls (Table S6), using a database of genes differentially expressed in 25 different GTEx tissues (14). We extracted tissue-specific enriched genes, based on a specificity index (pSI) threshold <0.01 and compared the number of tissue-specific enriched genes for each tissue that harbored private damaging mutations in cases and controls.

Synaptic genes

Tests of association between Xhosa-specific private mutations and schizophrenia were repeated for genes encoding synaptic proteins. A synaptic gene was defined using the SynGO database (36). We identified 425 private predicted-damaging variants in 230 mutation-intolerant

synaptic genes, of which 88 were “case only” and 72 were “control only”. The number of cases versus controls harboring benign private mutations in synaptic genes was also assessed (text Figure 2) to serve as a comparison for the analysis.

HIV status

Private damaging mutations, and private damaging mutations in synaptic genes, remained significantly enriched in cases versus controls when including HIV status as a covariate (Table S7). We also carried out these analyses excluding subjects who tested positive for HIV, which required redefining case- and control-specific genes. In all analyses, Xhosa persons with schizophrenia continued to demonstrate a significant enrichment of private damaging mutations, including for variants in synaptic genes (Table S7).

Integrating genetic and post-mortem gene expression data

Data from the psychENCODE databases (resource.psychencode.org/) was extracted for the 4096 protein coding genes with significantly altered expression levels in post-mortem brain tissue of in affected patients with schizophrenia, autism, or bipolar disorder versus controls (16). Of these 4096 genes, 547 genes harbored private damaging mutations in cases and the 493 genes with private damaging mutations in controls.

Gene level association analysis

Associations between schizophrenia and multiple rare damaging mutations in individual genes were evaluated using the Sequence Kernel Association Test (SKAT) (17). We tested all functional variants and all predicted damaging variants with three different minor allele frequency thresholds: <0.01, <0.005 and <0.001 (Table S8A); controlling for the total number of variants at each frequency threshold in each analysis, in addition to standard covariates

(gender, coverage, sequencing wave and principal components 1 to 10).

Functional variants were defined as all missense, truncating and splice mutations that are predicted to alter splicing. In-frame insertions (n=43) and deletions (n=271) were not included in the functional or damaging groups.

SKAT functions were used to create adjusted Q-Q plots (Figure S11) and to adjust for multiple tests using resampling (Table S8A). Adjustment for multiple comparisons was based on the family-wise error rate (FWER). The top 10 genes for each MAF threshold are listed in Table S8B.

Exome-wide association analyses

Logistic regression models were used to measure association between single SNPs and schizophrenia, corrected using PLINK for gender, sequencing wave, coverage, and first ten principal components. We excluded multi-allelic SNPs and SNPs failing Hardy-Weinberg equilibrium with $P \leq 1E\text{-}10$ in cases and $P \leq 1E\text{-}6$ in controls (48). Separate genome-wide association analyses were carried out for worldwide SNPs (MAF ≥ 0.01, N = 159,801)) and for African-specific SNPs (MAF 0.01-0.10, N = 73,374) (Figure S12). Only one SNP was significant after exome-wide multiple comparison adjustment: rs12600437 in the 3'UTR of zinc finger protein ZFP3. The four SNPs with MAF $\geq$ 0.01 with the lowest P-values are listed in Table S9. None of these SNPs lies within any of the 108 schizophrenia-associated regions reported in European populations (6).

The sample size of 909 cases and 917 controls is underpowered to detect modest effects. Given these sample sizes, estimated disease prevalence in the general population of 1%, and allele frequency >0.2, a relative risk of 1.9 or greater could be detected with power 0.82 at

significance threshold 5.0E-08 (49).

CNV analysis from exome sequence data

CoNIFER was used to normalize read depths and call CNVs in exome data (19). RPKM values were calculated and normalized to captured regions. CoNIFER was then used to calculate and transform Z-RPKM values, and to call CNVs. We focused on 15 CNV hotspots known to be associated with schizophrenia and other neuropsychiatric conditions (50). We evaluated each hotspot region for CNV calls by plotting the regions and manually evaluating the read depths. The CNV calls for each chromosomal hotspot are presented in Figure S13.

Analyses of a non-African Population

To compare findings in the Xhosa to those of other populations, we analyzed the data of Genovese et al (3; data at phs000473.v2) using the same metrics. The Genovese et al data are a case-control study of schizophrenia in the Swedish population. We applied our sample and variant level QC criteria to the Swedish dataset. Excluding individuals with high quality variant calls >3 standard deviation below median number of high quality variant calls of the cohort, first and second degree relatives and then, samples carrying >3 standard deviations above the median number of rare variants for the entire cohort, we obtained a dataset of 4436 cases and 5713 controls.

In order to evaluate a sample of Swedish cases and controls comparable to our Xhosa cohort, we selected subsets of 909 Swedish cases and 909 Swedish controls from the 4436 cases and 5713 controls that passed the filters above. Cases and controls in each Swedish subset were matched for batch and ancestry insofar as possible using information from phs000473.v2.

We randomly sampled the Swedish data 10,000 times to generate subsets of 909 cases and 909 controls. Cases and controls in each subset were matched for the number of private silent variants to correct batch and ancestry differences between samples. For each Swedish subset we counted the number of cases and the number of controls harboring qualifying private mutations exactly as described above for the Xhosa cases and controls. P-values for the Swedish data were calculated from the median numbers of qualifying cases and qualifying controls for each threshold using Fisher's exact test.

Results for Xhosa and Swedish populations are almost identical for the least constrained threshold for case-only and control-only genes and when all mutation-intolerant genes are included in analysis (Table S10). Results for the Xhosa are more influenced than results for the Swedes by higher thresholds on case-only and control-only genes and by constraining the analysis to include only synaptic genes. These differences suggest that the greater genetic variation of the Xhosa population yields more rigorous definitions of case-only and control-only genes at high thresholds, therefore fewer controls with qualifying mutations in such genes, and hence higher odds ratios and more significant P values.

In order to explore the effects of larger sample size on our analytic approach, we randomly selected 4000 Swedish cases and 4000 Swedish controls, each 10,000 times, from the samples passing our filters, then carried out the same analysis as above.

Higher odds ratios for the same hypotheses were found for the larger sample sizes (Table S10). While P-values are sensitive to larger sample sizes, this also suggests that greater genetic variation present in larger samples yields more rigorous definitions of case-only and control-only genes.

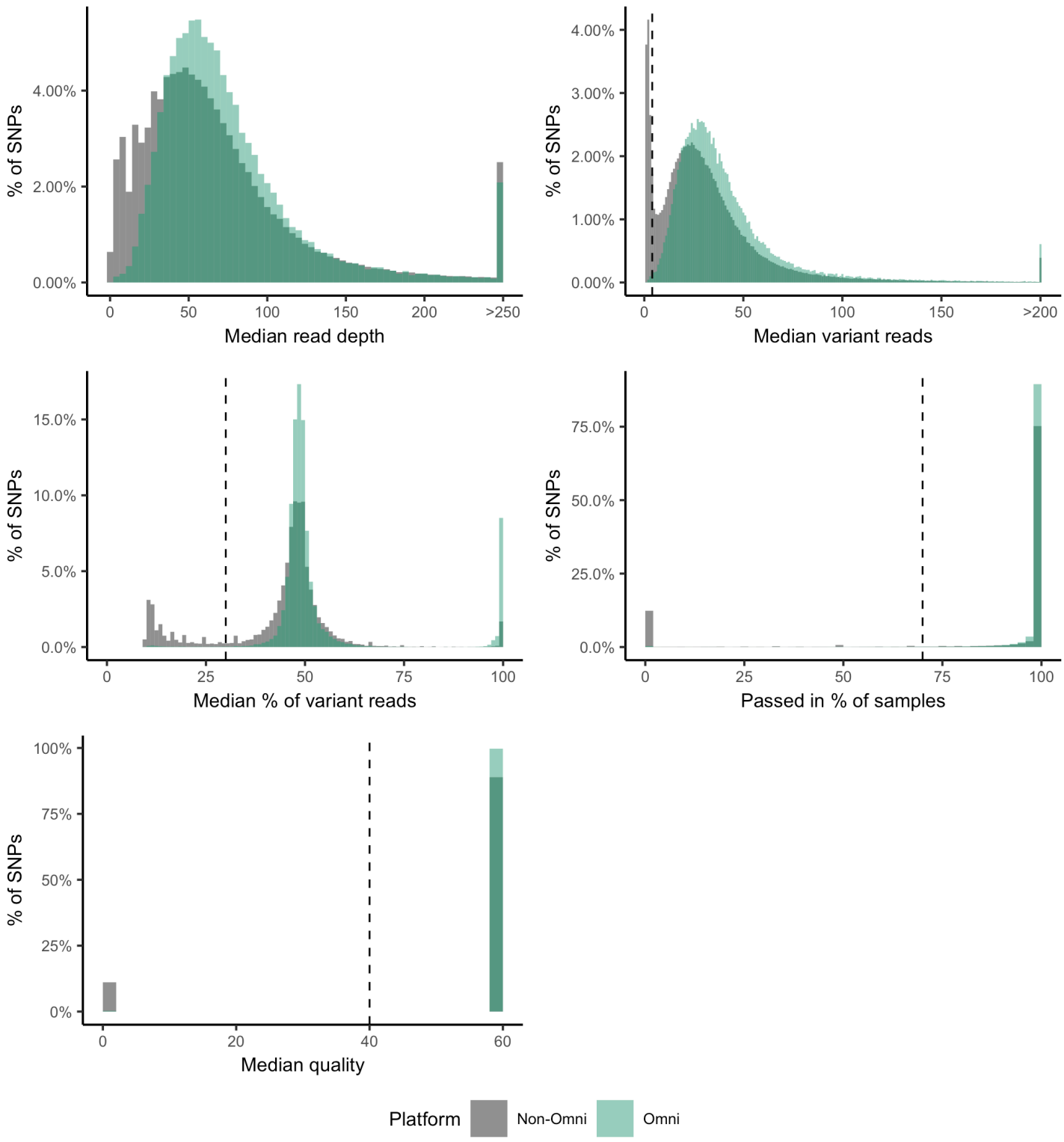


Figure S1: From sequenced exomes of all Xhosa participants, features of variant sites that also appear on Omni V2.5 (n=68,361) versus other SNP sites (n=498,356). “Median” indicates median number for each SNP site among all samples.

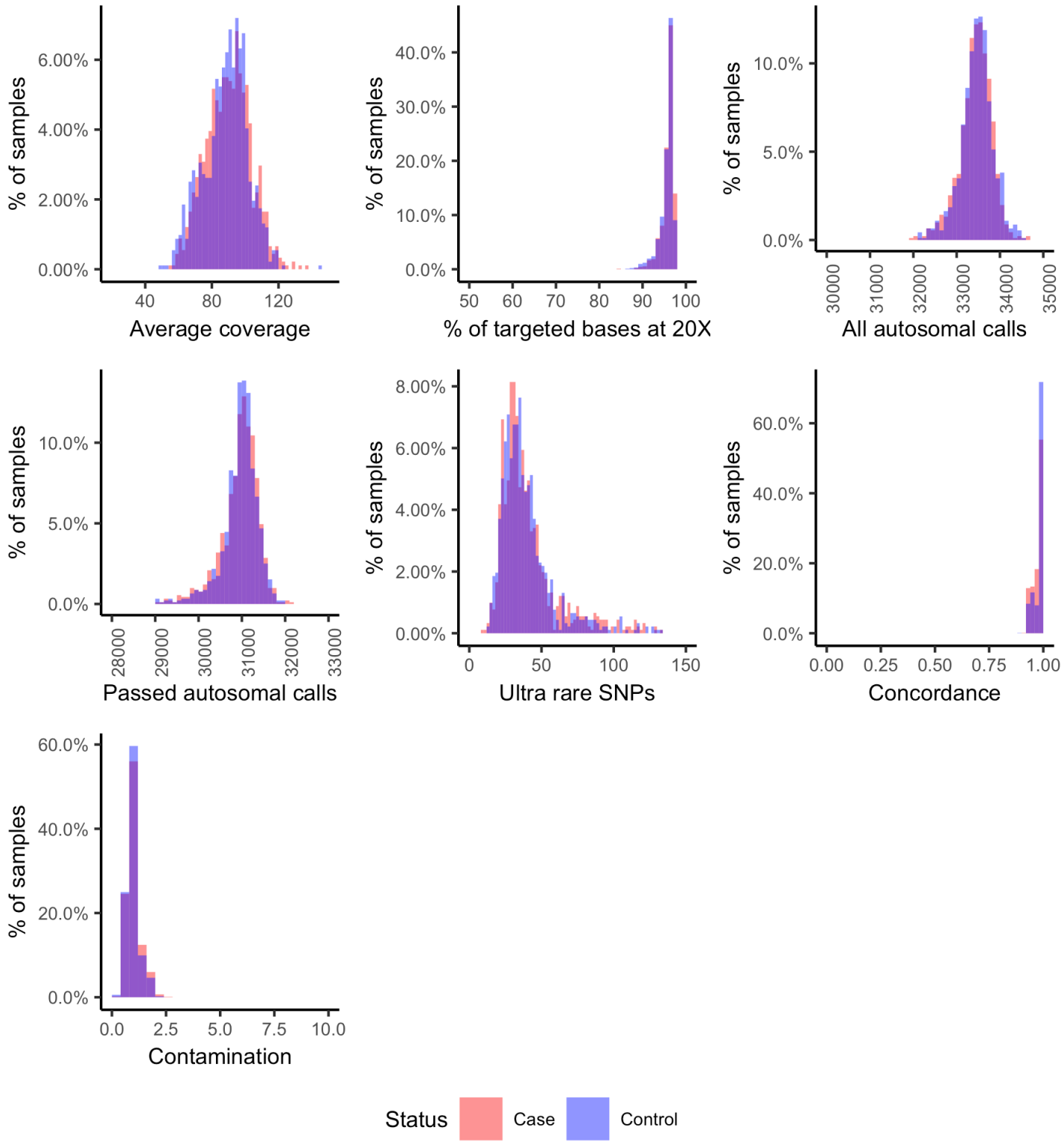


Figure S2. QC measurements in final sample set of 909 cases and 917 controls.

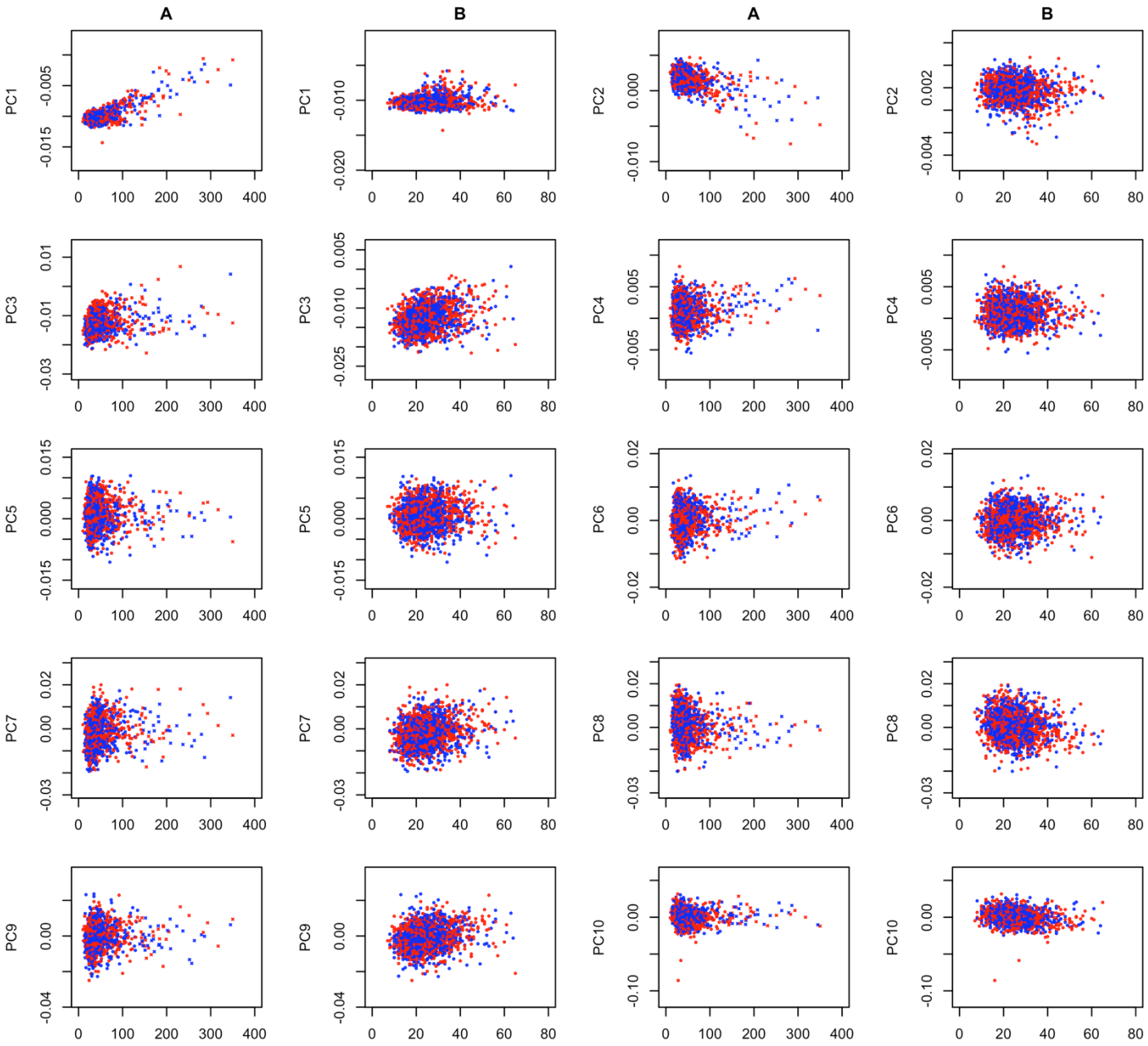


Figure S3. Relationship between first two principal components and number of rare variants per person. Panels labeled A illustrate all variants with only one Xhosa carrier (without population filtering), and reflect the effect of population outliers on the number of rare variants. Panels labeled B illustrate Xhosa-specific private variants used in the case-control analyses after removing private variant outliers (Red: cases; blue: controls; x: outliers).

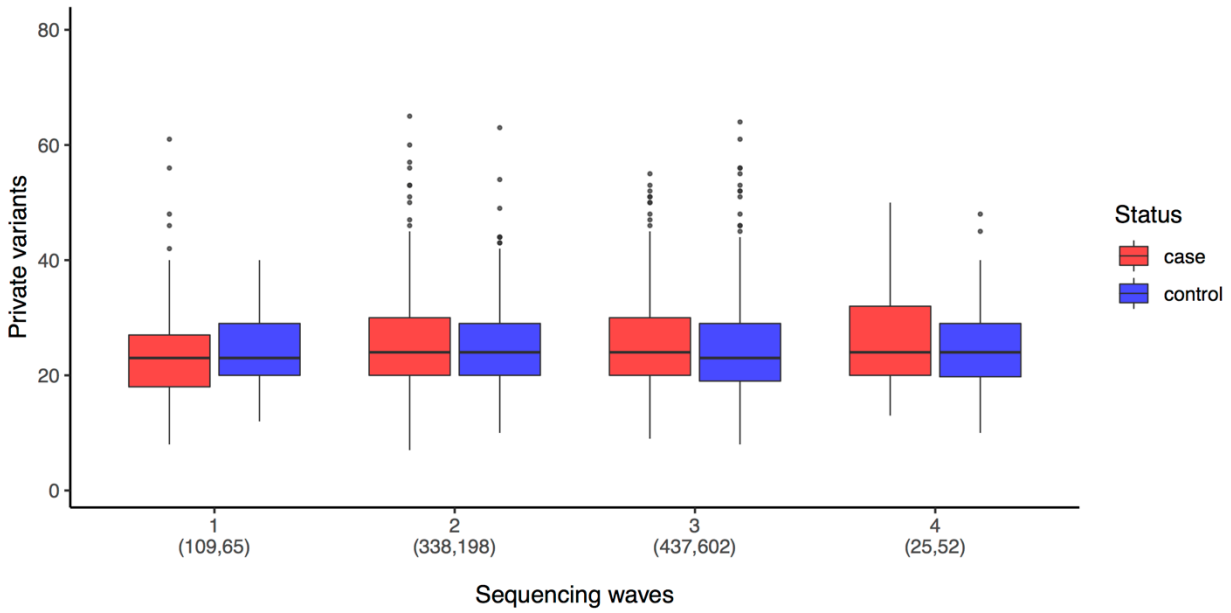


Figure S4. Distribution of private variants in cases and controls in each sequencing wave. Numbers of cases and controls in each wave are indicated in parenthesis. There were no significant differences between the number of private variants between the waves (ANOVA, F = 0.278, p = 0.60) or between private variants in cases and controls in each wave (Kolmogorov–Smirnov test).

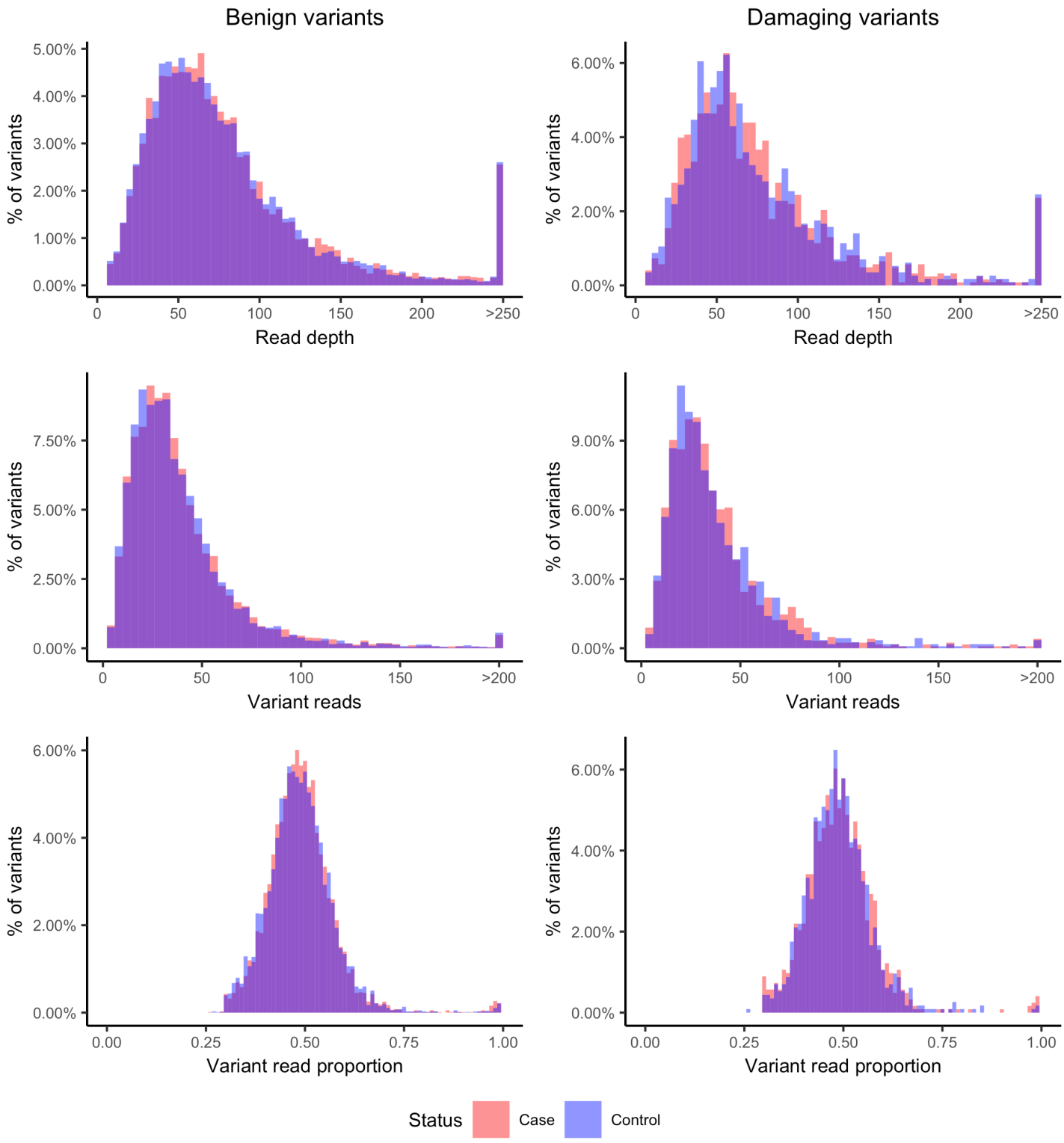


Figure S5. QC measurements of private benign variants and private damaging variants of intolerant genes in 909 cases and 917 controls.

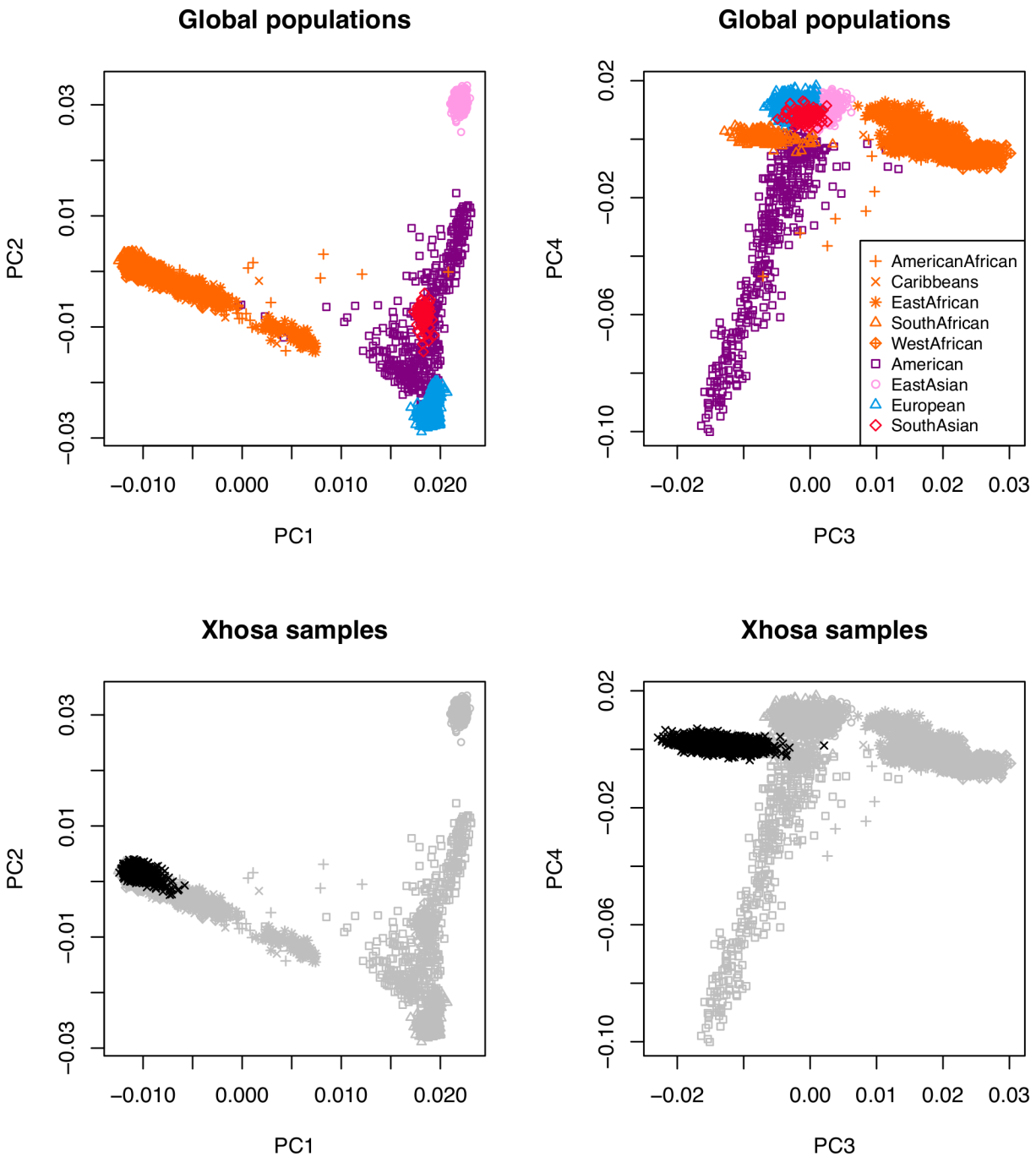


Figure S6. First four principal components from the 1000 Genomes and African Genome Variation Project (AGVP) (top) and from the Xhosa population (superimposed in black on bottom panels).

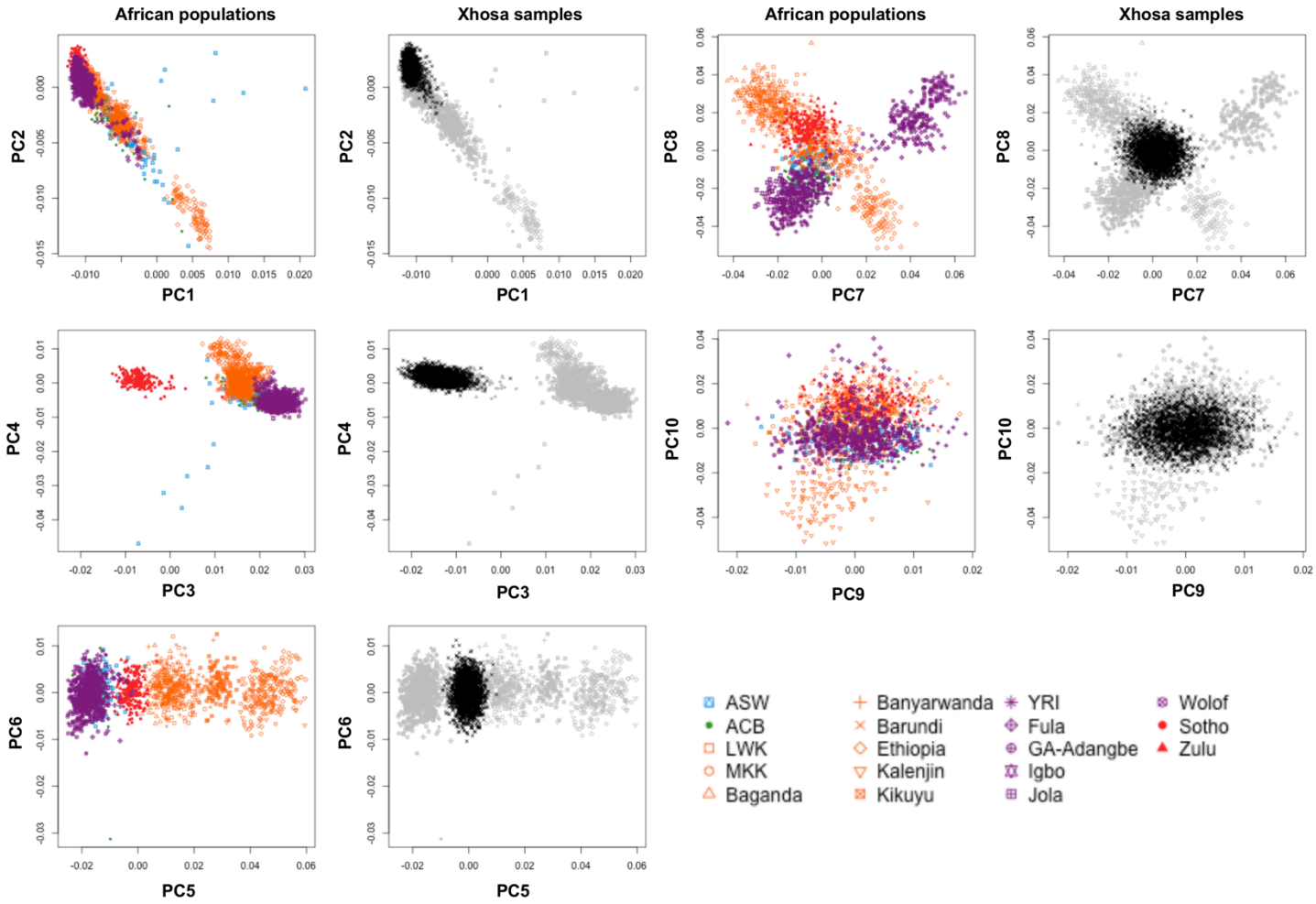
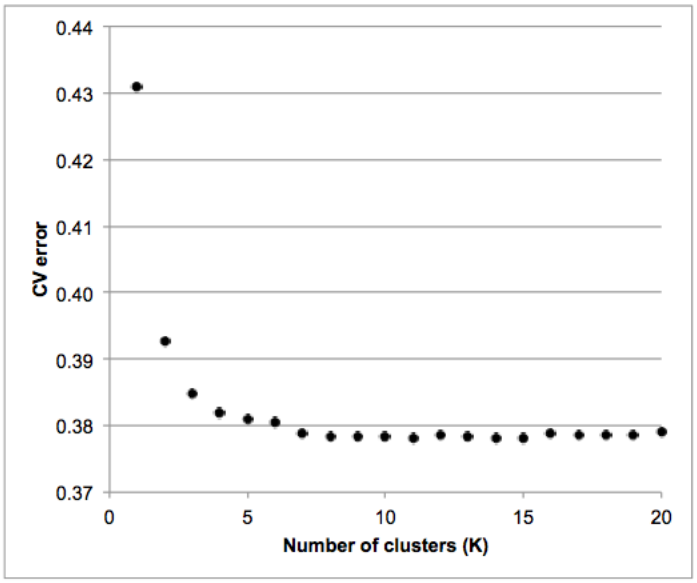


Figure S7. First 10 principal components of relationships between Xhosa and other African populations. (Abbreviations: ASW =American-African; ACB = Caribbean; LWK = Luhya in Webuye, Kenya; MKK = Kenya; YRI = Yoruba in Ibadan, Nigeria.)



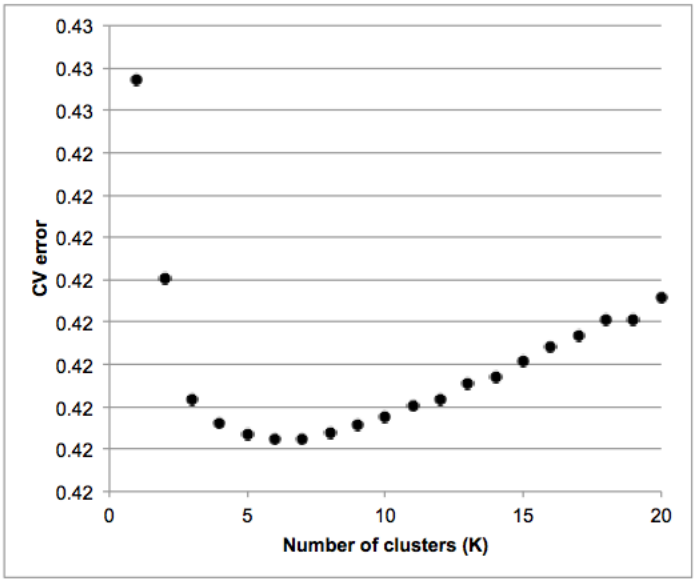


Figure S8. Cross-validation (CV) error of ADMIXTURE analysis of extended global (left) and African (right) data-sets for the different number of clusters (K). The CV errors are lowest for K=11 in global populations and K=6 in African populations, suggesting those numbers of clusters best fit the data.

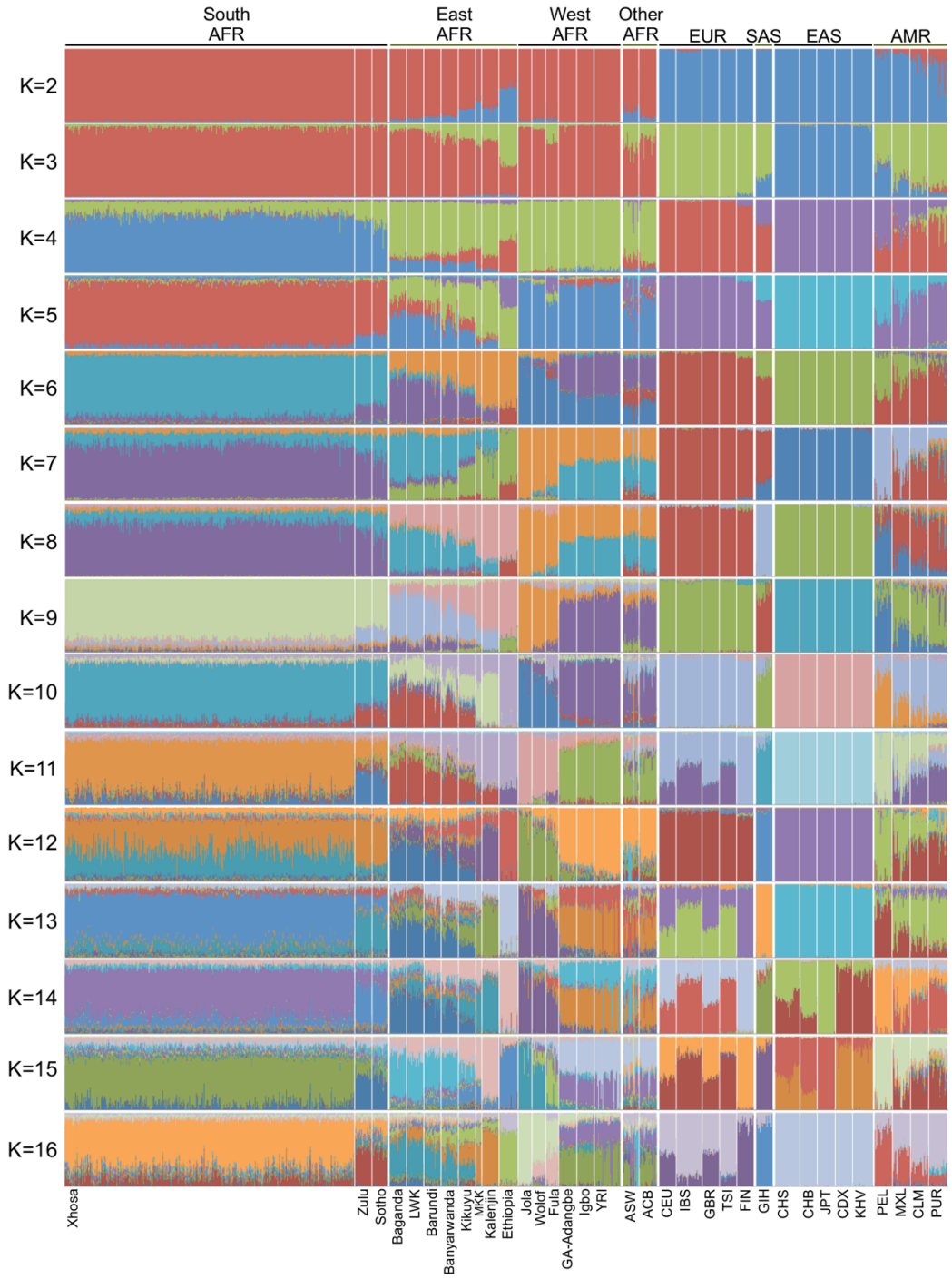


Figure S9. ADMIXTURE analysis of 34 populations for K=2 to K=16. The X-axis lists the populations (top) and sub-populations (bottom) that were included in the analysis. Each row is an independent ADMIXTURE run for the different number of clusters (K). The Y-axis represents the proportion of each color derived from all populations. K=4 and subsequent clusters demonstrate the separation of East, West and South African populations. At K=11 the Xhosa population is distinct from other South African populations (Zulu and Sotho) and characterized by at least 3 ancestral components with minimal components from East and West Africans.

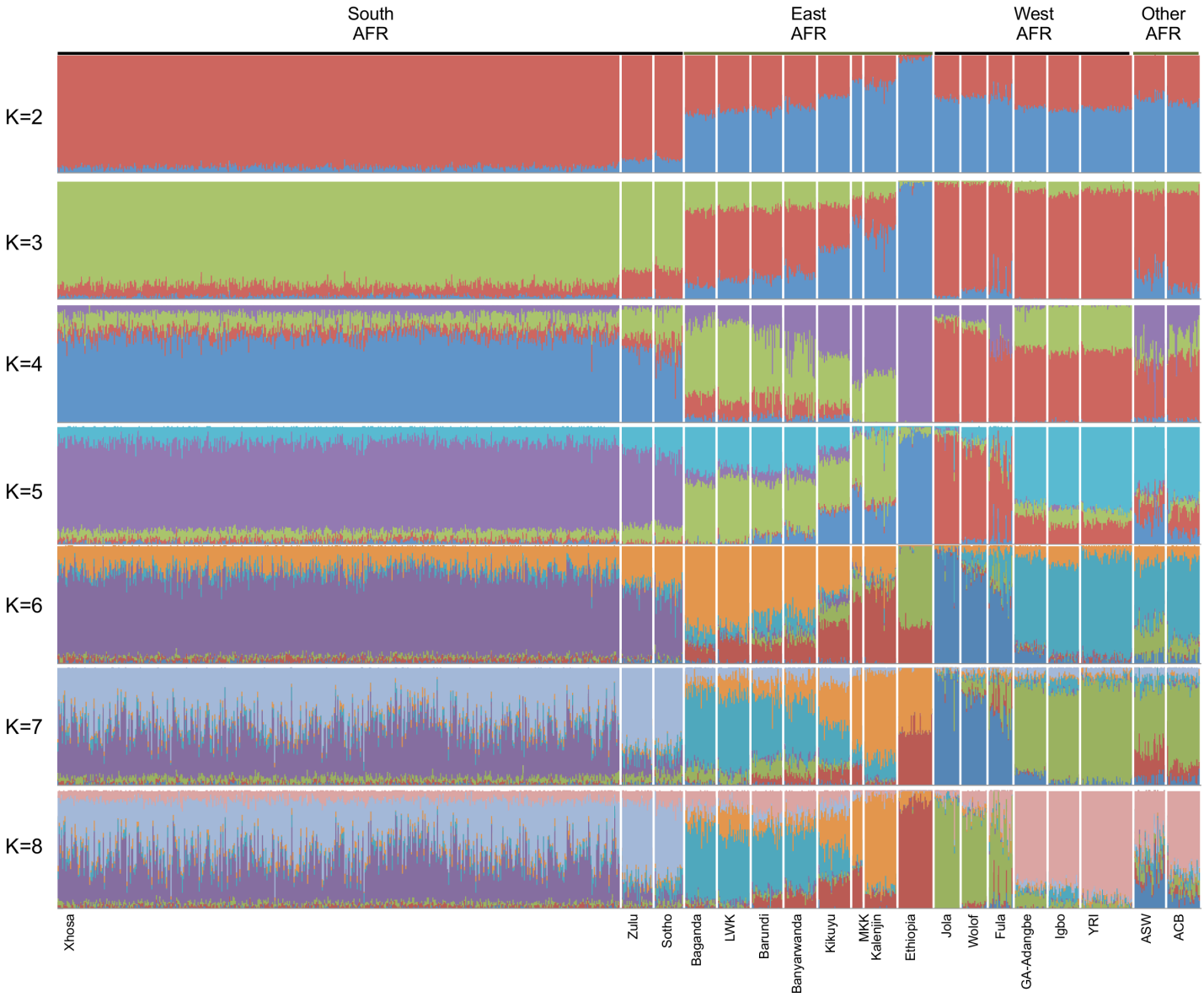


Figure S10. ADMIXTURE analysis of 19 African populations for clusters K=2 to K=8. At K=2 South African populations are distinct from East and West Africans. K=6 best distinguishes the Xhosa from each of the other African populations.

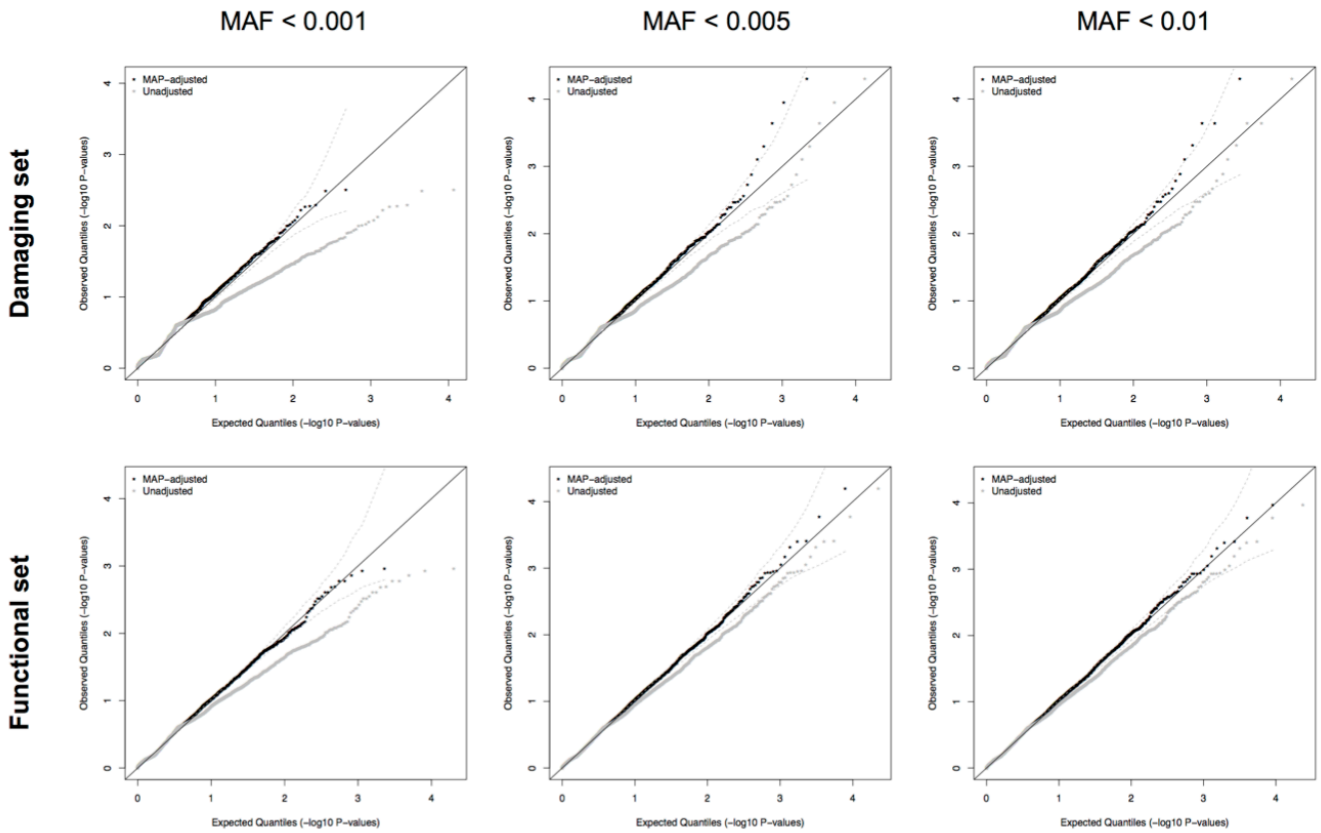


Figure S11. Q-Q plots for gene-by-gene analysis of rare alleles using SKAT.

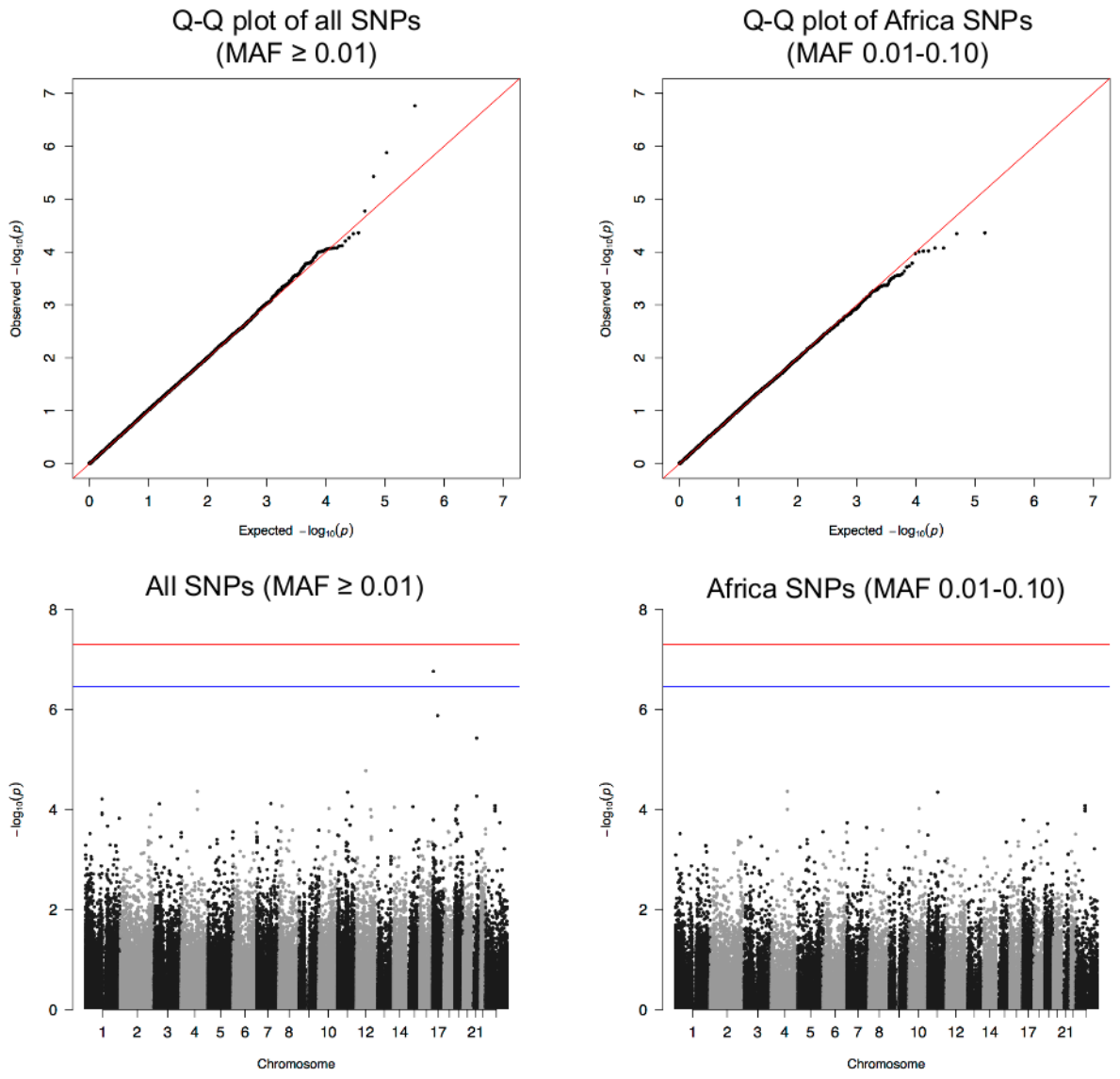


Figure S12. Q-Q plots (top) and Manhattan plots (bottom) for case-control analysis of all SNPs with minor allele frequency (MAF) ≥0.01 and for African-specific SNPs of 0.01 ≤ MAF ≤ 0.1. Genome-wide (5.00E-08) and exome-wide (3.50E-07) significance levels are shown as red and blue lines respectively.

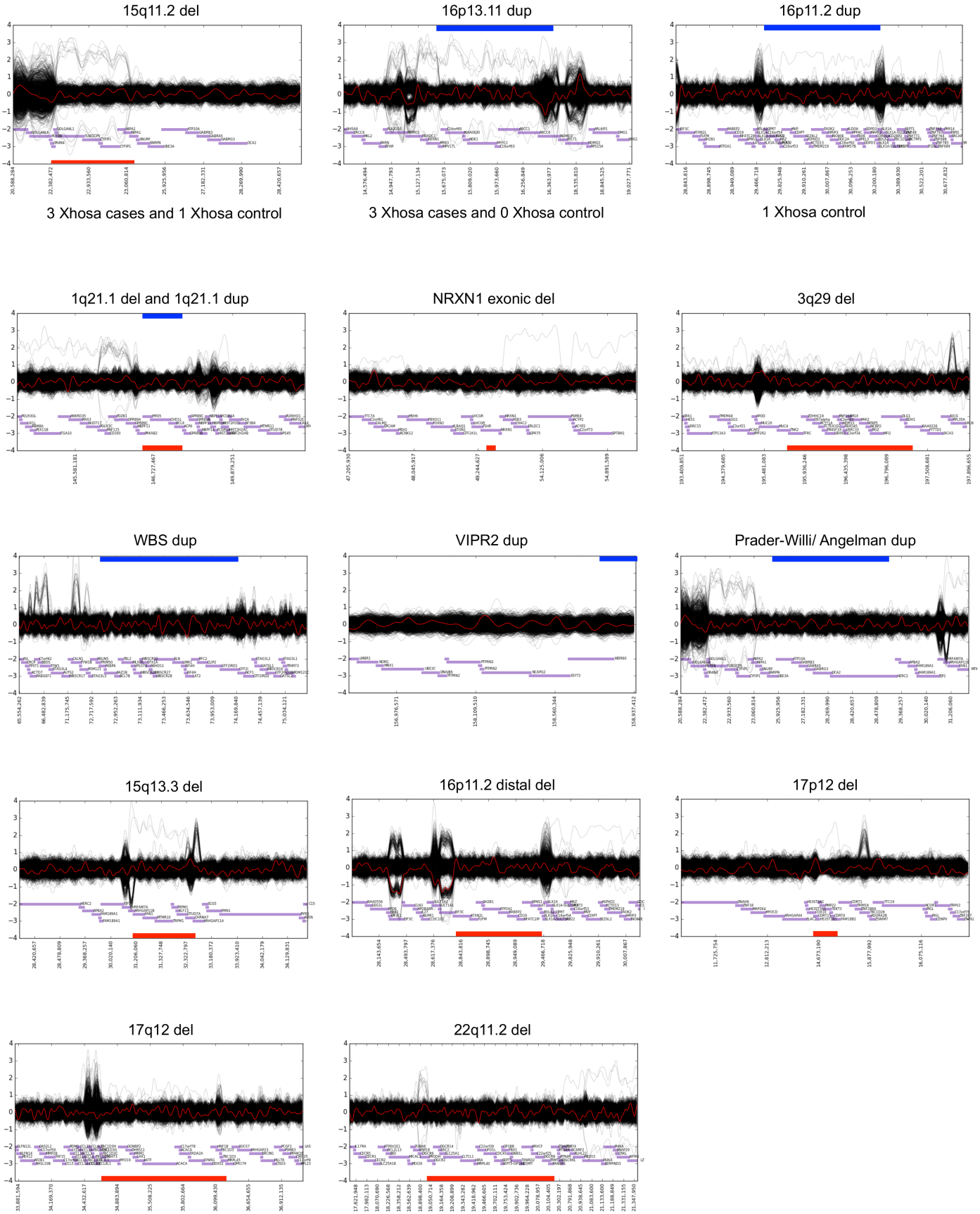


Figure S13. CNVs in cases and controls in genomic hotspots associated with severe mental illness.

Table S1. Demographic and Clinical Characteristics.

	Cases (n = 909)	**Controls (n = 917)**
Age (years)	36.0 ± 9.1	35.8 ± 9.0
Gender		
Males	799 (88 %)	797 (87 %)
Females	110 (12 %)	120 (13 %)
Site of Recruitment		
Western Cape Province	444 (49 %)	437 (48 %)
Eastern Cape Province	465 (51 %)	480 (52 %)
Treatment Setting		
Inpatient	518 (57 %)	12 (1 %)
Outpatient	388 (43%)	902 (99 %)
Unknown	3 (0.3 %)	3 (0.3 %)
HIV status		
Reactive	48 (5 %)	146 (16 %)
Non-reactive	795 (88 %)	727 (79 %)
Unknown/refused	66 (7 %)	44 (5 %)

Table S2. Recruitment Sites. H = Hospital, CHC = Community Health Centre.

Site	Type of care	Setting	City/Town
Western Cape			
Valkenberg H	In-patient	Urban	Cape Town area
Lentegeur H	In-patient	Urban	Cape Town area
Vangaurd CHC	Out-patient	Urban	Cape Town area
Khayelitsha site B CHC	Out-patient	Urban informal	Cape Town area
Mfuleni CHC	Out-patient	Urban informal	Cape Town area
Crossroads CHC	Out-patient	Urban informal	Cape Town area
Michael Mpongwana CHC	Out-patient	Urban informal	Cape Town area
Gugulethu CHC	Out-patient	Urban informal	Cape Town area
False Bay H OPD	Out-patient	Peri-urban	Simons Town, Cape Town
Eastern Cape			
Cecilia Makiwane H	Out-patient	Urban informal	East London
Nontyatyambo CHC	Out-patient	Urban informal	East London
Tower H	In-patient	Peri-urban	Fort Beaufort
Fort England H	In-patient	Peri-urban	Grahamstown (Makhanda)
Joza Clinic	Out-patient	Peri-urban	Grahamstown (Makhanda)
Settlers' Day H	Out-patient	Peri-urban	Grahamstown (Makhanda)
Elizabeth Donkin H	In-patient	Peri-urban	Port Elizabeth
Gqebera CHC	Out-patient	Urban informal	Port Elizabeth
KwaNobuhle CHC	Out-patient	Urban informal	Port Elizabeth
Dora Nginza H	In-patient	Urban informal	Port Elizabeth
Dora Nginza CHC	Out-patient	Urban informal	Port Elizabeth
Empilweni CHC	Out-patient	Urban informal	Port Elizabeth
KwaZakehlele CHC	Out-patient	Urban informal	Port Elizabeth
Motherwell CHC	Out-patient	Urban informal	Port Elizabeth
Komani H	In-patient	Peri-urban	Queenstown
Nomzano Clinic	Out-patient	Peri-urban	Queenstown
Philani CHC	Out-patient	Peri-urban	Queenstown

Table S3. Variants, by type, in Xhosa cases and controls.

Variant type	Cases	Controls	OR	P value
All Xhosa specific private variants				
Missense	15439 (16.98)	15165 (16.54)	1.01	0.50
Silent	7193 (7.91)	7216 (7.87)	0.98	0.32
Nonsense	502 (0.55)	495 (0.54)	1.00	1.00
Splice	278 (0.31)	248 (0.27)	1.10	0.28
Frameshift	156 (0.17)	147 (0.16)	1.04	0.70
Inframe	83 (0.09)	76 (0.08)	1.09	0.59
Stoploss	16 (0.02)	18 (0.02)	0.78	0.49
All variants	23,667 (26.04)	23,365 (25.48)	1.01	0.88
All damaging mutations in intolerant genes				
Missense	1052 (1.16)	998 (1.09)	1.07	0.13
Nonsense	94 (0.10)	97 (0.11)	0.93	0.66
Splice	56 (0.06)	31 (0.03)	1.96	0.003
Frameshift	27 (0.03)	15 (0.02)	1.97	0.03
All damaging variants	1229 (1.35)	1141 (1.24)	1.10	0.03
Damaging mutations in case only control only genes				
Missense	543 (0.60)	456 (0.50)	1.25	0.001
Nonsense	79 (0.09)	78 (0.09)	0.97	0.84
Splice	42 (0.05)	29 (0.03)	1.50	0.10
Frameshift	25 (0.03)	12 (0.01)	2.35	0.01
All qualified variants	689 (0.76)	575 (0.63)	1.25	0.0002

Table S4. Enrichment of private damaging mutations in cases and controls with different pLI thresholds.

pLI ≥ 0.5	All case genes or control genes			Synaptic case genes or control genes		
Threshold	1	2	3	1	2	3
Cases (N=909)	483	219	75	109	52	29
Controls (N=917)	436	135	45	78	14	6
Genes mutant in cases	547	106	22	88	21	9
Genes mutant in controls	493	62	14	72	6	2
Mutations in cases	689	248	79	120	53	29
Mutations in controls	575	143	47	80	14	6
O.R.	1.27	1.89	1.73	1.46	3.89	5.01
95% CI	[1.05 - 1.54]	[1.48 - 2.43]	[1.17 - 2.57]	[1.07 - 2.01]	[2.19 - 7.34]	[2.24 - 13.05]
P-value	0.015	2.53E-07	0.0054	0.02	1.28E-06	3.50E-05
O.R, per event, oligonic model		1.25			1.51	
95% CI		[1.11 - 1.41]			[1.14 - 2.02]	
P-value		0.0002			0.004	

pLI ≥ 0.9	All case genes or control genes			Synaptic case genes or control genes		
	1	2	3	1	2	3
Cases (N=909)	453	210	74	102	50	29
Controls (N=917)	419	140	48	74	14	6
Genes mutant in cases	498	101	22	82	20	9
Genes mutant in controls	453	64	15	68	6	2
Mutations in cases	634	237	78	113	51	29
Mutations in controls	538	148	50	76	14	6
O.R.	1.21	1.70	1.56	1.44	3.74	5.01
95% CI	[0.99 - 1.46]	[1.33 - 2.18]	[1.06 - 2.30]	[1.04 - 1.99]	[2.10 - 7.09]	[2.24 - 13.05]
P-value	0.057	1.84E-05	0.023	0.026	3.16E-06	3.50E-05
O.R, per event, oligonic model		1.23			1.49	
95% CI		[1.09 - 1.39]			[1.11 - 2.00]	
P-value		0.001			0.007	

Table S5. Private damaging mutations per person in cases and controls.

Number of mutations per person	Cases	Controls
5	1	0
4	3	0
3	31	22
2	129	94
1	319	320
0	426	481

Number of mutations in synaptic genes per person	Cases	Controls
3	1	0
2	9	2
1	99	76
0	800	839

Case status is significantly associated with harboring an increasing number of private rare damaging mutations in intolerant genes (odds ratio 1.25, 95% C.I. [1.11 - 1.41], P = 0.0002, for each additional mutation); and for harboring an increasing number of private damaging mutations in the subset of synaptic genes (odds ratio 1.51, 95% C.I. [1.14 - 2.02], P = 0.004, for each additional mutation). The analyses were corrected for the total number of private variants per person to control for outliers.

Table S6. Cases versus controls harboring private damaging mutations in genes enriched for brain expression (pSI < 0.01).

Tissues (pSI <0.01)	All genes	Cases	Controls	OR	95 % CI	P value
Brain	1094	80	47	1.81	[1.24 - 2.68]	0.002
Esophagus	229	2	9	0.25	[0.04 - 0.96]	0.043
Testis	1862	34	21	1.75	[1.00 - 3.13]	0.048
Uterus	155	10	3	2.92	[0.92 - 11.78]	0.070
Pituitary	665	40	24	1.61	[0.96 - 2.76]	0.072
Lung	336	1	7	0.25	[0.03 - 1.21]	0.089
Breast	141	1	6	0.25	[0.03 - 1.23]	0.093
Adipose tissue	174	1	5	0.32	[0.03 - 1.64]	0.18
Thyroid	252	3	8	0.44	[0.11 - 1.51]	0.20
Colon	401	7	13	0.56	[0.21 - 1.37]	0.21
Adrenal gland	289	10	5	1.93	[0.69 - 5.99]	0.21
Vagina	280	6	2	2.38	[0.58 - 13.37]	0.24
Pancreas	239	6	11	0.60	[0.21 - 1.59]	0.31
Fallopian tube	66	2	3	0.48	[0.08 - 2.64]	0.39
Muscle	351	7	12	0.67	[0.25 - 1.68]	0.40
Prostate	190	1	4	0.48	[0.05 - 2.63]	0.42
Kidney	372	6	5	1.47	[0.46 - 4.88]	0.51
Nerve	278	8	12	0.81	[0.32 - 2.00]	0.65
Skin	488	10	7	1.24	[0.47 - 3.42]	0.66
Blood	582	20	16	1.13	[0.58 - 2.25]	0.72
Liver	457	5	4	1.26	[0.35 - 4.80]	0.72
Stomach	216	3	4	0.83	[0.17 - 3.92]	0.82
Blood vessel	202	8	10	0.93	[0.35 - 2.40]	0.89
Heart	241	7	7	0.93	[0.32 - 2.75]	0.90
Ovary	184	8	8	0.96	[0.35 - 2.64]	0.93

Analysis is based on tissue enriched expression analysis (http://genetics.wustl.edu/jdlab/tsea/) in 25 GTEx tissues. “All genes” refers to the total number of genes that are more highly expressed in the specified tissue relative to other tissues. The statistical comparison is between the number of cases versus controls harboring one or more private damaging mutations in a tissue enriched gene.

Table S7. Enrichment of private damaging mutations in cases and controls, controlling for HIV status.

All samples		All case genes or control genes			Synaptic case genes or control genes		
	Threshold	1	2	3	1	2	3
Without HIV status as a covariate	Cases (N=909)	483	219	75	109	52	29
	Controls (N=917)	436	135	45	78	14	6
	O.R.	1.27	1.89	1.73	1.46	3.89	5.01
	95% CI	[1.05 - 1.54]	[1.48 - 2.43]	[1.17 - 2.57]	[1.07 - 2.01]	[2.19 - 7.34]	[2.24 - 13.05]
	P-value	0.015	2.53E-07	0.0054	0.02	1.28E-06	3.50E-05
	O.R (oligogenic)		1.25			1.51	
	95% CI		[1.11 - 1.41]			[1.14 - 2.02]	
	P-value		0.0002			0.004	
With HIV status as a covariate	O.R.	1.27	1.84	1.71	1.36	3.61	4.57
	95% CI	[1.05 - 1.55]	[1.43 - 2.36]	[1.15 - 2.55]	[0.99 - 1.87]	[2.02 - 6.84]	[2.04 - 11.86]
	P value	0.015	1.18E-06	0.007	0.058	5.63E-06	0.0001
	O.R (oligogenic)		1.24			1.40	
	95% CI		[1.10 - 1.40]			[1.05 - 1.88]	
	P-value		0.0005			0.02	

Excluding all HIV positive individuals		All case genes or control genes			Intolerant synaptic genes		
	Threshold	1	2	3	1	2	3
HIV non-reactive samples	Cases (n=861)	451	208	70	107	51	28
	Controls (n=771)	368	119	39	73	13	6
	O.R.	1.21	1.79	1.64	1.36	3.64	4.38
	95% CI	[0.99 - 1.48]	[1.39 - 2.33]	[1.09 - 2.50]	[0.99 - 1.89]	[2.02 - 7.01]	[1.95 - 11.42]
	P-value	0.07	6.96E-06	0.02	0.06	7.26E-06	0.0002
	O.R (oligogenic)		1.22			1.41	
	95% CI		[1.08 - 1.38]			[1.06 - 1.90]	
	P-value		0.002			0.02	

Table S8A

SKAT analysis of Xhosa-specific variants with threshold P-values

	MAF threshold	Variants	Genes	FWER 0.05	Bonferroni	Significant genes
Damaging	0.001	14,392	8231	3.30E-04	6.07E-06	none
Damaging	0.005	17,615	9247	4.61E-05	5.41E-06	none
Damaging	0.01	18,099	9383	3.46E-05	5.33E-06	none

Table S8B

SKAT Analysis: Top 10 genes for each minor allele frequency (MAF) threshold

Gene	P-value	Number of variants tested	Cases	Controls
MAF 0.001				
MEGF6	0.003	4	2	4
IGFN1	0.003	3	3	3
KAT6B	0.005	2	3	2
ATF7IP2	0.005	2	0	4
SH3BP1	0.005	4	2	3
TTC28	0.006	4	0	8
SEPSECS	0.007	2	1	3
MMP8	0.008	2	0	4
FREM2	0.009	7	4	3
VPS13A	0.009	5	3	4
MAF 0.005				
EXO1	0.00005	1	0	14
PTPN22	0.0001	1	15	1
SYT6	0.0002	1	15	1
GNRHR	0.0005	1	13	0
DLG1	0.0008	4	2	5
SPICE1	0.001	2	14	2
GPATCH1	0.002	3	6	12
FAT1	0.003	14	18	24
UBA7	0.003	5	9	19
OR10G3	0.003	2	1	7
MAF 0.01				
EXO1	0.00005	1	0	14
PTPN22	0.0002	1	15	1
SYT6	0.0002	1	15	1
GNRHR	0.0005	1	13	0
DLG1	0.0008	4	2	5
SPICE1	0.001	2	14	2
GPATCH1	0.002	3	6	12
COL14A1	0.002	9	26	11
FAT1	0.003	14	18	24
UBA7	0.003	5	9	19

Adjustment for multiple comparisons was based on the family-wise error rate (FWER).

Table S9. The common SNPs with smallest P-values in the exome-wide association analysis (MAF ≥ 0.01).

								MAF in Xhosa		MAF in 1000G				MAF in AGVP	
Coordinate	Alleles	SNP id	Gene	Variant	Context	Odds ratio	P value	Cases	Controls	AFR	EUR	EAS	SAS	South	East
chr17:4,997,433	G>A	rs12600437	ZFP3	c.*1125G>A	3' UTR	0.60	1.73E-07	0.12	0.19	0.10	0.16	0.47	0.55	0.18	0.13
chr17:35,538,304	G>C	rs6607364	ACACA	c.4777-7C>G	Intronic	0.63	1.33E-06	0.14	0.20	0.07	0.02	0.06	0.01	0.15	0.07
chr21:35,138,326	A>G	rs2248815	ITSN1	c.926+10A>G	Intronic	1.93	3.74E-06	0.09	0.05	0.14	0.18	0.14	0.11	0.09	0.11
chr12:67,706,466	G>A	rs1060350	CAND1	c.G3549A	Silent	1.47	1.68E-05	0.21	0.16	0.22	0.59	0.32	0.50	0.17	0.31

Table S10. Comparison of Xhosa and Swedish cohorts.

Xhosa population

	All case genes or control genes			Synaptic case genes or control genes		
Threshold	**1**	**2**	**3**	**1**	**2**	**3**
Cases (n=909)	483	219	75	109	52	29
Controls (n=917)	436	135	45	78	14	6
O.R.	1.25	1.84	1.74	1.47	3.91	5.00
95% CI	[1.04 - 1.51]	[1.44 - 2.35]	[1.17 - 2.61]	[1.07 - 2.02]	[2.12 - 7.70]	[2.03 - 14.80]
P-value (Fisher)	0.017	4.44E-07	0.0045	0.017	1.07E-06	5.61E-05

Swedish population

	All case genes or control genes			Synaptic case genes or control genes		
Threshold	**1**	**2**	**3**	**1**	**2**	**3**
Cases (n=909)	506	220	80	115	51	20
Controls (n=909)	458	184	65	93	37	14
O.R.	1.24	1.26	1.25	1.27	1.40	1.44
95% CI	[1.02 - 1.49]	[1.00 - 1.58]	[0.88 - 1.79]	[0.94 - 1.72]	[0.89 - 2.22]	[0.69 - 3.10]
P-value (Fisher)	0.03	0.05	0.23	0.12	0.16	0.39
Cases (n=4000)	1139	625	300	174	105	59
Controls (n=4000)	904	403	161	127	63	30
OR	1.36	1.65	1.93	1.39	1.68	1.98
95% CI	[1.23 - 1.51]	[1.44 - 1.89]	[1.58 - 2.37]	[1.09 - 1.77]	[1.22 - 2.35]	[1.25 - 3.19]
P value (Fisher)	1.92E-09	1.28E-13	2.59E-11	0.007	0.001	0.003

References and Notes

1. J. McClellan, M. C. King, Genetic heterogeneity in human disease. *Cell* **141**, 210–217 (2010). doi:10.1016/j.cell.2010.03.032 Medline

2. A. Ganna, F. K. Satterstrom, S. M. Zekavat, I. Das, M. I. Kurki, C. Churchhouse, J. Alfoldi, A. R. Martin, A. S. Havulinna, A. Byrnes, W. K. Thompson, P. R. Nielsen, K. J. Karczewski, E. Saarentaus, M. A. Rivas, N. Gupta, O. Pietiläinen, C. A. Emdin, F. Lescai, J. Bybjerg-Grauholm, J. Flannick, J. M. Mercader, M. Udler, M. Laakso, V. Salomaa, C. Hultman, S. Ripatti, E. Hämäläinen, J. S. Moilanen, J. Körkkö, O. Kuismin, M. Nordentoft, D. M. Hougaard, O. Mors, T. Werge, P. B. Mortensen, D. MacArthur, M. J. Daly, P. F. Sullivan, A. E. Locke, A. Palotie, A. D. Børglum, S. Kathiresan, B. M. Neale; GoT2D/T2D-GENES Consortium; SIGMA Consortium Helmsley IBD Exome Sequencing Project; FinMetSeq Consortium; iPSYCH-Broad Consortium, Quantifying the Impact of Rare and Ultra-rare Coding Variation across the Phenotypic Spectrum. *Am. J. Hum. Genet.* **102**, 1204–1211 (2018). doi:10.1016/j.ajhg.2018.05.002 Medline

3. G. Genovese, M. Fromer, E. A. Stahl, D. M. Ruderfer, K. Chambert, M. Landén, J. L. Moran, S. M. Purcell, P. Sklar, P. F. Sullivan, C. M. Hultman, S. A. McCarroll, Increased burden of ultra-rare protein-altering variants among 4,877 individuals with schizophrenia. *Nat. Neurosci.* **19**, 1433–1441 (2016). doi:10.1038/nn.4402 Medline

4. S. Gulsuner, T. Walsh, A. C. Watts, M. K. Lee, A. M. Thornton, S. Casadei, C. Rippey, H. Shahin, Consortium on the Genetics of Schizophrenia (COGS), PAARTNERS Study Group, V. L. Nimgaonkar, R. C. P. Go, R. M. Savage, N. R. Swerdlow, R. E. Gur, D. L. Braff, M.-C. King, J. M. McClellan, Spatial and temporal mapping of de novo mutations in schizophrenia to a fetal prefrontal cortical network. *Cell* **154**, 518–529 (2013). doi:10.1016/j.cell.2013.06.049 Medline

5. M. Fromer, A. J. Pocklington, D. H. Kavanagh, H. J. Williams, S. Dwyer, P. Gormley, L. Georgieva, E. Rees, P. Palta, D. M. Ruderfer, N. Carrera, I. Humphreys, J. S. Johnson, P. Roussos, D. D. Barker, E. Banks, V. Milanova, S. G. Grant, E. Hannon, S. A. Rose, K. Chambert, M. Mahajan, E. M. Scolnick, J. L. Moran, G. Kirov, A. Palotie, S. A. McCarroll, P. Holmans, P. Sklar, M. J. Owen, S. M. Purcell, M. C. O'Donovan, De novo mutations in schizophrenia implicate synaptic networks. *Nature* **506**, 179–184 (2014). doi:10.1038/nature12929 Medline

6. Schizophrenia Working Group of the Psychiatric Genomics Consortium, Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014). doi:10.1038/nature13595 Medline

7. J. M. McClellan, T. Lehner, M. C. King, Gene Discovery for Complex Traits: Lessons from Africa. *Cell* **171**, 261–264 (2017). doi:10.1016/j.cell.2017.09.037 Medline

8. L. L. Cavalli-Sforza, P. Menozzi, A. Piazza, *The History and Geography of Human Genes* (Princeton Univ. Press, 1994).

9. E. R. Chimusa, A. Meintjies, M. Tchanga, N. Mulder, C. Seoighe, H. Soodyall, R. Ramesar, A genomic portrait of haplotype diversity and signatures of selection in indigenous southern African populations. *PLOS Genet.* **11**, e1005052 (2015). doi:10.1371/journal.pgen.1005052 Medline

10. J. B. Peires, *The House of Phalo: A History of the Xhosa People in the Days of Their Independence* (Univ. of California Press, 1982).

11. J. L. Newman, *The Peopling of Africa: A Geographic Interpretation* (Yale Univ. Press, 1995).

12. D. W. Phillipson, *African Archaeology* (Cambridge Univ. Press, ed. 3, 2005).

13. Materials and methods are available as supplementary materials.

14. A. Wells, N. Kopp, X. Xu, D. R. O'Brien, W. Yang, A. Nehorai, T. L. Adair-Kirk, R. Kopan, J. D. Dougherty, The anatomical distribution of genetic associations. *Nucleic Acids Res.* **43**, 10804–10820 (2015). doi:10.1093/nar/gkv1262 Medline

15. F. Koopmans, P. van Nierop, M. Andres-Alonso, A. Byrnes, T. Cijsouw, M. P. Coba, L. N. Cornelisse, R. J. Farrell, H. L. Goldschmidt, D. P. Howrigan, N. K. Hussain, C. Imig, A. P. H. de Jong, H. Jung, M. Kohansalnodehi, B. Kramarz, N. Lipstein, R. C. Lovering, H. MacGillavry, V. Mariano, H. Mi, M. Ninov, D. Osumi-Sutherland, R. Pielot, K.-H. Smalla, H. Tang, K. Tashman, R. F. G. Toonen, C. Verpelli, R. Reig-Viader, K. Watanabe, J. van Weering, T. Achsel, G. Ashrafi, N. Asi, T. C. Brown, P. De Camilli, M. Feuermann, R. E. Foulger, P. Gaudet, A. Joglekar, A. Kanellopoulos, R. Malenka, R. A. Nicoll, C. Pulido, J. de Juan-Sanz, M. Sheng, T. C. Südhof, H. U. Tilgner, C. Bagni, À. Bayés, T. Biederer, N. Brose, J. J. E. Chua, D. C. Dieterich, E. D. Gundelfinger, C. Hoogenraad, R. L. Huganir, R. Jahn, P. S. Kaeser, E. Kim, M. R. Kreutz, P. S. McPherson, B. M. Neale, V. O'Connor, D. Posthuma, T. A. Ryan, C. Sala, G. Feng, S. E. Hyman, P. D. Thomas, A. B. Smit, M. Verhage; SynGo Consortium, SynGO: An evidence-based, expert-curated knowledge base for the synapse. *Neuron* **103**, 217–234.e4 (2019). doi:10.1016/j.neuron.2019.05.002 Medline

16. M. J. Gandal, P. Zhang, E. Hadjimichael, R. L. Walker, C. Chen, S. Liu, H. Won, H. van Bakel, M. Varghese, Y. Wang, A. W. Shieh, J. Haney, S. Parhami, J. Belmont, M. Kim, P. Moran Losada, Z. Khan, J. Mleczko, Y. Xia, R. Dai, D. Wang, Y. T. Yang, M. Xu, K. Fish, P. R. Hof, J. Warrell, D. Fitzgerald, K. White, A. E. Jaffe, M. A. Peters, M. Gerstein, C. Liu, L. M. Iakoucheva, D. Pinto, D. H. Geschwind; PsychENCODE Consortium, Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder. *Science* **362**, eaat8127 (2018). doi:10.1126/science.aat8127 Medline

17. I. Ionita-Laza, S. Lee, V. Makarov, J. D. Buxbaum, X. Lin, Sequence kernel association tests for the combined effect of rare and common variants. *Am. J. Hum. Genet.* **92**, 841–853 (2013). doi:10.1016/j.ajhg.2013.04.015 Medline

18. C. R. Marshall, D. P. Howrigan, D. Merico, B. Thiruvahindrapuram, W. Wu, D. S. Greer, D. Antaki, A. Shetty, P. A. Holmans, D. Pinto, M. Gujral, W. M. Brandler, D. Malhotra, Z. Wang, K. V. F. Fajarado, M. S. Maile, S. Ripke, I. Agartz, M. Albus, M. Alexander, F. Amin, J. Atkins, S. A. Bacanu, R. A. Belliveau, S. E. Bergen, M. Bertalan, E. Bevilacqua, T. B. Bigdeli, D. W. Black, R. Bruggeman, N. G. Buccola, R. L. Buckner, B. Bulik-Sullivan, W. Byerley, W. Cahn, G. Cai, M. J. Cairns, D. Campion, R. M. Cantor, V. J. Carr, N. Carrera, S. V. Catts, K. D. Chambert, W. Cheng, C. R. Cloninger, D. Cohen, P. Cormican, N. Craddock, B. Crespo-Facorro, J. J. Crowley, D. Curtis, M. Davidson, K. L. Davis, F. Degenhardt, J. Del Favero, L. E. DeLisi, D. Dikeos, T. Dinan, S. Djurovic, G. Donohoe, E. Drapeau, J. Duan, F. Dudbridge, P. Eichhammer, J.

Eriksson, V. Escott-Price, L. Essioux, A. H. Fanous, K.-H. Farh, M. S. Farrell, J. Frank, L. Franke, R. Freedman, N. B. Freimer, J. I. Friedman, A. J. Forstner, M. Fromer, G. Genovese, L. Georgieva, E. S. Gershon, I. Giegling, P. Giusti-Rodríguez, S. Godard, J. I. Goldstein, J. Gratten, L. de Haan, M. L. Hamshere, M. Hansen, T. Hansen, V. Haroutunian, A. M. Hartmann, F. A. Henskens, S. Herms, J. N. Hirschhorn, P. Hoffmann, A. Hofman, H. Huang, M. Ikeda, I. Joa, A. K. Kähler, R. S. Kahn, L. Kalaydjieva, J. Karjalainen, D. Kavanagh, M. C. Keller, B. J. Kelly, J. L. Kennedy, Y. Kim, J. A. Knowles, B. Konte, C. Laurent, P. Lee, S. H. Lee, S. E. Legge, B. Lerer, D. L. Levy, K.-Y. Liang, J. Lieberman, J. Lönnqvist, C. M. Loughland, P. K. E. Magnusson, B. S. Maher, W. Maier, J. Mallet, M. Mattheisen, M. Mattingsdal, R. W. McCarley, C. McDonald, A. M. McIntosh, S. Meier, C. J. Meijer, I. Melle, R. I. Mesholam-Gately, A. Metspalu, P. T. Michie, L. Milani, V. Milanova, Y. Mokrab, D. W. Morris, B. Müller-Myhsok, K. C. Murphy, R. M. Murray, I. Myin-Germeys, I. Nenadic, D. A. Nertney, G. Nestadt, K. K. Nicodemus, L. Nisenbaum, A. Nordin, E. O'Callaghan, C. O'Dushlaine, S.-Y. Oh, A. Olincy, L. Olsen, F. A. O'Neill, J. Van Os, C. Pantelis, G. N. Papadimitriou, E. Parkhomenko, M. T. Pato, T. Paunio, D. O. Perkins, T. H. Pers, O. Pietiläinen, J. Pimm, A. J. Pocklington, J. Powell, A. Price, A. E. Pulver, S. M. Purcell, D. Quested, H. B. Rasmussen, A. Reichenberg, M. A. Reimers, A. L. Richards, J. L. Roffman, P. Roussos, D. M. Ruderfer, V. Salomaa, A. R. Sanders, A. Savitz, U. Schall, T. G. Schulze, S. G. Schwab, E. M. Scolnick, R. J. Scott, L. J. Seidman, J. Shi, J. M. Silverman, J. W. Smoller, E. Söderman, C. C. A. Spencer, E. A. Stahl, E. Strengman, J. Strohmaier, T. S. Stroup, J. Suvisaari, D. M. Svrakic, J. P. Szatkiewicz, S. Thirumalai, P. A. Tooney, J. Veijola, P. M. Visscher, J. Waddington, D. Walsh, B. T. Webb, M. Weiser, D. B. Wildenauer, N. M. Williams, S. Williams, S. H. Witt, A. R. Wolen, B. K. Wormley, N. R. Wray, J. Q. Wu, C. C. Zai, R. Adolfsson, O. A. Andreassen, D. H. R. Blackwood, E. Bramon, J. D. Buxbaum, S. Cichon, D. A. Collier, A. Corvin, M. J. Daly, A. Darvasi, E. Domenici, T. Esko, P. V. Gejman, M. Gill, H. Gurling, C. M. Hultman, N. Iwata, A. V. Jablensky, E. G. Jönsson, K. S. Kendler, G. Kirov, J. Knight, D. F. Levinson, Q. S. Li, S. A. McCarroll, A. McQuillin, J. L. Moran, B. J. Mowry, M. M. Nöthen, R. A. Ophoff, M. J. Owen, A. Palotie, C. N. Pato, T. L. Petryshen, D. Posthuma, M. Rietschel, B. P. Riley, D. Rujescu, P. Sklar, D. St Clair, J. T. R. Walters, T. Werge, P. F. Sullivan, M. C. O'Donovan, S. W. Scherer, B. M. Neale, J. Sebat; CNV and Schizophrenia Working Groups of the Psychiatric Genomics Consortium, *Nat. Genet.* **49**, 27–35 (2017). doi:10.1038/ng.3725 Medline

19. N. Krumm, P. H. Sudmant, A. Ko, B. J. O'Roak, M. Malig, B. P. Coe, A. R. Quinlan, D. A. Nickerson, E. E. Eichler; NHLBI Exome Sequencing Project, Copy number variation detection and genotyping from exome sequence data. *Genome Res.* **22**, 1525–1532 (2012). doi:10.1101/gr.138115.112 Medline

20. D. Curtis, Polygenic risk score for schizophrenia is more strongly associated with ancestry than with schizophrenia. *Psychiatr. Genet.* **28**, 85–89 (2018). doi:10.1097/YPG.0000000000000206 Medline

21. J. Li, B. Wilkinson, V. A. Clementel, J. Hou, T. J. O'Dell, M. P. Coba, Long-term potentiation modulates synaptic phosphorylation networks and reshapes the structure of the postsynaptic interactome. *Sci. Signal.* **9**, rs8 (2016). doi:10.1126/scisignal.aaf6716 Medline

22. T. Singh, M. I. Kurki, D. Curtis, S. M. Purcell, L. Crooks, J. McRae, J. Suvisaari, H. Chheda, D. Blackwood, G. Breen, O. Pietiläinen, S. S. Gerety, M. Ayub, M. Blyth, T. Cole, D. Collier, E. L. Coomber, N. Craddock, M. J. Daly, J. Danesh, M. DiForti, A. Foster, N. B. Freimer, D. Geschwind, M. Johnstone, S. Joss, G. Kirov, J. Körkkö, O. Kuismin, P. Holmans, C. M. Hultman, C. Iyegbe, J. Lönnqvist, M. Männikkö, S. A. McCarroll, P. McGuffin, A. M. McIntosh, A. McQuillin, J. S. Moilanen, C. Moore, R. M. Murray, R. Newbury-Ecob, W. Ouwehand, T. Paunio, E. Prigmore, E. Rees, D. Roberts, J. Sambrook, P. Sklar, D. St Clair, J. Veijola, J. T. R. Walters, H. Williams, P. F. Sullivan, M. E. Hurles, M. C. O'Donovan, A. Palotie, M. J. Owen, J. C. Barrett; Swedish Schizophrenia Study; INTERVAL Study; DDD Study; UK10 K Consortium, Rare loss-of-function variants in SETD1A are associated with schizophrenia and developmental disorders. *Nat. Neurosci.* **19**, 571–577 (2016). doi:10.1038/nn.4267 Medline

23. S. J. Kaar, S. Natesan, R. McCutcheon, O. D. Howes, Antipsychotics: Mechanisms underlying clinical response and side-effects and novel treatment approaches based on pathophysiology. *Neuropharmacology*, 107704 (2019). doi:10.1016/j.neuropharm.2019.107704 Medline

24. W. C. Huang, K. Bennett, C. Gregg, Epigenetic and Cellular Diversity in the Brain through Allele-Specific Effects. *Trends Neurosci.* **41**, 925–937 (2018). doi:10.1016/j.tins.2018.07.005 Medline

25. C. Rotimi, A. Abayomi, A. Abimiku, V. M. Adabayeri, C. Adebamowo, E. Adebiyi, A. D. Ademola, A. Adeyemo, D. Adu, D. Affolabi, G. Agongo, S. Ajayi, S. Akarolo-Anthony, R. Akinyemi, A. Akpalu, M. Alberts, O. Alonso Betancourt, A. M. Alzohairy, G. Ameni, O. Amodu, G. Anabwani, K. Andersen, F. Arogundade, O. Arulogun, D. Asogun, R. Bakare, N. Balde, M. L. Baniecki, C. Beiswanger, A. Benkahla, L. Bethke, M. Boehnke, V. Boima, J. Brandful, A. I. Brooks, F. C. Brosius, C. Brown, B. Bucheton, D. T. Burke, B. G. Burnett, S. Carrington-Lawrence, N. Carstens, J. Chisi, A. Christoffels, R. Cooper, H. Cordell, N. Crowther, T. Croxton, J. de Vries, L. Derr, P. Donkor, S. Doumbia, A. Duncanson, I. Ekem, A. El Sayed, M. E. Engel, J. C. Enyaru, D. Everett, F. M. Fadlelmola, E. Fakunle, K. H. Fischbeck, A. Fischer, O. Folarin, J. Gamieldien, R. F. Garry, S. Gaseitsiwe, R. Gbadegesin, A. Ghansah, M. Giovanni, P. Goesbeck, F. X. Gomez-Olive, D. S. Grant, R. Grewal, M. Guyer, N. A. Hanchard, C. T. Happi, S. Hazelhurst, B. J. Hennig, C. Hertz-, Fowler, W. Hide, F. Hilderbrandt, C. Hugo-Hamman, M. E. Ibrahim, R. James, Y. Jaufeerally-Fakim, C. Jenkins, U. Jentsch, P. P. Jiang, M. Joloba, V. Jongeneel, F. Joubert, M. Kader, K. Kahn, P. Kaleebu, S. H. Kapiga, S. K. Kassim, I. Kasvosve, J. Kayondo, B. Keavney, A. Kekitiinwa, S. H. Khan, P. Kimmel, M. C. King, R. Kleta, M. Koffi, J. Kopp, M. Kretzler, J. Kumuthini, S. Kyobe, C. Kyobutungi, D. T. Lackland, K. A. Lacourciere, G. Landouré, R. Lawlor, T. Lehner, M. Lesosky, N. Levitt, K. Littler, Z. Lombard, J. F. Loring, S. Lyantagaye, A. Macleod, E. B. Madden, C. R. Mahomva, J. Makani, M. Mamven, M. Marape, G. Mardon, P. Marshall, D. P. Martin, D. Masiga, R. Mason, M. Mate-Kole, E. Matovu, M. Mayige, B. M. Mayosi, J. C. Mbanya, S. A. McCurdy, M. I. McCarthy, H. McIlleron, S. O. Mc'Ligeyo, C. Merle, A. O. Mocumbi, C. Mondo, J. V. Moran, A. Motala, M. Moxey-Mims, W. S. Mpoloka, C. L. Msefula, T. Mthiyane, N. Mulder, Gh. Mulugeta, D. Mumba, J. Musuku, M. Nagdee, O. Nash, D. Ndiaye, A. Q. Nguyen, M. Nicol, O. Nkomazana, S. Norris, B. Nsangi, A. Nyarko, M. Nyirenda, E. Obe, R. Obiakor, A.

Oduro, S. F. Ofori-Acquah, O. Ogah, S. Ogendo, K. Ohene-Frempong, A. Ojo, T. Olanrewaju, J. Oli, C. Osafo, O. Ouwe Missi Oukem-Boyer, B. Ovbiagele, A. Owen, M. O. Owolabi, L. Owolabi, E. Owusu-Dabo, G. Pare, R. Parekh, H. G. Patterton, M. B. Penno, J. Peterson, R. Pieper, J. Plange-Rhule, M. Pollak, J. Puzak, R. S. Ramesar, M. Ramsay, R. Rasooly, S. Reddy, P. C. Sabeti, K. Sagoe, T. Salako, O. Samassékou, M. S. Sandhu, O. Sankoh, F. S. Sarfo, M. Sarr, G. Shaboodien, I. Sidibe, G. Simo, M. Simuunza, L. Smeeth, E. Sobngwi, H. Soodyall, H. Sorgho, O. Sow Bah, S. Srinivasan, D. J. Stein, E. S. Susser, C. Swanepoel, G. Tangwa, A. Tareila, O. Tastan Bishop, B. Tayo, N. Tiffin, H. Tinto, E. Tobin, S. M. Tollman, M. Traoré, M. J. Treadwell, J. Troyer, M. Tsimako-Johnstone, V. Tukei, I. Ulasi, N. Ulenga, B. van Rooyen, A. P. Wachinou, S. P. Waddy, A. Wade, M. Wayengera, J. Whitworth, L. Wideroff, C. A. Winkler, S. Winnicki, A. Wonkam, M. Yewondwos, T. sen, N. Yozwiak, H. Zar; H3Africa Consortium, Enabling the genomic revolution in Africa., *Science* **344**, 1346–1348 (2014). doi:10.1126/science.1251546 Medline

26. American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders, Text Revision* (American Psychiatric Association, ed. 4, 2000).

27. M. M. Campbell, E. Susser, J. de Vries, A. Baldinger, G. Sibeko, M. M. Mndini, S. G. Mqulwana, O. A. Ntola, R. S. Ramesar, D. J. Stein, Exploring researchers' experiences of working with a researcher-driven, population-specific community advisory board in a South African schizophrenia genomics study. *BMC Med. Ethics* **16**, 45 (2015). doi:10.1186/s12910-015-0037-5 Medline

28. M. B. First, R. L. Spitzer, M. Gibbon, J. B. Williams, *Structured Clinical Interview for DSM-IV Axis I Disorders (SCID-I), Clinician Version, Administration Booklet* (American Psychiatric Press, 2012).

29. World Health Organization, Process of translation and adaptation of instruments; www.who.int/substance_abuse/research_tools/translation/en/ [accessed 8 July 2019].

30. F. J. Charlson, A. J. Ferrari, D. F. Santomauro, S. Diminic, E. Stockings, J. G. Scott, J. J. McGrath, H. A. Whiteford, Global Epidemiology and Burden of Schizophrenia: Findings From the Global Burden of Disease Study 2016. *Schizophr. Bull.* **44**, 1195–1203 (2018). doi:10.1093/schbul/sby058 Medline

31. R. E. Gur, R. G. Petty, B. I. Turetsky, R. C. Gur, Schizophrenia throughout life: Sex differences in severity and profile of symptoms. *Schizophr. Res.* **21**, 1–12 (1996). doi:10.1016/0920-9964(96)00023-0 Medline

32. M. N. Bainbridge, M. Wang, Y. Wu, I. Newsham, D. M. Muzny, J. L. Jefferies, T. J. Albert, D. L. Burgess, R. A. Gibbs, Targeted enrichment beyond the consensus coding DNA sequence exome reveals exons with higher variant densities. *Genome Biol.* **12**, R68 (2011). doi:10.1186/gb-2011-12-7-r68 Medline

33. J. G. Reid, A. Carroll, N. Veeraraghavan, M. Dahdouli, A. Sundquist, A. English, M. Bainbridge, S. White, W. Salerno, C. Buhay, F. Yu, D. Muzny, R. Daly, G. Duyk, R. A. Gibbs, E. Boerwinkle, Launching genomics into the cloud: Deployment of Mercury, a next generation sequence analysis pipeline. *BMC Bioinformatics* **15**, 30 (2014). doi:10.1186/1471-2105-15-30 Medline

34. M. A. DePristo, E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, A. A. Philippakis, G. del Angel, M. A. Rivas, M. Hanna, A. McKenna, T. J. Fennell, A. M. Kernytsky, A. Y. Sivachenko, K. Cibulskis, S. B. Gabriel, D. Altshuler, M. J. Daly, A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011). doi:10.1038/ng.806 Medline

35. S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. de Bakker, M. J. Daly, P. C. Sham, PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007). doi:10.1086/519795 Medline

36. A. Auton, L. D. Brooks, R. M. Durbin, E. P. Garrison, H. M. Kang, J. O. Korbel, J. L. Marchini, S. McCarthy, G. A. McVean, G. R. Abecasis; 1000 Genomes Project Consortium, A global reference for human genetic variation. *Nature* **526**, 68–74 (2015). doi:10.1038/nature15393 Medline

37. D. Gurdasani, T. Carstensen, F. Tekola-Ayele, L. Pagani, I. Tachmazidou, K. Hatzikotoulas, S. Karthikeyan, L. Iles, M. O. Pollard, A. Choudhury, G. R. S. Ritchie, Y. Xue, J. Asimit, R. N. Nsubuga, E. H. Young, C. Pomilla, K. Kivinen, K. Rockett, A. Kamali, A. P. Doumatey, G. Asiki, J. Seeley, F. Sisay-Joof, M. Jallow, S. Tollman, E. Mekonnen, R. Ekong, T. Oljira, N. Bradman, K. Bojang, M. Ramsay, A. Adeyemo, E. Bekele, A. Motala, S. A. Norris, F. Pirie, P. Kaleebu, D. Kwiatkowski, C. Tyler-Smith, C. Rotimi, E. Zeggini, M. S. Sandhu, The African Genome Variation Project shapes medical genetics in Africa. *Nature* **517**, 327–332 (2015). doi:10.1038/nature13997 Medline

38. A. Manichaikul, J. C. Mychaleckyj, S. S. Rich, K. Daly, M. Sale, W.-M. Chen, Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010). doi:10.1093/bioinformatics/btq559 Medline

39. D. H. Alexander, J. Novembre, K. Lange, Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009). doi:10.1101/gr.094052.109 Medline

40. K. Wang, M. Li, H. Hakonarson, ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010). doi:10.1093/nar/gkq603 Medline

41. I. A. Adzhubei, S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, P. Bork, A. S. Kondrashov, S. R. Sunyaev, A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010). doi:10.1038/nmeth0410-248 Medline

42. Y. Choi, G. E. Sims, S. Murphy, J. R. Miller, A. P. Chan, Predicting the functional effect of amino acid substitutions and indels. *PLOS ONE* **7**, e46688 (2012). doi:10.1371/journal.pone.0046688 Medline

43. G. Yeo, C. B. Burge, Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J. Comput. Biol.* **11**, 377–394 (2004). doi:10.1089/1066527041410418 Medline

44. M. Lek, K. J. Karczewski, E. V. Minikel, K. E. Samocha, E. Banks, T. Fennell, A. H. O'Donnell-Luria, J. S. Ware, A. J. Hill, B. B. Cummings, T. Tukiainen, D. P. Birnbaum, J. A. Kosmicki, L. E. Duncan, K. Estrada, F. Zhao, J. Zou, E. Pierce-Hoffman, J.

Berghout, D. N. Cooper, N. Deflaux, M. DePristo, R. Do, J. Flannick, M. Fromer, L. Gauthier, J. Goldstein, N. Gupta, D. Howrigan, A. Kiezun, M. I. Kurki, A. L. Moonshine, P. Natarajan, L. Orozco, G. M. Peloso, R. Poplin, M. A. Rivas, V. Ruano-Rubio, S. A. Rose, D. M. Ruderfer, K. Shakir, P. D. Stenson, C. Stevens, B. P. Thomas, G. Tiao, M. T. Tusie-Luna, B. Weisburd, H. H. Won, D. Yu, D. M. Altshuler, D. Ardissino, M. Boehnke, J. Danesh, S. Donnelly, R. Elosua, J. C. Florez, S. B. Gabriel, G. Getz, S. J. Glatt, C. M. Hultman, S. Kathiresan, M. Laakso, S. McCarroll, M. I. McCarthy, D. McGovern, R. McPherson, B. M. Neale, A. Palotie, S. M. Purcell, D. Saleheen, J. M. Scharf, P. Sklar, P. F. Sullivan, J. Tuomilehto, M. T. Tsuang, H. C. Watkins, J. G. Wilson, M. J. Daly, D. G. MacArthur; Exome Aggregation Consortium, *Nature* **536**, 285–291 (2016). doi:10.1038/nature19057 Medline

45. Genome Aggregation Database, gnomeAD; https://gnomad.broadinstitute.org/.

46. X. Wang, Firth logistic regression for rare variant association tests. *Front. Genet.* **5**, 187 (2014). Medline

47. K. E. Samocha, E. B. Robinson, S. J. Sanders, C. Stevens, A. Sabo, L. M. McGrath, J. A. Kosmicki, K. Rehnström, S. Mallick, A. Kirby, D. P. Wall, D. G. MacArthur, S. B. Gabriel, M. DePristo, S. M. Purcell, A. Palotie, E. Boerwinkle, J. D. Buxbaum, E. H. Cook Jr., R. A. Gibbs, G. D. Schellenberg, J. S. Sutcliffe, B. Devlin, K. Roeder, B. M. Neale, M. J. Daly, A framework for the interpretation of de novo mutation in human disease. *Nat. Genet.* **46**, 944–950 (2014). doi:10.1038/ng.3050 Medline

48. A. T. Marees, H. de Kluiver, S. Stringer, F. Vorspan, E. Curis, C. Marie-Claire, E. M. Derks, A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *Int. J. Methods Psychiatr. Res.* **27**, e1608 (2018). doi:10.1002/mpr.1608 Medline

49. J. L. Johnson, G. R. Abecasis, GAS Power Calculator: web-based power calculator for genetic association. bioRxiv 164343 [Preprint]. Xx xxxx xxxx. https://doi.org/10.1101/164343.

50. E. Rees, J. T. R. Walters, L. Georgieva, A. R. Isles, K. D. Chambert, A. L. Richards, G. Mahoney-Davies, S. E. Legge, J. L. Moran, S. A. McCarroll, M. C. O'Donovan, M. J. Owen, G. Kirov, Analysis of copy number variations at 15 schizophrenia-associated loci. *Br. J. Psychiatry* **204**, 108–114 (2014). doi:10.1192/bjp.bp.113.131052 Medline