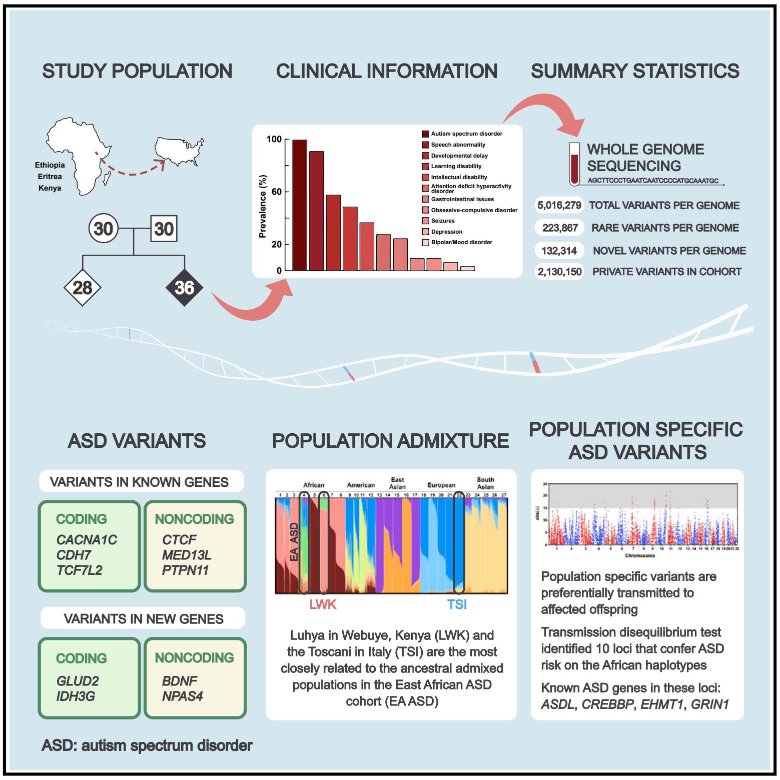
Cell Genomics
Article

The genetics of autism spectrum disorder in an East African familial cohort

Graphical abstract


STUDY POPULATION
Ethiopia
Eritrea
Kenya
30
30
28
36
CLINICAL INFORMATION
100
60
20
0
Prevalence (%)
Autism spectrum disorder
Speech abnormality
Developmental delay
Learning disability
Intellectual disability
Attention deficit hyperactivity disorder
Gastrointestinal issues
Obsessive-compulsive disorder
Seizures
Depression
Bipolar/Mood disorder
SUMMARY STATISTICS
WHOLE GENOME SEQUENCING
AGCTTCCCTGAATCAATCCCCATGCAAATGC
5,016,279 TOTAL VARIANTS PER GENOME
223,867 RARE VARIANTS PER GENOME
132,314 NOVEL VARIANTS PER GENOME
2,130,150 PRIVATE VARIANTS IN COHORT
ASD VARIANTS
VARIANTS IN KNOWN GENES
CODING
CACNA1C
CDH7
TCF7L2
NONCODING
CTCF
MED13L
PTPN11
VARIANTS IN NEW GENES
CODING
GLUD2
IDH3G
NONCODING
BDNF
NPAS4
POPULATION ADMIXTURE
African
American
East Asian
European
South Asian
EA ASD
LWK
TSI
Luhya in Webuye, Kenya (LWK) and the Toscani in Italy (TSI) are the most closely related to the ancestral admixed populations in the East African ASD cohort (EA ASD)
POPULATION SPECIFIC ASD VARIANTS
Chromosome
Population specific variants are preferentially transmitted to affected offspring
Transmission disequilibrium test identified 10 loci that confer ASD risk on the African haplotypes
Known ASD genes in these loci: ASDL, CREBBP, EHMT1, GRIN1
ASD: autism spectrum disorder

Highlights

- The study investigates the genetics of autism in an African population
- Whole-genome sequencing discovered 2.13 million novel private variants
- Detected rare variants in known and in new candidate autism genes
- Admixture mapping identified 10 autism risk loci on the African haplotypes

Authors

Islam Oguz Tuncay, Darlene DeVries, Ashlesha Gogate, ..., Kimberly Goodspeed, Leah Seyoum-Tesfa, Maria H. Chahrour

Correspondence

maria.chahrour@utsouthwestern.edu

In brief

Tuncay et al. investigate the genetics of autism in an East African cohort. They discover 2.13 million private variants and detect rare variants that segregate with autism in known genes and in new candidates. Through admixture mapping, they identify loci that confer autism risk on the African haplotypes.

Tuncay et al., 2023, Cell Genomics *3*, 100322
July 12, 2023 © 2023 The Author(s).
https://doi.org/10.1016/j.xgen.2023.100322

Cell Genomics

CellPress
OPEN ACCESS

Article

The genetics of autism spectrum disorder in an East African familial cohort

Islam Oguz Tuncay,[1] Darlene DeVries,[2] Ashlesha Gogate,[2] Kiran Kaur,[2] Ashwani Kumar,[2] Chao Xing,[2,3,4] Kimberly Goodspeed,[5,6,7] Leah Seyoum-Tesfa,[8] and Maria H. Chahrour[1,2,7,9,10,11,*]

[1]Department of Neuroscience, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA
[2]Eugene McDermott Center for Human Growth and Development, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA
[3]Department of Population and Data Sciences, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA
[4]Lyda Hill Department of Bioinformatics, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA
[5]Department of Pediatrics, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA
[6]Department of Neurology, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA
[7]Department of Psychiatry, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA
[8]Reaching Families Advocacy and Support Group, Dallas, TX 75243, USA
[9]Center for the Genetics of Host Defense, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA
[10]Peter O'Donnell Jr. Brain Institute, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA
[11]Lead contact
*Correspondence: maria.chahrour@utsouthwestern.edu

https://doi.org/10.1016/j.xgen.2023.100322

SUMMARY

Autism spectrum disorder (ASD) is a group of complex neurodevelopmental conditions affecting communication and social interaction in 2.3% of children. Studies that demonstrated its complex genetic architecture have been mainly performed in populations of European ancestry. We investigate the genetics of ASD in an East African cohort (129 individuals) from a population with higher prevalence (5%). Whole-genome sequencing identified 2.13 million private variants in the cohort and potentially pathogenic variants in known ASD genes (including *CACNA1C*, *CHD7*, *FMR1*, and *TCF7L2*). Admixture analysis demonstrated that the cohort comprises two ancestral populations, African and Eurasian. Admixture mapping discovered 10 regions that confer ASD risk on the African haplotypes, containing several known ASD genes. The increased ASD prevalence in this population suggests decreased heterogeneity in the underlying genetic etiology, enabling risk allele identification. Our approach emphasizes the power of African genetic variation and admixture analysis to inform the architecture of complex disorders.

INTRODUCTION

Autism spectrum disorder (ASD) is a constellation of individually rare neurodevelopmental disorders characterized by stereotyped behaviors and impairments in social interaction and communication. Despite the high heritability estimates of ASD (~83%–90%[1–3]), extreme genetic and phenotypic heterogeneity have posed major obstacles to gene discovery.[4] Rare variants, both spontaneously arising (*de novo*) and inherited, have been demonstrated to contribute to ASD, highlighting its complex genetic architecture.[5] Significant advances over the past decade enabled the identification of hundreds of genes that underlie ASD. However, the genetic variants identified to date account for only ~30% of the disease burden.[6]

ASD affects 1 in 44 children in the United States (US).[7] The reported prevalence is geographically variable within the US, presumably due to decreased surveillance where access to assessment, diagnosis, and services are limited. Several studies have shown a higher prevalence of ASD in children born to East African parents, including increased prevalence in children of Somali and Ethiopian parents living in Sweden, and in the Somali community in Minnesota.[8–10] Furthermore, an epidemiological study conducted across Texas public schools found that 1 in 20 children (5%) of Ethiopian ancestry is affected, which is two times higher than the prevalence of ASD in the general population.[11] Since large-scale prevalence studies for ASD have not been carried out in Africa, data from the aforementioned studies in immigrant populations can provide reliable insights into population prevalence. While there is a potential for ascertainment bias in small studies and for other confounding factors such as the stress of immigration or assessment bias due to cultural differences, these studies nonetheless strongly suggest increased ASD prevalence in East African children. An important motivation for our study is the paucity of genomic studies that capture the diversity on the African continent, and, as genetic findings move into the clinic, this translates to healthcare disparities.[12–14] Here, we set out to investigate the genetic contribution to the increased ASD prevalence in children from East African origin.

Populations in Africa are the most genetically diverse in the world, carrying up to three times as many rare variants as

European or East Asian populations.[15] To date there is no comprehensive catalog of African genetic diversity. The African Genome Variation Project, an international collaboration aimed at characterizing African genetic diversity, reported that the Ethiopian population has the greatest proportion of all the novel and private genomic variation found in Africa (~24% of genomic variants).[16] Given the higher incidence of ASD in the East African population, we sought to investigate the underlying genetic susceptibility factor(s) in this population. We recruited a cohort of East African families in the US ascertained on probands with ASD. The parents who enrolled in the study were all born in East Africa, predominantly Ethiopia and Eritrea, and the children were all born to East African parents. Thus, there is no present-day admixture in our cohort with individuals of other ancestries. We leveraged ancestral genomic information to identify ASD genes through whole-genome sequencing (WGS) and admixture analysis in a familial cohort from a population with a high rate of ASD.

RESULTS

WGS and variant identification in the East African ASD cohort

East African families who have at least one child affected with ASD were enrolled in our study in the US. Of the total 33 families who enrolled, the majority are Ethiopian, Eritrean, or both, and one family is Kenyan. Although the populations in East Africa are highly genetically diverse, the majority of families enrolled in our study come from two geographically adjacent regions of Ethiopia, the Amhara and Tigray regions, and belong to the Ethiosemitic linguistic group, a subset of the Afroasiatic family of languages. Interestingly, genome-wide genotyping data suggest that populations of Amhara and Tigray share common genetic ancestry.[17,18] The cohort consisted of 30 trios with one affected offspring (simplex families) and three multiplex families. Parental age, which is a possible risk factor for ASD, was not significantly different at the time of birth of affected compared with unaffected offspring (Figure S1). All affected individuals received a clinical diagnosis of ASD, and the affected male to female ratio in the cohort was 3.5 to 1. Demographics and clinical information for the cohort are provided in Figure S2, Table S1, and under ''clinical information'' (STAR Methods).

We performed WGS on samples from 129 individuals in the East African ASD cohort including 36 affected children. The average read depth was 37×, with no difference in sequencing depth between samples from affected and unaffected individuals (Figure S3A), male and female (Figure S3B), or parent and offspring (Figure S3C). On average, 97.55% and 86.87% of bases were covered at a mean read depth of at least 10× and 20×, respectively (Figure S3D). An average of 5,015,279 total variants were identified per genome; of those, 4,024,260 were single nucleotide variants (SNVs) and 991,020 were insertions or deletions (indels) (Table S2). After filtering for rare variants with a minor allele frequency (MAF) <1% in all annotated populations (the 1000 Genomes project [1000G],[19] the Genome Aggregation Database [gnomAD],[20] the Greater Middle East Variome project [GME],[21] and the Human Heredity and Health in Africa project [H3Africa][22,23]), there were, on average, 223,867 rare variants per genome, of which 213,178 were heterozygous and 10,690 were homozygous (Table S2). We discovered an average of 132,314 novel variants per genome that have not been reported in any of the populations in the public databases that we used for annotation (Table S2). Furthermore, we found an average of 13,875 novel variants per individual that were private, meaning they have not been reported in any of the annotated populations and they were not present in any other individual in the cohort (Table S3). In total, there were 2,130,158 novel private variants in the cohort (Table S3). As expected, more private variants were present in parents compared with offspring (Figure S4). It is important to note that the presence of private variants may relate to the fact that a substantial amount of genetic information in African populations is not represented in the current human reference genome. As recently reported, the pan-African genome contains approximately 10% more DNA than the most current version of the human reference genome.[24] An average of 23 private variants per individual were predicted to be possibly pathogenic (Table S3). Given the genetic diversity present on the African continent and the lack of comprehensive sequencing of samples from diverse African populations,[22,25] this is in line with expectations that we would discover a meaningful number of novel variants in our cohort.

To determine whether there was an excess of potentially pathogenic variants in affected compared with unaffected individuals in the cohort, we performed a burden analysis. We compared rare and rare homozygous variants under four categories: total coding, loss of function (LoF), missense damaging, or LoF and missense damaging. We found no difference in the burden of these variants between affected and unaffected individuals (Figure S5). This result is unsurprising given that ASD is a collection of individually rare diseases and is genetically heterogeneous, caused by rare alleles of large effect; therefore, an excess of rare alleles will not be observed except in very large cohorts that can capture this heterogeneity.[26] Given the increased prevalence of ASD in East African families, we expect inherited variation to drive the increase in prevalence with a decrease in heterogeneity compared with populations with lower prevalence.

Given the contribution of copy number variants (CNVs) to ASD,[27] we called CNVs in affected individuals using within-family controls by CNVkit.[28] Out of the total CNVkit calls, we identified 84 loci that overlapped with known ASD CNVs, as defined by the Simons Foundation Autism Research Initiative (SFARI) Gene database[29] (one to eight CNVs per affected individual, with an average of two CNVs) (Table S4A). To identify genomic regions that were significantly deleted or amplified across multiple affected individuals compared with controls, we analyzed the CNVkit segmentation files with GISTIC.[30] We identified eight regions on chromosomes 2, 8, 15, 17, and 22 (false discovery rate <0.01), six deleted and two amplified, that overlapped with known ASD CNVs (Table S4B). One particular region that stood out in our CNV analysis was the 17p11.2 locus. Deletions and duplications at this locus are associated with Smith-Magenis[31] and Potocki-Lupski syndromes,[32] respectively, and have also been reported in ASD. Both syndromes are characterized by intellectual disability, speech delay, and a host of developmental abnormalities. Although African population frequency data for the four 17p11.2 CNVs that we identified (Table S4B) are lacking, we examined their frequency in the gnomAD structural variant (SV) database, which contains data from 10,847 genomes of different ancestries, including African American.[33] We did not find the four CNVs (with the same boundaries) in gnomAD SV; however, we found CNVs that overlapped with them. The population allele frequencies for the overlapping CNVs ranged from 0.0046% to 29.8% across all populations and from 0% to 21.3% in the African American population in gnomAD SV (Table S4C). Further investigation including sequencing larger African cohorts and functional analyses are necessary to determine whether the identified CNVs play a pathogenic role in ASD in affected individuals in the cohort.

Discovery of candidate ASD variants in known and in new disease genes

We initially focused our analysis on rare coding variants (MAF < 1% in 1000G,[19] gnomAD,[20] GME,[21] and H3Africa[22,23]) in affected individuals in our cohort that were either *de novo* or segregated with ASD in the family under homozygous, compound heterozygous, or X-linked inheritance. We identified, on average, one *de novo* and three inherited homozygous rare coding variants, as well as four compound heterozygous variants in two genes, per affected individual (Table S5). We also identified, on average, five X-linked rare coding variants per affected male (Table S5). In addition, we found no change in the number of rare *de novo* variants with parental age, regardless of affection status (Figure S6). In total, we identified 156 genes that carried 284 rare potentially pathogenic coding variants (see STAR Methods for definition) (Table S6). Out of these genes, 16 are reported in the SFARI Gene database[29] and 49 are known disease genes, some of which are associated with neurodevelopmental phenotypes including intellectual disability, brain abnormalities, and epilepsy, as reported in the Online Mendelian Inheritance in Man (OMIM) database[34] and the Gene2Phenotype (G2P) Developmental Disorders (DD) panel[35] (Table S6).

Through homozygosity analysis, we mapped rare inherited homozygous variants to genomic runs of homozygosity (ROHs) in affected individuals. We identified an average of 169 rare inherited homozygous variants that were located within an ROH, the vast majority of which were noncoding variants (Table S5). Three variants were coding and potentially pathogenic (Table S6). Furthermore, to predict any functional impact for the rare noncoding inherited homozygous variants within ROHs, we annotated them using publicly available chromatin immunoprecipitation sequencing (ChIP-seq) and assay for transposase-accessible chromatin with sequencing (ATAC-seq) datasets,[36–38] as previously described[39] (see STAR Methods for details). We identified 229 and 240 variants located within promoter and enhancer regions, respectively, that are active in the human brain (Table S5). Of these, 39 variants (zero to four per genome) were within human brain-specific promoter or enhancer regions (Table S7).

The identified rare variants were further prioritized based on their predicted functional impact (see STAR Methods for details). Table 1 summarizes potentially pathogenic variants in known ASD or neurodevelopmental disease genes for each affected individual. We identified 45 variants in 18 affected individuals (approximately one to four variants per individual), including coding variants in 20 genes and noncoding variants in nine brain-specific regulatory regions. ASD and neurodevelopmental disease genes with potentially pathogenic coding variants included *CACNA1C* (ASD, Timothy syndrome, MIM: 601005),[40] *CDH23* (Usher syndrome, MIM: 601067),[41] *CHD7* (ASD, CHARGE syndrome, MIM: 214800),[42,43] *DLL1* (neurodevelopmental disease, MIM: 618709),[44] *FMR1* (ASD, fragile X syndrome, MIM: 300624),[45,46] *PCDHAC1* (ASD),[47,48] *TCF7L2* (*TCF7L2*-related neurodevelopmental disorder, ASD),[49,50] and *ZNF407* (SIMHA syndrome, MIM: 619557).[51] Noncoding regulatory variants included ones in the promotors of *CTCF*,[52] *MED13L*,[53] and *PTPN11*[54] (Table 1). In addition, there were 66 coding and 17 noncoding variants in 63 genes that have not been previously associated with ASD or other neurodevelopmental disorders, and they warrant further investigation and functional characterization (Table 2). Variants included frameshift deletions in *GLUD2* and *IDH3G* (Table 2), both of which encode mitochondrial enzymes with major roles in the tricarboxylic acid cycle, a glutamate dehydrogenase[55] and an isocitrate dehydrogenase,[56] respectively. Two of the noncoding variants occurred in promoters of *BDNF* and *NPAS4* (Table 2), two extensively studied genes with established roles in neuronal development and function.[57,58]

Genetic diversity and population history of the East African ASD cohort

The African continent, home to the oldest human populations, harbors the greatest genetic diversity.[15,59–61] This diversity can be represented by the percentage of the genome that is in a heterozygous state. We measured heterozygosity in the East African ASD cohort compared with populations from the 1000G[19] (Table S8). As expected, we found that the genetic variation is much higher in the East African cohort (0.12% heterozygosity) compared with all non-African populations (heterozygosity ranging from 0.08% to 0.09%) (Figure 1A). Heterozygosity across the genome in the East African cohort (0.12%) is slightly higher than that for all other six African populations analyzed (0.11%) (Figure 1A).

Using principal-component analysis (PCA) to explore the relationships between the East African ASD cohort and populations from the 1000G,[19] we found that the first principal component distinguished the East African cohort from non-African populations (Figure S7A). Our cohort clustered in an intermediate position between European and African populations with clear separation between our cohort and other African populations. Given that our cohort is predominantly Ethiopian and Eritrean, this finding is consistent with expectations since the Ethiopian population is known to have Eurasian admixture.[17,62] Five individuals from one family in our cohort are Kenyan, and these samples clustered together with other African groups. We then analyzed the cohort in relationship to other African groups using publicly available data, which included genotypes from the 1000G[19] African samples and additional African samples. Samples from 22 African countries were analyzed. The number of samples and the geographic regions of origin are listed in Table S9.[19,63–70] Samples from countries in Central Africa, East Africa, North Africa, South Africa, and West Africa clustered together in their respective regions (Figure S7B). Our ASD cohort clustered

Table 1. Potentially pathogenic variants in known ASD and neurodevelopmental disease genes identified in affected individuals from the East African ASD cohort

Affected individual	Inheritance	Variant(s)	Variant type	Gene(s)	Variant location	Mutation	Relevant OMIM or G2PDD phenotype	SFARI score	pLI score	LOEUF score
MCD-01-4	inherited homozygous (ROH)	chr7:91,321,289:C:A	SNV	*FZD1*, *MTERF1*, *AKAP9*	enhancer	–	–	2 (*AKAP9*)	0.04; 0; 0	0.60; 1.26; 0.40
MCD-02-3	X-linked	chrX:135,960,149:G:A	missense	*RBMX*	exonic	p.P105S	intellectual disability (XLR)	–	0.83	0.43

MCD-04-5	inherited homozygous (ROH)	chr12:116,591,972:C:G;
		chr12:116,592,155:T:A;
		chr12:116,714,156:T:C
			SNV	*MED13L*	enhancer, promoter	–	impaired intellectual development (AD)	1	1	0.06

MCD-04-5	inherited homozygous (ROH)	chr7:98,477,886:T:-	Indel	*TRRAP*	promoter	–	developmental delay (AD)	2S	1	0.06
MCD-05-3	compound heterozygous	chr17:78,078,662:G:A; chr17:78,083,769:C:G	missense	*GAA*	exonic	p.A93T; p.P451R	glycogen storage disease II (AR)	–	0	0.98
MCD-05-3	inherited homozygous	chr10:114,849,211:C:A	missense	*TCF7L2*	exonic	p.P179H	*TCF7L2*-related neurodevelopmental disorder	1	1	0.27
MCD-07-3	inherited homozygous (ROH)	chr16:71,062,908:A:C	SNV	*HYDIN*, *VAC14*	enhancer	–	ciliary dyskinesia (AR) (*HYDIN*); striatonigral degeneration (AR), progressive neurological disorder and regression (*VAC14*)	2 (*HYDIN*)	0; 0.19	0.51; 0.42
MCD-07-3	compound heterozygous	chr17:78,081,608:A:G; chr17:78,086,394:G:A	missense	*GAA*	exonic	p.N290D; p.R591Q	glycogen storage disease II (AR)	–	0	0.98
MCD-08-3*	*De novo*	chr6:170,597,575:T:C	missense	*DLL1*	exonic	p.E141G	neurodevelopmental disorder (AD)	2S	1	0.10
MCD-08-3*	*De novo*	chr8:61,757,960:C:A	missense	*CHD7*	exonic	p.H1734Q	CHARGE syndrome (AD)	1	1	0.08
MCD-08-3*	*De novo*	chr5:140,308,275:.:T	Frameshift	*PCDHAC1*	exonic	p.S601Lfs*4	–	2	0	1.02
MCD-08-3*	*De novo*	chr12:57,570,830:C:T	missense	*LRP1*	exonic	p.T1333I	–	2	1	0.06
MCD-13-3	inherited homozygous (ROH)	chr12:112,927,353:C:T	SNV	*PTPN11*	enhancer	–	LEOPARD syndrome (AD)	1	1	0.14
MCD-13-3	X-linked	chrX:77,245,178:A:T	missense	*ATP7A*	exonic	p.T354S	Menkes disease (XLR)	–	1	0.22
MCD-15-3	compound heterozygous	chr8:2,886,901:G:A; chr8:3,087,702:C:T	missense	*CSMD1*	exonic	p.L2599F; p.R1402H	–	2	1	0.21

(*Continued on next page*)

Table 1. ***Continued***

Affected individual	Inheritance	Variant(s)	Variant type	Gene(s)	Variant location	Mutation	Relevant OMIM or G2PDD phenotype	SFARI score	pLI score	LOEUF score
MCD-15-3	compound heterozygous	chr8:17,815,232:T:C; chr8:17,830,089:C:G	missense	*PCM1*	exonic	p.M663T; p.T1279R	–	2	0	0.58
MCD-16-3	compound heterozygous	chr12:2,794,928:G:A; chr12:2,797,853:C:T	missense	*CACNA1C*	exonic	p.R1875Q; p.R2017W	Timothy syndrome (AD)	1	1	0.10

MCD-17-3	compound heterozygous	chr5:89,923,448:G:T;
		chr5:89,949,452:A:T;
		chr5:90,368,293:T:C
			missense	*ADGRV1*	exonic	p.D365Y;
						p.N1354I;
						p.L6061S
							Usher syndrome (AR)	–	0	0.52

MCD-17-3	inherited homozygous (ROH)	chr1:154,842,199:-:GCTGCT	indel	*KCNN3*	promoter	–	Zimmermann-Laband syndrome 3 (AD)	–	0.97	0.32
MCD-18-3	compound heterozygous	chr1:215,901,460:C:T; chr1:215,933,087:G:T	missense	*USH2A*	exonic	p.G3993D; p.Q3716K	Usher syndrome (AR)	2	0	0.86
MCD-18-3	inherited homozygous	chr7:4,830,879:G:A	missense	*AP5Z1*	exonic	p.V763M	spastic paraplegia (AR)	–	0	1.47
MCD-19-3	compound heterozygous	chr18:50,432,527:A:G; chr18:51,013,227:C:T	missense	*DCC*	exonic	p.N176D; p.P1266L	developmental split-brain syndrome (AR); Mirror movements 1 (AD)	2	0.99	0.28
MCD-19-3	compound heterozygous	chr18:72,775,399:G:A; chr18:72,776,411:T:C	missense	*ZNF407*	exonic	p.D1908N; p.L2245P	SIMHA syndrome (AR)	–	1	0.09
MCD-19-3	inherited homozygous (ROH)	chr3:123,168,229:C:A	SNV	*ADCY5*	promoter	–	neurodevelopmental disorder with dyskinesia (AD or AR)	2	1	0.25
MCD-20-3, MCD-20-4	inherited homozygous (ROH)	chr16:67,596,146:G:A	SNV	*CTCF*	promoter	–	intellectual development disorder (AD)	1	1	0.15
MCD-22-3	inherited homozygous	chr10:73,485,206:C:T	missense	*CDH23*	exonic	p.R1170W	Usher syndrome (AR)	–	0	0.57
MCD-24-3	compound heterozygous	chr4:187,538,263:T:A; chr4:187,629,414:T:C	missense	*FAT1*	exonic	p.N2991Y; p.E523G	–	2	0	0.43
MCD-25-3	inherited homozygous (ROH)	chr2:25,140,835:G:A; chr2:25,142,282:C:T	SNV	*ADCY3*	promoter	–	–	2	0	0.68
MCD-33-4	compound heterozygous	chr5:150,886,883:G:T; chr5:150,901,082:C:T	missense	*FAT2*	exonic	p.P4117T; p.G3691E	spinocerebellar ataxia (AD)	–	0	0.51
MCD-33-4	X-linked	chrX:147,027,118:A:G	missense	*FMR1*	exonic	p.Q462R	fragile X syndrome (XLR)	1	0.65	0.42

List of deleterious coding and brain-specific regulatory noncoding variants affecting known ASD or neurodevelopmental disease genes identified for each affected individual. ROH indicates inherited homozygous variants that are within runs of homozygosity. For SFARI score, S denotes syndromic genes. AD, autosomal dominant; AR, autosomal recessive; indel, insertion or deletion; LOEUF, loss-of-function observed/expected upper bound fraction; SNV, single nucleotide variant; XLR, X-linked recessive. *Sample with a missing parent sample where compound heterozygous variant calling was not possible and *de novo*, inherited homozygous, and X-linked variant calling relied on one parent only.

Table 2. Potentially pathogenic variants in novel candidate ASD genes identified in affected individuals from the East African ASD cohort

Affected individual	Inheritance	Variant(s)	Variant type	Gene(s)	Variant location	Mutation	pLI score	LOEUF score
MC-35-3	compound heterozygous	chr1:3,329,057:G:A; chr1:3,331,149:G:T	missense	*PRDM16*	exonic	p.G766S; p.D877Y	1	0.19
MC-35-3	X-linked	chrX:9,863,880:C:G	missense	*SHROOM2*	exonic	p.S644R	0	0.53
MCD-01-3	inherited homozygous (ROH)	chr11:27,743,920:C:A	SNV	*BDNF*	promoter	–	0.66	1.52
MCD-02-3	X-linked	chrX:112,022,894:C:T	missense	*AMOT*	exonic	p.G830R	1	0.27
MCD-02-3	X-linked	chrX:153,051,602: GGCTACAGGA:-	Frameshift	*IDH3G*	exonic	p.L379Pfs*91	0.09	0.71
MCD-04-5	compound heterozygous	chr2:108,924,874:G:C; chr2:108,910,151:C:T	missense; Stop gain	*SULT1C2*	exonic	p.R282T; p.Q10X	0	1.34
MCD-04-5	X-linked	chrX:67,731,809:G:A	missense	*YIPF6*	exonic	p.R59H	0.90	0.41
MCD-05-3	compound heterozygous	chr11:92,577,302:C:T; chr11:92,533,549:G:A	missense	*FAT3*	exonic	p.S3590L; p.R2457Q	1	0.25

MCD-06-3*	inherited homozygous (ROH)	chr15:79,143,366:A:T	SNV	*ADAMTS7*,
				MORF4L1,
				CTSH
					enhancer	–	0; 0.99;
							0
								0.65;
								0.27;
								1.01

MCD-07-3	inherited homozygous (ROH)	chr10:16,392,185:G:A	SNV	*MINDY3*, *PTER*	enhancer	–	0.01; 0	0.054; 1.82
MCD-07-3	inherited homozygous (ROH)	chr16:1,040,862:G:A	SNV	*SOX8*, *SSTR5-AS1*, *SSTR5*	enhancer	–	0.67; –; 0	0.51; –; 1.59

MCD-08-3*	inherited homozygous (ROH)	chr6:96,506,344:-: AAAAA	Indel	*FUT9*, *MANEA*,
				UFL1
					enhancer	–	0.08;
							0; 0
								0.71;
								0.86;
								0.84

MCD-10-3	compound heterozygous	chr3:97,686,151:T:C; chr3:97,677,992:T:C	missense	*RIOX2*	exonic	p.Y96C; p.H195R	0	1.07
MCD-11-3	inherited homozygous (ROH)	chr4:15,937,843:A:G	missense	*FGFBP1*	exonic	p.V138A	0.59	1.06
MCD-11-3	X-linked	chrX:101,909,395:A:G	missense	*GPRASP1*	exonic	p.E185G	0.31	0.42
MCD-13-3	compound heterozygous	chr6:75,893,146:A:G; chr6:75,838,134:A:G	missense	*COL12A1*	exonic	p.I504T; p.V2073A	0.97	0.28
MCD-13-3	compound heterozygous	chr15:59,373,446:A:G; chr15:59,359,263:G:C	missense	*RNF111*	exonic	p.M754V; p.G556A	1	0.21
MCD-14-3*	X-linked	chrX:9,693,868:C:T	missense	*GPR143*	exonic	p.S378N	0.93	0.37
MCD-14-3*	X-linked	chrX:153,035,691:C:T	missense	*PLXNB3*	exonic	p.P592S	0.23	0.37
MCD-14-3*	*De novo*	chr8:144,688,700:A:T	stop gain	*PYCR3*	exonic	p.C154X	0	1.43
MCD-15-3	compound heterozygous	chr22:50,315,388:G:A; chr22:50,315,966:C:G	missense	*CRELD2*	exonic	p.E191K; p.S205C	0	1.05

(*Continued on next page*)

Table 2. ***Continued***

Affected individual	Inheritance	Variant(s)	Variant type	Gene(s)	Variant location	Mutation	pLI score	LOEUF score
MCD-15-3	X-linked	chrX:71,873,343:A:G	missense	*PHKA1*	exonic	p.I360T	0	0.55
MCD-16-3	X-linked	chrX:110,491,920:C:T	missense	*CAPN6*	exonic	p.R454H	0.82	0.39
MCD-16-3	*De novo*	chr5:31,317,594:A:T	missense	*CDH6*	exonic	p.N542I	0.92	0.34
MCD-16-3	X-linked	chrX:83,128,919:G:T	missense	*CYLC1*	exonic	p.K401N	0.93	0.37
MCD-16-3	compound heterozygous	chr13:76,414,552:C:T; chr13:76,409,449:G:T	missense	*LMO7*	exonic	p.P935L; p.D870Y	0	0.41
MCD-17-3	X-linked	chrX:120,182,861:GT:.	frameshift	*GLUD2*	exonic	p.W442Afs*16	0.03	1.27
MCD-17-3	inherited homozygous (ROH)	chr10:25,463,474:G:C	SNV	*GPR158-AS1*	promoter	–	–	–
MCD-17-3	X-linked	chrX:17,820,025:C:T	missense	*RAI2*	exonic	p.E36K	0.61	0.66
MCD-17-3	X-linked	chrX:153,714,877:T:C	missense	*UBL4A*	exonic	p.Q16R	0.30	1.18
MCD-18-3	compound heterozygous	chr5:74,364,457:G:A; chr5:74,532,495:T:-	missense; Frameshift	*ANKRD31*	exonic	p.R1837C; p.Q6Rfs*10	0	0.87
MCD-18-3	X-linked	chrX:149,680,360:C:T	missense	*MAMLD1*	exonic	p.L672F	0.63	0.45
MCD-18-3	compound heterozygous	chr12:110,943,480:C:T; chr12:110,952,911:C:T	missense	*RAD9B*	exonic	p.P59L; p.A110V	0	1.37
MCD-19-3	compound heterozygous	chr2:209,210,794:A:G; chr2:209,190,942:C:G	missense	*PIKFYVE*	exonic	p.K1711R; p.T1136S	0	0.39
MCD-21-3	compound heterozygous	chr3:111,603,790:A:G; chr3:111,603,963:T:A	missense	*PHLDB2*	exonic	p.K289R; p.S347T	0	0.59
MCD-21-3	inherited homozygous (ROH)	chr15:55,881,474:TTTTTG:-	indel	*PYGO1*	promoter	–	0.99	0.23
MCD-22-3	compound heterozygous	chr5:138,857,917:C:T; chr5:138,858,039:C:A	missense	*TMEM173*	exonic	p.A233T; p.G192V	0	0.90
MCD-23-3	inherited homozygous (ROH)	chr8:28,166,424:-: TGTGTGTGTGTGTGTGT	indel	*ELP3*, *PNOC*	enhancer	–	0; 0.02	1.02; 1.10
MCD-24-3	compound heterozygous	chr3:52,555,908:G:A; chr3:52,557,482:G:A	missense	*STAB1*	exonic	p.R2071H; p.A2394T	0	0.81
MCD-24-3	compound heterozygous	chr15:54,792,341:C:T; chr15:54,919,033:G:A	missense	*UNC13C*	exonic	p.H1709Y; p.G2123R	0	0.54
MCD-24-3	inherited homozygous (ROH)	chr20:62,406,258:G:A	SNV	*ZBTB46*	promoter	–	0.81	0.40
MCD-25-3	compound heterozygous	chr1:24,389,694:C:T; chr1:24,387,787:G:C	missense	*MYOM3*	exonic	p.G1231E; p.A1316G	0	1.11

MCD-25-3	inherited homozygous (ROH)	chr7:88,387,647:T:G;
		chr7:88,387,990:T:C;
		chr7:88,388,566:G:A
			SNV	*ZNF804B*	promoter	–	0	1

MCD-26-3	compound heterozygous	chr19:55,748,036:G:A; chr19:55,742,199:G:C	missense	*PPP6R1*	exonic	p.R655C; p.P838R	0.98	0.30

(*Continued on next page*)

Table 2. ***Continued***

Affected individual	Inheritance	Variant(s)	Variant type	Gene(s)	Variant location	Mutation	pLI score	LOEUF score
MCD-27-3	compound heterozygous	chr15:68,609,616:C:T; chr15:68,624,693:C:T	missense	*ITGA11*	exonic	p.R901Q; p.V517I	0	0.55
MCD-27-3	inherited homozygous	chr11:74,570,283:C:T; chr11:74,638,465:C:T	missense	*XRRA1*	exonic	p.V356M; p.V157M	0	0.78
MCD-28-3	compound heterozygous	chr14:105,414,461:C:G; chr14:105,416,755:T:G	missense	*AHNAK2*	exonic	p.E2343Q; p.E1578A	0	1.01
MCD-28-3	inherited homozygous	chr4:101,331,507:G:A	missense	*EMCN*	exonic	p.H240Y	0	0.96
MCD-28-3	compound heterozygous	chr20:20,493,587:C:A; chr20:20,552,253:G:A	missense	*RALGAPA2*	exonic	p.V1476F; p.A1002V	0	0.59

MCD-29-3*	inherited homozygous (ROH)	chr8:132,048,583:C:T;
		chr8:132,052,935:G:C;
		chr8:132,053,736:C:T
			SNV	*ADCY8*	promoter	–	0	0.54

MCD-29-3*	inherited homozygous (ROH)	chr3:159,560,791: TTTTTTTTTTTTT:-	Indel	*IQCJ-SCHIP1*, *SCHIP1*	promoter	–	0.03; 0.99	0.53; 0.23
MCD-30-3*	inherited homozygous (ROH)	chr11:66,190,893:T:G	SNV	*NPAS4*	promoter	–	0.97	0.32
MCD-32-4	X-linked	chrX:77,378,818:A:C	missense	*PGK1*	exonic	p.N295H	0.77	0.47
MCD-32-4	X-linked	chrX:131,205,199:G:A	missense	*STK26*	exonic	p.A234T	0.31	0.54
MCD-32-4	X-linked	chrX:37,931,320:G:C	missense	*SYTL5*	exonic	p.G117A	0	0.65

MCD-33-3;
MCD-33-4
	compound heterozygous	chr12:58,220,823:C:T;
		chr12:58,220,841:C:T;
		chr12:58,220,831:C:G
			missense	*CTDSP2*	exonic	p.V104M;
						p.D98N;
						p.R101T
							0.11	0.67

List of deleterious coding and brain-specific regulatory noncoding variants affecting novel candidate ASD genes identified for each affected individual. ROH indicates inherited homozygous variants that are within runs of homozygosity. Indel, insertion or deletion; LOEUF, loss-of-function observed/expected upper bound fraction; SNV, single nucleotide variant. * Samples with a missing parent sample where compound heterozygous variant calling was not possible and *de novo*, inherited homozygous, and X-linked variant calling relied on one parent only.

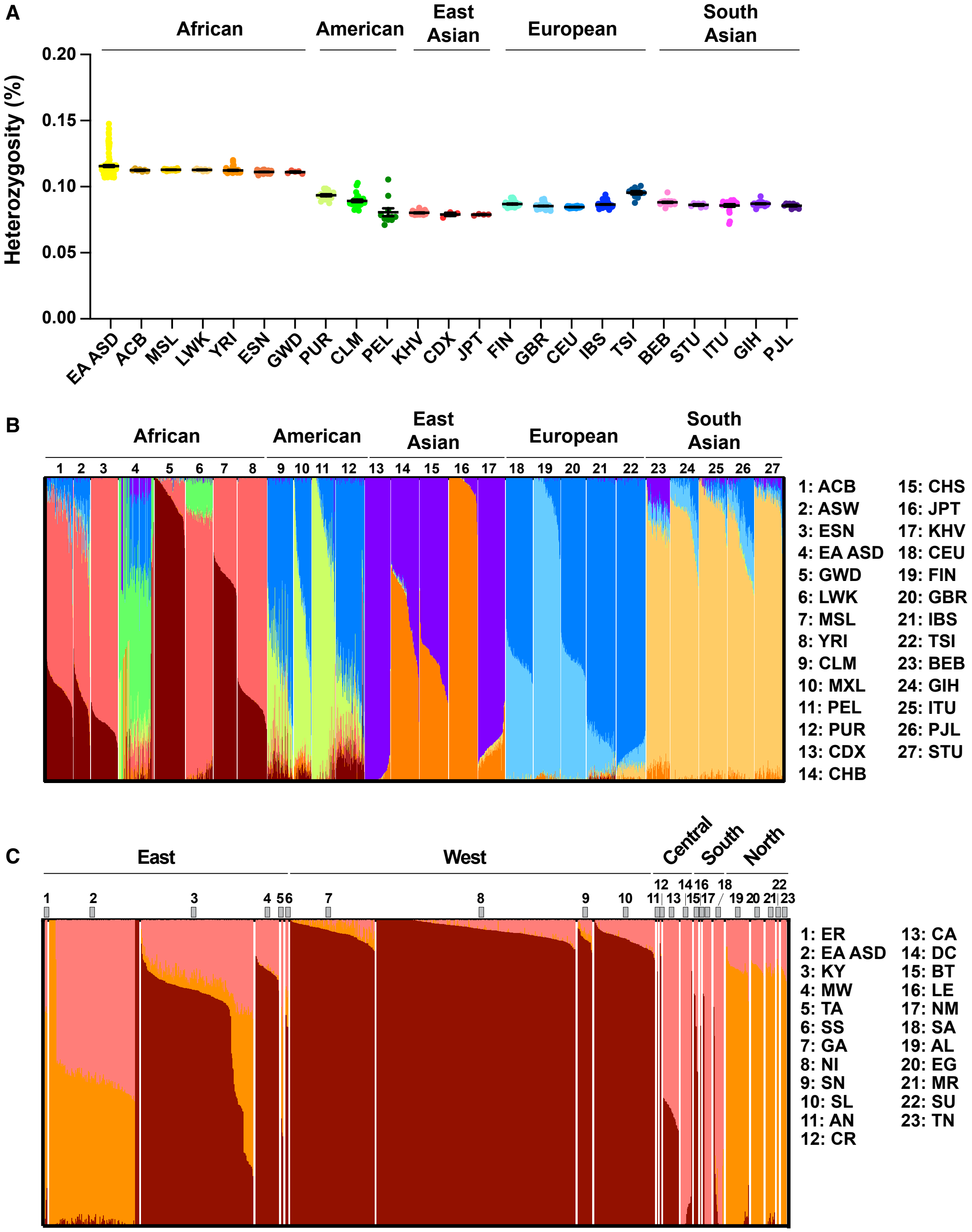


Figure 1. Heterozygosity and population structure of the East African ASD cohort

(A) Heterozygosity across the human genome calculated in the East African ASD (EA ASD) cohort and the 1000 Genomes project populations from all continents. The East African genome is enriched for heterozygous variants compared with non-African genomes. Mean ± SEM are shown in black. Population abbreviations are defined in Table S8.

(legend continued on next page)

away from other groups on the first principal component in a position adjacent to other East African groups, particularly Kenya, and intermediate between groups from West Africa and North Africa (Figure S7B). While more sequencing on the African continent is needed to capture regional genetic diversity, the separation of clusters from various geographic regions supports our observation that our predominantly Ethiopian and Eritrean cohort is distinct from other African groups.

Population admixture in the East African ASD cohort

East African populations, and in particular the Ethiopians, have a proportion of Eurasian ancestry.[18,62] We carried out global admixture analysis first on our cohort alone to determine the number of presumed admixed populations and then analyzed our cohort in combination with data from the 1000G[19] to determine which modern-day populations from the 1000G[19] (Table S8) are most closely related to the admixed ancestral populations in our cohort. We ran the analysis using values of K from 2 to 20, where K is the number of presumed ancestral populations (Figures 1B and S8). For the East African ASD cohort, K = 2 yielded the smallest value for cross-validation error indicating two admixed ancestral populations in our cohort (Figure S9A). For the dataset combined with the 1000G,[19] K = 9 yielded the lowest cross-validation error, indicating nine presumed ancestral populations in the combined dataset (Figure S9B). Global admixture analysis with nine presumed ancestral populations in the combined East African ASD cohort and 1000G[19] dataset indicated that the modern-day reference populations most closely related to the ancestral admixed populations in our cohort are the Luhya in Webuye, Kenya (LWK) and the Toscani in Italy (TSI) (Figure 1B; Table S10).

The proportion of ancestry from each presumed population was determined for every sample and averaged across the cohort (Table S10). For samples in the East African ASD cohort, the predominant ancestry was related to population 6 (50.35%), which is most closely related to the LWK group, and population 8 (16.58%), which is most closely related to the TSI group (Table S10). Consistent with our PCA analysis, which showed the ASD cohort as being distinct from other African groups, population six, which represents the most closely related African population to our cohort, is not the predominant population in the LWK group, representing 11.26% of the LWK ancestry (Table S10). The LWK are also admixed and show a higher proportion of admixture related to other African groups (Table S10). In line with our PCA results, the East African ASD cohort clusters most closely with the LWK group but the LWK group clusters more closely with other African groups. Thus, the East African ASD cohort is distinct from other African groups, with the LWK being the most closely related modern-day population.

We carried out the same admixture analysis on a combined dataset of the ASD cohort and samples from the 22 African groups (Table S9). The lowest cross-validation error was observed at K = 3 (Figure S9C). Despite the small number of samples from some of the regions (Table S9), which affected the ability to discern admixture proportions, the analysis showed that the East African ASD cohort is distinct from other African groups (Figures 1C and S10).

Genomic variants within differential ancestry peaks

To leverage the full information from WGS with our local admixture analysis, we calculated the number of alleles in the entire cohort that mapped to the LWK reference population across all markers. We expect that risk alleles would appear in regions enriched for African ancestry given the high prevalence of ASD in the East African population compared with the prevalence in the US and Europe. Since by far the predominant ancestry in the cohort mapped to the LWK reference, it was not possible to discern regions with overrepresentation of African ancestry as most of the genome for all individuals was 100% LWK reference. We then looked for genomic regions that have differential ancestry representation between affected and unaffected individuals (Figures 2 and S11). Given that transmitted ancestral segments can vary within a family as a result of recombination, we took a familial analysis approach in our cohort. We compared the percentage of LWK alleles at each position in affected versus unaffected individuals (Table S11A). We defined admixture peaks as regions where the difference in the percentage of LWK mapped alleles (absolute value of delta [abs(Δ)]) was ≥15% (Figure 2; Table S11B). We extracted variants mapping to these differential ancestry peaks and found an average of 13,575 variants per individual, 4,364 of which were genic (Table S11C). We restricted our analysis to rare variants with an MAF < 1% in all annotated populations (1000G,[19] gnomAD,[20] GME,[21] and H3Africa[22,23]) and filtered for variants that segregate with phenotype within families (Table S11D). We identified a total of 205 variants, the majority affecting noncoding regions, one nonframeshift insertion, and five nonsynonymous SNVs resulting in missense mutations (Table S11D). These five variants occurred in two genes, *CFAP46* and *ZFHX3*, that have not been implicated in ASD or other neurodevelopmental disorders (Table S11D). *CFAP46* encodes a cilia- and flagella-associated protein with high expression in the brain, and *ZFHX3* encodes a transcription factor with roles in the circadian system.[71,72] One of the noncoding variants mapped to an intron of *TUBGCP2*, which encodes a component of the microtubule-organizing center. Mutations in *TUBGCP2* result in neurodevelopmental phenotypes, including pachygyria, microcephaly, and developmental delay (MIM: 618737).[73] In addition, we did not find any variants that were specific to our cohort and common to the entire cohort that could be less than fully penetrant and acting to sensitize the genetic background.

(B) Global admixture analysis in the East African ASD cohort and populations from the 1000 Genomes project with K = 9 indicated two admixed populations in the ASD cohort. The most closely related groups to the ancestral groups for the ASD cohort are the Luhya in Webuye, Kenya (LWK) and the Toscani in Italy (TSI). Population abbreviations are defined in Table S8.

(C) Global admixture analysis in the ASD cohort and African populations (listed in Table S9) with K = 3. While the LWK population is the most closely available African reference population, the East African ASD cohort shows distinct ancestry that clusters away from other African populations. Population abbreviations are defined in Table S9.

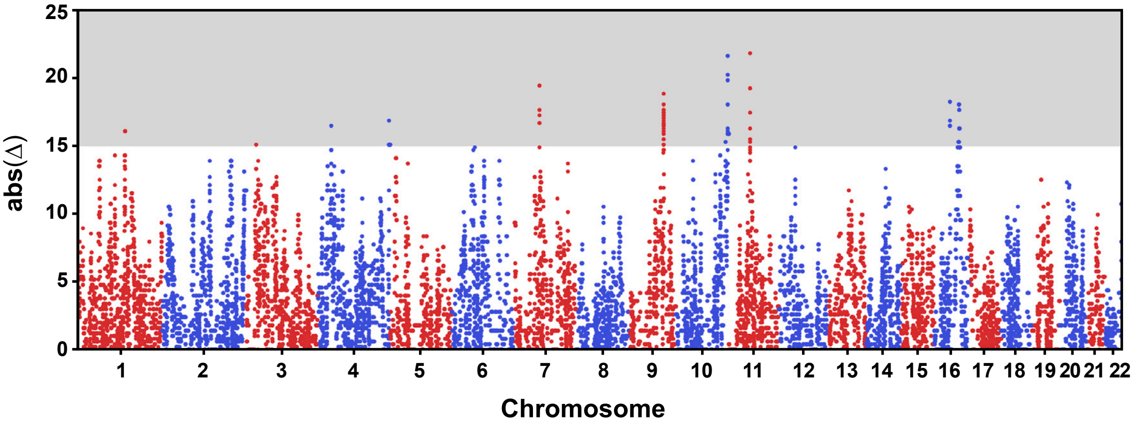


Figure 2. Differential ancestry peaks between affected and unaffected individuals across the genome

Haplotypes were assigned to either LWK or TSI ancestry, and local ancestry analysis was performed to identify genomic regions that exhibit differential LWK/TSI ancestry between affected and unaffected individuals. The absolute delta (abs(Δ)) represents the absolute value of the difference in percentage alleles of LWK ancestry comparing affected and unaffected individuals. Variants under admixture peaks with delta ≥ 15% (shaded area) are presented in Table S11. Data points are colored in red and blue for alternating chromosomes.

Local admixture mapping identifies population-specific ASD risk variants

We sought to investigate whether population-specific variants were preferentially transmitted to affected offspring in our cohort. We mapped alleles to either the LWK or the TSI reference populations and generated haplotype segments that are assigned to one of the two reference populations for each individual across all chromosomes. These haplotypes were used to construct segments with markers that were assigned a value of 0, 1, or 2 to indicate the count of LWK-assigned alleles at each position. The encoded alleles at each marker position were treated as genotypes and analyzed using the transmission disequilibrium test (TDT) in our cohort of trio and quad families. A quantile-quantile (Q-Q) plot showed that the data are normally distributed (Figure S12). Correcting for multiple testing using the number of ancestry haplotype switches (Table S12), the Bonferroni corrected significance threshold was $p = 3.32 \times 10^{-4}$. We identified 755 genomic regions with differential transmission of alleles that met this threshold ($p < 3.32 \times 10^{-4}$) (Table S13). Out of these regions, there were 10 that had the highest TDTae statistic (LRT >65), and where the LWK-assigned alleles were more often transmitted to affected offspring (Table 3). This suggests that the LWK-assigned haplotypes for these regions, which are more often transmitted to affected offspring, confer ASD risk. The 10 regions included several known ASD and neurodevelopmental disease genes, including *ADSL*, *CREBBP*, *EHMT1*, and *GRIN1* (Table 3). Of note, one of these regions located on chromosome 8 (chr8: 32,045,867-35,209,816) has an odds ratio (OR) = 11.31 (Tables 3 and S13). The region contains 10 genes, two of which, *TTI2* and *UNC5D*, have been previously reported in neurodevelopmental disease, including ASD and intellectual disability. *TTI2* encodes Telo2-Interacting Protein 2, which functions in DNA damage response and is a member of the Triple T complex, which regulates telomere length and the abundance of phosphoinositide 3-kinase-related protein kinases, key signaling molecules in brain development and function.[74] Pathogenic recessive variants in *TTI2* cause a developmental disorder characterized by intellectual disability and severe speech delay (MIM: 615541).[48,74,75] Variants in the regulatory region upstream of *UNC5D* have been identified in ASD, although the gene itself has not been strongly linked to the disorder (SFARI Gene score of 3).[76–78] UNC5D is a receptor for the axon guidance molecule Netrin 1.[79] Other genes in the region include *NRG1*, which has been associated with schizophrenia risk[80,81] and encodes for the signaling protein Neuregulin 1 with critical roles in cortical development, function, and plasticity.[82]

DISCUSSION

We performed WGS in a cohort of 33 families of Ethiopian, Eritrean, and Kenyan descent. We discovered over 2.1 million private variants in 129 individuals that have not been previously reported. This is almost twice the number of variants per individual compared with a recent report that identified 3.4 million novel variants from WGS of 426 African individuals.[23] By prioritizing rare variants that were either *de novo* or segregated with ASD in the family under different modes of inheritance, we identified potentially causative variants in known ASD genes, neurodevelopmental disorder genes, as well as genes not previously associated with any disorder. The known genes included high-confidence ASD genes *CACNA1C*, *CHD7*, *FMR1*, and *TCF7L2*, among other neurodevelopmental disease genes (e.g., *CDH23*, *DLL1*, *RBMX*). The missense variant identified in *FMR1* maps to a region on the encoded protein FMRPrequired for nuclear export[83,84] and for its interaction with RANBP9,[85] a protein involved in many cellular processes, including nuclear trafficking and microtubule nucleation,[86,87] and potentially modulating FMRP RNA-binding properties.[85] In addition, we identified noncoding variants in regulatory elements of known ASD genes (*CTCF*, *MED13L*, and *PTPN11*) (Table 1). Sequencing studies in larger cohorts and additional experimental validation will be required to establish causality

Table 3. Genomic regions where the African-assigned LWK alleles confer ASD risk

Region number	Start marker	End marker	Chromosome	Start	End	Size (Mb)	Genes	Known ASD gene(s)	SFARI score	Known neurodevelopmental disease gene(s)
1	10773	10905	2	153,362,927	155,354,991	1.99	6	*GALNT13*	2	–
2	4074	4167	9	73,274,607	74,655,760	1.38	5	–	–	–
3	590	978	11	5,199,208	8,602,452	3.40	98	*APBB1*	2	*DCHS1*, *EIF3F*
4	6177	6239	12	84,038,708	85,090,747	1.05	–	–	–	–
5	2236	2384	22	38,038,060	40,816,749	2.78	68	*ADSL*, *CACNA1I*, *CSNK1E*, *SGSM3*, *TNRC6B*	1; 2; 2; 2; 2	*PDGFB*, *PLA2G6*, *SOX10*
6	2823	2998	14	54,334,781	57,025,840	2.69	23	–	–	–
7	298	320	16	3,747,204	4,045,272	0.30	4	*CREBBP*	1	–
8	8808	8816	1	142,685,709	142,705,564	0.02	1	–	–	–
9	8973	9185	9	138,862,766	141,112,375	2.25	104	*CACNA1B*, *EHMT1*, *GRIN1*, *PNPLA7*	2; 1; 1; 2	*ABCA2*, *CACNA1B*, *INPP5E*, *MAN1B1*, *PMPCA*
10	3504	3686	8	32,045,867	35,209,816	3.16	10	*TTI2*, *UNC5D*	S; 3	–

TDT using alleles derived from local admixture analysis encoding African (LWK) or European (TSI) ancestry identified regions of the genome in which the African-assigned LWK alleles were transmitted to affected offspring more frequently than the European-assigned TSI alleles (Bonferroni corrected $p < 3.32 \times 10^{-4}$). The top 10 regions with TDTae statistic (LRT) >65 are shown. For SFARI score, S denotes syndromic genes.

for the candidate genes that have not been previously linked to disease (Table 2).

While individuals from other regions in Africa were sequenced as part of the 1000G,[19] no samples from Ethiopia were included in either the 1000G[19] or the HapMap projects.[19,88] Previous work genotyping individuals from Ethiopia has detected admixture with a significant Eurasian ancestry contribution that is estimated to result from genetic backflow into Africa, possibly from the Levant region approximately 3,000 years ago.[17,18] East Africa, including Ethiopia, is an important region in the study of human migration where ancient radiation events may have contributed to the genetic makeup of other groups across the continent[63] and more recent Eurasian admixture is likely to have contributed to the Eurasian ancestry component in southern Africa.[17] Since East African populations, and in particular the Ethiopians, have a proportion of Eurasian ancestry,[18,62] we analyzed the WGS data to determine genetic admixture in our cohort. We then performed admixture mapping, which is a method for identifying disease-causing variants in populations where the disease risk varies by ancestry. It assumes that the frequency of the causal variant(s) differs between ancestral populations.[89,90] Since the prevalence of ASD is higher in East African populations than in other populations, we hypothesized that individuals from our cohort may harbor unique ASD risk variants that are not found in other populations. Although we hypothesized that this increased prevalence is due to inherited genetic factors, we expect that the East African population is still subject to the background factors driving baseline prevalence of ASD in all populations, including *de novo* mutation. To investigate an inherited genetic component to ASD in our cohort, we performed TDT analysis using markers coded for local admixture ancestry. We identified 10 loci on chromosomes 1, 2, 8, 9, 11, 12, 14, 16, and 22 that are preferentially transmitted to affected offspring (Table 3). Several genes in these regions are known ASD and neurodevelopmental disease genes (e.g., *ADSL*, *CREBBP*, *EHMT1*, and *GRIN1*) (Table 3). The finding indicates that the alleles assigned to the LWK African reference population confer risk, consistent with the idea that admixture mapping can identify risk alleles that will be derived from the group with higher prevalence for a phenotype. The LWK group is a modern-day proxy for the ancestral admixed African genetic contribution, which we cannot sample directly. Although prevalence studies have not been carried out in Kenya, several studies point to an increased prevalence of ASD in children born to East African parents,[8–11] consistent with our finding that the LWK-assigned segments confer risk.

Most of the diversity present in the human genome remains on the continent of Africa, since populations that migrated out of Africa over 50,000 years ago were comparatively small and carried only a subset of the standing variation of the parent populations.[60,61] The identification of variants in African populations will lead to an understanding of common biological processes that affect health and disease in all populations.[25] Perhaps most importantly, as genetics findings make their way into the clinic, equity in healthcare depends on inclusion of all human populations.[12–14] Current databases are biased toward populations of European ancestry,[91] giving an incomplete picture of human diversity and inevitably failing to capture population-specific variants and genetic backgrounds that are medically relevant. Multiple efforts have been undertaken and are currently underway to sequence individuals from the thousands of diverse populations in Africa in order to provide a more complete understanding of human genetic diversity.[12,16,22,92] Our study contributes to unraveling the genetic underpinnings of ASD in an African population.

Limitations of the study

A key starting point in our investigation is the increased prevalence of ASD in children of East African ancestry. Another major driver for our study is the diversity gap in existing genetic studies.

Thus, we set out to investigate the genetic underpinnings of this increased prevalence and to contribute to the ongoing efforts to bridge the diversity gap. While an inherent limitation to our study is the small size of the cohort we analyzed, we were able to leverage the power of African genomics to inform a complex disorder. In future efforts, a larger sample size will enable further analyses, including better cataloging of structural variation. In addition, we grappled with the complete lack of ASD prevalence studies in the East African countries represented by our study participants (Ethiopia, Eritrea, and Kenya). We therefore relied on data from prevalence studies in immigrant populations, and we discuss the potential limitation of this approach in the section "introduction." The current work further emphasizes the need for ASD genetics and prevalence studies in Africa, and it contributes to the growing efforts aimed at understanding the genetics of ASD and other neurodevelopmental disorders in Africa.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Subjects and specimens
  - Clinical information
- METHOD DETAILS
  - Whole genome sequencing and data processing
  - Variant filtration
  - Noncoding variant annotation
  - Variant prioritization
  - Burden analysis
  - Copy number variant (CNV) analysis
  - Assessment of runs of homozygosity
  - Heterozygosity
  - Principal component analysis
  - Admixture analysis

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xgen.2023.100322.

ACKNOWLEDGMENTS

We are grateful to the families for their participation in our study. We acknowledge the role of the Human Heredity and Health in Africa (H3Africa) Consortium in making this research possible through the sharing of data. This publication is the work of the authors and does not represent the views of the H3Africa Consortium or their funders. We thank Dr. Derek Gordon for help and support with the TDTae program, Ms. Chelsea Burroughs for assistance in preparing the figures, Ms. Emma Bergman for assistance in preparing the graphical abstract, and Ms. Naviya Ravi for assistance in database searches. This work was supported by a generous gift from LaVerne M. Dutton, the University of Texas Southwestern Medical Center, and the Walter and Lillian Cantor Foundation to M.H.C.

AUTHOR CONTRIBUTIONS

M.H.C. conceived the study, designed the experiments, and oversaw the project. I.O.T., D.D., A.G., K.K., and M.H.C. performed experiments and analyzed data. K.G. reviewed clinical data. A.K. performed and C.X. supervised the CNV analysis. L.S.-T. referred subjects. M.H.C. wrote the manuscript. All authors participated in reviewing and editing of the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

INCLUSION AND DIVERSITY

One or more of the authors of this paper self-identifies as an underrepresented ethnic minority in their field of research or within their geographical location. One or more of the authors of this paper self-identifies as a gender minority in their field of research. One or more of the authors of this paper self-identifies as a member of the LGBTQIA+ community.

Received: November 16, 2022
Revised: March 9, 2023
Accepted: April 16, 2023
Published: May 11, 2023

REFERENCES

1. Colvert, E., Tick, B., McEwen, F., Stewart, C., Curran, S.R., Woodhouse, E., Gillan, N., Hallett, V., Lietz, S., Garnett, T., et al. (2015). Heritability of autism spectrum disorder in a UK population-based twin sample. JAMA Psychiatr. *72*, 415–423. https://doi.org/10.1001/jamapsychiatry.2014.3028.
2. Sandin, S., Lichtenstein, P., Kuja-Halkola, R., Hultman, C., Larsson, H., and Reichenberg, A. (2017). The heritability of autism spectrum disorder. JAMA *318*, 1182–1184. https://doi.org/10.1001/jama.2017.12141.
3. Tick, B., Bolton, P., Happé, F., Rutter, M., and Rijsdijk, F. (2016). Heritability of autism spectrum disorders: a meta-analysis of twin studies. J. Child Psychol. Psychiatry *57*, 585–595. https://doi.org/10.1111/jcpp.12499.
4. Betancur, C. (2011). Etiological heterogeneity in autism spectrum disorders: more than 100 genetic and genomic disorders and still counting. Brain Res. *1380*, 42–77. https://doi.org/10.1016/j.brainres.2010.11.078.
5. Geschwind, D.H., and State, M.W. (2015). Gene hunting in autism spectrum disorder: on the path to precision medicine. Lancet Neurol. *14*, 1109–1120. https://doi.org/10.1016/S1474-4422(15)00044-7.
6. de la Torre-Ubieta, L., Won, H., Stein, J.L., and Geschwind, D.H. (2016). Advancing the understanding of autism disease mechanisms through genetics. Nat. Med. *22*, 345–361. https://doi.org/10.1038/nm.4071.
7. Maenner, M.J., Shaw, K.A., Bakian, A.V., Bilder, D.A., Durkin, M.S., Esler, A., Furnier, S.M., Hallas, L., Hall-Lande, J., Hudson, A., et al. (2021). Prevalence and characteristics of autism spectrum disorder among children aged 8 Years - autism and developmental disabilities monitoring network, 11 sites, United States, 2018. MMWR. Surveill. Summ. *70*, 1–16. https://doi.org/10.15585/mmwr.ss7011a1.
8. Barnevik-Olsson, M., Gillberg, C., and Fernell, E. (2008). Prevalence of autism in children born to Somali parents living in Sweden: a brief report. Dev. Med. Child Neurol. *50*, 598–601. https://doi.org/10.1111/j.1469-8749.2008.03036.x.
9. Magnusson, C., Rai, D., Goodman, A., Lundberg, M., Idring, S., Svensson, A., Koupil, I., Serlachius, E., and Dalman, C. (2012). Migration and autism spectrum disorder: population-based study. Br. J. Psychiatry *201*, 109–115. https://doi.org/10.1192/bjp.bp.111.095125.
10. (2018). The Minnesota-autism and developmental disabilities monitoring network. https://addm.umn.edu/.

11. Seyoum-Tesfa, L., Bowyer, P., Dickerson, A., and Jackson, P. (2015). Prevalence of autism in children of East African descent in Texas. In Lonestar LEND Conference, the LoneStar LEND is a Collaboration Between the University of Texas HSC, University of Houston, Texas Woman's University, Baylor College of Medicine, and MHMRA of Harris County.

12. Lumaka, A., Carstens, N., Devriendt, K., Krause, A., Kulohoma, B., Kumuthini, J., Mubungu, G., Mukisa, J., Nel, M., Olanrewaju, T.O., et al. (2022). Increasing African genomic data generation and sharing to resolve rare and undiagnosed diseases in Africa: a call-to-action by the H3Africa rare diseases working group. Orphanet J. Rare Dis. *17*, 230. https://doi.org/10.1186/s13023-022-02391-w.

13. Hanchard, N.A., Chahrour, M., and de Vries, J. (2022). Tailored community engagement to address the genetics diversity gap. Med *3*, 369–370. https://doi.org/10.1016/j.medj.2022.05.010.

14. Martin, A.R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B.M., and Daly, M.J. (2019). Clinical use of current polygenic risk scores may exacerbate health disparities. Nat. Genet. *51*, 584–591. https://doi.org/10.1038/s41588-019-0379-x.

15. The 1000 Genomes Project Consortium; Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56–65. https://doi.org/10.1038/nature11632.

16. Gurdasani, D., Carstensen, T., Tekola-Ayele, F., Pagani, L., Tachmazidou, I., Hatzikotoulas, K., Karthikeyan, S., Iles, L., Pollard, M.O., Choudhury, A., et al. (2015). The African genome variation project shapes medical genetics in Africa. Nature *517*, 327–332. https://doi.org/10.1038/nature13997.

17. Pagani, L., Kivisild, T., Tarekegn, A., Ekong, R., Plaster, C., Gallego Romero, I., Ayub, Q., Mehdi, S.Q., Thomas, M.G., Luiselli, D., et al. (2012). Ethiopian genetic diversity reveals linguistic stratification and complex influences on the Ethiopian gene pool. Am. J. Hum. Genet. *91*, 83–96. https://doi.org/10.1016/j.ajhg.2012.05.015.

18. Pickrell, J.K., Patterson, N., Loh, P.R., Lipson, M., Berger, B., Stoneking, M., Pakendorf, B., and Reich, D. (2014). Ancient west Eurasian ancestry in southern and eastern Africa. Proc. Natl. Acad. Sci. USA *111*, 2632–2637. https://doi.org/10.1073/pnas.1313787111.

19. 1000 Genomes Project Consortium; Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., and Abecasis, G.R. (2015). A global reference for human genetic variation. Nature *526*, 68–74.

20. Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature *581*, 434–443. https://doi.org/10.1038/s41586-020-2308-7.

21. Scott, E.M., Halees, A., Itan, Y., Spencer, E.G., He, Y., Azab, M.A., Gabriel, S.B., Belkadi, A., Boisson, B., Abel, L., et al. (2016). Characterization of Greater Middle Eastern genetic variation for enhanced disease gene discovery. Nat. Genet. *48*, 1071–1076. https://doi.org/10.1038/ng.3592.

22. H3Africa Consortium; Rotimi, C., Abayomi, A., Abimiku, A., Adabayeri, V.M., Adebamowo, C., Adebiyi, E., Ademola, A.D., Adeyemo, A., Adu, D., et al. (2014). Research capacity. Enabling the genomic revolution in Africa. Science *344*, 1346–1348.

23. Choudhury, A., Aron, S., Botigué, L.R., Sengupta, D., Botha, G., Bensellak, T., Wells, G., Kumuthini, J., Shriner, D., Fakim, Y.J., et al. (2020). High-depth African genomes inform human migration and health. Nature *586*, 741–748. https://doi.org/10.1038/s41586-020-2859-7.

24. Sherman, R.M., Forman, J., Antonescu, V., Puiu, D., Daya, M., Rafaels, N., Boorgula, M.P., Chavan, S., Vergara, C., Ortega, V.E., et al. (2019). Assembly of a pan-genome from deep sequencing of 910 humans of African descent. Nat. Genet. *51*, 30–35. https://doi.org/10.1038/s41588-018-0273-y.

25. McClellan, J.M., Lehner, T., and King, M.C. (2017). Gene discovery for complex traits: lessons from Africa. Cell *171*, 261–264. https://doi.org/10.1016/j.cell.2017.09.037.

26. Fu, J.M., Satterstrom, F.K., Peng, M., Brand, H., Collins, R.L., Dong, S., Wamsley, B., Klei, L., Wang, L., Hao, S.P., et al. (2022). Rare coding variation provides insight into the genetic architecture and phenotypic context of autism. Nat. Genet. *54*, 1320–1331. https://doi.org/10.1038/s41588-022-01104-0.

27. Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T., Yamrom, B., Yoon, S., Krasnitz, A., Kendall, J., et al. (2007). Strong association of de novo copy number mutations with autism. Science *316*, 445–449. https://doi.org/10.1126/science.1138659.

28. Talevich, E., Shain, A.H., Botton, T., and Bastian, B.C. (2016). CNVkit: genome-wide copy number detection and visualization from targeted DNA sequencing. PLoS Comput. Biol. *12*, e1004873. https://doi.org/10.1371/journal.pcbi.1004873.

29. Basu, S.N., Kollu, R., and Banerjee-Basu, S. (2009). AutDB: a gene reference resource for autism research. Nucleic Acids Res. *37*, D832–D836. https://doi.org/10.1093/nar/gkn835.

30. Mermel, C.H., Schumacher, S.E., Hill, B., Meyerson, M.L., Beroukhim, R., and Getz, G. (2011). GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol. *12*, R41. https://doi.org/10.1186/gb-2011-12-4-r41.

31. Greenberg, F., Lewis, R.A., Potocki, L., Glaze, D., Parke, J., Killian, J., Murphy, M.A., Williamson, D., Brown, F., Dutton, R., et al. (1996). Multidisciplinary clinical study of Smith-Magenis syndrome (deletion 17p11.2). Am. J. Med. Genet. *62*, 247–254. https://doi.org/10.1002/(SICI)1096-8628(19960329)62:3<247::AID-AJMG9>3.0.CO;2-Q.

32. Potocki, L., Chen, K.S., Park, S.S., Osterholm, D.E., Withers, M.A., Kimonis, V., Summers, A.M., Meschino, W.S., Anyane-Yeboa, K., Kashork, C.D., et al. (2000). Molecular mechanism for duplication 17p11.2- the homologous recombination reciprocal of the Smith-Magenis microdeletion. Nat. Genet. *24*, 84–87. https://doi.org/10.1038/71743.

33. Collins, R.L., Brand, H., Karczewski, K.J., Zhao, X., Alföldi, J., Francioli, L.C., Khera, A.V., Lowther, C., Gauthier, L.D., Wang, H., et al. (2020). A structural variation reference for medical and population genetics. Nature *581*, 444–451. https://doi.org/10.1038/s41586-020-2287-8.

34. OMIM, Online Mendelian Inheritance in Man. http://omim.org/.

35. Thormann, A., Halachev, M., McLaren, W., Moore, D.J., Svinti, V., Campbell, A., Kerr, S.M., Tischkowitz, M., Hunt, S.E., Dunlop, M.G., et al. (2019). Flexible and scalable diagnostic filtering of genomic variants using G2P with Ensembl VEP. Nat. Commun. *10*, 2373. https://doi.org/10.1038/s41467-019-10016-3.

36. Cheung, I., Shulha, H.P., Jiang, Y., Matevossian, A., Wang, J., Weng, Z., and Akbarian, S. (2010). Developmental regulation and individual differences of neuronal H3K4me3 epigenomes in the prefrontal cortex. Proc. Natl. Acad. Sci. USA *107*, 8824–8829. https://doi.org/10.1073/pnas.1001702107.

37. Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shoresh, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M., et al. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. Nature *473*, 43–49. https://doi.org/10.1038/nature09906.

38. Markenscoff-Papadimitriou, E., Whalen, S., Przytycki, P., Thomas, R., Binyameen, F., Nowakowski, T.J., Kriegstein, A.R., Sanders, S.J., State, M.W., Pollard, K.S., and Rubenstein, J.L. (2020). A chromatin accessibility atlas of the developing human telencephalon. Cell *182*, 754–769.e18. https://doi.org/10.1016/j.cell.2020.06.002.

39. Tuncay, I.O., Parmalee, N.L., Khalil, R., Kaur, K., Kumar, A., Jimale, M., Howe, J.L., Goodspeed, K., Evans, P., Alzghoul, L., et al. (2022). Analysis of recent shared ancestry in a familial cohort identifies coding and noncoding autism spectrum disorder variants. NPJ Genom. Med. *7*, 13. https://doi.org/10.1038/s41525-022-00284-2.

40. Splawski, I., Timothy, K.W., Sharpe, L.M., Decher, N., Kumar, P., Bloise, R., Napolitano, C., Schwartz, P.J., Joseph, R.M., Condouris, K., et al. (2004). Ca(V)1.2 calcium channel dysfunction causes a multisystem disorder including arrhythmia and autism. Cell *119*, 19–31. https://doi.org/10.1016/j.cell.2004.09.011.

41. Bolz, H., von Brederlow, B., Ramírez, A., Bryda, E.C., Kutsche, K., Nothwang, H.G., Seeliger, M., del C-Salcedó Cabrera, M., Vila, M.C., Molina, O.P., et al. (2001). Mutation of CDH23, encoding a new member of the cadherin gene family, causes Usher syndrome type 1D. Nat. Genet. *27*, 108–112. https://doi.org/10.1038/83667.

42. Vissers, L.E.L.M., van Ravenswaaij, C.M.A., Admiraal, R., Hurst, J.A., de Vries, B.B.A., Janssen, I.M., van der Vliet, W.A., Huys, E.H.L.P.G., de Jong, P.J., Hamel, B.C.J., et al. (2004). Mutations in a new member of the chromodomain gene family cause CHARGE syndrome. Nat. Genet. *36*, 955–957. https://doi.org/10.1038/ng1407.

43. O'Roak, B.J., Vives, L., Girirajan, S., Karakoc, E., Krumm, N., Coe, B.P., Levy, R., Ko, A., Lee, C., Smith, J.D., et al. (2012). Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. Nature *485*, 246–250. https://doi.org/10.1038/nature10989.

44. Fischer-Zirnsak, B., Segebrecht, L., Schubach, M., Charles, P., Alderman, E., Brown, K., Cadieux-Dion, M., Cartwright, T., Chen, Y., Costin, C., et al. (2019). Haploinsufficiency of the notch ligand DLL1 causes variable neurodevelopmental disorders. Am. J. Hum. Genet. *105*, 631–639. https://doi.org/10.1016/j.ajhg.2019.07.002.

45. Rogers, S.J., Wehner, D.E., and Hagerman, R. (2001). The behavioral phenotype in fragile X: symptoms of autism in very young children with fragile X syndrome, idiopathic autism, and other developmental disorders. J. Dev. Behav. Pediatr. *22*, 409–417. https://doi.org/10.1097/00004703-200112000-00008.

46. Hatton, D.D., Sideris, J., Skinner, M., Mankowski, J., Bailey, D.B., Jr., Roberts, J., and Mirrett, P. (2006). Autistic behavior in children with fragile X syndrome: prevalence, stability, and the impact of FMRP. Am. J. Med. Genet. *140A*, 1804–1813. https://doi.org/10.1002/ajmg.a.31286.

47. Anitha, A., Thanseem, I., Nakamura, K., Yamada, K., Iwayama, Y., Toyota, T., Iwata, Y., Suzuki, K., Sugiyama, T., Tsujii, M., et al. (2013). Protocadherin alpha (PCDHA) as a novel susceptibility gene for autism. J. Psychiatry Neurosci. *38*, 192–198. https://doi.org/10.1503/jpn.120058.

48. Ruzzo, E.K., Pérez-Cano, L., Jung, J.Y., Wang, L.K., Kashef-Haghighi, D., Hartl, C., Singh, C., Xu, J., Hoekstra, J.N., Leventhal, O., et al. (2019). Inherited and de novo genetic risk for autism impacts shared networks. Cell *178*, 850–866.e26. https://doi.org/10.1016/j.cell.2019.07.015.

49. Satterstrom, F.K., Kosmicki, J.A., Wang, J., Breen, M.S., De Rubeis, S., An, J.Y., Peng, M., Collins, R., Grove, J., Klei, L., et al. (2020). Large-scale exome sequencing study implicates both developmental and functional changes in the neurobiology of autism. Cell *180*, 568–584.e23. https://doi.org/10.1016/j.cell.2019.12.036.

50. Dias, C., Pfundt, R., Kleefstra, T., Shuurs-Hoeijmakers, J., Boon, E.M.J., van Hagen, J.M., Zwijnenburg, P., Weiss, M.M., Keren, B., Mignot, C., et al. (2021). De novo variants in TCF7L2 are associated with a syndromic neurodevelopmental disorder. Am. J. Med. Genet. *185*, 2384–2390. https://doi.org/10.1002/ajmg.a.62254.

51. Zahra, Q., Çakmak, Ç., Koprulu, M., Shuaib, M., Sobreira, N., Kalsner, L., Sobreira, J., Guillen Sacoto, M.J., Malik, S., and Tolun, A. (2020). Biallelic ZNF407 mutations in a neurodevelopmental disorder with ID, short stature and variable microcephaly, hypotonia, ocular anomalies and facial dysmorphism. J. Hum. Genet. *65*, 1115–1123. https://doi.org/10.1038/s10038-020-0812-0.

52. Gregor, A., Oti, M., Kouwenhoven, E.N., Hoyer, J., Sticht, H., Ekici, A.B., Kjaergaard, S., Rauch, A., Stunnenberg, H.G., Uebe, S., et al. (2013). De novo mutations in the genome organizer CTCF cause intellectual disability. Am. J. Hum. Genet. *93*, 124–131. https://doi.org/10.1016/j.ajhg.2013.05.007.

53. Iossifov, I., Ronemus, M., Levy, D., Wang, Z., Hakker, I., Rosenbaum, J., Yamrom, B., Lee, Y.H., Narzisi, G., Leotta, A., et al. (2012). De novo gene disruptions in children on the autistic spectrum. Neuron *74*, 285–299. https://doi.org/10.1016/j.neuron.2012.04.009.

54. Tartaglia, M., Mehler, E.L., Goldberg, R., Zampino, G., Brunner, H.G., Kremer, H., van der Burgt, I., Crosby, A.H., Ion, A., Jeffery, S., et al. (2001). Mutations in PTPN11, encoding the protein tyrosine phosphatase SHP-2, cause Noonan syndrome. Nat. Genet. *29*, 465–468. https://doi.org/10.1038/ng772.

55. Li, Q., Guo, S., Jiang, X., Bryk, J., Naumann, R., Enard, W., Tomita, M., Sugimoto, M., Khaitovich, P., and Pääbo, S. (2016). Mice carrying a human GLUD2 gene recapitulate aspects of human transcriptome and metabolome development. Proc. Natl. Acad. Sci. USA *113*, 5358–5363. https://doi.org/10.1073/pnas.1519261113.

56. Brenner, V., Nyakatura, G., Rosenthal, A., and Platzer, M. (1997). Genomic organization of two novel genes on human Xq28: compact head to head arrangement of IDH gamma and TRAP delta is conserved in rat and mouse. Genomics *44*, 8–14. https://doi.org/10.1006/geno.1997.4822.

57. Miranda, M., Morici, J.F., Zanoni, M.B., and Bekinschtein, P. (2019). Brain-derived neurotrophic factor: a key molecule for memory in the Healthy and the pathological brain. Front. Cell. Neurosci. *13*, 363. https://doi.org/10.3389/fncel.2019.00363.

58. Fu, J., Guo, O., Zhen, Z., and Zhen, J. (2020). Essential functions of the transcription factor Npas4 in neural circuit development, plasticity, and diseases. Front. Neurosci. *14*, 603373. https://doi.org/10.3389/fnins.2020.603373.

59. Li, J.Z., Absher, D.M., Tang, H., Southwick, A.M., Casto, A.M., Ramachandran, S., Cann, H.M., Barsh, G.S., Feldman, M., Cavalli-Sforza, L.L., and Myers, R.M. (2008). Worldwide human relationships inferred from genome-wide patterns of variation. Science *319*, 1100–1104. https://doi.org/10.1126/science.1153717.

60. Cann, R.L., Stoneking, M., and Wilson, A.C. (1987). Mitochondrial DNA and human evolution. Nature *325*, 31–36. https://doi.org/10.1038/325031a0.

61. Tishkoff, S.A., Reed, F.A., Friedlaender, F.R., Ehret, C., Ranciaro, A., Froment, A., Hirbo, J.B., Awomoyi, A.A., Bodo, J.M., Doumbo, O., et al. (2009). The genetic structure and history of Africans and African Americans. Science *324*, 1035–1044. https://doi.org/10.1126/science.1172257.

62. Gallego Llorente, M., Jones, E.R., Eriksson, A., Siska, V., Arthur, K.W., Arthur, J.W., Curtis, M.C., Stock, J.T., Coltorti, M., Pieruccini, P., et al. (2015). Ancient Ethiopian genome reveals extensive Eurasian admixture throughout the African continent. Science *350*, 820–822. https://doi.org/10.1126/science.aad2879.

63. Skoglund, P., Thompson, J.C., Prendergast, M.E., Mittnik, A., Sirak, K., Hajdinjak, M., Salie, T., Rohland, N., Mallick, S., Peltzer, A., et al. (2017). Reconstructing prehistoric African population structure. Cell *171*, 59–71.e21. https://doi.org/10.1016/j.cell.2017.08.049.

64. Lazaridis, I., Patterson, N., Mittnik, A., Renaud, G., Mallick, S., Kirsanow, K., Sudmant, P.H., Schraiber, J.G., Castellano, S., Lipson, M., et al. (2014). Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature *513*, 409–413. https://doi.org/10.1038/nature13673.

65. Mallick, S., Li, H., Lipson, M., Mathieson, I., Gymrek, M., Racimo, F., Zhao, M., Chennagiri, N., Nordenfelt, S., Tandon, A., et al. (2016). The Simons genome diversity project: 300 genomes from 142 diverse populations. Nature *538*, 201–206. https://doi.org/10.1038/nature18964.

66. Meyer, M., Kircher, M., Gansauge, M.T., Li, H., Racimo, F., Mallick, S., Schraiber, J.G., Jay, F., Prüfer, K., de Filippo, C., et al. (2012). A high-coverage genome sequence from an archaic Denisovan individual. Science *338*, 222–226. https://doi.org/10.1126/science.1224344.

67. Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T., and Reich, D. (2012). Ancient admixture

in human history. Genetics *192*, 1065–1093. https://doi.org/10.1534/genetics.112.145037.

68. Prüfer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., Heinze, A., Renaud, G., Sudmant, P.H., de Filippo, C., et al. (2014). The complete genome sequence of a Neanderthal from the Altai Mountains. Nature *505*, 43–49. https://doi.org/10.1038/nature12886.

69. Skoglund, P., Mallick, S., Bortolini, M.C., Chennagiri, N., Hünemeier, T., Petzl-Erler, M.L., Salzano, F.M., Patterson, N., and Reich, D. (2015). Genetic evidence for two founding populations of the Americas. Nature *525*, 104–108. https://doi.org/10.1038/nature14895.

70. Vyas, D.N., Al-Meeri, A., and Mulligan, C.J. (2017). Testing support for the northern and southern dispersal routes out of Africa: an analysis of Levantine and southern Arabian populations. Am. J. Phys. Anthropol. *164*, 736–749. https://doi.org/10.1002/ajpa.23312.

71. Wilcox, A.G., Vizor, L., Parsons, M.J., Banks, G., and Nolan, P.M. (2017). Inducible knockout of mouse Zfhx3 emphasizes its key role in setting the pace and amplitude of the adult circadian clock. J. Biol. Rhythms *32*, 433–443. https://doi.org/10.1177/0748730417722631.

72. Parsons, M.J., Brancaccio, M., Sethi, S., Maywood, E.S., Satija, R., Edwards, J.K., Jagannath, A., Couch, Y., Finelli, M.J., Smyllie, N.J., et al. (2015). The regulatory factor ZFHX3 modifies circadian function in SCN via an AT motif-driven Axis. Cell *162*, 607–621. https://doi.org/10.1016/j.cell.2015.06.060.

73. Mitani, T., Punetha, J., Akalin, I., Pehlivan, D., Dawidziuk, M., Coban Akdemir, Z., Yilmaz, S., Aslan, E., Hunter, J.V., Hijazi, H., et al. (2019). Bi-Allelic pathogenic variants in TUBGCP2 cause microcephaly and lissencephaly spectrum disorders. Am. J. Hum. Genet. *105*, 1005–1015. https://doi.org/10.1016/j.ajhg.2019.09.017.

74. Langouët, M., Saadi, A., Rieunier, G., Moutton, S., Siquier-Pernet, K., Fernet, M., Nitschke, P., Munnich, A., Stern, M.H., Chaouch, M., and Colleaux, L. (2013). Mutation in TTI2 reveals a role for triple T complex in human brain development. Hum. Mutat. *34*, 1472–1476. https://doi.org/10.1002/humu.22399.

75. Najmabadi, H., Hu, H., Garshasbi, M., Zemojtel, T., Abedini, S.S., Chen, W., Hosseini, M., Behjati, F., Haas, S., Jamali, P., et al. (2011). Deep sequencing reveals 50 novel genes for recessive cognitive disorders. Nature *478*, 57–63. https://doi.org/10.1038/nature10423.

76. Schmitz-Abe, K., Sanchez-Schmitz, G., Doan, R.N., Hill, R.S., Chahrour, M.H., Mehta, B.K., Servattalab, S., Ataman, B., Lam, A.T.N., Morrow, E.M., et al. (2020). Homozygous deletions implicate non-coding epigenetic marks in Autism spectrum disorder. Sci. Rep. *10*, 14045. https://doi.org/10.1038/s41598-020-70656-0.

77. Gamsiz, E.D., Viscidi, E.W., Frederick, A.M., Nagpal, S., Sanders, S.J., Murtha, M.T., Schmidt, M., Simons Simplex Collection Genetics Consortium; Triche, E.W., Geschwind, D.H., et al. (2013). Intellectual disability is associated with increased runs of homozygosity in simplex autism. Am. J. Hum. Genet. *93*, 103–109. https://doi.org/10.1016/j.ajhg.2013.06.004.

78. Walker, S., and Scherer, S.W. (2013). Identification of candidate intergenic risk loci in autism spectrum disorder. BMC Genom. *14*, 499. https://doi.org/10.1186/1471-2164-14-499.

79. Zhu, Y., Li, Y., Haraguchi, S., Yu, M., Ohira, M., Ozaki, T., Nakagawa, A., Ushijima, T., Isogai, E., Koseki, H., et al. (2013). Dependence receptor UNC5D mediates nerve growth factor depletion-induced neuroblastoma regression. J. Clin. Invest. *123*, 2935–2947. https://doi.org/10.1172/JCI65988.

80. Stefansson, H., Sigurdsson, E., Steinthorsdottir, V., Bjornsdottir, S., Sigmundsson, T., Ghosh, S., Brynjolfsson, J., Gunnarsdottir, S., Ivarsson, O., Chou, T.T., et al. (2002). Neuregulin 1 and susceptibility to schizophrenia. Am. J. Hum. Genet. *71*, 877–892. https://doi.org/10.1086/342734.

81. Stefansson, H., Sarginson, J., Kong, A., Yates, P., Steinthorsdottir, V., Gudfinnsson, E., Gunnarsdottir, S., Walker, N., Petursson, H., Crombie, C., et al. (2003). Association of neuregulin 1 with schizophrenia confirmed in a Scottish population. Am. J. Hum. Genet. *72*, 83–87. https://doi.org/10.1086/345442.

82. Shi, L., and Bergson, C.M. (2020). Neuregulin 1: an intriguing therapeutic target for neurodevelopmental disorders. Transl. Psychiatry *10*, 190. https://doi.org/10.1038/s41398-020-00868-5.

83. Sittler, A., Devys, D., Weber, C., and Mandel, J.L. (1996). Alternative splicing of exon 14 determines nuclear or cytoplasmic localisation of fmr1 protein isoforms. Hum. Mol. Genet. *5*, 95–102. https://doi.org/10.1093/hmg/5.1.95.

84. Kim, M., Bellini, M., and Ceman, S. (2009). Fragile X mental retardation protein FMRP binds mRNAs in the nucleus. Mol. Cell Biol. *29*, 214–228. https://doi.org/10.1128/MCB.01377-08.

85. Menon, R.P., Gibson, T.J., and Pastore, A. (2004). The C terminus of fragile X mental retardation protein interacts with the multi-domain Ran-binding protein in the microtubule-organising centre. J. Mol. Biol. *343*, 43–53. https://doi.org/10.1016/j.jmb.2004.08.024.

86. Nakamura, M., Masuda, H., Horii, J., Kuma, K.i., Yokoyama, N., Ohba, T., Nishitani, H., Miyata, T., Tanaka, M., and Nishimoto, T. (1998). When overexpressed, a novel centrosomal protein, RanBPM, causes ectopic microtubule nucleation similar to gamma-tubulin. J. Cell Biol. *143*, 1041–1052. https://doi.org/10.1083/jcb.143.4.1041.

87. Greenbaum, L., Katcoff, D.J., Dou, H., Gozlan, Y., and Malik, Z. (2003). A porphobilinogen deaminase (PBGD) Ran-binding protein interaction is implicated in nuclear trafficking of PBGD in differentiating glioma cells. Oncogene *22*, 5221–5228. https://doi.org/10.1038/sj.onc.1206723.

88. International HapMap Consortium (2003). The international HapMap project. Nature *426*, 789–796.

89. Rife, D.C. (1954). Populations of hybrid origin as source material for the detection of linkage. Am. J. Hum. Genet. *6*, 26–33.

90. Shriner, D. (2013). Overview of admixture mapping. Curr. Protoc. Hum. Gene. *Chapter 1*, Unit 1 23. https://doi.org/10.1002/0471142905.hg0123s76.

91. Landry, L.G., Ali, N., Williams, D.R., Rehm, H.L., and Bonham, V.L. (2018). Lack of diversity in genomic databases is A Barrier to translating precision medicine research into practice. Health Aff. *37*, 780–785. https://doi.org/10.1377/hlthaff.2017.1595.

92. Fan, S., Kelly, D.E., Beltrame, M.H., Hansen, M.E.B., Mallick, S., Ranciaro, A., Hirbo, J., Thompson, S., Beggs, W., Nyambo, T., et al. (2019). African evolutionary history inferred from whole genome sequence data of 44 indigenous African populations. Genome Biol. *20*, 82. https://doi.org/10.1186/s13059-019-1679-2.

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

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

95. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324.

96. Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., et al. (2013). From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr. Protoc. Bioinformatics *43*, 11.10.1–11.10.33. https://doi.org/10.1002/0471250953.bi1110s43.

97. Baran, Y., Pasaniuc, B., Sankararaman, S., Torgerson, D.G., Gignoux, C., Eng, C., Rodriguez-Cintron, W., Chapela, R., Ford, J.G., Avila, P.C., et al. (2012). Fast and accurate inference of local ancestry in Latino populations. Bioinformatics *28*, 1359–1367. https://doi.org/10.1093/bioinformatics/bts144.

98. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., and

Sham, P.C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575. https://doi.org/10.1086/519795.

99. Gel, B., Díez-Villanueva, A., Serra, E., Buschbeck, M., Peinado, M.A., and Malinverni, R. (2016). regioneR: an R/Bioconductor package for the association analysis of genomic regions based on permutation tests. Bioinformatics *32*, 289–291. https://doi.org/10.1093/bioinformatics/btv562.

100. Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., and Li, H. (2021). Twelve years of SAMtools and BCFtools. GigaScience *10*, giab008. https://doi.org/10.1093/gigascience/giab008.

101. Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. (2002). The human genome browser at UCSC. Genome Res. *12*, 996–1006. https://doi.org/10.1101/gr.229102.

102. Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., et al. (2011). The variant call format and VCFtools. Bioinformatics *27*, 2156–2158. https://doi.org/10.1093/bioinformatics/btr330.

103. Ng, P.C., and Henikoff, S. (2003). SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res. *31*, 3812–3814. https://doi.org/10.1093/nar/gkg509.

104. Kumar, P., Henikoff, S., and Ng, P.C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat. Protoc. *4*, 1073–1081. https://doi.org/10.1038/nprot.2009.86.

105. Adzhubei, I., Jordan, D.M., and Sunyaev, S.R. (2013). Predicting functional effect of human missense mutations using PolyPhen-2. Curr. Protoc. Hum. Gene. *Chapter 7*, Unit7 20. https://doi.org/10.1002/0471142905.hg0720s76.

106. Chun, S., and Fay, J.C. (2009). Identification of deleterious mutations within three human genomes. Genome Res. *19*, 1553–1561. https://doi.org/10.1101/gr.092619.109.

107. Schwarz, J.M., Cooper, D.N., Schuelke, M., and Seelow, D. (2014). MutationTaster2: mutation prediction for the deep-sequencing age. Nat. Methods *11*, 361–362. https://doi.org/10.1038/nmeth.2890.

108. Reva, B., Antipin, Y., and Sander, C. (2011). Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res. *39*, e118. https://doi.org/10.1093/nar/gkr407.

109. Shihab, H.A., Gough, J., Cooper, D.N., Stenson, P.D., Barker, G.L.A., Edwards, K.J., Day, I.N.M., and Gaunt, T.R. (2013). Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. Hum. Mutat. *34*, 57–65. https://doi.org/10.1002/humu.22225.

110. Choi, Y., and Chan, A.P. (2015). PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. Bioinformatics *31*, 2745–2747. https://doi.org/10.1093/bioinformatics/btv195.

111. Dong, C., Wei, P., Jian, X., Gibbs, R., Boerwinkle, E., Wang, K., and Liu, X. (2015). Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. Hum. Mol. Genet. *24*, 2125–2137. https://doi.org/10.1093/hmg/ddu733.

112. GTEx Consortium; Laboratory, Data Analysis &Coordinating Center LDACC—Analysis Working Group; Statistical Methods groups—Analysis Working Group; Enhancing GTEx eGTEx groups; NIH Common Fund; NIH/NCI; NIH/NHGRI; NIH/NIMH; NIH/NIDA; Biospecimen Collection Source Site—NDRI (2017). Genetic effects on gene expression across human tissues. Nature *550*, 204–213. https://doi.org/10.1038/nature24277.

113. Carter, H., Douville, C., Stenson, P.D., Cooper, D.N., and Karchin, R. (2013). Identifying Mendelian disease genes with the variant effect scoring tool. BMC Genomics *14* (*Suppl 3*), S3. https://doi.org/10.1186/1471-2164-14-S3-S3.

114. Douville, C., Masica, D.L., Stenson, P.D., Cooper, D.N., Gygax, D.M., Kim, R., Ryan, M., and Karchin, R. (2016). Assessing the pathogenicity of insertion and deletion variants with the variant effect scoring tool (VEST-Indel). Hum. Mutat. *37*, 28–35. https://doi.org/10.1002/humu.22911.

115. Rentzsch, P., Witten, D., Cooper, G.M., Shendure, J., and Kircher, M. (2019). CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. *47*, D886–D894. https://doi.org/10.1093/nar/gky1016.

116. Arbiza, L., Gronau, I., Aksoy, B.A., Hubisz, M.J., Gulko, B., Keinan, A., and Siepel, A. (2013). Genome-wide inference of natural selection on human transcription factor binding sites. Nat. Genet. *45*, 723–729. https://doi.org/10.1038/ng.2658.

117. Gronau, I., Arbiza, L., Mohammed, J., and Siepel, A. (2013). Inference of natural selection from interspersed genomic elements based on polymorphism and divergence. Mol. Biol. Evol. *30*, 1159–1171. https://doi.org/10.1093/molbev/mst019.

118. Davydov, E.V., Goode, D.L., Sirota, M., Cooper, G.M., Sidow, A., and Batzoglou, S. (2010). Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput. Biol. *6*, e1001025. https://doi.org/10.1371/journal.pcbi.1001025.

119. Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., et al. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. *15*, 1034–1050. https://doi.org/10.1101/gr.3715005.

120. Garber, M., Guttman, M., Clamp, M., Zody, M.C., Friedman, N., and Xie, X. (2009). Identifying novel constrained elements by exploiting biased substitution patterns. Bioinformatics *25*, i54–i62. https://doi.org/10.1093/bioinformatics/btp190.

121. Landrum, M.J., Lee, J.M., Benson, M., Brown, G.R., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Jang, W., et al. (2018). ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res. *46*, D1062–D1067. https://doi.org/10.1093/nar/gkx1153.

122. Nott, A., Holtman, I.R., Coufal, N.G., Schlachetzki, J.C.M., Yu, M., Hu, R., Han, C.Z., Pena, M., Xiao, J., Wu, Y., et al. (2019). Brain cell type-specific enhancer-promoter interactome maps and disease-risk association. Science *366*, 1134–1139. https://doi.org/10.1126/science.aay0793.

123. Doan, R.N., Lim, E.T., De Rubeis, S., Betancur, C., Cutler, D.J., Chiocchetti, A.G., Overman, L.M., Soucy, A., Goetze, S., et al.; Autism Sequencing Consortium (2019). Recessive gene disruptions in autism spectrum disorder. Nat. Genet. *51*, 1092–1098. https://doi.org/10.1038/s41588-019-0433-8.

124. Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience *4*, 7. https://doi.org/10.1186/s13742-015-0047-8.

125. Chahrour, M.H., Yu, T.W., Lim, E.T., Ataman, B., Coulter, M.E., Hill, R.S., Stevens, C.R., Schubert, C.R., ARRA Autism Sequencing Collaboration; and Greenberg, M.E., et al. (2012). Whole-exome sequencing and homozygosity analysis implicate depolarization-regulated neuronal genes in autism. PLoS Genet. *8*, e1002635. https://doi.org/10.1371/journal.pgen.1002635.

126. McKinney, W. (2010). Data structures for statistical computing in python. In Proceedings of the 9th Python in Science Conference, *445*, pp. 51–56.

127. Hunter, J.D. (2007). Matplotlib: a 2D graphics environment. Comput. Sci. Eng. *9*, 90–95. https://doi.org/10.1109/Mcse.2007.55.

128. Gordon, D., Haynes, C., Johnnidis, C., Patel, S.B., Bowcock, A.M., and Ott, J. (2004). A transmission disequilibrium test for general pedigrees

that is robust to the presence of random genotyping errors and any number of untyped parents. Eur. J. Hum. Genet. *12*, 752–761. https://doi.org/10.1038/sj.ejhg.5201219.

129. Gordon, D., Heath, S.C., Liu, X., and Ott, J. (2001). A transmission/disequilibrium test that allows for genotyping errors in the analysis of single-nucleotide polymorphism data. Am. J. Hum. Genet. *69*, 371–380.

130. Shriner, D., Adeyemo, A., and Rotimi, C.N. (2011). Joint ancestry and association testing in admixed individuals. PLoS Comput. Biol. *7*, e1002325. https://doi.org/10.1371/journal.pcbi.1002325.

131. Thode, H.C. (2002). Testing for Normality, 1st Edition (CRC Press).

132. Aldor-Noiman, S., Brown, L.D., Buja, A., Rolke, W., and Stine, R.A. (2013). The power to see: a new graphical test of normality. Am. Statistician *67*, 249–260.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
1000 Genomes Project	The 1000 Genomes Project Consortium	ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/
ClinVar database	ClinVar	https://www.ncbi.nlm.nih.gov/clinvar/
Gene2Phenotype	Thormann et al.[35]	https://www.ebi.ac.uk/gene2phenotype/
Genome Aggregation Database (gnomAD)	Karczewski et al.[20]	https://gnomad.broadinstitute.org/
Genotype-Tissue Expression (GTEx) portal	GTEx Consortium	https://gtexportal.org/home/
Greater Middle East Variome (GME) project	Scott et al.[21]	http://igm.ucsd.edu/gme/
Human Heredity and Health in Africa (H3Africa) project	H3Africa Consortium	https://h3africa.org/
Online Mendelian Inheritance in Man (OMIM)	OMIM	https://omim.org/
Reich Laboratory Dataset	Reich Laboratory	https://reich.hms.harvard.edu/datasets
Simons Foundation Autism Research Initiative (SFARI) Gene database	Basu et al.[29]	https://www.sfari.org/resource/sfari- gene/
The Encyclopedia of DNA Elements (ENCODE)	ENCODE Project	https://www.encodeproject.org/
Software and algorithms		
ADMIXTURE	Alexander et al.[93]	https://dalexander.github.io/admixture/d ownload.html
ANNOVAR	Wang et al.[94]	https://annovar.openbioinformatics.org/e n/latest/user-guide/download/
Burrows-Wheeler Aligner (BWA)	Li and Durbin[95]	https://github.com/lh3/bwa
CNVkit	Talevich et al.[28]	https://github.com/etal/cnvkit
Genome Analysis Toolkit (GATK)	Van der Auwera et al.[96]	https://software.broadinstitute.org/gatk/
GISTIC2.0	Mermel et al.[30]	https://github.com/broadinstitute/gistic2
LAMP-LD	Baran et al.[97]	https://bogdan.dgsom.ucla.edu/pages/la mp/
Picard	Broad Institute of MIT and Harvard	https://broadinstitute.github.io/picard/
PLINK	Purcell et al.[98]	http://www.cog-genomics.org/plink2/
regioneR	Gel et al.[99]	https://www.bioconductor.org/packages/release/bioc/html/regioneR.html
Samtools	Danecek et al.[100]	http://www.htslib.org/download/
UCSC Genome Browser	Kent et al.[101]	http://genome.ucsc.edu
Vcftools	Danecek et al.[102]	https://vcftools.sourceforge.net/

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Maria H. Chahrour (maria.chahrour@utsouthwestern.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

Data reported in this paper will be shared by the lead contact upon request.
This paper does not report original code.
Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Subjects and specimens

All human studies were reviewed and approved by the institutional review board of the University of Texas Southwestern Medical Center (UTSW). Families were recruited either from the Dallas Fort Worth area or nationally in the US, all belonging to a community of individuals with East African ancestry, and written informed consent was obtained from all study participants. The majority of enrolled families were Ethiopian (20/33), Eritrean (9/33), or both (3/33), and one family was Kenyan (1/33). In addition to East African ancestry, inclusion criteria included a diagnosis of autism spectrum disorder (ASD) by a neurologist, child psychiatrist, or psychologist. Patients with genetically defined syndromes, specifically Fragile X syndrome, Angelman syndrome, Rett syndrome, or Tuberous sclerosis complex, were excluded from study participation. All patients enrolled in the study received a diagnosis of ASD from their referring clinicians who performed physical and behavioral assessments and administered the standard autism diagnostic measures (ADOS, ADI-R, and DSM-V). With the exception of one father with possible signs of ASD (coded as unknown phenotype in our analyses), none of the parents had a diagnosis of ASD. Blood samples were collected from all available family members by peripheral venipuncture. Genomic DNA was isolated from circulating leukocytes using AutoPure (Qiagen, Hilden, Germany) according to the manufacturer's instructions.

Clinical information

A standardized medical questionnaire was collected from each participating family (33). For the following clinical description, three records were excluded due to incomplete information (3/36 probands). Of the 33 probands where the medical questionnaire was complete, all but one family had prenatal care, and 8/33 reported pregnancy-related complications including gestational diabetes (4/33), hyperemesis gravidarum (1/33), infections (2/33), and twin gestation with hypertension (1/33). Complications within the immediate postnatal period included jaundice (2/33), feeding problems or vomiting (4/33), and one case with atrial and ventricular septal defects and heart murmur (1/33). All probands were born full term except for three, which included the twins (born at 29 weeks gestation) and one late preterm birth at 36 weeks gestation (delivery complicated by placental abruption). The mean birth weight for the cohort is 3,442.1 g (SD 653.3 g) and most were born via Cesarean section (18/33). Three of the probands also had seizures, two with seizures reportedly controlled on monotherapy and one who continues to have seizures in sleep most nights. One proband also reported alopecia. Hearing problems (1/33) and vision problems (2/33) were rarely reported. Constipation was frequently reported (7/33) and six probands were reportedly on restrictive diets, mostly gluten-free and casein-free. Most were not taking any prescription medications (22/33) and only four were prescribed psychotropic medications commonly used to manage behavioral problems in ASD. All 36 probands were diagnosed with ASD, and comorbidities included language and speech abnormalities, developmental delay, and cognitive deficits among others (Figure S2). The age at diagnosis was available for 20/36 probands, and it ranged from 6 months to 7 years (mean of 2.825 years, median of 3 years, and mode of 3 years), with the exception of two probands diagnosed at 15 and 19 years of age. Furthermore, 26/33 probands were receiving special education services through the public school system. Of the 33 families reviewed, six families reported a family history of ASD (inclusive of the three multiplex families), one family reported a mother with language delay, and one family reported a father with signs of ASD but with no clinical diagnosis of ASD.

METHOD DETAILS

Whole genome sequencing and data processing

Whole genome sequencing was performed on an Illumina NovaSeq 6000 platform (San Diego, California). DNA quality and quantity were assessed using a Qubit High Sensitivity Assay (Thermo Fisher Scientific, Waltham, Massachusetts) and gel electrophoresis. Between 100 ng and 1 μg of DNA was used for genomic library preparation using the Illumina TruSeq DNA Library Prep Kit according to the manufacturer's protocol and libraries were pair-end sequenced (150 bp read lengths). The genomes were processed following the best practices recommended by the Broad Institute.[96] Reads were aligned to the human reference genome version GRCh37/hg19 using the Burrows-Wheeler Aligner (BWA, version 0.7.10).[95] Duplicate reads were removed using Picard (version 1.117). Local realignment, quality recalibration, and variant (single nucleotide variants (SNVs) and insertions or deletions (indels)) detection were performed using the Genome Analysis Toolkit (GATK; version 3.3).[96] All reported variants passed filtering according to GATK[96] best practices. Depth was calculated using samtools[100] depth and coverage was assessed using custom scripts. The percent coverage at 1X, 4X, 10X, 20X, 30X, and 40X was calculated as the number of base pair positions sequenced to a given depth divided by the total number of bases sequenced.

Variant call format (VCF) files for SNVs and indels were annotated with ANNOVAR[94] using allele frequencies from the 1000 Genomes project (1000G),[19] the Genome Aggregation Database (gnomAD; v2.1.1),[20] the Greater Middle East Variome project (GME),[21] and the Human Heredity and Health in Africa project (H3Africa).[22,23] Annotated VCF files were uploaded into an SQL database for working storage and analysis. Genome data was stored and analyses were performed on the Texas Advanced Computing Center (TACC) high-performance computing servers, a resource of the University of Texas (Austin, TX).

Variant filtration

Variants were quality filtered in SQL with a PASS designation in the GATK pipeline,[96] a genotype quality (GQ) score of $\geq$30, and 10$\leq$ total read depth $\leq$100. Rare variants were defined as those with minor allele frequencies (MAF) < 1% in 1000G,[19] gnomAD,[20] GME,[21] and H3Africa.[22,23] Novel variants were defined as variants that are not found in the four aforementioned public datasets. Private variants were defined as novel variants that occurred only in a single individual in our cohort.

De novo variants were defined as heterozygous private variants present in affected individuals (absent from the genome of either the father, the mother, or the sibling(s) when available). To minimize potential false positive *de novo* calls, we applied additional filtering steps, requiring that *de novo* variants have the following criteria: (i) GQ = 99, (ii) alternate allele depth (AD-Alt) $\geq$ 10, (iii) reference allele depth (AD-Ref) $\geq$ 10, (iv) 0.3 $\leq$ AD-Alt/read depth (DP) $\leq$ 0.7, (v) SOR $\leq$1.5, QD $\geq$ 10, Mapping Quality (MQ) $\geq$ 59, Quality $\geq$999, (vi) $-1.4\leq$ ReadPosRankSum $\leq$3.0, (vii) length(Alt) $\leq$ 50, and length(Ref) $\leq$ 50. Compound heterozygous variants in offspring were defined as inherited heterozygous coding (exonic or splice site) variants that occurred within the same gene and that were present in heterozygous form in one parent but not the other. Compound heterozygous variants occurring in unaffected siblings in the same compound form as the affected individual were excluded. All compound heterozygous variants were filtered for AD-Alt $\geq$10, AD-Ref $\geq$10, and 0.3 $\leq$ AD-Alt/DP $\leq$ 0.7. Inherited homozygous variants were required to be present in heterozygous form in both the father and the mother, excluding variants that are homozygous in either one of the parents or unaffected siblings when available, on the assumption of full penetrance. X-linked variants were required to be present in a male offspring, heterozygous in the mother, and absent from the father.

Noncoding variant annotation

Custom SQL and Python scripts were used to annotate noncoding variants with three datasets: (1) chromatin state segmentation from nine human cell lines,[37] (2) maps of histone H3K4me3 mark in human prefrontal cortex (PFC) from 11 individuals,[36] and (3) predicted developmental brain enhancers from fetal brain samples.[38] The columns in Table S7 derived from each dataset were denoted as ENCODE, uMass, and CBA, respectively. Additional details are presented in Table S14. Variants that were found within a peak in the uMass dataset were marked as "predicted human brain promoter" variants. Predicted human brain promoter variants that were absent from regions with "1_Active_Promoter" prediction in any one of the 9 non-neuronal cell lines in the ENCODE dataset were marked as "predicted human brain-specific promoter" variants. Variants that were found within a predicted regulatory element (pRE) region in the CBA dataset were marked as "predicted human brain enhancer" variants. Predicted human brain enhancer variants that were absent from regions with "4_Strong_Enhancer" or "5_Strong_Enhancer" prediction in any of the 9 non-neuronal cell lines in the ENCODE dataset were marked as "predicted human brain-specific enhancer" variants.

Variant prioritization

Rare variants that are *de novo*, compound heterozygous, inherited homozygous, or X-linked, were considered to be possibly pathogenic if they met the following criteria: (1) splice site variants, (2) exonic variants with a predicted protein effect of frameshift indels, nonframeshift indels, stopgain, stoploss, or unknown effect, (3) exonic nonsynonymous SNVs that were predicted to be damaging by at least 1 of the following 9 algorithms used: SIFT,[103,104] PolyPhen-2 HumVar,[105] LRT,[106] MutationTaster2,[107] MutationAssessor,[108] FATHMM,[109] PROVEAN,[110] MetaSVM,[111] and MetaLR.[111] PolyPhen-2 HumVar was chosen over PolyPhen-2 HumDiv because the former is more appropriate for Mendelian variants with drastic effect as we expect for ASD, while the latter is appropriate for common variants of smaller effect size.

Possibly pathogenic variants were compared to the list of genes implicated in ASD from the Simons Foundation Autism Research Initiative (SFARI) Gene 2018 database[29] (using the 2022 Q2 release), the Gene2Phenotype (G2P) Developmental Disorders (DD) Panel,[35] and a list of neurodevelopmental disease genes.[4] Variants were also screened for any phenotypic association in G2PDD (confidence categories "definitive" and "strong")[35] and the Online Mendelian Inheritance in Man (OMIM) database.[34] Gene expression data was obtained from the Genotype-Tissue Expression (GTEx) portal.[112] The GTEx Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. Gene constraint was assessed using pLI, LOEUF, and Z scores from gnomAD.[20]

To prioritize candidate disease variants, we considered the following criteria: (1) segregation with phenotype in the family: we excluded variants that were present with the same genotype in unaffected siblings and prioritized variants that were present with the same genotype in affected siblings; (2) protein effect: we prioritized loss of function (LoF) variants and nonsynonymous SNVs with high probability of deleteriousness based on scores from prediction tools, including the 9 algorithms mentioned above, in addition to VEST,[113,114] CADD,[115] fitCons,[116,117] GERP++,[118] phyloP,[119] phastCons,[119] and SiPhy[120]; (3) gene constraint: we prioritized variants within genes with higher pLI, lower LOEUF, and higher Z scores; (4) gene expression: we prioritized variants within genes with higher expression in the brain; (5) disease association: we prioritized variants within genes that had a SFARI Gene[29] score of 1, 2, 3, or S, or were associated with a neurodevelopmental phenotype as annotated by OMIM[34]; (6) ClinVar[121] clinical significance annotation when available.

For noncoding variant prioritization, we sought additional verification of regulatory element prediction. Brain-specific enhancer and promoter variants were visualized using a UCSC genome browser[101] track of brain cell-type specific proximity ligation-assisted ChIP-seq (PLAC-seq) data from Nott et al.[122] PLAC-seq identifies long-range chromatin interactions at promoters and enhancers. We marked variants as linked to a certain gene if the enhancer region where the variant is located (based on the ChIP-seq and

ATAC-seq data from the aforementioned UCSC genome browser tracks) was linked to the promoter of the target gene in the PLAC-seq data. Noncoding variants that were verified to be within brain-specific enhancers or promoters of genes with known brain-relevant functions were prioritized as candidate disease causing.

Burden analysis

The burden of rare LoF and predicted damaging missense variants was analyzed by comparing categories of variants identified in affected versus unaffected individuals. LoF variants were defined as variants that are exonic or splice site predicted to result in a frameshift indel, a stopgain or stoploss, or splicing error. Missense variants were defined as nonsynonymous exonic or splice site. Missense damaging variants were defined as nonsynonymous SNVs that were predicted to be damaging by at least 5 of the 9 algorithms described above under variant prioritization. Comparisons were made between affected and unaffected genomes in the above categories for all rare variants, including homozygous variants to evaluate biallelic damaging allele burden.[123]

Copy number variant (CNV) analysis

We used CNVkit[28] to detect CNVs based on read depth in affected samples relative to the average read depth in unaffected samples in the same family as controls. Since CNVkit[28] only uses sequencing read depth information without considering variant heterozygosity information, it is relatively insensitive to call heterozygous CNVs and cannot confidently infer whether a CNV is *de novo* or not. We used GISTIC2.0[30] on segmented files generated from CNVkit[28] to further evaluate the significance of the amplified and deleted segments between the affected and unaffected samples. The criteria included a threshold for copy number amplification and deletion of 0.1, confidence level of 99%, and FDR of 0.05. CNVs in affected individuals that overlap with ASD CNVs annotated in SFARI Gene[29] (2022 Q2 release) were identified using R Bioconductor package regioneR.[99] Significance of the overlap was tested by performing an overlap permutation test, also using regioneR.[99] CNVs were annotated with population allele frequencies from gnomAD structural variant (SV) v2.1 database.[33]

Assessment of runs of homozygosity

PLINK version 1.90b6.11[98,124] was used for all analyses. VCF files were converted into PLINK format using vcftools version 0.1.13.[102] The cohort was assessed for relatedness using PLINK-genome. Autosomal variants were filtered for Hardy-Weinberg equilibrium ($p < 0.001$), MAF >5%, and maximum missing genotype rate of 25%. Runs of homozygosity (ROHs) were identified in PLINK using a sliding window analysis with a 100 base pair window size, allowing for 30 heterozygous variants and 30 missing genotypes per window in accordance with previously described methods.[125] The resulting segments were then filtered using a percent homozygosity (PHOM) threshold of 75%. ROH summary statistics were calculated for 125 samples, excluding samples from three individuals who were outside a family unit and one proband with <90% coverage. Percentage of genome within ROHs was estimated as the ratio of total ROH length to total autosomal bases sequenced at 1X.

Heterozygosity

Percent heterozygosity in the cohort was determined by dividing the number of heterozygous calls as determined by GATK[96] variant calling by the number of sequenced bases. The number of sequenced bases was determined by using samtools[100] depth. For the 1000G[19] samples PLINK files were created (PLINK version 1.90b6.11)[98,124] and the number of heterozygous calls was determined by using –het in PLINK. The output gives the observed number of homozygous calls and the total number of genotypes for each sample. The number of heterozygous calls was determined by subtracting the number of homozygous calls from the total number of sequenced genotypes. This was divided by the total number of sequenced bases which was determined using samtools[100] depth. For subpopulations ASW, CHB, CHS, and MXL, we were unable to calculate heterozygosity since only exome sequencing data are available for these four subpopulations.

Principal component analysis

Principal component analysis (PCA) was carried out in PLINK version 1.90b6.11[98,124] using Phase 3 1000G[19] data. PCA input files from our samples were pruned to remove variants with MAF <5%, missing genotype rate greater than 5%, and pruned for linkage disequilibrium (LD) with an r2 threshold of 0.2 using PLINK –indep-pairwise 50 5 0.2. Triallelic and palindromic variants were also removed. The set of variants that remained was extracted from the 1000G[19] dataset and these were merged with our cohort dataset. PCA was run in PLINK using the –pca flag and the first two principal components were plotted in R. PCA for African groups was carried out using publicly available data which included genotypes from the 1000G[19] African samples and additional African samples (see Table S9 for details). Analysis was performed for unrelated individuals using data from either pedigree founders (parents) or unrelated offspring. Ancient samples and samples that could not be confirmed as modern-day were removed from analysis.

Admixture analysis

Global ancestry admixture analysis was carried out using ADMIXTURE version 1.3.0.[93] The variant dataset was pruned to remove variants with MAF <5%, missing genotype rate greater than 5%, and pruned for LD with an r2 threshold of 0.1 using PLINK –indep-pairwise 50 10 0.1. ADMIXTURE was run on our cohort for values of K from 2 to 20 and the minimum cross validation error was identified at K = 2 (Figure S9A). The value of K which yields the smallest cross validation error value is taken to be the best approximation of the actual number of ancestral populations. Our pruned dataset was merged with the 1000G[19] genotypes and ADMIXTURE was run on the merged dataset, again using values of K from 2 to 20 and the minimum cross validation error for the merged dataset was identified at K = 9 (Figure S9B). We then merged our pruned dataset with publicly available data which included genotypes from the 1000G[19] African samples and additional African samples (see Table S9 for details). ADMIXTURE was run on the resulting dataset for values of K from 2 to 20. The minimum cross validation error for this dataset was identified at K = 3 (Figure S9C). For each analyzed sample, the proportion of ancestry from each presumed population was determined from the Q file output from ADMIXTURE.

Local ancestry was analyzed using LAMP-LD version 1.0[97] assuming two-way admixture based on the minimum cross validation error determined with ADMIXTURE. The two most closely related reference populations from the 1000G[19] were the LWK (Luhya in Webuye, Kenya) and the TSI (Toscani in Italy). PLINK files with genotypes for the LWK and TSI groups were merged with genotypes from our cohort and the resulting set was pruned to remove variants with missing genotype rate greater than 5%. From this pruned dataset, our cohort samples, an LWK reference dataset, and a TSI reference dataset were extracted for input into LAMP-LD. Chromosomes were analyzed independently. A total of 209,999 markers were analyzed. Output files were analyzed using a custom Python pipeline.

LAMP-LD assigns segments to one of the reference populations determined in global admixture analysis to be closest to the ancestral admixed population. The number of ancestral switches is a function of the number of recombination events since the two admixed populations encountered each other. More recent admixture will result in larger blocks of ancestry and thereby fewer ancestry switches. Older admixture will result in decay of LD blocks and a larger number of ancestry switches. The mean number of ancestry switches per chromosome was determined by counting the number of ancestral segments per chromosome for each individual and taking the average for the cohort. LAMP-LD output was analyzed using a custom Python pipeline.

The percentage of African and European ancestry as determined by assignments to LWK or TSI reference groups, respectively, was calculated for affected and unaffected offspring by summing the number of LWK assigned alleles for each marker and dividing by the number of offspring in each group. To catalog admixture peaks, segments were identified that had a difference in percent LWK ancestry (absolute value of delta, abs(Δ)) of 15 percentage points or greater. The Manhattan plots were generated with a custom Python script using the Pandas[126] and Matplotlib[127] libraries. For extracting variants from admixture peaks (Tables S11C and S11D), since the analysis was run with a sparse marker set, we expanded the regions of interest up until the position of the first neighboring marker on either side. All variants within these peaks were extracted in SQL for further analysis.

LAMP-LD output at each allele was converted to a genotype-like format where each marker is coded based on assigned ancestry. Markers were coded as having 0, 1, or 2 alleles from the LWK reference population. We used TDTae[128] to perform a likelihood-based transmission disequilibrium test (TDT) in our cohort of families that consist of parents and at least one affected offspring per family. We selected TDTae to run this analysis because it can extract allele transmission information even if one parent is missing, as is the case in some of the families in our cohort. Unaffected siblings and second-degree relatives were included in the analysis when available. TDTae analysis was performed under dominant, recessive, and multiplicative models using default parameters and the GLHO (Gordon Liu Heath Ott) error model.[129] For each marker, the TDTae statistics (also known as LRT values, given by the formula -2[LogLike(H1)-LogLike(H0)]) obtained from the three models were compared, and the model that resulted with the highest LRT value was selected for further analysis. Since the markers for the analysis were generated from haplotypes and are not independent, we used the mean number of ancestry switches (Table S12) to approximate the number of independent tests in order to correct for multiple testing.[130] Using this method, the Bonferroni corrected significance threshold was $p = 3.32\times10^{-4}$. A quantile-quantile (Q-Q) analysis of the TDTae *p* values was performed and plotted using the R package qqplotr.[131,132]