Genetic structure correlates with ethnolinguistic diversity in eastern and southern Africa

Elizabeth G. Atkinson[1,2*], Shareefa Dalvie[3,4,†], Yakov Pichkar[5,†], Allan Kalungi[6,7†], Lerato Majara[8,9†], Anne Stevenson[2,10], Tamrat Abebe[11], Dickens Akena[6], Melkam Alemayehu[12], Fred K. Ashaba[13], Lukoye Atwoli[14,15], Mark Baker[2], Lori B. Chibnik[2,16], Nicole Creanza[5], Mark J. Daly[1,2], Abebaw Fekadu[12,17], Bizu Gelaye[10,2], Stella Gichuru[14], Wilfred E. Injera[18], Roxanne James[9], Symon M. Kariuki[19,20], Gabriel Kigen[21], Nastassja Koen[3,4], Karestan C. Koenen[2,10], Zan Koenig[2], Edith Kwobah[14], Joseph Kyebuzibwa[6], Henry Musinguzi[13], Rehema M. Mwema[19], Benjamin M. Neale[1,2], Carter P. Newman[2,10], Charles R.J.C. Newton[19,20], Linnet Ongeri[19], Sohini Ramachandran[22], Raj Ramesar[23], Welelta Shiferaw[12], Dan J. Stein[3,4], Rocky E. Stroud[2,10], Solomon Teferra[12], Zukiswa Zingela[24], Alicia R. Martin[1,2], NeuroGAP-Psychosis Study Team

Affiliations

1 Analytic and Translational Genetics Unit and Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA

2 Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA

3 Department of Psychiatry and Mental Health, University of Cape Town, Cape Town, South Africa

4 South African Medical Research Council (SAMRC) Unit on Risk and Resilience in Mental Disorders, Neuroscience Institute, University of Cape Town, Cape Town, South Africa

5 Department of Biological Sciences, Vanderbilt University, Nashville, TN

6 Department of Psychiatry, School of Medicine, College of Health Sciences, Makerere University, Kampala, Uganda

7 Mental Health Section of MRC/UVRI & LSHTM Uganda Research Unit, Entebbe, Uganda

8 South African Medical Research Council (SAMRC) Human Genetics Research Unit, Division of Human Genetics, Institute of Infectious Disease and Molecular Medicine, University of Cape Town, Cape Town, South Africa

9 Department of Psychiatry and Mental Health, University of Cape Town, Cape Town, South Africa

10 Department of Epidemiology, Harvard T. H. Chan School of Public Health, Boston, MA, USA

11 Department of Microbiology, Immunology, and Parasitology, School of Medicine, College of Health Sciences, Addis Ababa University, Addis Ababa, Ethiopia

12 Department of Psychiatry, School of Medicine, College of Health Sciences, Addis Ababa University, Addis Ababa, Ethiopia

13 Department of Immunology & Molecular Biology, College of Health Sciences, Makerere University, Kampala, Uganda

14 Department of Mental Health, School of Medicine, Moi University College of Health Sciences, Eldoret, Kenya

15 Brain and Mind Institute and Department of Internal Medicine, Medical College East Africa, the Aga Khan University, Nairobi, Kenya

16 Department of Neurology, Massachusetts General Hospital, Boston, MA, USA

17 Centre for Innovative Drug Development & Therapeutic Trials for Africa, Addis Ababa University, Addis Ababa, Ethiopia

18 Department of Immunology, School of Medicine, Moi University College of Health Sciences, Eldoret, Kenya

19 Neurosciences Unit, Clinical Department, KEMRI-Wellcome Trust Research Programme-Coast, Kilifi, Kenya

20 Department of Psychiatry, University of Oxford, Oxford, UK

21 Department of Pharmacology and Toxicology, School of Medicine, Moi University College of Health Sciences, Eldoret, Kenya

22 Department of Ecology and Evolutionary Biology and Center for Computational Molecular Biology, Brown University, Providence, RI

23 South African Medical Research Council (SAMRC) Unit on Risk and Resilience in Mental Disorders, University of Cape Town and Neuroscience Institute, Cape Town, South Africa

24 Department of Psychiatry and Human Behavioral Sciences, Walter Sisulu University, Mthatha, South Africa

Author list footnotes

†These authors contributed equally to this work.

*Correspondence/lead contact: eatkinso@broadinstitute.org

Corresponding author contact information

Elizabeth Atkinson, PhD

Email: eatkinso@broadinstitute.org

Phone: (202) 246-0666

Summary

African populations are the most diverse in the world yet are sorely underrepresented in medical genetics research. Here, we examine the structure of African populations using genetic and comprehensive multigenerational ethnolinguistic data from the Neuropsychiatric Genetics of African Populations-Psychosis study (NeuroGAP-Psychosis) consisting of 900 individuals from Ethiopia, Kenya, South Africa, and Uganda. We find that self-reported language classifications meaningfully tag underlying genetic variation that would be missed with consideration of geography alone, highlighting the importance of culture in shaping genetic diversity. Leveraging our uniquely rich multi-generational ethnolinguistic metadata, we track language transmission through the pedigree, observing the disappearance of several languages in our cohort as well as notable shifts in frequency over three generations. We further find significantly higher language transmission rates for matrilineal groups as compared to patrilineal. We highlight both the diversity of variation within the African continent, as well as how within-Africa variation can be informative for broader variant interpretation; many variants appearing rare elsewhere are common in parts of Africa. The work presented here improves the understanding of the spectrum of genetic variation in African populations and highlights the enormous and complex genetic and ethnolinguistic diversity within Africa.

Keywords

Diverse populations; genotype; population genetics; linguistics; population structure

Introduction

Humans originated in Africa, resulting in more genetic variation on the African continent than anywhere else in the world; the average African genome has nearly a million more genetic variants than the average non-African person[1]. Africa is also immensely culturally and ethno-linguistically diverse; while the rest of the world averages 3.2 to 4.7 ethnic groups per country, African countries have an average of greater than 8 each and account in total for 43% of the world's ethnic groups[2]. Despite this diversity, African ancestry individuals are sorely underrepresented in genomic studies, making up only about 2% of GWAS participants[3,4]. Furthermore, the vast majority of African ancestry populations currently represented in genetic studies are African Americans or Afro-Caribbeans (72-93% in the GWAS catalog and ≥ 90% in gnomAD) with primarily West African ancestral origins[5]. These resources thus currently leave out the substantial diversity from regions of Africa that are disproportionately informative for human genetics.

Populations underrepresented in genetic studies contribute disproportionately to our understanding of biomedical phenotypes relative to European ancestry populations. Despite their paltry representation in GWAS, African ancestry populations contribute 7% of genome-wide significant associations[5,6]. African population genetic studies are especially informative given their unique evolutionary history, high level of genetic variation, and rapid linkage disequilibrium decay[7]. This Eurocentric bias in current genomics studies and resources also makes African descent individuals less likely to benefit from key genomic findings that do not translate fully across populations, contributing to health disparities[8]. In this study, we better characterize the immense genetic and ethnolinguistic diversity in four countries in eastern and southern Africa, offering insights into population history and structure in diverse African populations. Data are from 900 genotype samples that are part of the Neuropsychiatric Genetics of African Populations-Psychosis study (NeuroGAP-Psychosis), a major research and capacity building initiative in Ethiopia, Kenya, South Africa, and Uganda[9,10]

Genetic variation in Africa has been previously described as following not only isolation-by-distance expectations, but as being influenced by multiple interconnected ecological, historical, environmental, cultural, and linguistic factors[11–16]. These factors capture distinct variation from that tagged by genetics and can be informative for understanding population substructure. Better characterization of the ethnolinguistic composition of these samples is a key initial step towards running well-calibrated statistical genomics analyses including association studies. If ethnolinguistic variation tags additional structure than that captured by geography, explicit incorporation of relevant cultural information into such analyses tests may be the optimal strategy. In this study, we explore the genetics of Africa and how peoples' cultural affiliations and languages are related to genetic variation on the continent. We also explore ongoing cultural changes and consider the impact they will have on the genetics of Africa.

Results

Genetic Population Structure and Admixture

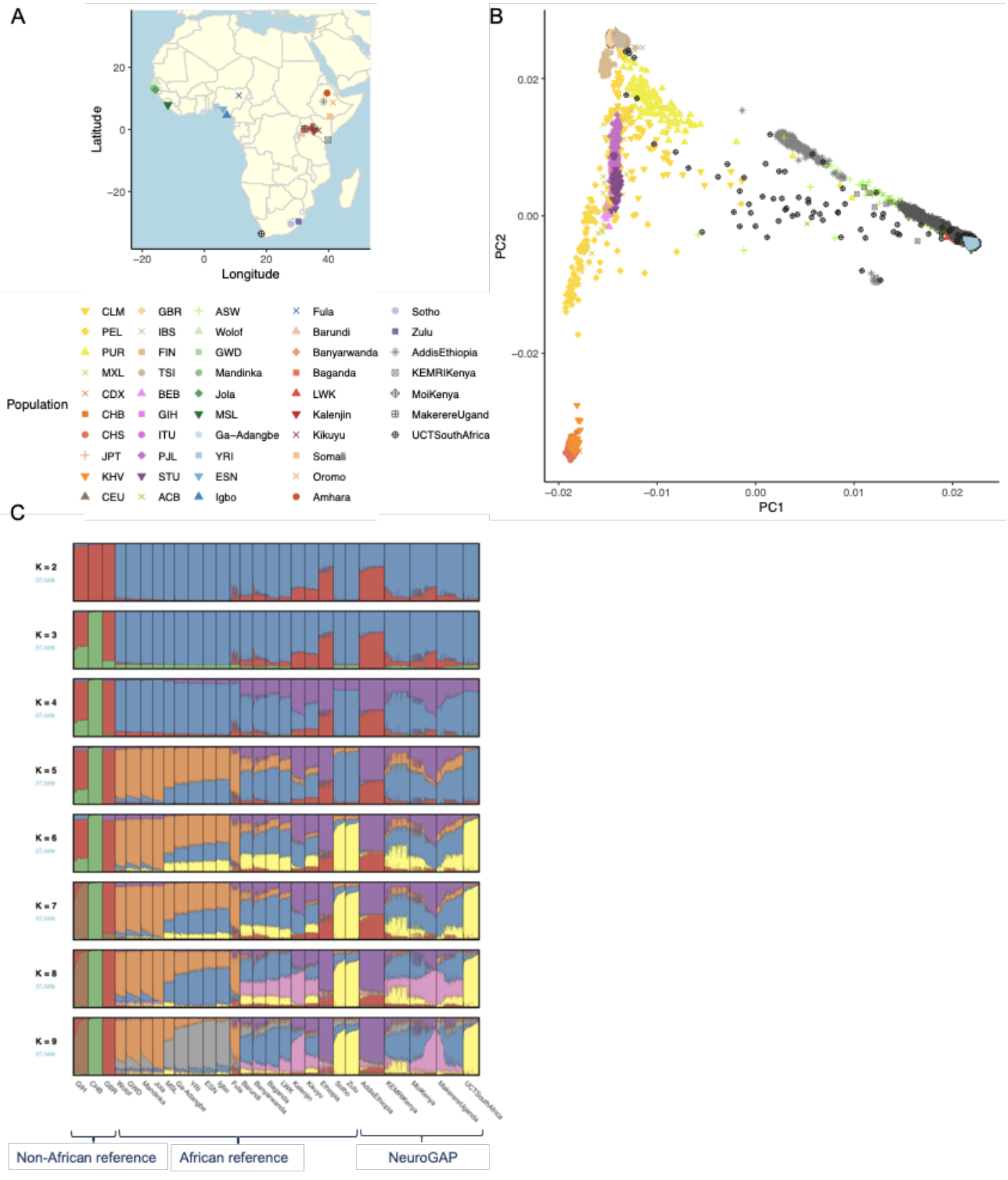
We compared the ancestral composition of our samples relative to global reference data from the 1000 Genomes Project and the African Genome Variation Project (AGVP) to see the full breadth of genetic diversity. Most NeuroGAP-Psychosis samples appear genetically similar to their geographically closest reference samples when compared to global datasets (**Figure 1**). However, large amounts of admixture is visible for some individuals, particularly among South African individuals (*Supplemental Information*). In South Africa, some individuals cluster wholly within the European reference cluster; this is expected based on the demographic composition of Cape Town, where these samples were collected, which is home to a substantial fraction of people of Dutch ancestry (Afrikaners) and individuals of mixed ancestry[12,15,17–19].

We additionally investigated the degree of admixture within samples and how genetic groups cluster in the data. We ran ADMIXTURE analyses, which partition genetic variation into a given number of distinct genetic clusters. This helps to visualize the groups that are most genetically distinct from one another, as each additional component can be thought of as representing the next most differentiated ancestry component in the

data, akin to principal components analysis (PCA). We identified the best fit *k* value, using five-fold cross validation, to be 9 using a tailored global reference.

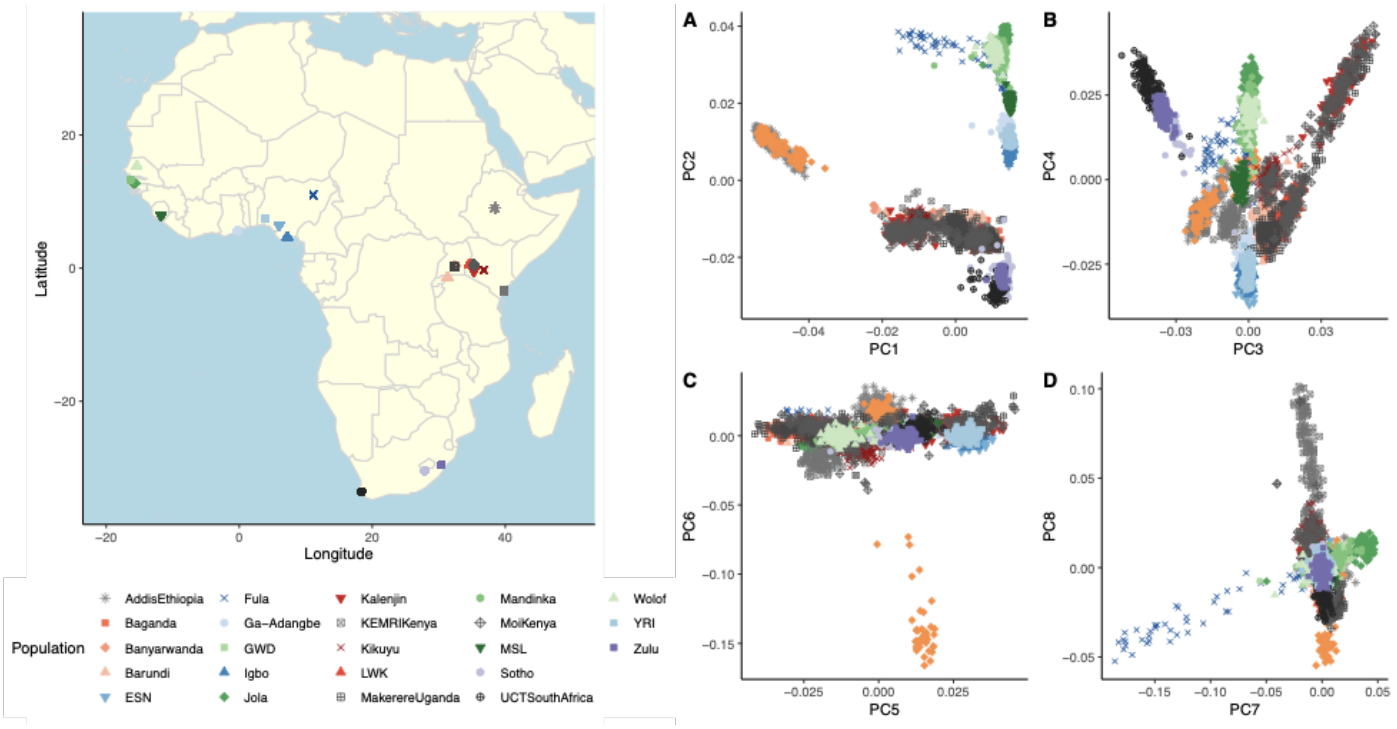
Examining the ancestry composition at the best fit *k*, we identify several ancestry components unique to areas within continental Africa (**Figure 1**C). Notably, several such components, including those unique to Ethiopia (purple), West Africa (orange), and South Africa (yellow) appear at earlier values of *k* than that separating South Asians from East Asians and Europeans (brown). While sample sizes affect the ordering of components identified in ADMIXTURE analyses, this suggests a high level of genetic differentiation between areas of the African continent rivaling that between those out-of-Africa continental ancestries, as has been previously demonstrated. We also note that Ethiopian participants have evidence of Eurasian admixture, possibly related to back-migration into the African continent[17,20–22].

Figure 1*. Genetic and admixture composition of the NeuroGAP-Psychosis samples against a global reference. A) Map showing the geographic location of African populations included, color coded by the ancestry components found to be unique to that region. B) First 2 principal components showing NeuroGAP-Psychosis samples as projected onto global variation of the full 1000 Genomes and AGVP. While most samples fall on a cline, some South African samples exhibit high amounts of admixture and European genetic ancestry. C) ADMIXTURE analysis for k=2 through 9 of all African reference and cohort samples as well as three representative non-African populations from the 1000 Genomes Project. GIH are the Gujarati Indian from Houston, Texas, CHB are the Han Chinese in Beijing, China, and the GBR are British in England and Scotland, which were included to capture South Asian, East Asian, and European admixture, respectively. Individuals are represented as bar charts sorted by population, and ancestry components for each person are visualized with different colors. The best supported k value with cross-validation was k=9.*



Projecting our samples onto PC space generated from only African reference samples, the top two principal components (PCs) separate geography, and more specifically East-West and North-South patterns of variation within Africa (**Figure 2**), mirroring our expectation of isolation by geographic distance in human genetic data. At higher PCs, however, there is fine-scale structure in the data separating different geographically proximal groups within the East African individuals, shown in red. We thus focus our deeper examinations into the East African samples to assess potential drivers of this differentiation (**Supplementary Figure 1**). For a detailed discussion of genetic variation within each country see the *Supplementary Information*.

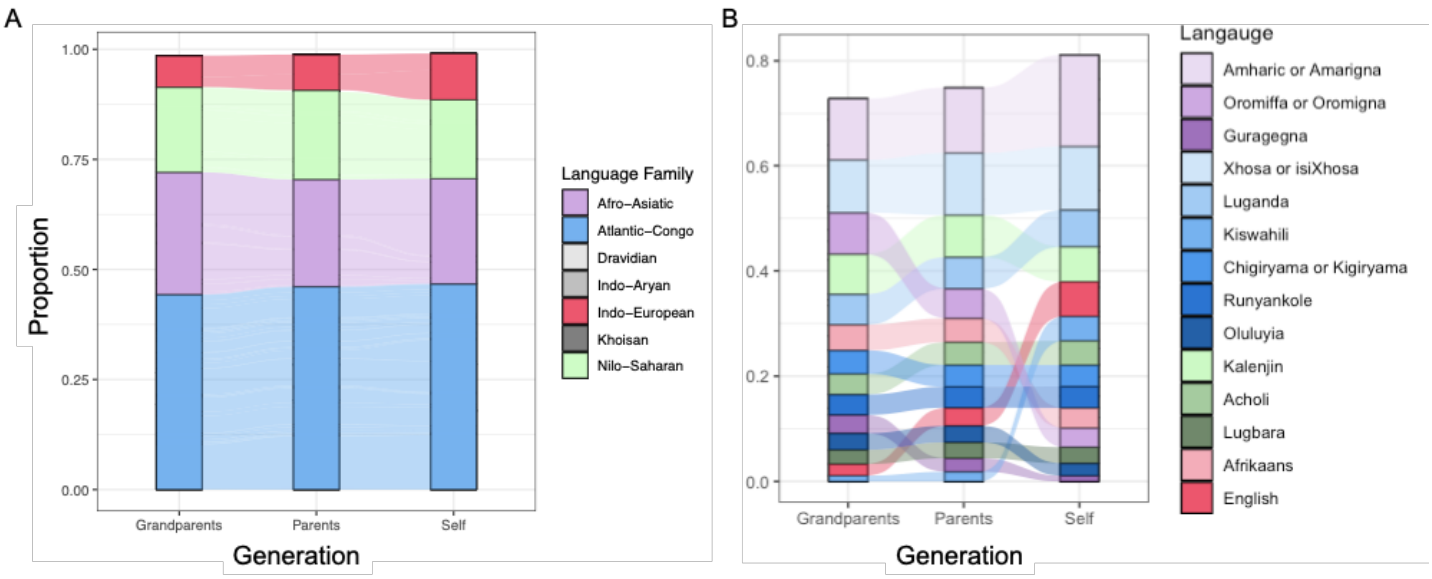
Figure 2*. Genetic composition of subcontinental African structure in the NeuroGAP-Psychosis samples. A-D: PCA biplots for PCs 1-8 with an African reference panel of 1000 Genomes Project AFR populations and the AGVP dataset. A map of collection locations is shown to the left of PCA plots.*



Self-reported Population Composition

Across samples with self-reported ethnolinguistic information, we observe 62 primary ethnicities and 107 primary languages in the 960 NeuroGAP-Psychosis samples, including missing data (Figure 3). We also find that languages have shifted in frequency over time, with English increasing in reporting frequency in the current generation, and several grandparental languages disappearing in our dataset (**Figure 3; Supplementary Figures 2-5**).

Figure 3*. Primary self-reported language shifts over three generations. A) Individual languages were re-classified into broader language families for comparable granularity. B) All languages reported with at least 3% frequency in any generation are shown across the generations. Note the increase in endorsement of English and drop in Oromiffa/Oromigna in the present generation.*



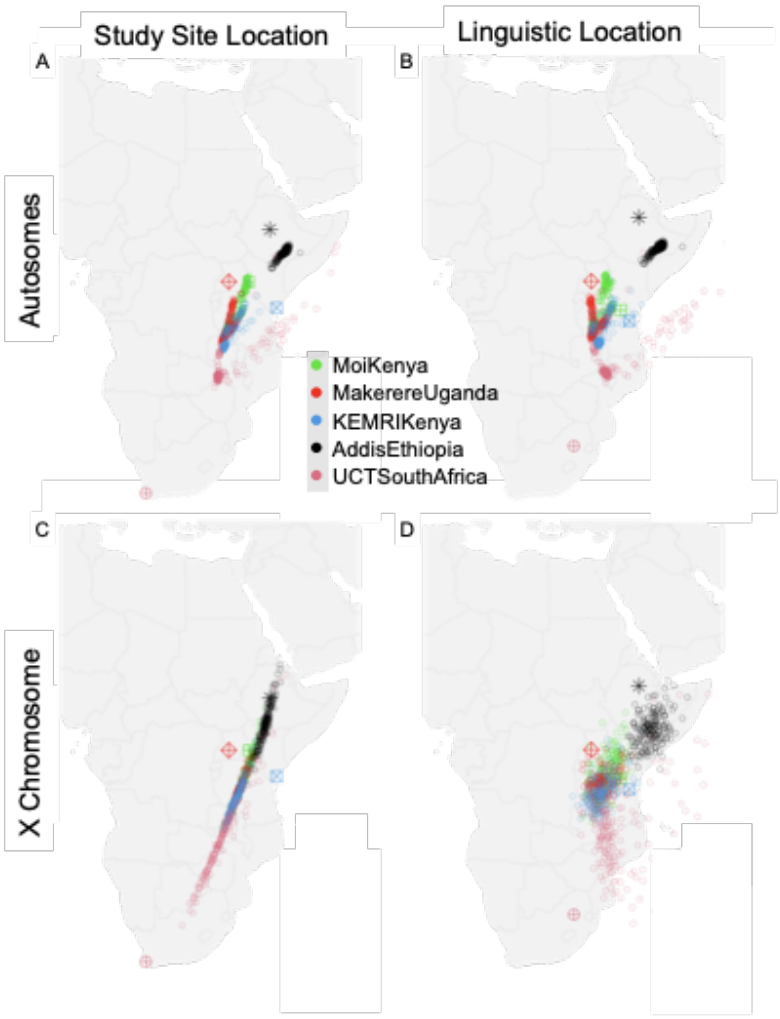
Genetic Variation Partitions with Language

To assess the correlation between the language that an individual reports to be their primary and the genetic partitioning that we observed, we conducted Procrustes analyses to measure the correlation between genetic, linguistic, and geographic variation. Procrustes analysis minimizes the distance between two sets of coordinates, so we can compare genetic variation reduced to two PCs to the location of each population. Using

the phonemes (units of sound) found in the self-reported languages of individuals and their families, along with the first two PCs of autosomal and X chromosome variation, we found consistent correlations between genetic, linguistic, and geographic variation throughout Africa (**Table 1**). Because the autosomes and X chromosomes have considerably different numbers of single nucleotide polymorphisms (SNPs), we additionally compared X chromosome variation to that from chromosome 22, which is most similar in SNP count to X (variant counts without/with reference panel: X = 603/1348, chr22 = 705/1455; **Supplementary Figure 6**). To measure linguistic variation, we queried the PHOIBLE 2.0 phonemic database[23], which contains phoneme inventories and phoneme qualities for many languages around the world. The resulting matrices of mean phoneme presences were used in a PCA to create three sets of linguistic PCs: from personally spoken languages of the participant, a combined score from those spoken by matrilineal relatives (mother and maternal grandmother), and a combined score from those of patrilineal relatives (father and paternal grandfather).

The first two PCs of both autosomal and X chromosome variation correlate more closely to geography (p=0.643 and 0.625 respectively; p<5E-5) than the first two PCs of linguistic variation (p=0.481; p<5E-5). Genetics are also correlated to linguistic variation to a lesser extent, and autosomal variation is consistently more strongly correlated to this linguistic variation than is X chromosome variation. When considering individuals from the entire dataset—Eastern and South Africa—patrilineal languages are more closely correlated to genetics than are matrilineal languages (by ~15%).

Figure 4*. Procrustes analyses indicate that autosomal genetic diversity is better correlated with geography than is X chromosome diversity. Plots represent the first two genetic PCs after a procrustes-transformation. The upper panels use PCs generated using autosomal variation, and the lower panels use X chromosome variation. The left column uses the locations of the study site at which each individual was sampled; the right column uses each individual's self-reported languages and the centroids of these languages to identify a geographic midpoint of that individual's languages. Individuals are colored by primary field site. For each primary field site, the midpoint of individuals' locations (by study site or languages spoken) is represented by a large point.*



Language Transmission Through Families

As we have detailed multi-generational ethnolinguistic information (see STAR *Methods* "*Ethnolinguistic Phenotypes*"), we computed overall transmission rates of language families over three generations. We initially examined the raw self-reported information of the participant with respect to the primary, second and third language spoken. We assessed the frequency with which the primary language reported by the participant matched each of their older relatives' (i.e. maternal and paternal grandparents, mother and father) as well as the frequency with which the participants' primary reported language matched that of the languages reported for their relative (**Table 2**). We find that transmission rates are similar between family members of the same generation when looking at primary language matching any language whether including or excluding English.

Partitioning East African individuals by the presence of matri- vs patri-lineneal transmission in their traditional societies (from Murdock's Ethnographic Atlas, code *EA076*[24,25]), we see a significantly higher transmission rate from individuals assigned to a matrilineal classification (p=0.028).

Testing for Evidence of Sex-biased Demography

To examine if there was evidence for sex-biased gene flow in our samples, we ran more Procrustes analyses comparing genetic and linguistic variation on the X chromosome as compared to the autosomes. We also assessed the similarity of ancestry proportions on the X chromosome versus autosomes. Ancestry fractions were highly correlated across these genomic regions, indicating no evidence for sex-biased demography at this scale (**Supplementary Figure 7**). Similarly, the Procrustes tests showed significant correlation between PCs 1 and 2 of X and autosomal variation (p=0.880 for all of Africa and p=0.884 for East Africa alone). Compared to chr22 instead, results were similar (p=0.836 for all Africa and p=0.841 for East Africa). Wilcoxon signed rank tests comparing the fractions of ancestry on X versus autosomes from ADMIXTURE at k=4 did not find a significant difference in the means, nor for PC1 vs PC2 (p = 0.3754).

Reference Panels Miss Meaningful Allele Frequency Resolution within Africa

We visualized allele frequencies for functionally important variants across our 5 collection sites as compared to reference data from the 1000 Genomes Project. One example variant, key in beta-thalassemia, dramatically varies in frequency depending on the precise location in Africa (**Supplementary Figure 8**). As this variant has direct consequences on human health, consideration of the difference in frequency across the continent is useful. For another example, rs72629486, a missense coding single nucleotide variant in the gene *ACTRT2,* ranges in minor allele frequency (MAF) in NeuroGAP-Psychosis from 5% in Ethiopia down to 1.3% in Uganda. This is nearly the full range of the frequency distribution for all global populations in the gnomAD database[26], which lists the variant in the AFR as 5.5%, missing finer resolution. rs72629486 is predicted to be deleterious and probably damaging by SIFT and PolyPhen, respectively, and has a combined annotation dependent depletion score of 22.9, highlighting that this variant is likely to be highly functionally important[27–29].

Discussion

Africa is a highly diverse continent, home to immense genetic, linguistic, and cultural diversity. This ethnolinguistic variation is extremely complex and is meaningful to disentangle prior to statistical genetics analyses. Here, we measured the correlation between genetic, linguistic, and geographic variation, finding that genetic and linguistic variation are closely correlated to each other as well as to geography across the African continent. This is consistent with previous work examining global patterns of diversity as well as the 'Bantu expansion', one of the largest demographic events in African history[11,12,16,30–32]. However, we find that in East Africa, language better separates genetic structure in our dataset than does geography (**Figure 1, Supplementary Figure 1**), a phenomenon that has been noted in Europe and Ethiopia previously[20,22,33,34]. This is notable, as most studies currently operate under the expectation of perfect isolation by distance. We find here that individuals collected from the same geographic location show significant genetic differentiation by language family, particularly in East Africa where there is immense linguistic diversity. This finding should influence how population substructure is controlled for in genetic tests, suggesting that a more nuanced treatment of genetic clusters with incorporation of ethnolinguistic classifications may sometimes be the most suitable approach. For example, future work exploring the direct incorporation of ethnolinguistic affiliations into linear mixed models would be useful, e.g. in the context of a kinship matrix equivalent[35].

As there is such immense genetic variation across the African continent[19,36–39], we highlight cases where such variability may be particularly informative. Africa is not simply one monolithic location, as it is sometimes treated in major genomics resources such as gnomAD allele frequency reports and the TOPMed dataset (data that include primarily or exclusively African Americans)[26,40]. Rather, there is an inordinate amount of genetic variability within it. These examples highlight both the diversity of variation within the African continent, as well as the fact that within-Africa variation can be informative for broader variant interpretation; many variants appearing rare elsewhere are common in parts of Africa.

As part of the NeuroGAP-Psychosis study's recruitment process, multi-generational self-reported ethnolinguistic data was collected from participants, including individual ethnicity and at least primary, second and third language from participants for themselves, as well as for each of their parents and grandparents. This provides us with an unusually rich depth of multigenerational demographic information from participants, a unique strength of our dataset that affords us the opportunity to investigate language transmission through the pedigree and shifts in language frequencies over time. First, we examined the overall change in self-reported language frequencies over three generations. Perhaps most striking is the increase in the reporting frequency of English by participants as their primary language as compared to their reports for older generations of their family. We also find that twelve languages reported for earlier generations were not spoken by the participants, indicating that they have disappeared from our dataset. Khoekhoe, Somali, and Urdu disappeared in the parental generation, and Amba, Afar, Argobba, Gumuz, Harar, Hindi, Soddo, Soo, and Tamil were no longer reported languages in the participants' generation. Interestingly, these languages represent a mix of both historically spoken and imported languages for the countries that enrolled participants in the NeuroGAP-Psychosis study. While these results are intriguing, we wish to stress that our participants are not necessarily representative of the local populations from which they come. A further consideration is a potential upwards bias towards reporting of English and Amharic as a primary language due to a preference towards reporting the language of consent as primary (consent form languages options increased over time; for example, in Ethiopia, initially only English and Amharic were offered), as well as towards languages taught in local educational systems. This additionally highlights the importance of careful consideration of items on self-report forms to ensure accurate and representative phenotype collection.

To take a closer look at language transmission across the pedigree, we calculated frequencies of transmission between various relatives in our family tree. In these calculations, we ran tests both including as well as excluding English in the event of such a potential upwards bias, and to get a better sense of transmission of languages that have been present in the continent for a longer period of time than recently imported languages. We additionally reclassified groups as being matrilineal or patrilineal using the database Ethnographic Atlas (EA)[24] and recalculated the transmission rates within those two classifications.

Matri/patrilineal implies the pattern of inheritance or the tracing of kinship and whether a child generally identifies more with the social system of the mother's or father's line. Interestingly, though our sample size for matrilineal groups is quite small (N=105 and 674 for matrilineal and patrilineal respectively), we find that there is a significantly higher language transmission rate for individuals assigned to matrilineal groups.

In summary, better understanding the composition of samples is a key first step to calibrating subsequent statistical genetics analyses. Cultural factors such as language can dramatically impact the structure of cohort data; we find that self-reported language classifications meaningfully tag underlying genetic variation that would be missed with consideration of geography alone. The work presented here improves the understanding of the immense spectrum of genetic and ethnolinguistic variation found across multiple African populations and sheds light on the shifts in language endorsement over the past three generations in five collection sites.

Tables

Table 1*. Procrustes correlations between genetics, geography, and language. All p < 5E-5. The first two PCs of autosomal and X chromosome variation were used for comparisons. Linguistic variation was calculated as a function of mean phoneme presence across all languages reported by the individual across their pedigree.*

			Languages spoken by		
Subset of individuals	PCs 1 & 2: Genetic	Geography	Self	Mother & Maternal Grandmother	Father & Paternal Grandfather
All individuals	Autosomal	0.6327	0.450	0.3167	0.3764
	X chr.	0.6248	0.423	0.3046	0.3713
East Africa	Autosome	0.7734	0.627	0.5616	0.5648
	X chr.	0.6810	0.585	0.5423	0.5304

***Table 2**. Language transmission rates from relatives.*

Frequency of a participants' reported primary language matching one of the top three reported languages spoken by relatives. Rates were calculated both with and without excluding English. In East Africa, individuals were additionally partitioned by their affiliation with either a matrilineal vs patrilineal ethnolinguistic group.

	Transmission Rate			
Family Member	**All**	**Excl. English**	**Patrilineal (E. Africa)**	**Matrilineal (E. Africa)**
Father	0.810	0.818	0.837	0.871
Mother	0.802	0.809	0.811	0.800
Paternal grandfathers	0.778	0.775	0.726	0.926
Paternal grandmothers	0.773	0.767	0.738	0.939
Maternal grandfathers	0.762	0.758	0.708	0.903
Maternal grandmothers	0.758	0.753	0.726	0.812

STAR Methods

Collection Strategy

As described in more detail in the published protocol[9], NeuroGAP-Psychosis is a case-control study recruiting participants from more than two dozen hospitals and medical clinics in Ethiopia, Kenya, South Africa, and Uganda. Participants are recruited in languages in which they are fluent, including Acholi, Afrikaans, Amharic, English, Kiswahili, Luganda, Lugbara, Oromiffa/Oromigna, Runyankole, and isiXhosa. After consenting to be in the study, participants give a saliva sample using an Oragene kit (OG-500.005) for DNA extraction. Study staff then ask a range of questions on demographics, mental health, and physical health and take the participants' blood pressure, heart rate, height, and weight. The whole study visit lasts approximately 60-90 minutes.

344

345 *Ethnolinguistic Phenotypes*

346 Multiple phenotypes related to self-reported ethnolinguistic categorizations have been collected as part of the
347 recruitment process. This includes multi-generational data including each participants' primary, secondary and
348 tertiary language and ethnicity, and birth country. All linguistic data were collected from participants both for
349 themselves as well as for each of their parents and grandparents, giving an unusually rich depth of information.
350 The specific phrasing of questions collected are as follows:
351 Primary language: "What primary language do you speak?"
352 2nd language: "What 2nd language do you speak?"
353 3rd language: "What 3rd language do you speak?"
354 Primary ethnicity: "What is your ethnicity or tribe?"
355 2nd ethnicity: "What is your ethnicity or tribe?"
356 3rd ethnicity: "What is your ethnicity or tribe?"
357
358 Reports for other relatives followed similar phrasing. The primary language question for each is listed, with
359 primary swapped for '2nd' or '3rd' for the second and third reported languages for that family member.
360 Mother: "What was the primary language that your biological mother spoke?"
361 Father: "What was the primary language that your biological father spoke?"
362 Maternal grandmother: "What primary language did your biological mother's mother speak?"
363 Maternal grandfather: "What primary language did your biological mother's father speak?"
364 Paternal grandmother: "What primary language did your biological father's mother speak?
365 Paternal grandfather: "What primary language did your biological father's father speak?
366

367 *Genetic Data Quality Control*

368 Quality control (QC) procedures for NeuroGAP-Psychosis data were done using the Hail python library
369 (www.Hail.is). All of the data was stored on Google Cloud. The QC steps and filters used were adapted from

Ricopili[41] and Anderson et al. 2011[42]. The data was genotyped using the Illumina Global Screening Array. For each of the five NeuroGAP-Psychosis sites, a vcf with genotyping data was stored on Google Cloud. Before QC, each data vcf contained 192 samples and 687537 variants. When looking at the data pre-QC we discovered elevated deviations in Hardy Weinberg Equilibrium. We found that the metric which outlined the individuals causing these deviations was called autocall call rate, Illumina's custom genotype calling algorithm (See Supplementary Information). The QC filtering steps took place after removing individuals with an autocall call rate less than .95. 937 of the original 960 individuals remained. These 960 individuals were used for the linguistic transmission analyses presented here, while for genetic analyses further QC on variants was conducted.

The site vcfs were imported as Hail matrix tables and annotated with appropriated data from the metadata file before being merged. The resulting matrix table had 937 samples and 687537 variants. Prior to QC, the joint dataset was split into autosomes, PAR, and nonPAR regions of the X chromosome. QC filtering was conducted separately for the autosome and X chromosome regions. Pre-QC, the autosomal dataset had 937 samples and 669346 variants. The following is a list of the QC steps and parameters used for autosomal QC. (1) Removing variants with a call rate less than 95%. After filtering, 638235 variants remained. (2) Removing individuals with a call rate less than 98%. After filtering, 930 individuals remained. (3) Removing individuals whose reported sex did not match their genotypic sex. After filtering, 923 individuals remained. (4) Removing variants with a minor allele frequency less than 0.5%. After filtering, 360,321 variants remained. (5) Removing variants with a Hardy Weinberg Equilibrium p-value less than 1×10^{-3}. After filtering, 331667 variants remained. (6) Using PC-Relate with 10 PCs, removing individuals with a kinship coefficient greater than .125. After filtering, 900 individuals remained. After autosomal QC, 900 individuals and 331667 variants remained.

The PAR and nonPAR regions of the X chromosome were subset to the 900 samples which passed autosomal QC before going through variant QC. The same variant thresholds used for autosomal QC were used to conduct QC on the PAR region. Pre-QC the PAR region dataset had 900 samples and 518 variants. (1) After SNP call rate filtering, 515 variants remained. (2) After MAF filtering, 411 variants remained. (3) After HWE

filtering, 402 variants remain. Post-QC, the PAR region had 900 samples and 402 variants. For the nonPAR region, the dataset was split by sex. The female nonPAR dataset had 441 samples and 17673 variants. Variant QC was carried out on the females using the following metrics. (1) Removing variants with a call rate less than 98%. After filtering, 16261 variants remained. (2) Removing variants with a minor allele frequency less than 1%. After filtering, 11113 variants remained. (3) Removing variants with a Hardy Weinberg Equilibrium p-value less than 1×10^{-6}. After filtering, 11104 variants remained. After nonPAR QC on the females, the male nonPAR dataset was merged with the female QC'd nonPAR dataset. The final nonPAR dataset had 900 samples and 11104 variants. After filtering, the three datasets were merged into one matrix table. The final merged, post-QC dataset contained 900 samples and 343173 variants and was written out to vcf and plink format for further analyses. The counts of variants/individuals per site after autosomal and X QC can be found in **Supplementary Tables 1-2**.

After QC, the dataset was merged with two different reference panel datasets, the 1000 Genomes Project (TGP)[43] and the AGVP[38]. Before merging the datasets, AGVP had 1297 samples and 1778578 variants while TGP had 2504 samples and 17892192 variants. Before these two datasets were merged the variants in the AGVP dataset were flipped using the plink command --flip. In addition, indels were removed from the TGP dataset, and variants with more than 3 alleles were removed from the AGVP dataset. After removing triallelic sites from the AGVP dataset, there were 1297 samples and 1771279 variants. After removing indels from the TGP dataset, there were 16101868 variants and 2504 samples. After merging the two reference panels, there were 3801 samples and 16194904 variants. After the two reference datasets were merged, --geno filter from plink was run with .05 threshold to remove variants which had missing genotype call rates greater than 95%. After this filter, 1677440 variants and 3801 samples remained. Lastly, related individuals were removed from the merged AGVP and 1000 Genome dataset. The final dataset had 3784 samples and 1677440 variants.

The reference dataset was then merged with the postQC NeuroGAP-Psychosis dataset containing both autosomal and X chromosome data. Before merging with the reference panel the NeuroGAP-Psychosis dataset had 900 samples and 343173 variants. Variants with more than 3 alleles were removed from the

NeuroGAP-Psychosis dataset. After this the dataset had 343166 variants. After merging the NeuroGAP-Psychosis dataset with the AGVP+TGP reference panel the dataset contained 4684 samples and 1814839 variants. A --geno filter with .05 was run on the merged dataset. After the filter, 4684 samples and 205767 variants remained. Our processing pipeline is freely available at: https://github.com/atgu/NeuroGAP/tree/master/PilotDataQC.

Population Structure and Admixture Analyses

Cohort data from the five NeuroGAP-Psychosis plates were merged with African reference populations from the 1000 Genomes Project[43] and the African Genome Variation Project[38]. These populations provide reasonably comprehensive geographic coverage across the African continent from currently available reference panels and contain populations which are co-located in the same countries as all NeuroGAP-Psychosis samples. PCA was run using flashPCA[44]. Detailed examination of admixture was conducted using the program ADMIXTURE[45] with five-fold cross validation error to inform the correct number of clusters. Plots from ADMIXTURE output were generated with pong[46]. ADMIXTURE was run using a tailored representation of global genetic data consisting of all continental African populations, the CHB population from China to capture East Asian admixture, the GBR from Britain to capture European admixture, and the GIH from India to capture South Asian ancestry. Fst estimates across populations were generated using smartPCA[47]. Fst heatmaps were generated in R using the package *corrplot*. The relationship between ancestry composition on the autosomes vs X chromosome was examined using Pearson correlation and mantel tests in R with the package *ade4*. Frequency plots of variants across the globe were created with the GGV browser[48].

Relationship between Genetics and Language

To measure linguistic variation, we made use of the PHOIBLE 2.0 phonemic database[23], which contains phoneme inventories and phoneme qualities for languages around the world. For every individual, we identified all languages spoken—excluding English—which were present in the PHOIBLE database (84.5% of languages spoken by the individuals themselves, and 81.1% of languages spoken by their relatives). Using the phoneme

inventories (including both primary phonemes and their allophones) from PHOIBLE, we found the mean phoneme presence for each individual's or each relative's spoken languages. The resulting matrices (of individuals or their relatives, and mean phoneme presences) were used for PCAconducted in R to create three sets of principal components (PCs): from personally spoken languages, from those spoken by matrilineal relatives (mother and maternal grandmother), and from those of patrilineal relatives (father and patrilineal grandfather).

First, all languages were assigned the highest-level classifications available in Glottolog 4.2.1[49]. These classifications were modified to minimize the number of high-level classifications while maintaining an element of geographic origin. Several classifications were consolidated into Nilo-Saharan (made up of Nilotic, Central Sudanic, Kuliak and Gamuz classifications) and Khoisan (Khoe-Kwadi, Kxa, and Tuu), and Afro-Asiatic was expanded (with Ta-Ne-Omotic and Dizoid). Indo-European was split to account for the recent history of its speakers: Afrikaans and Oorlams were placed into a unique category, the languages of Europe into another, and those of the Indian subcontinent (Hindi and Urdu) into a third. We excluded languages that were unclassified or identified as speech registers.

Every individual was associated with a survey location, meaning the geographic coordinates where the sample was collected, and we used the spoken languages to assign a different, linguistic location. To do this, using all languages an individual spoke, and these languages' locations from Glottolog, we calculated the mean location of each individual's languages.

To compare linguistic, genetic, and geographic variation, we used a set of Procrustes analyses implemented in R[50]. For linguistic and genetic variation, the first two PCs of variation were used. Since Procrustes minimizes the sum of squared euclidean distances, the geographic coordinates of each individual were converted to points on a sphere. To measure the correlation between geographic variation and linguistic or genetic variation, the latter were transformed (via rotation and scaling) to minimize the sum of squared distance

476 between individuals' locations and the transformed genetic or linguistic PCs. The first two PCs of procrustes-
477 transformed linguistic and genetic variation—representing their similarity to geographic variation—were then
478 plotted onto a map.

479

480 *Anthropological variables*

481 To identify relevant anthropological data, we accessed data from the Ethnographic Atlas (EA)[24] using D-
482 Place[25]. We associated each ethnicity reported in the NeuroGAP-Psychosis survey data to a society in the EA
483 (if possible), and used two available variables (*EA012: Marital residence with kin*, and *EA076: Inheritance rule*
484 *for movable property*). For ethnicities with data, individuals whose ethnicities were associated with consistent
485 inheritance rules or marital residence patterns were assigned that rule or pattern. Of the 907 NeuroGAP-
486 Psychosis individuals, 751 were assigned a marital residence pattern (patrilocal, neolocal, or virilocal-like) and
487 779 were assigned an inheritance rule (matrilineal or patrilineal).

488

489

References

1. Consortium T 1000 GP, The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes [Internet]. Vol. 491, Nature. 2012. p. 56–65. Available from: http://dx.doi.org/10.1038/nature11632

2. Fearon JD. Ethnic and Cultural Diversity by Country*. J Econ Growth. 2003 Jun 1;8(2):195–222.

3. Sirugo G, Williams SM, Tishkoff SA. The Missing Diversity in Human Genetic Studies. Cell. 2019 May 2;177(4):1080.

4. Popejoy AB, Fullerton SM. Genomics is failing on diversity. Nature. 2016 Oct 13;538(7624):161–4.

5. Martin AR, Teferra S, Möller M, Hoal EG, Daly MJ. The critical needs and challenges for genetic architecture studies in Africa. Curr Opin Genet Dev. 2018 Dec;53:113–20.

6. Morales J, Welter D, Bowler EH, Cerezo M, Harris LW, McMahon AC, et al. A standardized framework for representation of ancestry data in genomics studies, with application to the NHGRI-EBI GWAS Catalog. Genome Biol. 2018 Feb 15;19(1):21.

7. Tishkoff SA, Verrelli BC. Patterns of human genetic diversity: implications for human evolutionary history and disease. Annu Rev Genomics Hum Genet. 2003;4:293–340.

8. Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ. Clinical use of current polygenic risk scores may exacerbate health disparities. Nat Genet. 2019 Apr;51(4):584–91.

9. Stevenson A, Akena D, Stroud RE, Atwoli L, Campbell MM, Chibnik LB, et al. Neuropsychiatric Genetics of African Populations-Psychosis (NeuroGAP-Psychosis): a case-control study protocol and GWAS in Ethiopia, Kenya, South Africa and Uganda. BMJ Open. 2019 Feb 19;9(2):e025469.

10. van der Merwe C, Mwesiga EK, McGregor NW, Ejigu A, Tilahun AW, Kalungi A, et al. Advancing neuropsychiatric genetics training and collaboration in Africa. The Lancet Global Health. 2018 Mar 1;6(3):e246–7.

11. Baker JL, Rotimi CN, Shriner D. Human ancestry correlates with language and reveals that race is not an objective genomic classifier. Sci Rep. 2017 May 8;7(1):1572.

12. Uren C, Kim M, Martin AR, Bobo D, Gignoux CR, van Helden PD, et al. Fine-Scale Human Population Structure in Southern Africa Reflects Ecogeographic Boundaries. Genetics. 2016 Sep;204(1):303–14.

13. Henn BM, Botigué LR, Gravel S, Wang W, Brisbin A, Byrnes JK, et al. Genomic ancestry of North Africans supports back-to-Africa migrations. PLoS Genet. 2012 Jan;8(1):e1002397.

14. Henn BM, Botigué LR, Peischl S, Dupanloup I, Lipatov M, Maples BK, et al. Distance from sub-Saharan Africa predicts mutational load in diverse human genomes. Proc Natl Acad Sci U S A. 2016 Jan 26;113(4):E440–9.

15. Sikora M, Laayouni H, Calafell F, Comas D, Bertranpetit J. A genomic analysis identifies a novel component in the genetic structure of sub-Saharan African populations. Eur J Hum Genet. 2011 Jan;19(1):84–8.

16. Creanza N, Ruhlen M, Pemberton TJ, Rosenberg NA, Feldman MW, Ramachandran S. A comparison of worldwide phonemic and genetic variation in human populations. Proc Natl Acad Sci U S A. 2015 Feb 3;112(5):1265–72.

17. Pickrell JK, Patterson N, Loh P-R, Lipson M, Berger B, Stoneking M, et al. Ancient west Eurasian ancestry in southern and eastern Africa [Internet]. Vol. 111, Proceedings of the National Academy of Sciences. 2014. p. 2632–7. Available from: http://dx.doi.org/10.1073/pnas.1313787111

18. Chimusa ER, Meintjies A, Tchanga M, Mulder N, Seoighe C, Soodyall H, et al. A genomic portrait of haplotype diversity and signatures of selection in indigenous southern African populations. PLoS Genet. 2015 Mar;11(3):e1005052.

19. Bergström A, McCarthy SA, Hui R, Almarri MA, Ayub Q, Danecek P, et al. Insights into human genetic variation and population history from 929 diverse genomes. Science [Internet]. 2020 Mar 20 [cited 2020 Mar 19];367(6484). Available from: https://science.sciencemag.org/content/367/6484/eaay5012/tab-pdf

20. López S, Tarekegn A, Band G, van Dorp L, Bird N. The genetic landscape of Ethiopia: diversity, intermixing and the association with culture. bioRxiv [Internet]. 2021; Available from: https://www.biorxiv.org/content/10.1101/756536v2.abstract

21. Pagani L, Schiffels S, Gurdasani D, Danecek P, Scally A, Chen Y, et al. Tracing the route of modern humans out of Africa by using 225 human genome sequences from Ethiopians and Egyptians. Am J Hum Genet. 2015 Jun 4;96(6):986–91.

22. Pagani L, Kivisild T, Tarekegn A, Ekong R, Plaster C, Gallego Romero I, et al. Ethiopian genetic diversity reveals linguistic stratification and complex influences on the Ethiopian gene pool. Am J Hum Genet. 2012 Jul 13;91(1):83–96.

23. Moran S, McCloy D. PHOIBLE 2.0. Jena: Max Planck Institute for the Science of Human History. 2019;

24. Murdock GP, Textor R, Barry H III, White DR, Gray JP, Divale W. 2000. Ethnographic atlas. World Cultures. 1999;10(1):24–136.

25. Kirby KR, Gray RD, Greenhill SJ, Jordan FM, Gomes-Ng S, Bibiko H-J, et al. D-PLACE: A Global Database of Cultural, Linguistic and Environmental Diversity. PLoS One. 2016 Jul 8;11(7):e0158391.

26. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020 May;581(7809):434–43.

27. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res. 2003 Jul 1;31(13):3812–4.

28. Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Hum Genet. 2013 Jan;Chapter 7:Unit7.20.

29. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. 2019 Jan 8;47(D1):D886–94.

30. de Filippo C, Bostoen K, Stoneking M, Pakendorf B. Bringing together linguistic and genetic evidence to

test the Bantu expansion. Proc Biol Sci. 2012 Aug 22;279(1741):3256–63.

31. Beleza S, Gusmão L, Amorim A, Carracedo A, Salas A. The genetic legacy of western Bantu migrations. Hum Genet. 2005 Aug;117(4):366–75.

32. Li S, Schlebusch C, Jakobsson M. Genetic variation reveals large-scale population expansion and migration during the expansion of Bantu-speaking peoples. Proc Biol Sci [Internet]. 2014 Oct 22;281(1793). Available from: http://dx.doi.org/10.1098/rspb.2014.1448

33. Longobardi G, Ghirotto S, Guardiano C, Tassi F, Benazzo A, Ceolin A, et al. Across language families: Genome diversity mirrors linguistic variation within Europe: Genome Diversity Across Language Families. Am J Phys Anthropol. 2015 Aug;157(4):630–40.

34. Piazza A, Rendine S, Minch E, Menozzi P, Mountain J, Cavalli-Sforza LL. Genetics and the origin of European languages. Proc Natl Acad Sci U S A. 1995 Jun 20;92(13):5836–40.

35. Heckerman D, Gurdasani D, Kadie C, Pomilla C, Carstensen T, Martin H, et al. Linear mixed model for heritability estimation that explicitly addresses environmental variation. Proc Natl Acad Sci U S A. 2016 Jul 5;113(27):7377–82.

36. 1000 Genomes Project Consortium, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, et al. A map of human genome variation from population-scale sequencing. Nature. 2010 Oct 28;467(7319):1061–73.

37. Choudhury A, Aron S, Botigué LR, Sengupta D, Botha G, Bensellak T, et al. High-depth African genomes inform human migration and health. Nature. 2020 Oct;586(7831):741–8.

38. Gurdasani D, Carstensen T, Tekola-Ayele F, Pagani L, Tachmazidou I, Hatzikotoulas K, et al. The African Genome Variation Project shapes medical genetics in Africa. Nature. 2015 Jan 15;517(7534):327–32.

39. Gurdasani D, Carstensen T, Fatumo S, Chen G, Franklin CS, Prado-Martinez J, et al. Uganda Genome Resource Enables Insights into Population History and Genomic Discovery in Africa. Cell. 2019 Oct

583 31;179(4):984–1002.e36.

584 40. Taliun D, Harris DN, Kessler MD, Carlson J, Szpiech ZA, Torres R, et al. Sequencing of 53,831 diverse
585 genomes from the NHLBI TOPMed Program. Nature. 2021 Feb 10;590(7845):290–9.

586 41. Lam M, Awasthi S, Watson HJ, Goldstein J, Panagiotaropoulou G, Trubetskoy V, et al. RICOPILI: Rapid
587 Imputation for COnsortias PIpeLIne. Bioinformatics [Internet]. 2019 Aug 8; Available from:
588 http://dx.doi.org/10.1093/bioinformatics/btz633

589 42. Anderson CA, Pettersson FH, Clarke GM, Cardon LR, Morris AP, Zondervan KT. Data quality control in
590 genetic case-control association studies. Nat Protoc. 2010 Sep;5(9):1564–73.

591 43. Auton A, Salcedo T. The 1000 Genomes Project [Internet]. Assessing Rare Variation in Complex Traits.
592 2015. p. 71–85. Available from: http://dx.doi.org/10.1007/978-1-4939-2824-8_6

593 44. Abraham G, Qiu Y, Inouye M. FlashPCA2: principal component analysis of Biobank-scale genotype
594 datasets. Bioinformatics. 2017 Sep 1;33(17):2776–8.

595 45. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals.
596 Genome Res. 2009 Sep;19(9):1655–64.

597 46. Behr AA, Liu KZ, Liu-Fang G, Nakka P, Ramachandran S. pong: fast analysis and visualization of latent
598 clusters in population genetic data. Bioinformatics. 2016 Sep 15;32(18):2817–23.

599 47. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS Genet. 2006
600 Dec;2(12):e190.

601 48. Marcus JH, Novembre J. Visualizing the geography of genetic variants. Bioinformatics. 2017 Feb
602 15;33(4):594–5.

603 49. Hammarström H, Forkel R, Haspelmath M, Bank S. glottolog/glottolog: Glottolog database 4.2.1 [Internet].
604 2020. Available from: https://zenodo.org/record/3754591

605 50. Dixon P. VEGAN, a package of R functions for community ecology. J Veg Sci. 2003 Dec 9;14(6):927–30.

Acknowledgements

We thank all participants for their willingness to contribute their data to this effort. John Mugane and Ana Maria Olivares provided advice and assistance related to this project. This work was supported by funding from the NIH/National Institute of Mental Health (K01MH121659 and T32MH017119 to E.G.A.; K99/R00MH117229 to A.R.M.; L.B.C., K.C.K, B.G., D.J.S, S.T., and D.A. are supported in part by R01MH120642).

Author Contributions

Conceptualization, E.GA.; A.R.M.; K.C.K.; Formal Analysis, E.G.A., S.D., Y.P., A.K., L.M.; Data Curation, M.B., Z.K., C.P.N., A.S., R.E.S; Investigation: S.G., J.K., R.J., R.M.M., M.A., W.E.I., G.K., W.S., F.K.A., H.M., L.M.; Resources, T.A., D.A., M.A., F.K.A., L.A., A.F., S.G., W.E.I., R.J., S.M.K., G.K., E.K., J.K., H.M., R.M.M., C.R.J.N., R.R., W.S., D.J.S., S.T., Z.Z.; Writing - Original Draft, E.G.A., A.R.M.; Writing - Review & Editing, all authors; Visualization, E.G.A., S.D., Y.P., A.K., L.M., Z.K.; Supervision, N.C., M.J.D, B.M.N., K.C.K., S.R., A.S., T.A., D.A., L.A., A.F., W.E.I., S.M.K., G.K., E.K., C.P.N., C.R.J.N., R.R., D.J.S., R.E.S., S.T., Z.Z., L.B.C., B.G.; Project Administration, A.S., R.E.S., M.A., S.G., R.J., J.K., L.O., R.M.M., C.P.N.; Funding Acquisition, M.J.D, B.M.N., K.C.K.

Human Subjects Approval

Ethical clearances to conduct this study have been obtained from all participating sites, including:

· Ethiopia: Addis Ababa University College of Health Sciences (#014/17/Psy) and the Ministry of Science and Technology National Research Ethics Review Committee (#3.10/14/2018).

· Kenya: Moi University School of Medicine Institutional Research and Ethics Committee (#IREC/2016/145, approval number: IREC 1727), Kenya National Council of Science and Technology (#NACOSTI/P/17/56302/19576), KEMRI Centre Scientific Committee (CSC# KEMRI/CGMRC/CSC/070/2016), KEMRI Scientific and Ethics Review Unit (SERU# KEMRI/SERU/CGMR-C/070/3575)

· South Africa: The University of Cape Town Human Research Ethics Committee (#466/2016)

· Uganda: The Makerere University School of Medicine Research and Ethics Committee (SOMREC #REC REF 2016-057) and the Uganda National Council for Science and Technology (UNCST #HS14ES)

· USA: The Harvard T.H. Chan School of Public Health (#IRB17-0822)

Data and Code Availability Statement

The genetic data generated during this study for NeuroGAP-Psychosis samples will be made available on dbGAP (in process). Code used to process and analyze data is freely available here: https://github.com/atgu/NeuroGAP.

Declaration of Interests

A.R.M. has consulted for 23andMe and Illumina and received speaker fees from Genentech, Pfizer, and Illumina. B.M.N. is a member of the Deep Genomics Scientific Advisory Board. He also serves as a consultant for the Camp4 Therapeutics Corporation, Takeda Pharmaceutical and Biogen. M.J.D. is a founder of Maze Therapeutics. The remaining authors declare no competing interests.