The American Journal of Human Genetics, Volume *109*

Supplemental information

Genetic structure correlates with ethnolinguistic diversity in eastern and southern Africa

Elizabeth G. Atkinson, Shareefa Dalvie, Yakov Pichkar, Allan Kalungi, Lerato Majara, Anne Stevenson, Tamrat Abebe, Dickens Akena, Melkam Alemayehu, Fred K. Ashaba, Lukoye Atwoli, Mark Baker, Lori B. Chibnik, Nicole Creanza, Mark J. Daly, Abebaw Fekadu, Bizu Gelaye, Stella Gichuru, Wilfred E. Injera, Roxanne James, Symon M. Kariuki, Gabriel Kigen, Nastassja Koen, Karestan C. Koenen, Zan Koenig, Edith Kwobah, Joseph Kyebuzibwa, Henry Musinguzi, Rehema M. Mwema, Benjamin M. Neale, Carter P. Newman, Charles R.J.C. Newton, Linnet Ongeri, Sohini Ramachandran, Raj Ramesar, Welelta Shiferaw, Dan J. Stein, Rocky E. Stroud, Solomon Teferra, Mary T. Yohannes, Zukiswa Zingela, Alicia R. Martin, and NeuroGAP-Psychosis Study Team

Supplemental Information

Supplemental Notes

Finer scale description of genetic structure across NeuroGAP-Psychosis countries

Ethiopia: The pilot data from Addis Ababa University (AAU) falls cleanly within the Ethiopian reference panel cluster, as would be expected by the collection location in Addis. This also matched with the fact that the majority of the participants' self-reported languages were Amhara and Oromo, and we have reference panels from these corresponding ethnic groups from the AGVP. Individuals from Ethiopia tend to be quite genetically distinct from people from other areas of Africa, pulling out a unique ancestral component at K=4, immediately after the separation of European and east Asian individuals from Africa. They also appear to have some European admixture, visible as the red component in ADMIXTURE plots (Figure 1A). This may be related to back-migration into the continent [1–4].

Kenya: The pilot data from Moi University falls within the East African cluster, as would be expected by the collection location in Eldoret (Figure 1B). Furthermore, it seems to fall with the Kalenjin and Luhya ("LWK") groups primarily, which are the most common self-reported ancestry that participants reported in these 192 samples (Figure 1). Interestingly, two geographically close East African populations (shown in red) dispersed into distinct clusters, which by PC5 define that axis of variation. We next investigated features that might explain this differentiation between closely geographically oriented groups. The two distinct red East African groups appear to speak different languages, one speaking an Afro-Asiatic language and one a Niger-Congo, such they function as reasonably independent groups genetically even though they are in very close geographic proximity to one another.

The pilot data from the KEMRI-Wellcome Trust also overlaps roughly with the East African reference panels, but the core of the pilot samples do not lie squarely on the reference panels. There are a couple of reasons why this might be happening: 1) the reference panels for Kenya are from the Kalenjin and Luhya ("LWK") groups, which are from western Kenya and geographically far away from Kilifi where the participants were recruited. 2) Due to the history of coastal Kenya, there is likely a lot of admixture between people who originated from the coast and people of Arabic ancestry. Unfortunately, there are no reference panels from East African coastal populations or from the Arabian Peninsula. 3) There could be a technical error with the data.

Uganda: The pilot data from Makerere University in Uganda also falls cleanly within the East African cluster, as would be expected by the collection location in Kampala. Furthermore, it seems to fall with the Baganda ethnic group primarily, which is the most common self-reported ancestry that participants reported in these 192 samples. We note the breakdown of Ugandan samples by language group in a similar fashion to what we observed in Kenya, and have included them in more detailed analyses of the correlation between genetic similarity and language family divergence.

South Africa: The pilot data from the University of Cape Town (UCT) falls most closely to the South African reference panels (in purple) on PC space. However, the core of the pilot samples do not lie squarely on the reference panels. There are several possible explanations for this: 1) the reference panels for South Africa are from the Zulu and the Sotho groups, which are in eastern South Africa and geographically far away from Cape Town and other locations, where the participants were recruited. 2) Cape Town is inhabited by people all over Africa and the world and there are many immigrants living there. Since NeuroGAP-Psychosis does not exclude participants based on ancestry or where they were born, there are likely to be people who were born outside of South Africa taking

part in the study, leading to several individuals falling in other geographic areas of Africa. 3) Due to the history of South Africa, with immigration from East Africa, Europe, Malaysia, among other places, and with intermarriage with the indigenous Khoi and San groups, there is a lot of admixture in the Western Cape [2,5–8]. Indeed, we see indications of admixture in our NeuroGAP-Psychosis UCT samples, both within different African continental groups as well as contributions from other continental groups.

Supplemental Figures

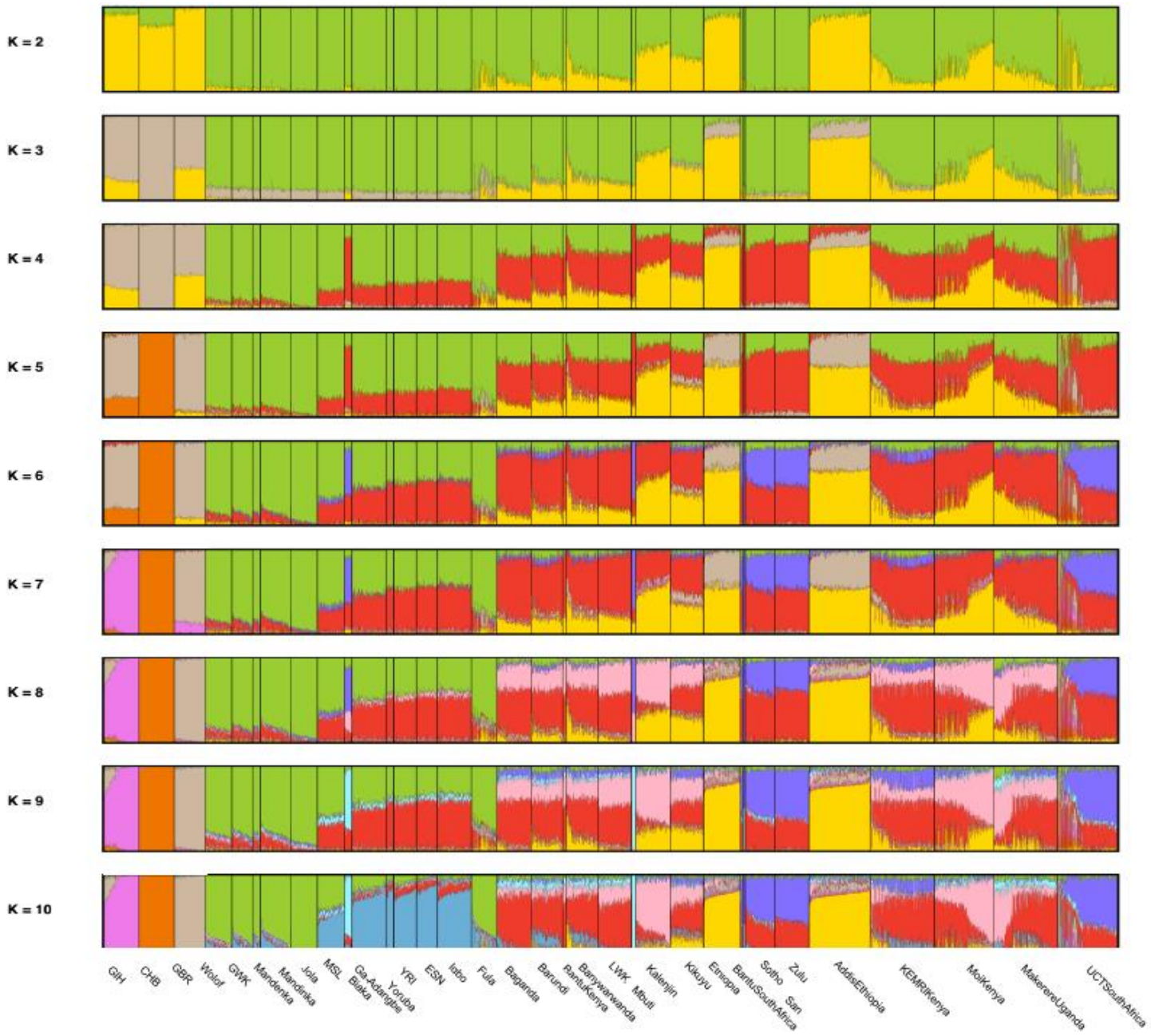


Figure S1. Population composition of the NeuroGAP-Psychosis dataset. *ADMIXTURE plot showing k=2 through k=10 for all African populations as well as a tailored non-African reference panel comprising representation from a south Asian (GIH, fuschia), east Asian (CHB, orange), and European (GBR, tan). A full description of populations, their source datasets, their geographic locations, and their linguistic assignments can be found in Table S1. ADMIXTURE results represent the unanimous consensus after 9 runs at each value of k.*

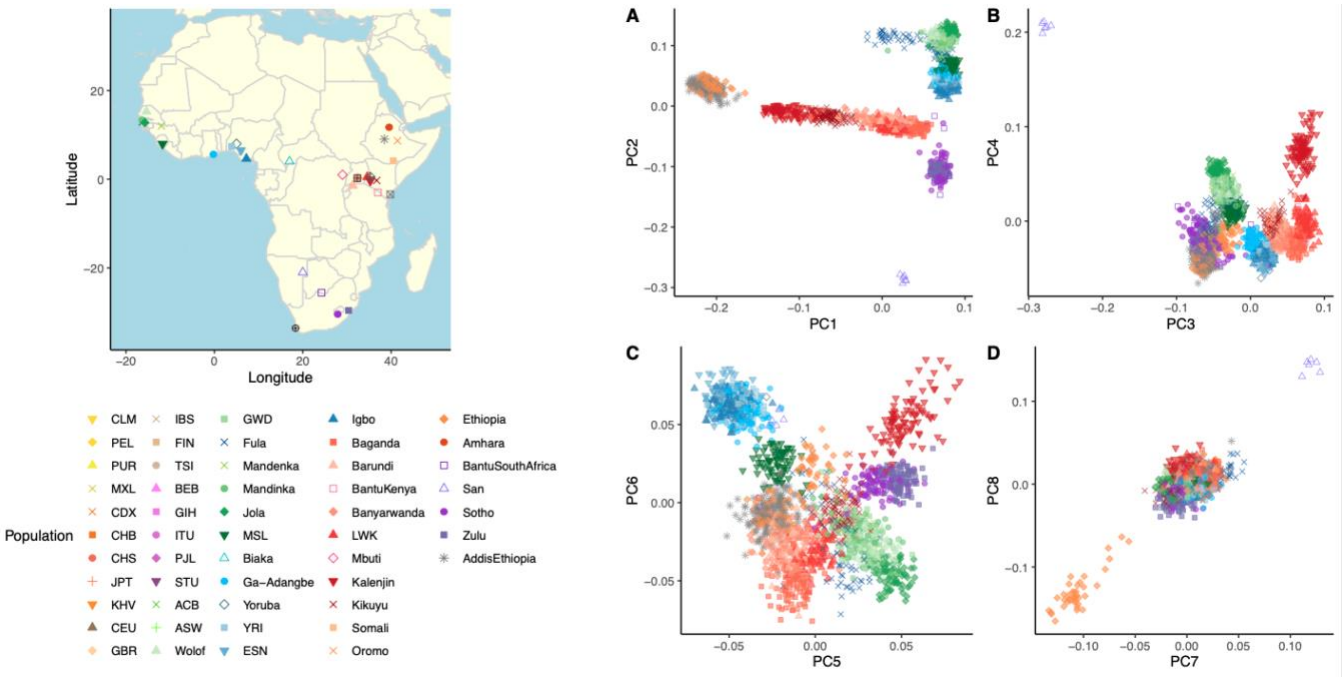


Figure S2. Fine-scale structure of genetic variation in East Africa. *A map showing the location of populations plotted is shown on the left.* ***A-D****) PCA plots for PCs 1-8 showing clustering of AddisEthiopia NeuroGAP-Psychosis samples across PC space with an African reference panel.*

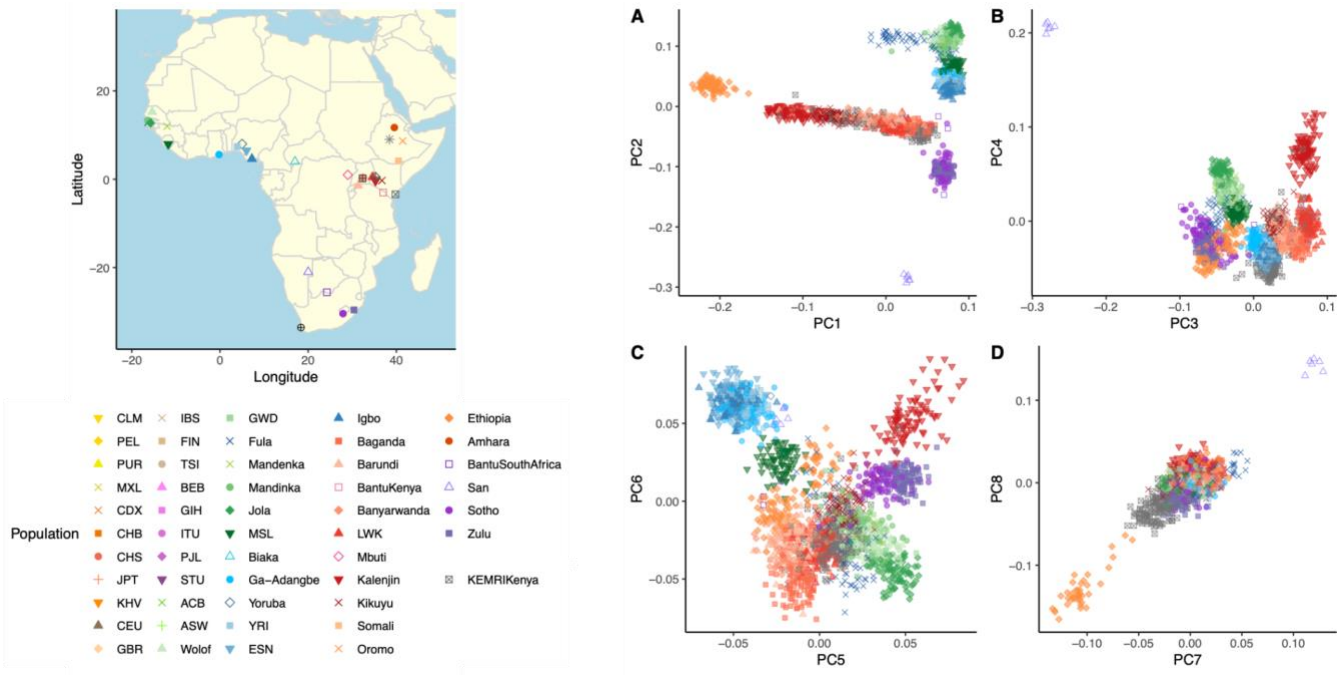


Figure S3. Fine-scale structure of genetic variation in the KEMRIKenya NeuroGAP-Psychosis collection site. *A map showing the location of populations plotted is shown on the left.* ***A-D****) PCA plots for PCs 1-8 showing clustering of KEMRIKenya NeuroGAP-Psychosis samples across PC space with an African reference panel.*

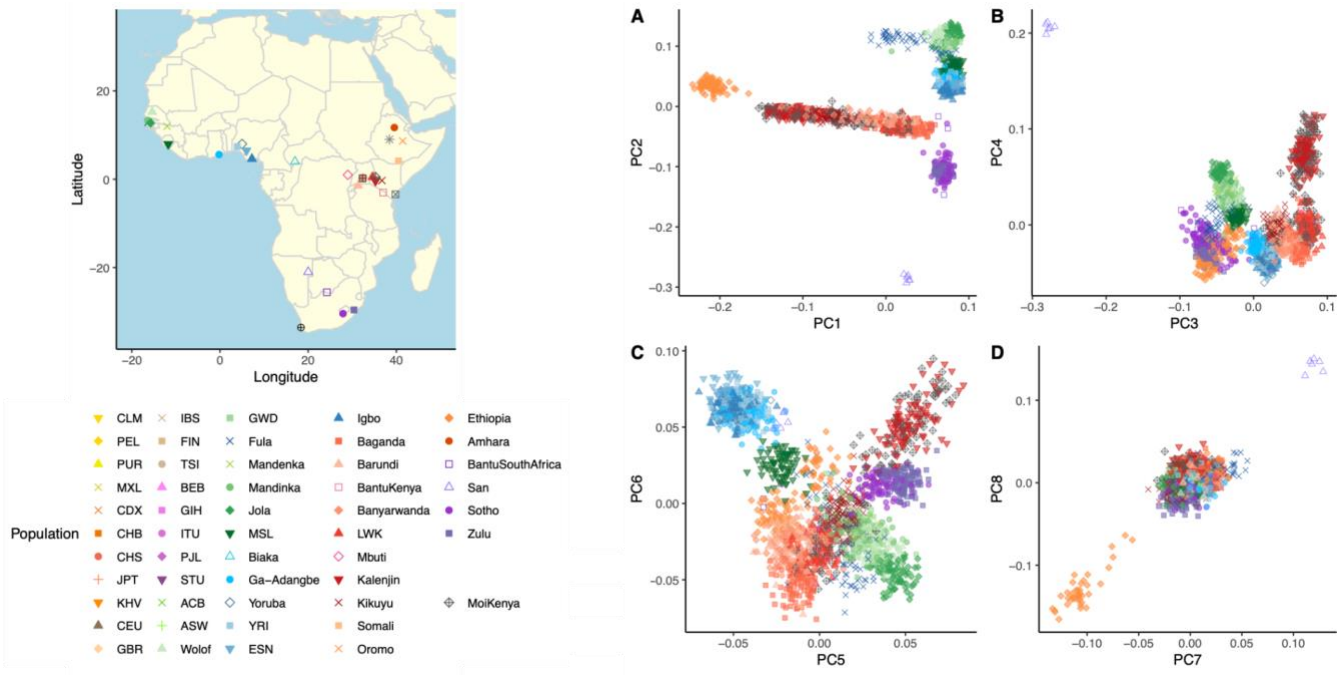


Figure S4. Fine-scale structure of genetic variation in the MoiKenya NeuroGAP-Psychosis collection site. *A map showing the location of populations plotted is shown on the left.* ***A-D****) PCA plots for PCs 1-8 showing clustering of MoiKenya NeuroGAP-Psychosis samples across PC space with an African reference panel.*

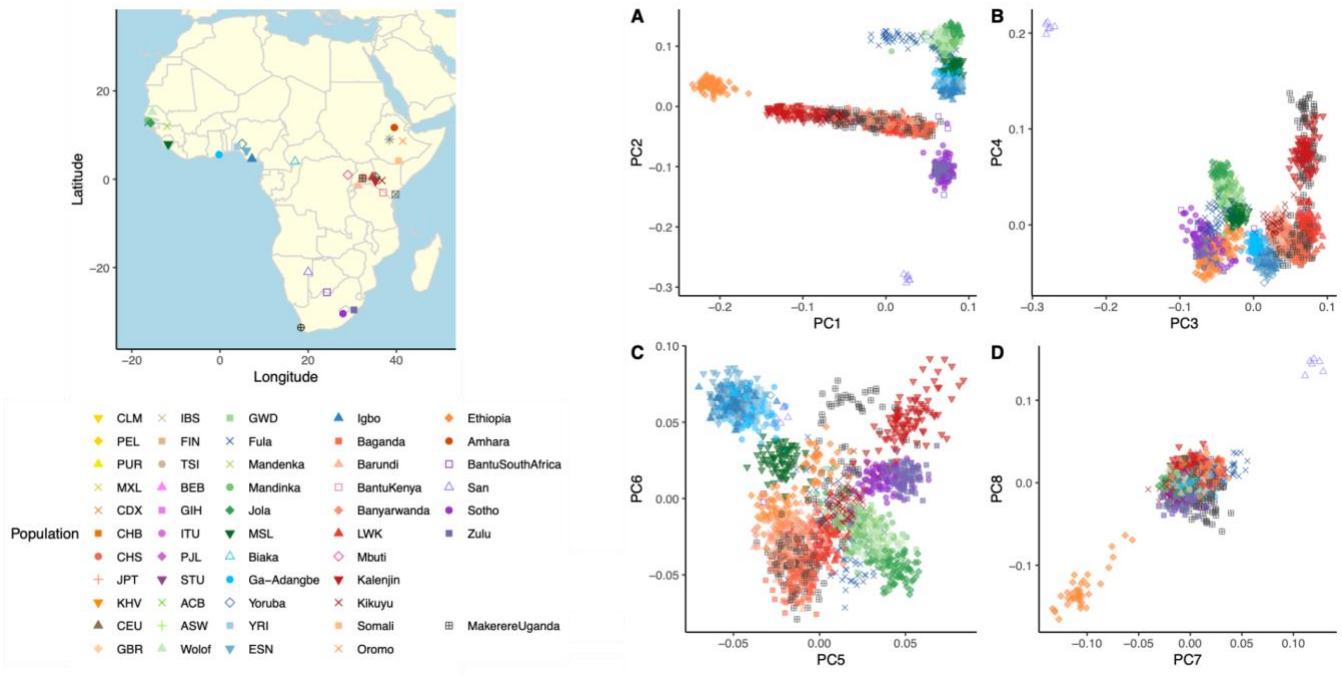


Figure S5. Fine-scale structure of genetic variation in the MakerereUganda NeuroGAP-Psychosis collection site. *A map showing the location of populations plotted is shown on the left.* ***A-D****) PCA plots for PCs 1-8 showing clustering of MakerereUganda NeuroGAP-Psychosis samples across PC space with an African reference panel.*

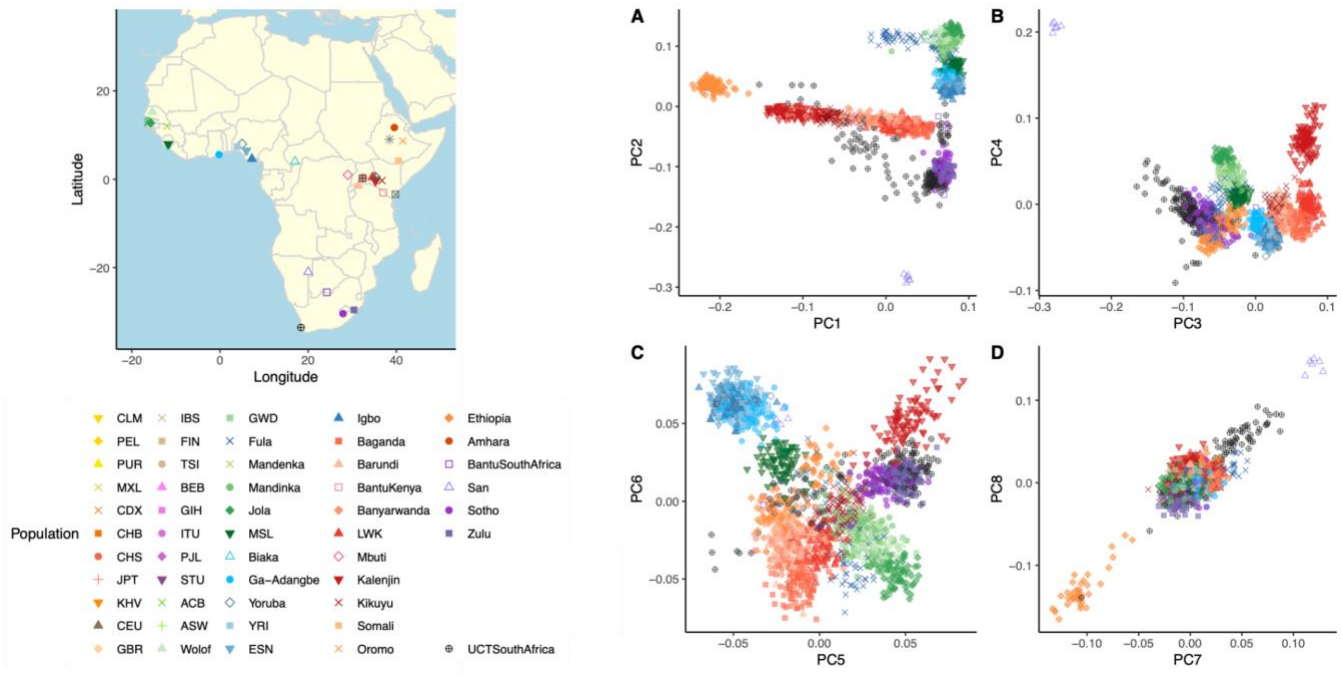


***Figure S6**. **Fine-scale structure of genetic variation in the UCTSouthAfrica NeuroGAP-Psychosis collection site.** A map showing the location of populations plotted is shown on the left. **A-D**) PCA plots for PCs 1-8 showing clustering of UCTSouthAfrica NeuroGAP-Psychosis samples across PC space with an African reference panel.*

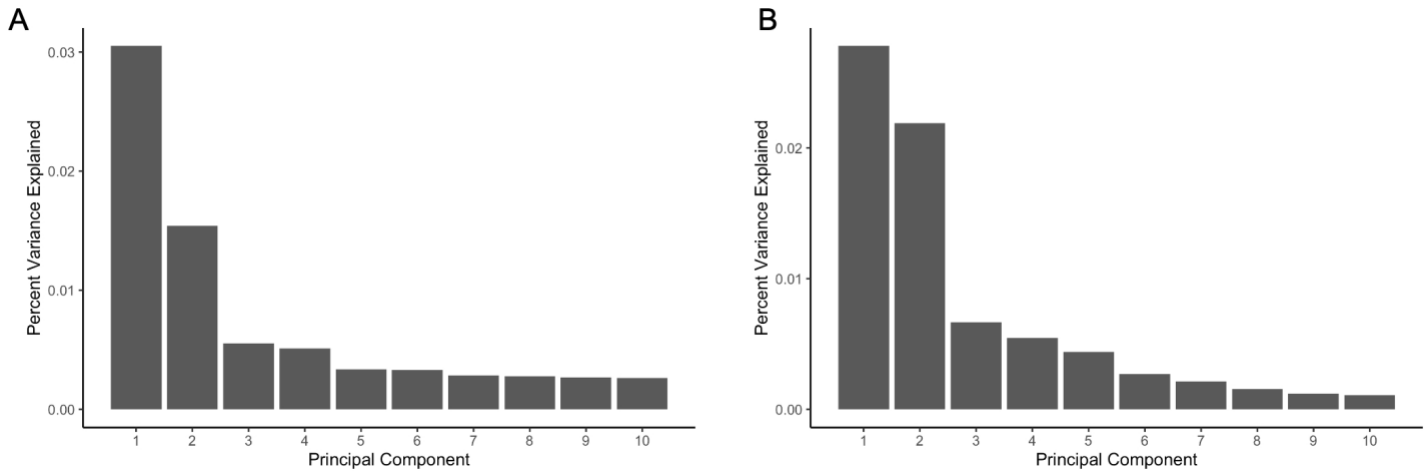


Figure S7. ***Percent variance explained by PCA.*** *Percent variance explained by the first ten Principal Components for analysis including (**A**) the global reference panel, as presented in Figure 1; and (**B**) for the African tailored reference panel, as presented in Figure 2.*

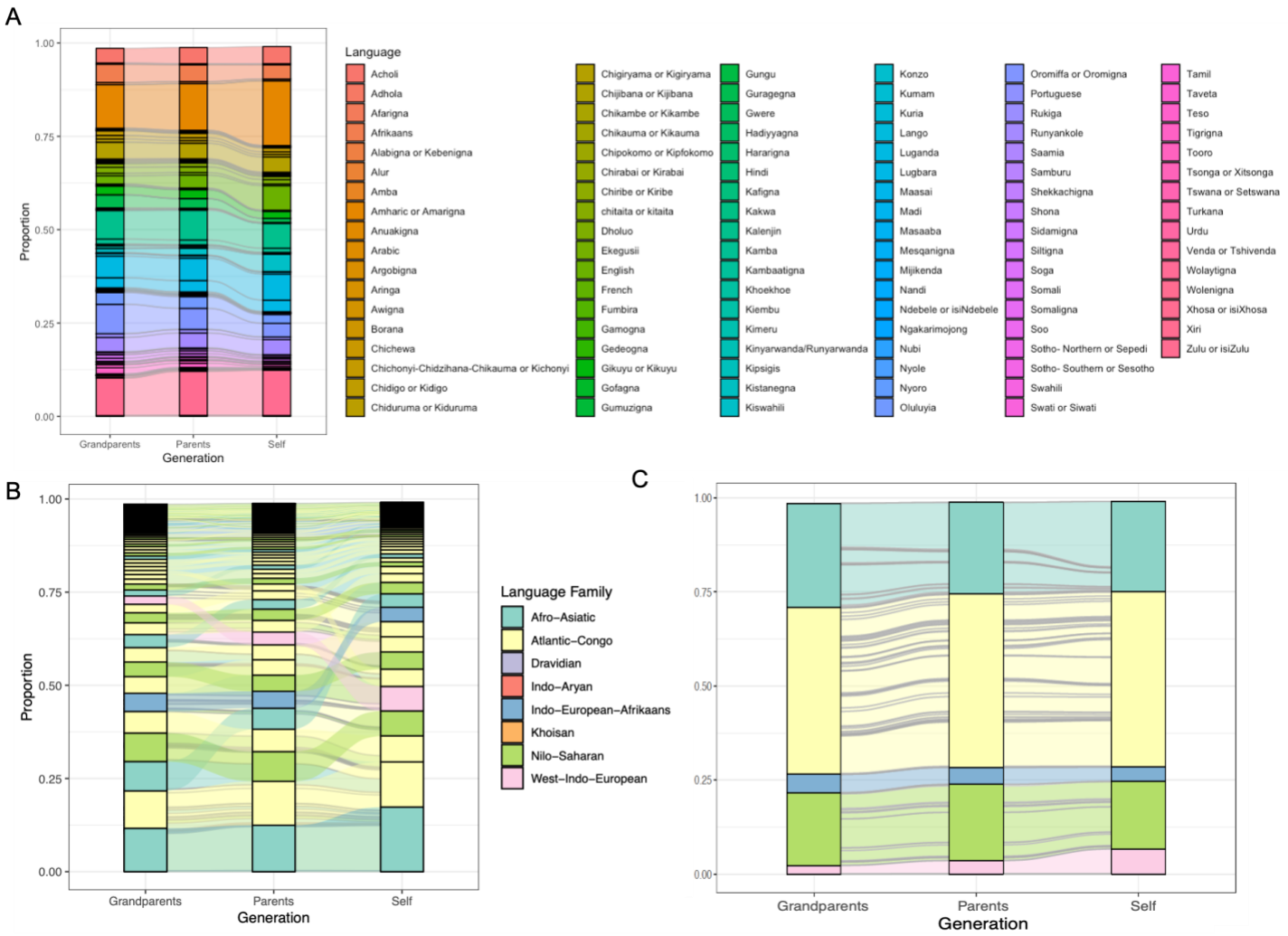


Figure S8. Phenotypic composition of NeuroGAP-Psychosis samples. *Alluvial plot showing the full self-reported primary language reports from participants. A) Primary languages shown individually across the pedigree. B) Primary languages sorted by frequency in each generation and colored by language family. C) Primary language frequency change over generations. Individual strata (separated by gray lines) show specific languages within each language group.*

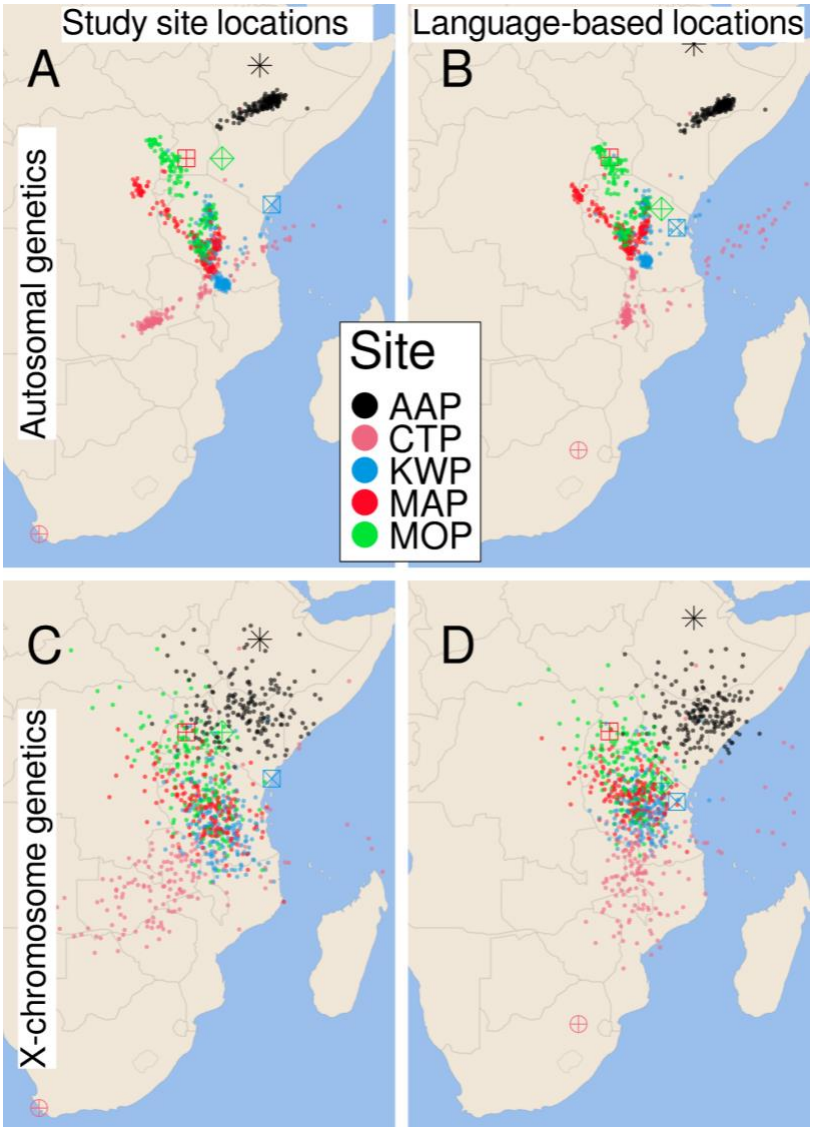


Figure S9. Procrustes analyses indicate that autosomal genetic diversity is better correlated with geography than is X chromosome diversity*. Plots represent the first three genetic PCs after a Procrustes transformation. The upper panels use PCs generated using autosomal variation, and the lower panels 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. Key for NeuroGAP-Psychosis sites: AAP= Addis Ababa University, CTP= University of Cape Town, KWP= KEMRI-Wellcome Trust, MAP=Makerere University, MOP= Moi University.*

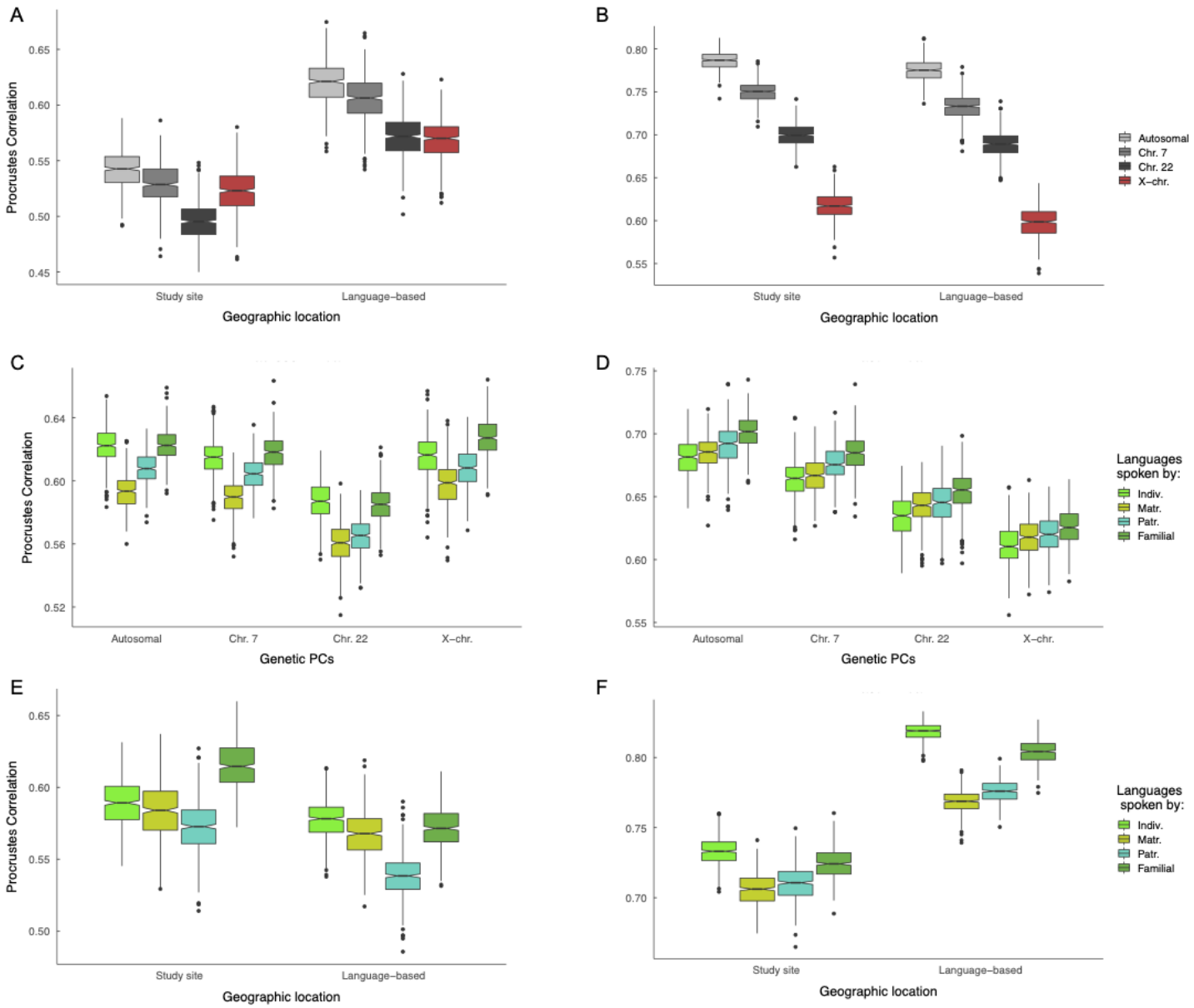


Figure S10. Procrustes correlations between genetics, geography, and language *(all p < 5E-5). Procrustes correlations are shown between:* ***A,B****) geography and genetics.* ***C,D)*** *genetics and language, and* ***E,F)*** *geography and language. The left column includes results for the entire NeuroGAP-Psychosis collection. The right column contains results subset to the four cohorts in East Africa. Genetics analyses were conducted for the complete autosomes, the X chromosome, and two autosomal comparisons to the X: chr7 (similar length) and chr22 (similar SNP count). For linguistic analyses, linguistic variation is measured by the first three PCs of phoneme inventories from languages reported by individuals as spoken by themselves and their*

relatives. Matrilineal relatives include the mother and maternal grandmother. Patrilineal relatives include the father and paternal grandfather. Familial refers to a weighted average of all reported family members. Note that Y-axis labels vary between plots.

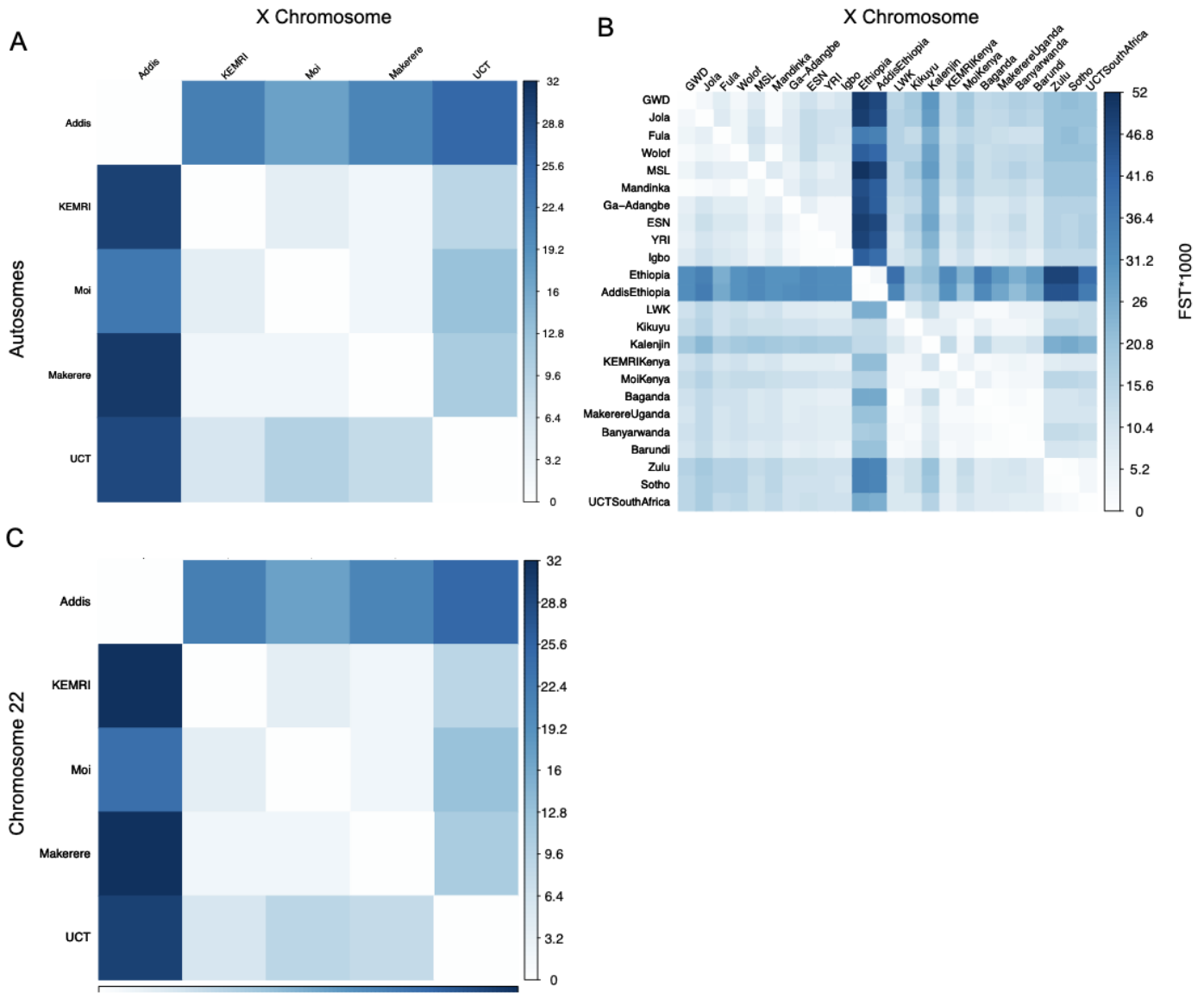


Figure S11. *Genetic differentiation across the autosomes compared to the X chromosome. Heatmap showing the* F_{ST} *estimates calculated between pairwise populations' autosomes (below the diagonal) as compared to the X chromosome (above the diagonal).* F_{ST} *values are multiplied by 1000 for easier interpretation.* ***A, C)*** F_{ST} *estimates just between NeuroGAP-Psychosis collection sites. Panel A includes the entire autosomes while panel C is only chromosome 22 for comparison.* ***B****)* F_{ST} *estimates between NeuroGAP-Psychosis collection sites as well as all African populations in our reference panel.*

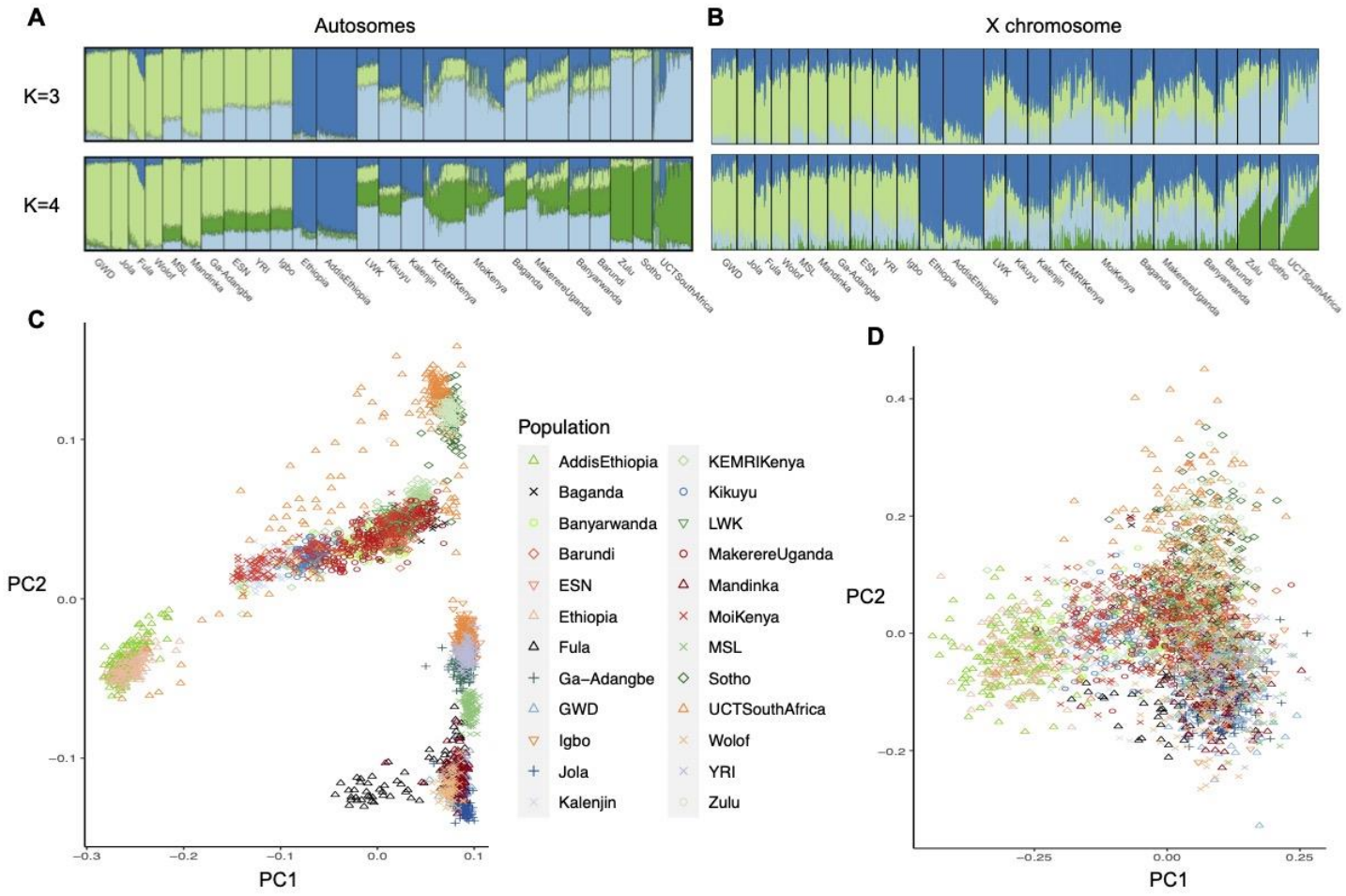
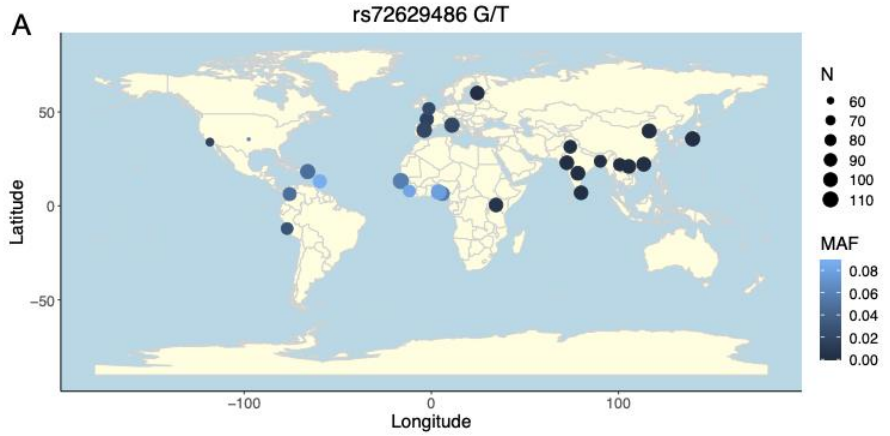


***Figure S12**. **Comparison of ancestry proportions on the autosomes as compared to the X chromosome.** Autosomes are shown in the left column, X chromosome on the right. **A-B**) ADMIXTURE runs at k=3 and 4. Colors are matched with light green tagging east African genetic variation, dark blue tagging Ethiopian variation, light blue tagging west African component, and dark green tagging a south African component. **C-D)** PC plots for the first two principal components of genetic variation in the autosomes and X chromosome.*



B

Population Frequencies

Population	Allele Count	Allele Number	Number of Homozygotes	Allele Frequency
African/African-American	1361	24624	40	0.05527
Ashkenazi Jewish	317	10294	4	0.03079
Other	170	7194	3	0.02363
Latino/Admixed American	766	35412	16	0.02163
European (non-Finnish)	1324	127584	11	0.01038
South Asian	139	30612	4	0.004541
European (Finnish)	58	25112	1	0.002310
East Asian	0	19890	0	0.000
Female	2018	127872	39	0.01578
Male	2117	152850	40	0.01385
Total	**4135**	**280722**	**79**	**0.01473**

Figure S13. African genetic variation is broadly informative. A) *the frequency of rs2071348, previously demonstrated to influence beta thalassemia, varies within the African continent dramatically, even across only our 5 pilot NeuroGAP-Psychosis sites. In Africa alone, missense variant rs72629486 spans the entire range of global frequencies reported in the gnomad database.* ***B)*** *Screenshot of the population frequencies of rs72629486 in gnomAD; Feb 28, 2021.*

Supplemental Tables

Table S1. Key for location/country and dataset of origin for all populations included in analyses. *'Region' indicates the continental assignment for non-African populations and the geographic region assigned within Africa as according to the UN Statistics division geoscheme[9].*

Table S2. Raw data for language phenotypes reported for each familial relationship across the NeuroGAP-Psychosis dataset*. See Methods section 'Ethnolinguistic Phenotypes' for a detailed description of the specific phenotypes collected. Data is sorted alphabetically by language name.*

Table S3. Classification of self-reported primary ethnicities included in the surveys, with associated data collected from the Ethnographic Atlas. *See Methods section 'Ethnolinguistic Phenotypes' for a detailed description of the specific phenotypes collected. Data is sorted alphabetically by language name.*

	Moi, Kenya	**Ethiopia**	**KEMRI, Kenya**	**South Africa**	**Uganda**
Autocall Call Rate (samples)	189	183	188	185	192
Variant Call Rate (variants)	638235	638235	638235	638235	638235
Individual Call Rate (samples)	189	181	188	182	190
Sex Violations (samples)	187	181	188	179	188
Minor Allele Frequency (variants)	360321	360321	360321	360321	360321
Hardy Weinberg Equilibrium (variants)	331667	331667	331667	331667	331667
Sample Relatedness (samples)	173	179	187	175	186
Final Counts (variants / samples)	331667 / 173	331667 / 179	331667 / 187	331667 / 175	331667 / 186

Table S4. Variant and individual counts throughout the Autosomal QC process. *Rows show the number of samples/variants remaining after that round of QC for each study site.*

	PAR Region	Female non-PAR Region
Variant Call Rate	515	16261
MAF	411	11113
HWE	402	11104
Final Counts	900 Samples 402 Variants	900 Samples 11104 Variants

***Table S5**. **Variant counts throughout X Chromosome QC.** Rows show the number of samples/variants remaining after that round of QC for each study site. PAR indicates the pseudo-autosomal region of the X chromosome, non-PAR includes all other regions.*

Supplemental References

1. López, S., Tarekegn, A., Band, G., van Dorp, L., Bird, N., Morris, S., Oljira, T., Mekonnen, E., Bekele, E., Blench, R., et al. (2021). The genetic landscape of Ethiopia: diversity, intermixing and the association with culture. *Nat Commun* **12,** 3581 (2021).

2. 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. Proceedings of the National Academy of Sciences *111*, 2632–2637.

3. Pagani, L., Schiffels, S., Gurdasani, D., Danecek, P., Scally, A., Chen, Y., Xue, Y., Haber, M., Ekong, R., Oljira, T., et al. (2015). Tracing the route of modern humans out of Africa by using 225 human genome sequences from Ethiopians and Egyptians. Am. J. Hum. Genet. *96*, 986–991.

4. 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.

5. Uren, C., Kim, M., Martin, A.R., Bobo, D., Gignoux, C.R., van Helden, P.D., Möller, M., Hoal, E.G., and Henn, B.M. (2016). Fine-Scale Human Population Structure in Southern Africa Reflects Ecogeographic Boundaries. Genetics *204*, 303–314.

6. Sikora, M., Laayouni, H., Calafell, F., Comas, D., and Bertranpetit, J. (2011). A genomic analysis identifies a novel component in the genetic structure of sub-Saharan African populations. Eur. J. Hum. Genet. *19*, 84–88.

7. Chimusa, E.R., Meintjies, A., Tchanga, M., Mulder, N., Seoighe, C., Soodyall, H., and Ramesar, R. (2015). A genomic portrait of haplotype diversity and signatures of selection in indigenous southern African populations. PLoS Genet. *11*, e1005052.

8. Bergström, A., McCarthy, S.A., Hui, R., Almarri, M.A., Ayub, Q., Danecek, P., Chen, Y., Felkel, S., Hallast, P., Kamm, J., et al. (2020). Insights into human genetic variation and population history from 929 diverse genomes. Science *367*,.

9. United Nations Statistics Division Methodology. Standard country or area codes for statistical use (M49). https://unstats.un.org/unsd/methodology/m49/