Supplementary Information: Genetic structure correlates with ethnolinguistic diversity in eastern and southern Africa

Supplementary Information

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 a likely lot of admixture between people who originated from the coast and people of Arabic ancestry. Admixture is when two historically separate groups of people mix with each other. 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.

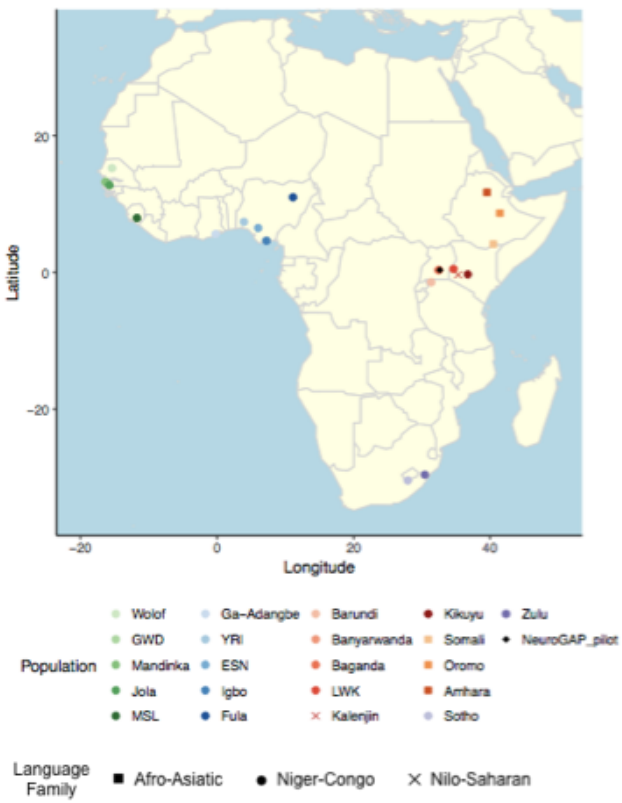
Uganda: The pilot data from Uganda also falls cleanly within the East African cluster, as would be expected by the collection location in Kampala. Furthermore, it seems to fall with the Bagandan ethnic group primarily, which is the most common self-reported ancestry that participants reported in these 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 a small contribution from other continental groups.

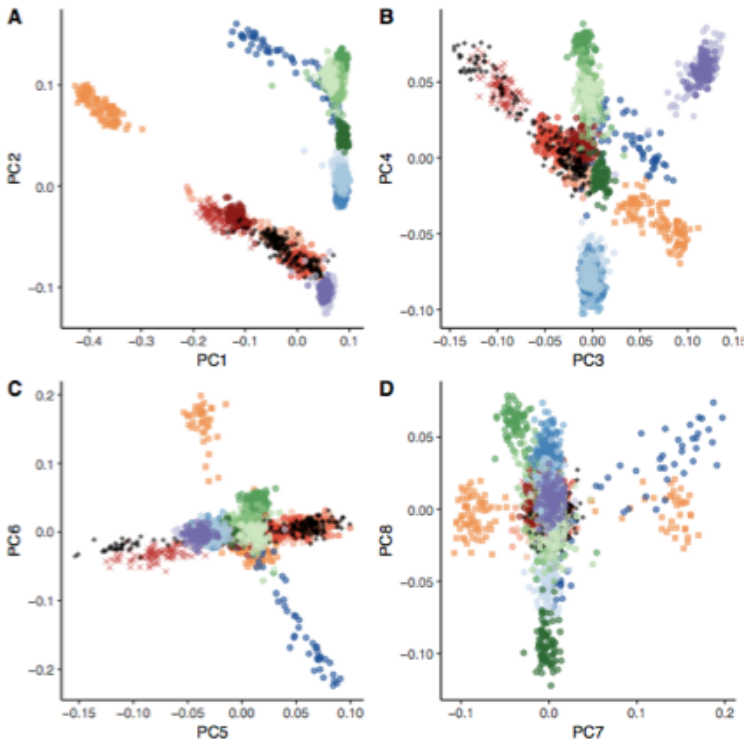
Supplementary Figures

Supplementary Figure 1*. Fine-scale structure of genetic variation in East Africa partitions with language. A map showing the location of populations plotted is shown on the left.* ***A****) PCA plot showing partitioning of Kenyan samples from Moi university across PC space with an African reference panel. These serve as an example of the trend in population structure observed across our three East African sites. Note the spread of participants defining PC5.*

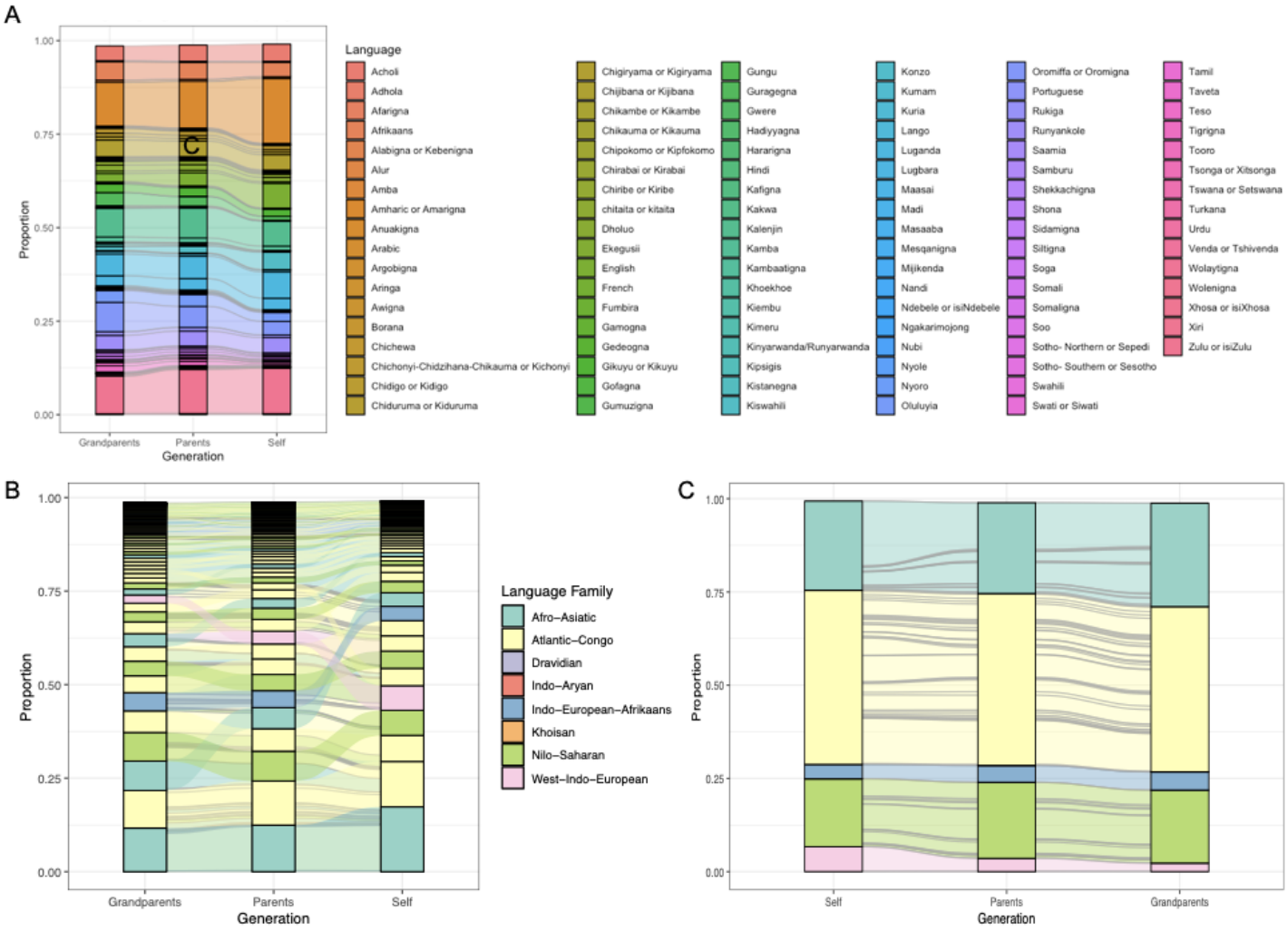

Latitude
Longitude
20
0
−20
−20
0
20
40

Population
Wolof
GWD
Mandinka
Jola
MSL
Ga−Adangbe
YRI
ESN
Igbo
Fula
Barundi
Banyarwanda
Baganda
LWK
Kalenjin
Kikuyu
Somali
Oromo
Amhara
Sotho
Zulu
NeuroGAP_pilot

Language Family
Afro-Asiatic
Niger-Congo
Nilo-Saharan

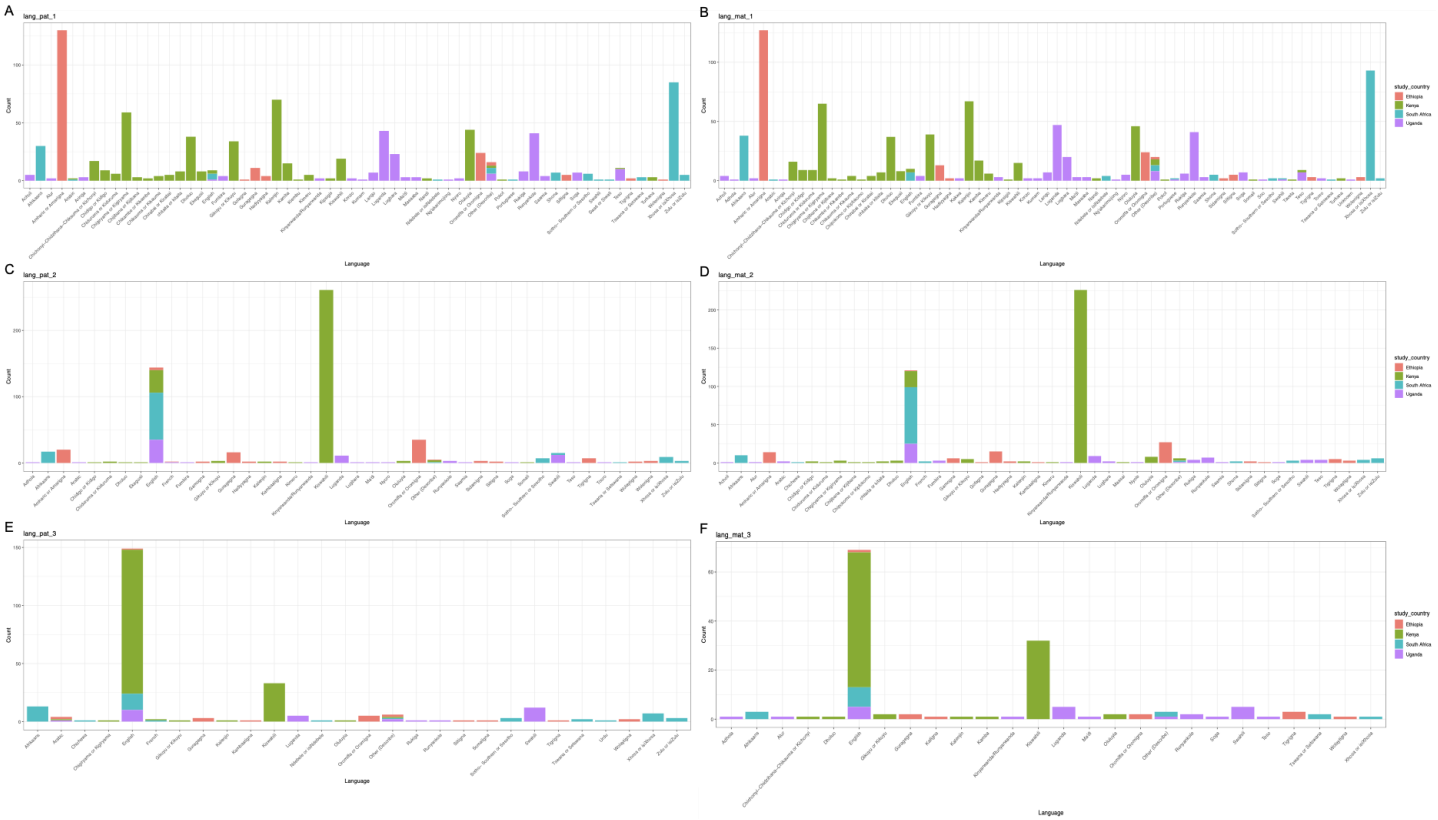

A
PC2
PC1
B
PC4
PC3
C
PC6
PC5
D
PC8
PC7

Supplementary Figure 2. *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.*

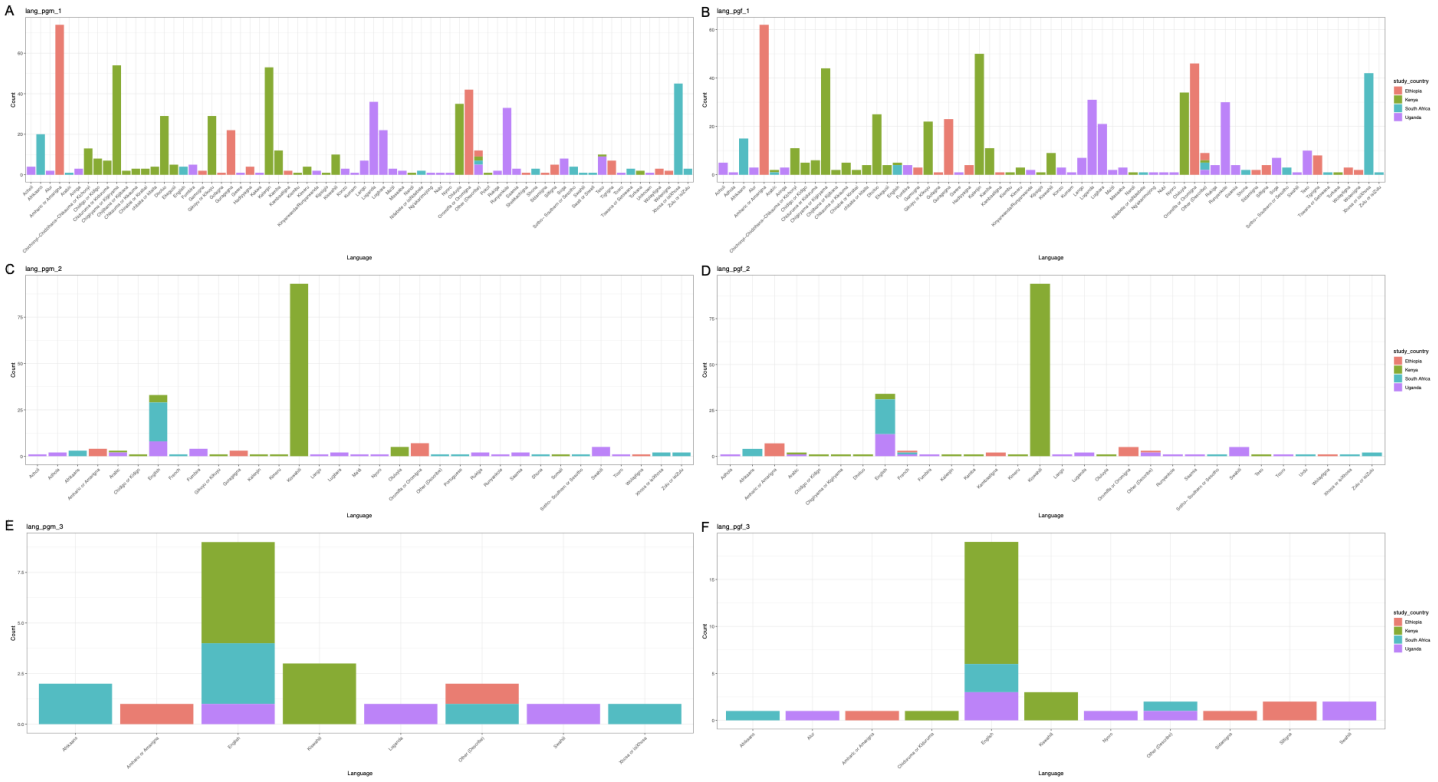


Supplementary Figure 3. *NeuroGAP-Psychosis self-reported languages and ethnicities. The left column shows languages the person speaks, the right their identified ethnicities. The rows show the primary, secondary and tertiary self reports.*

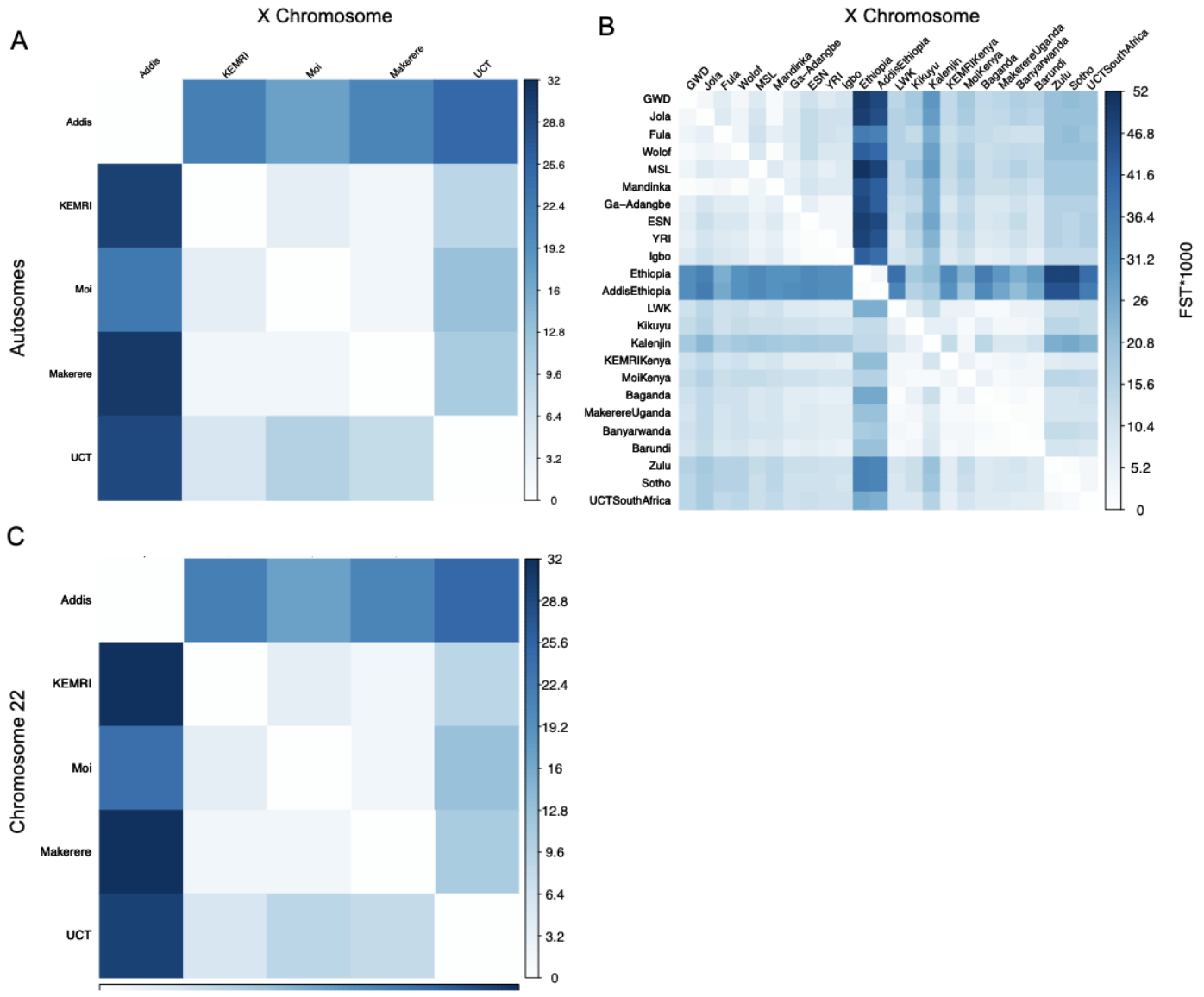
Supplementary Figure 4. *NeuroGAP-Psychosis self-reported parental languages. The left column shows paternal languages, the right maternal. The rows show the primary, secondary and tertiary self reports.*



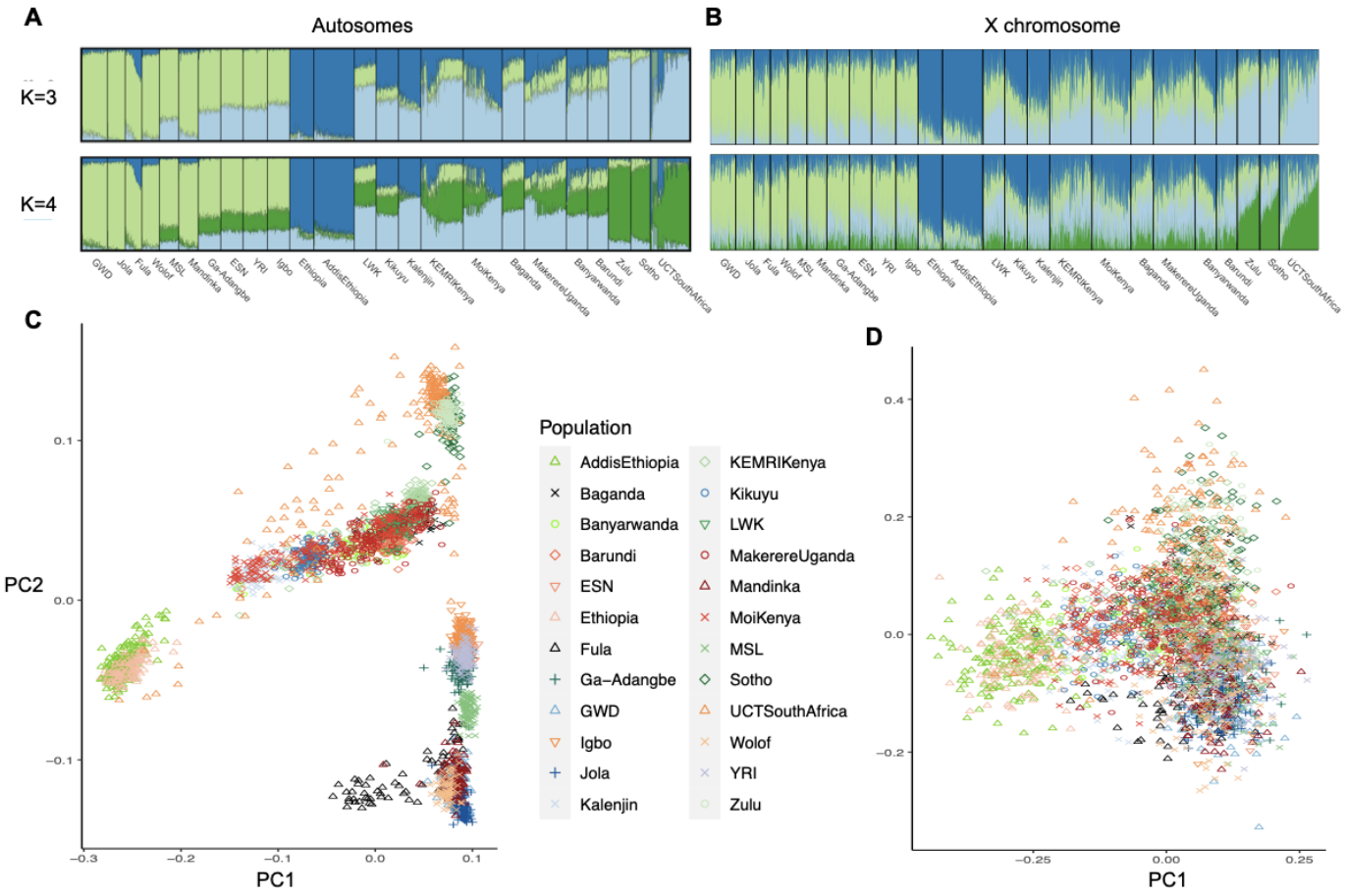
Supplementary Figure 5. *NeuroGAP-Psychosis self-reported grandparental languages. The left column shows grandpaternal languages, the right grandmaternal. The rows show the primary, secondary and tertiary self reports.*



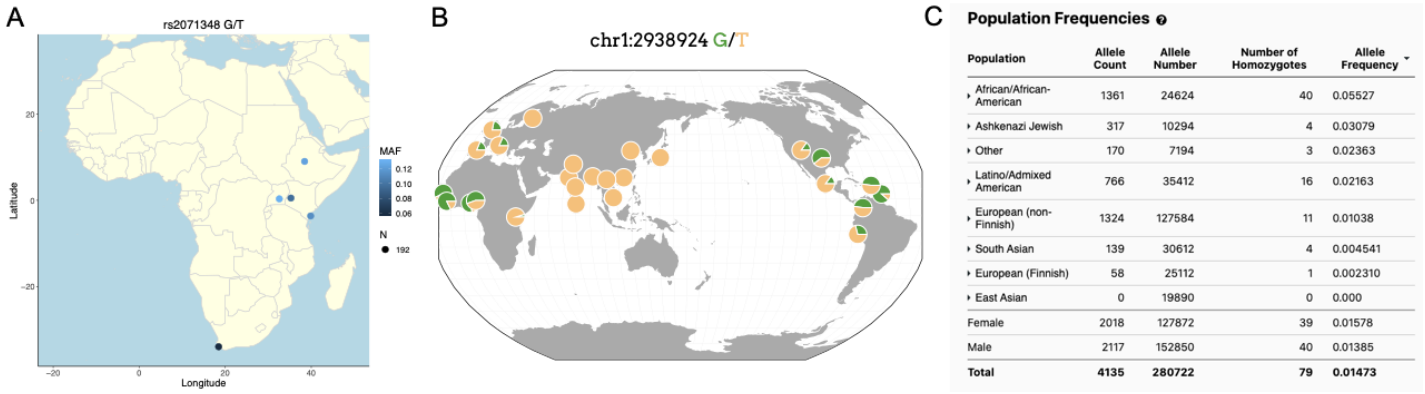
Supplementary Figure 6. *Genetic differentiation across the autosomes compared to the X chromosome. Heatmap showing the Fst estimates calculated between pairwise populations' autosomes (y axis) as compared to the X chromosome (x axis). Fst values are multiplied by 1000 for easier interpretation. A) Fst estimates just between NeuroGAP-Psychosis collection sites. B) Fst estimates between NeuroGAP-Psychosis collection sites as well as all African populations in our reference panel. C) ADMIXTURE plot for the autosomes and (D) X chromosome. (E) PCA plots for the autosomes and (F) X chromosome.*



Supplementary Figure 7. *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 biplots for the first two principal components of genetic variation in the autosomes and X chromosome.*



Supplementary Figure 8. *African genetic variation is broadly informative. (A) the frequency of rs2071348, previously demonstrated to influence beta thalassemia, varies in frequency within the African continent dramatically, even across only our 5 pilot NeuroGAP-Psychosis sites. (B) In Africa alone, missense variant rs72629486 spans the entire range of global frequencies reported in the gnomad database. (C) Screenshot of the population frequencies of rs72629486 in gnomAD; Feb 28, 2021.*



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**

Supplementary Tables

Supplementary Table 1*. Variant and individual counts throughout the Autosomal QC process.*

Autosomal QC Filter Results					
Filter Name	**Variants or Individuals Remaining After Filter per Pilot Site**				
	Moi, Kenya	Ethiopia	KEMRI	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

***Supplementary Table 2**. Variant counts throughout X Chromosome QC.*

Filter Name	**Variants Remaining After Filter**	
	PAR Region	Female nonPAR Region
Variant Call Rate	515	16261
MAF	411	11113
HWE	402	11104
Final Counts	900 Samples 402 Variants	900 Samples 11104 Variants

References

1. 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

2. 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

3. 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.

4. 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.

5. 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.

6. 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.

7. 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.

8. 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