SUPPLEMENTARY INFORMATION

doi:10.1038/nature13997

TABLE OF CONTENTS

SUPPLEMENTARY METHODS

1.0 CURATION OF AGVP GENOTYPE DATA

1.1 SAMPLES AND POPULATIONS

We genotyped 2,185 samples from 16 different African populations from SSA on the Illumina HumanOmni 2.5M BeadChip array and sequenced 320 individuals at 4x coverage from 7 ethno-linguistic groups (**SM Table 1 and 2**). As the samples were genotyped at different times, 720 samples were genotyped on the Illumina HumanOmni 2.5M-quad BeadChip array (hereafter referred as quad) and 1,465 on the Illumina HumanOmni 2.5M-8 BeadChip array (hereafter referred as octo) (**SM Table 3**), which replaced the quad chip on the market during the course of the study. The octo and quad arrays were noted to be very similar, with 99.7% overlap between genotyped sites. To assess concordance of genotypes between chips, 29 samples from one population (Baganda) were genotyped on both platforms. Additionally, four populations from East, West and South Africa (Baganda, Banyarwanda, Ga-Adangbe and Zulu) were genotyped partially on the quad chip and partially on the octo chips, providing us with a further opportunity to rigorously examine and control for any chip effects. We describe this process in more detail in **Supplementary Note 1**.

In order to characterise genetic variation among populations representative of the most common ethno-linguistic groups in Africa, we chose populations belonging to the Niger-Congo, Nilo-Saharan and Afro-Asiatic groups from SSA, including those that are part of the H3Africa initiative.[1] Ethno-linguistic grouping was based on self-identification, and should be considered a broad construct that encompasses shared cultural heritage, ancestry, history, homeland, language or ideology.

In order to supplement the African diversity panel, we also included 2.5M-quad data for Yoruba (YRI) and Luhya (LWK) individuals from the 1000 Genomes Project. Masaai (MKK) was not included, due to the small number of samples. In total, 18 population groups were studied (**SM Table 1**, **Figure 1**). These populations primarily belonged to the Niger-Congo linguistic group, except for the Kalenjin who speak Nilo-Saharan languages, and the Ethiopian populations-Oromo, Amhara and Somali, who speak Afro-Asiatic languages. Herding and farming were the primary traditional modes of subsistence for all populations. Details of each population are presented in **SM Table 1.** The three Ethiopian ethno-linguistic groups were collapsed for downstream analyses due to small sample size. Informed consent was obtained from all study participants. Relevant study protocols for the use of biological samples in genetic studies and

data sharing were approved by Institutional Research Boards and research ethics committees in the relevant countries and/or in the UK for each study contributing samples to the project. Additional approvals were obtained from the Wellcome Trust Sanger Institute's Human Materials and Data Management Committee (HMDMC).

SM Table 1: Details of populations included in the AGVP

Population name	Region (UN Statistics Division geoscheme)	Country	Language family	Language subgroup
Baganda	Eastern Africa	Uganda	Niger-Congo	Bantoid
Banyarwanda	Eastern Africa	Uganda	Niger-Congo	Bantoid
Barundi	Eastern Africa	Uganda	Niger-Congo	Bantoid
Luhya†	Eastern Africa	Kenya	Niger-Congo	Bantoid
Kikuyu	Eastern Africa	Kenya	Niger-Congo	Bantoid
Sotho	Southern Africa	South Africa	Niger-Congo	Bantoid
Zulu	Southern Africa	South Africa	Niger-Congo	Bantoid
Yoruba†	Western Africa	Nigeria	Niger-Congo	Defoid
Igbo	Western Africa	Nigeria	Niger-Congo	Igboid
Ga-Adangbe	Western Africa	Ghana	Niger-Congo	Kwa
Jola	Western Africa	Gambia	Niger-Congo	Bak
Fula	Western Africa	Gambia	Niger-Congo	Senegambian
Wolof	Western Africa	Gambia	Niger-Congo	Senegambian
Mandinka	Western Africa	Gambia	Niger-Congo	Mande
Amhara	Eastern Africa	Ethiopia	Afro-Asiatic	Semitic
Oromo	Eastern Africa	Ethiopia	Afro-Asiatic	Cushitic
Somali*	Eastern Africa	Ethiopia*	Afro-Asiatic	Cushitic
Kalenjin	Eastern Africa	Kenya	Nilo-Saharan	Eastern Sudanic

† these populations were included from the 1000 Genomes Project[2].

* Although these samples were collected from Ethiopia, many of these individuals are from Somalia (Mogadishu).

SM Table 2: Populations sequenced as part of study

Population	Number	Average coverage	Overlap with genotype sample	No. of variants called
Baganda	100	4x	94	20,461,747
Zulu	100	4x	95	20,267,592
Ethiopia	120	4x, 8x	63	20,452,231
Amhara	24		24	
Oromo	24		19	
Somali	24		20	
Wolayta	24			
Gumuz	24			
Total	320		252	29,809,603

A separate global dataset was also generated, to contextualise population data from Africa, including 2.5M-quad data for 1,556 samples from 17 additional global populations (GBR, ACB, ASW, CDX, CEU, CHB, CHS, CLM, FIN, GIH, IBS, JPT, KHV, MXL, PEL, PUR, and TSI) from the 1000 Genomes project (**SM Table 3**). To avoid differential calling and batch effects, we called genotypes from intensity data for all samples on each chip together using the Illuminus algorithm [3]. 1000 Genomes Project genotype raw intensity files were obtained with permission from the Broad Institute. Genotypes for these populations were also re-called with other samples on the quad chip.

1.2 Sample and SNP Quality control

At the start, low quality variants that mapped to multiple regions within the human genome or did not map to any region were removed. Duplicate variants genotyped on the chip were also excluded during filtering. Only variants overlapping between quad and octo chips (2,251,315 variants) were included in subsequent analyses.

Stringent quality control filtering was carried out within each population. Each population genotyped over two chips was treated as two separate populations for the purposes of quality control and filtering. Samples with a call rate below 98% and heterozygosity greater than 3 SD from the mean were filtered sequentially. Sex check was carried out in PLINK using default F values of <0.2 for males and >0.8 for females. Samples with discordant genetic sex and reported sex were removed. Following this, SNP filtering was carried out across the remaining samples, and sites called in <98% of samples were removed from each population. Sites in Hardy Weinberg disequilibrium ($p<10^{-7}$) were also removed from each population. Identity by descent (IBD) was measured within each population and related individuals (IBD>0.05) were removed using an algorithm that retained the maximum number of individuals by removing individuals related to the largest number in the dataset iteratively. For individuals related to equal numbers in the dataset, samples with lower call rates were preferentially removed. Data were then merged into two datasets: the African dataset comprising genotype data from 16 populations, and the global dataset comprising data from 33 populations (**SM Tables 4 and 5**). Only the intersection of all variants that passed QC in all populations merged were included in each of these datasets (**SM Tables 4 and 5**). Following the merger, the global dataset included data on 2,864 individuals and 1,399,027 variants, while the African dataset included data on 1,481 individuals and 1,577,224 variants (**SM Tables 4 and 5**). Following the above QC process, principal component analysis (PCA) was carried out in EIGENSOFT v 3.0 for each population and across all populations to inspect the data for chip effects and outliers (see **Supplementary Note 1**).

SM Table 3: Global samples and distribution over BeadChips at various stages of analyses

	Genotyped samples[‡]		Post-calling samples		Post-QC samples		Sub-sampling sets		
	quad	*octo*	*quad*	*octo*	*quad*	*octo*	*quad*	*octo*	
Europe									
GBR	-	-	101	-	91	-	91	-	91
FIN	-	-	100	-	97	-	97	-	97
CEU	-	-	104	-	95	-	95	-	95
TSI	-	-	100	-	92	-	92	-	92
IBS	-	-	150	-	99	-	99	-	99
Asia									
CHS	-	-	150	-	86	-	86	-	86
CDX	-	-	100	-	83	-	83	-	83
CHB	-	-	100	-	98	-	98	-	98
JPT	-	-	100	-	96	-	96	-	96
KHV	-	-	121	-	96	-	96	-	96
America									
GIH	-	-	100	-	95	-	95	-	95
MXL	-	-	100	-	47	-	47	-	47
CLM	-	-	107	-	65	-	65	-	65
PEL	-	-	105	-	50	-	50	-	50
PUR	-	-	111	-	72	-	72	-	72
ASW	-	-	97	-	49	-	49	-	49
ACB	-	-	102	-	72	-	72	-	72
Africa									
Baganda[§]	100	229	100	228	90	197	44	56	100
Banyarwanda	104	360	103	358	83	223	20	80	100
Ga-Adangbe	99	11	97	11	90	11	89	11	100
Zulu	100	13	100	13	9	95	95	5	100
Barundi	-	191	-	189	-	97	-	97	97
Ethiopia*	-	129	-	123	-	108	-	107	107
Fula	-	98	-	95	-	74	-	74	74
Jola	-	102	-	95	-	79	-	79	79
Mandinka	-	120	-	105	-	88	-	88	88
Sotho	-	104	-	103	-	86	-	86	86
Wolof	-	108	-	103	-	78	-	78	78
Igbo	104	-	102	-	99	-	99	-	99
Kalenjin	110	-	110	-	100	-	100	-	100
Kikuyu	103	-	102	-	99	-	99	-	99
Luhya (LWK)	-	-	100	-	74	-	74	-	74
Yoruba (YRI)	-	-	161	-	100	-	100	-	100
TOTAL	**720**	**1465**	**2823**	**1423**	**2127**	**1136**			**2864**

[‡] This information was not available for the '1000 Genome Project' populations.

[§] For Baganda a set of 29 samples was genotyped in duplicate on both chips. Of these 26 passed QC on both chips.

* The Ethiopian group comprises 3 populations, which were grouped together for the QC due to small sample numbers and shared Semitic-Cushitic languages: Amhara (46), Oromo (31), Somali (52).

SM Table 4: Number of SNPs retained in each population after QC and after chip-effect removal

Population	Post-QC SNPs		Post-removal of chip-effect SNPs
	quad	*octo*	
Europe			
GBR	2,173,225	-	2,173,225
FIN	2,170,696	-	2,170,696
CEU	2,214,433	-	2,214,433
TSI	2,191,635	-	2,191,635
IBS	2,201,626	-	2,201,626
Asia			
CHS	2,214,433	-	2,214,433
CDX	2,191,635	-	2,191,635
CHB	2,201,626	-	2,201,626
JPT	2,214,433	-	2,214,433
KHV	2,191,635	-	2,191,635
America			
GIH	2,196,526	-	2,196,526
MXL	2,197,501	-	2,197,501
CLM	2,207,652	-	2,207,652
PEL	2,226,856	-	2,226,856
PUR	2,211,146	-	2,211,146
ASW	2,170,758	-	2,170,758
ACB	2,134,208	-	2,134,208
Africa			
Baganda	2,186,500	2,185,277	2,124,005
Banyarwanda	2,221,259	2,173,753	2,144,229
Ga-Adangbe	2,173,260	2,178,518	2,178,911
Zulu	2,172,152	2,117,266	2,050,451
Barundi	-	2,178,911	2,178,911
Ethiopia*	-	2,143,095	2,143,095
Fula	-	2,103,594	2,103,594
Jola	-	2,092,279	2,092,279
Mandinka	-	2,074,615	2,074,615
Sotho	-	2,139,912	2,139,912
Wolof	-	2,085,695	2,085,695
Igbo	2,165,570	-	2,165,570
Kalenjin	2,212,582	-	2,212,582
Kikuyu	2,210,814	-	2,210,814
Luhya (LWK)	2,182,223	-	2,182,223
Yoruba (YRI)	2,208,067	-	2,208,067
African dataset	**-**	**-**	**1,399,027**
Global dataset	**-**	**-**	**1,577,224**

* The Ethiopian group comprises 3 populations, which were grouped together for the QC due to small sample numbers and shared Semitic-Cushitic languages: Amhara (46), Oromo (31), Somali (52).

1.3 Removal of chip effects from datasets

Chip effects were assessed and removed from datasets using a variety of approaches. Chip effects between the quad and octo chip were evident on PCA among African and global, and PCs representing chip effects were identified for each dataset. SNPs highly weighted along these PCs were systematically removed as outlined in **Supplementary Note 1**. Removal of chip effects and homogenisation of data was confirmed by PCA, ADMIXTURE analysis, and PCA projections for Baganda duplicate samples that were genotyped on both chips. These methods are detailed in **Supplementary Note 1**. Following removal of chip effects, and curation of datasets, all populations were subsampled to include a maximum of 100 individuals from a given population group (**SM Table 3**).

1.4 Addition of publicly available data

In order to contextualise our data with regard to publicly available datasets including African and global populations, we generated a number of datasets using publicly available data from Khoe-San, North African, 1000 Genomes Project, Human Origins and HGDP populations (**SM Table 5**). All data downloaded was filtered using stringent QC per population, applying the same process outlined for AGVP data, for consistency. Only SNPs retained following QC among all populations for each dataset were included in each final dataset. A summary of all datasets curated can be found in **SM Table 5**.

1.10 Data availability

Raw and curated genetic data has been deposited at the European Genome-phenome Archive (EGA, http://www.ebi.ac.uk/ega/), which is hosted by the EBI, under accession numbers (EGAS00001000959 (genotype data), EGAS00001000363 **(**Uganda 4x WGS), EGAS00001000238 (Ethiopia 8x WGS), EGAS00001000286(Zulu 4x WGS) and EGAS00001000960 (curated WGS vcf files)). All source code for analyses is available on correspondence with authors.

SM Table 5: A summary of datasets used in different analyses

Dataset	Component populations	Sample no.	SNP density
AGV dataset	AGVP populations	1,481	1,577,224
Global+AGV dataset	1000 Genomes Project[2]* +AGVP populations	2,864	1,399,027
AGV extended dataset	AGVP populations + Khoe San (Schlebusch)[4] + MKK (1000 Genomes Project[2])	1,605	905,145
AGV extended + HGDP African + North African + Khoe San datasets	AGVP populations + Khoe San (Schlebusch)[4] + MKK (1000 Genomes Project[2]) + North African (Henn)[5] † + Khoe San (Henn)[6] ‡ + HGDP African populations§	1,819	21,448
Global extended dataset	1000 Genomes Project[2] + AGVP populations + Khoe San (Schlebusch)[4] + MKK (1000 Genomes Project)	2,988	826,965
Global extended+ Human origins array data	1000 Genomes Project[2] + AGVP populations + global populations on the Human origins array (Pickrell)[7]	3,904	139,950
Global extended + HGDP + North African + Khoe San datasets	1000 Genomes Project[2] + AGVP populations + Khoe San (Schlebusch)[4] + MKK (1000 Genomes Project) + North African (Henn)[5] + Khoe San (Henn)[6] + HGDP global populations	3,202	19,675

*1000 Genomes Project genotype raw intensity data can be found at ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20120131_omni_genotypes_and_intensities/
† Data downloaded from http://bhusers.upf.edu/dcomas/north-african-affy-6-0-data-henn-et-al-submitted/
§ Raw genotype data downloaded from http://www.hagsc.org/hgdp/files.html
‡ Data downloaded from http://www-evo.stanford.edu/repository/paper0002/
MKK: Masaai

2.0 CURATION OF AGVP SEQUENCE DATA

2.1 WHOLE GENOME LOW COVERAGE SEQUENCING

We examined low coverage (mean coverage 4x) whole genome sequence from 320 individuals from three distinct population groups **(SM Table 2**). Sequencing was carried out on the Illumina HiSeq 2000 platform. Genomic DNA (approximately 1 ug) was fragmented to an average size of 500 bp and subjected to DNA library creation using established Illumina paired-end protocols. Adapter-ligated libraries were amplified and indexed via PCR. A portion of each library was used to create an equimolar pool comprising 8 indexed libraries. Libraries were subjected to 100 base paired-end sequencing (HiSeq 2000; Illumina) following manufacturer's instructions. Sequencing was carried out to achieve a mean coverage of approximately 4x across the genome. For the Ethiopian population, further sequencing was carried out to achieve an average coverage of 8x across the genome.

2.2 DATA PROCESSING AND CURATION OF SEQUENCE DATA

Following generation of raw reads, duplicate reads were marked, and mapping was carried out using BWA v0.5.10 to the human reference genome (GRCh37). The GRCh37 version used was the same as that used by phase II of the 1000 Genomes Project, including decoy configurations and the Ebstein Barr Virus sequence. The BWA Backtrack Algorithm was used for mapping. Lanes were merged into samples, and sample level bam improvement was carried out using GATK, as described in **Supplementary Note 9**. Two samples from the Complete Genomics and the Platinum Genomes highly curated set (http://www.illumina.com/platinumgenomes/) were included for validation of the data processing pipeline (NA12878 and NA19240) (see **Supplementary Note 9**). Genotype calling was carried out across all 320 samples using Unified Genotyper v 2.4. We used the emit_variants_only setting, and did not call any insertions or deletions in this dataset. Annotation of variants was carried out using various annotations with VariantAnnotator (**Supplementary Note 9**). Following this we carried out Variant Quality Score Recalibration (VQSR) using the HapMap and 1000 Genomes Omni 2.5 M data as truth and training datasets. A tranche sensitivity of 99% was applied for filtering, as this corresponded to the highest point on the ROC curve generated for the gold standard samples used for validation (**Supplementary Note 9, SN9 Fig 2**). Following this we carried out genotype refinement of probabilities using Beagle v3.3.2. Accuracy of the final dataset was noted to be high with a mean correlation r^2 of > 0.95 with genotype data for all three populations. Using the test samples, we obtained a final sensitivity and specificity for capture of variants of 97% and 91% respectively. Finally, we phased the data generated with SHAPEIT2 v 2 release 727 in order to generate a reference panel for imputation.

3.0 ANALYSIS OF POPULATION STRUCTURE AND GENE FLOW

3.1 EVALUATION OF ALLELE FREQUENCY SPECTRA

We plotted the allele frequency spectra for all African and global populations (**Supplementary Note 2**). Derived allele frequencies were calculated for each population, and the frequency of these was plotted within allele frequency bins. The curation of derived alleles is described in **section 3.8**. The number of monomorphic sites (those with derived allele frequency=0) was also calculated for each population. We compared allele frequency spectra generated from the chip with those generated from sequence data (**Supplementary Note 2**). In order to assess allele frequency spectra across the populations sequenced, we used direct approaches used to assess spectra from reads,[8] rather than calculating the allele frequency proportions from multi-sample called data, which has been shown to be biased and underestimate rare variants.[9] For this, we used the maximum likelihood method calculated by EM optimisation implemented in ANGSD (http://popgen.dk/angsd/index.php/SFS_Estimation).[8]

3.2 PRINCIPAL COMPONENT ANALYSES

Principal component analyses were carried out using EIGENSOFT v4.2. PCA was carried out after LD pruning to a threshold of r^2=0.5 using a sliding window approach with a window size of 50 SNPs, sliding 5 SNPs sequentially. Regions of known long range LD were removed from analysis, as described previously.[10] PCA was carried out for all datasets including extended datasets described in **Extended Data Figures 2-5** and **Supp. Figures 1-2**. Additional PC analyses for evaluation of clines were carried out using YRI and CEU/Ju/'hoansi for calculation of PCs with projection onto remaining individuals (**Extended Data Figure 5**).

3.3 POPULATION DIFFERENTIATION

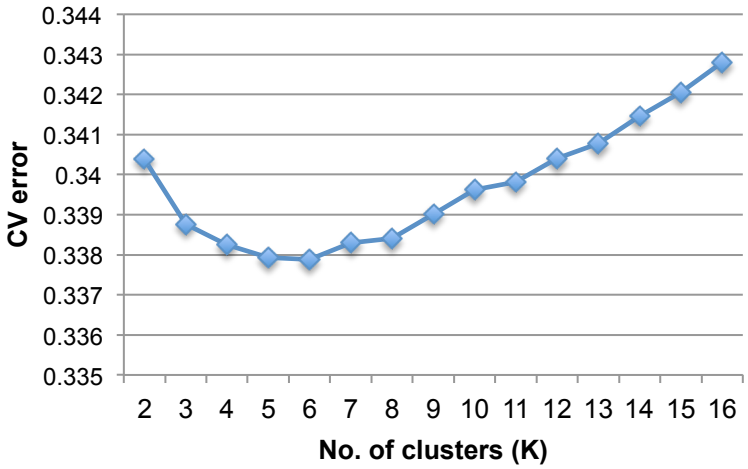
F_{ST} was calculated using the Hudson estimator as a measure of population differentiation using EIGENSOFT v4.2, as this is not dependent on sample size ratios and does not systematically overestimate F_{ST}.[11] Calculation was carried out in the genotype data: the African dataset, global dataset and for the African extended datasets, and the sequence data, in order to contextualise population differentiation in a global and diverse African context, as well as compare estimates to assess the degree of ascertainment bias affecting F_{ST} estimation on the chip. Although F_{ST} estimates from the chip were correlated with those from sequence data, genotype data tended to over-estimate these parameters, suggesting some degree of ascertainment bias (**Supp Note 2**). However, we must note that rare variants called in sequence data are also likely to be underestimated with multi-sample calling, making it possible that even estimates from low

coverage sequence data are biased, as has been noted before.[9] Mean pairwise F_{ST} metrics were calculated for broad linguistic groups, and by region. Our estimates of Europe-Africa differentiation were comparable to the Hudson F_{ST} estimates from sequence data from the 1000 Genomes Project reported.[11] In order to evaluate the effect of Eurasian admixture on differentiation among African populations, we carried out masking of Eurasian ancestry among all populations (**Supplementary Note 6**) and re-calculated F_{ST} statistics.

3.4 ADMIXTURE CLUSTERING ANALYSIS

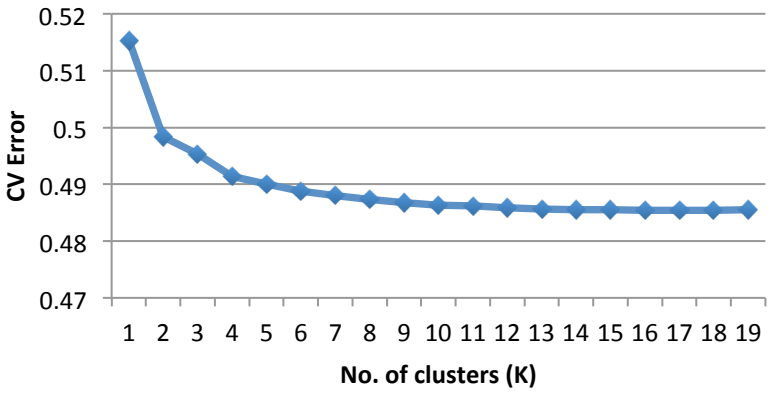
Clustering of genetic data from the AGVP was carried out using ADMIXTURE specifiying K=2 to 20 clusters in each model for all datasets in **SM Table 5**. Cross validation was used to assess the number of clusters with the best fit for each analysis (**SM Figures 1-3**). ADMIXTURE analysis was carried out on the AGVP set (**Figure 1b**), the African and global extended datasets including Khoe-San and HGDP populations (**Extended Data Figure 6**), and the global dataset including data from the Human Origins Array (**Figure 1c**).[12] ADMIXTURE analyses were repeated 20 times using a seed derived from the time of analysis, and results were combined using the *LargeKGreedy* algorithm in CLUMPP with 1000 repeats.[13] LD pruning to an r^2 of 0.2 was carried out prior to analysis, and known regions of long range LD were removed, as previously described.[10]

SM Figure 1: Cross validation error of the African dataset on ADMIXTURE analysis



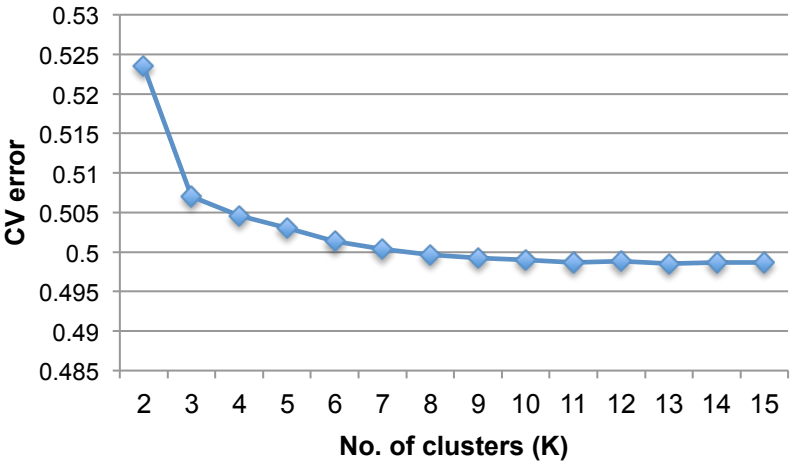
SM Fig 1 represents the five fold cross-validation (CV) error on ADMIXTURE analysis of the African dataset for different numbers of clusters specified. The CV error is the lowest for K=6, suggesting that this best fits the data.

SM Figure 2: Cross validation error of ADMIXTURE analysis of the global extended dataset including 1000 Genome Project populations + AGVP + Human origins array dataset



SM. Figure 2 represents the five fold cross-validation (CV) error on ADMIXTURE analysis of an global extended dataset for different numbers of clusters specified. The CV error is the lowest for K=18, suggesting that this best fits the data.

SM Figure 3: Cross validation error of ADMIXTURE analysis of the global extended dataset including 1000 Genome Project populations + AGVP + North African + HGDP+ Khoe-San populations



SM. Figure 3 represents the five fold cross-validation (CV) error on ADMIXTURE analysis of the global extended dataset for different numbers of clusters specified. The CV error is the lowest for K=13, suggesting that this best fits the data.

3.5 Formal tests for gene flow (*f3* tests)

As neither principal component analysis, nor ADMIXTURE analysis are formal tests of admixture, we sought to formally assess admixture in the AGVP populations. In order to examine Eurasian ancestry in AGVP populations, we tested a model with admixture between populations related to CEU and YRI by using these as reference populations testing the tree (CEU, YRI; X), X being each of the AGVP populations. Here, Eurasian ancestry/gene flow refers to ancient gene flow from an ancestral population that is closely related to populations currently living in Western Europe. However, as it is difficult to identify the precise source of this ancestry, which may be the result of multiple population movements—including through Europe, the Middle-East, or from other parts of Africa, we will hereby broadly refer to this as

'Eurasian gene flow/ancestry'. *f*3 tests are robust to complex ancestry in the admixing populations, ascertainment bias and the choice of reference populations, as has been described previously.[12] The test statistic is negative if X has complex history and admixture from populations related to the reference populations, as this topology that would lead to a negative term in the *f*3 parameter. We similarly tested for hunter-gatherer admixture in each of the AGVP populations, by carrying out *f3* tests of the form (Ju/'hoansi, YRI; X), as Ju/'hoansi is well known to have the lowest known non-Khoe-San admixture, thereby providing a good reference for *f3* tests.[14] However, as Khoe-San populations could be outgroups to actual admixing populations, especially in East and West Africa, where the admixing population may be other HG populations in the region (e.g. Pygmies), we also evaluated *f3* tests using other reference populations, including Biaka Pygmy, Mbuti Pygmy and Hadza. We compared the *f3* statistics using different references, to try and identify which reference populations were most closely related to the ancestral mixing populations. A Z score of below -3 (equivalent to a $p<0.001$) was considered statistically significant). In sensitivity analyses, we additionally evaluated the possible effects of ascertainment on *f3* statistics (**Supplementary Note 2**).

3.6. Addition of chimpanzee sequence data for outgroup tests

In order to carry out *f4* ratio tests for analysis of admixture, we merged our global extended dataset with chimpanzee sequence data.[15] Chimpanzee sequence data was downloaded from the ftp link provided (ftp://birch.well.ox.ac.uk/BAMs/HG18/). Genotype calling was carried out for all sites on the 2.5M quad chip on the hg18 build using the hg18 human reference including EBV sequence. Genotype calling was carried out from bam files from the ten chimpanzee sequences with Unified Genotyper (GATK v2.15) using the options -stand_call_conf 30.0 -stand_emit_conf 10.0 -dcov 100. We then used vcftools to filter all sites with a sample called proportion <80% and sites with a QUAL score of below 30. Liftover was used to map these data onto the hg19 build. Data from these ten samples were subsequently used for outgroup analyses of *f4* ratio tests, and to curate derived alleles, as described in section 2.8.

3.7 F4 ratio tests

We further sought to quantify admixture using *f4* ratio tests.[12] We were interested in quantifying the proportion of Eurasian ancestry among the AGVP populations, and Khoe-San populations, as well as quantifying the proportion of HG (Khoe-San/Hadza/Pygmy) ancestry among the AGVP populations. We used different topologies to examine the proportion of admixture from non-SSA and Khoe-San populations (**Supp. Notes 3 and 4**).

i. Quantifying Eurasian ancestry

To calculate the proportion of Eurasian admixture, we calculated the ratio of *f4* tests (Han, Orcadian; YRI,X) and (Han,Orcadian;YRI,Druze), as these would provide an estimate of proportional West Eurasian ancestry in test population X (**SN3 Figure 1**). This ratio is complicated by a small proportion of African ancestry among Druze. We statistically corrected for this, by following a similar procedure to Pickrell et al.[7] Similarly, this ratio may also be affected by the small proportion of Eurasian ancestry among YRI. In order to reduce bias due to this, we only included YRI individuals with <0.0025% of Eurasian ancestry on ADMIXTURE clustering (K=18), designated as 'pure YRI'. We compared ratios obtained using this method, with those obtained with masking of Eurasian ancestry among YRI and African ancestry among Druze. We describe these analyses in more detail in **Supplementary Note 3**.

ii. Quantifying hunter gatherer ancestry

To calculate the proportion of HG ancestry among the AGVP populations, we calculated the ratio of the *f4* tests *f4*(Jola, Chimp; Pygmy, X) and *f4*(Jola, Chimp; Pygmy, YRI) for Pygmy-like ancestry, and the ratio of *f4*(Jola, Chimp; Ju/'hoan North, X) and *f4*(Jola, Chimp, Ju/'hoan North, YRI) for Khoe-San like ancestry among admixed populations. In order to avoid bias due to Eurasian and Bantu ancestry among Ju/'hoan North, and Eurasian and HG ancestry among YRI and Jola, we applied several approaches including using only unadmixed individuals in analysis, and masking non-HG ancestry among Ju/'hoan North/Pygmy and Eurasian ancestry among YRI and Jola. We describe these analyses in detail in **Supplementary Note 4**.

3.8 Curation of derived alleles

We used a combination of the 1000 Genomes ancestral reference sequence, and sequence data called from 10 chimpanzee sequences, to curate a set of derived alleles for the 2.5 Omni chip array. For all sites on the chip, we noted that 6.2% of sites from the 1000 Genomes human ancestral sequence had poor support (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/supporting/ancestral_alignments/human_ancestor_GRCh37_e59.README). In order to assess the accuracy of these derived alleles, we compared the 1000 Genomes human ancestral sequence data for the sites on the chimp, with the consensus sequence that we had generated from ten chimpanzee sequences. There was high concordancy (97%). However, among discordant sites, 94% of 1000 Genomes ancestral alleles were low confidence, indicating that low confidence alleles were enriched among discordant sites (fisher exact test, p value <0.00001). We used an approach whereby we

merged these two sources of information, prioritising the consensus chimpanzee sequence when there was discordancy between the 1000 Genomes human reference ancestral sequence, where the alleles had poor support or were missing. We excluded 3,411 sites where both sources of information provided high confidence but discordant alleles.

3.9 Dating and quantification of admixture using admixture LD based approaches

Although admixture is a complex and dynamic process among populations, in order to understand gene flow between populations, we sought to model this as discrete point admixture events occurring between two populations at a time. We also modelled more complex admixture with multiple mixing events using more sophisticated modelling methods outlined in **Supplementary Note 5**. In order to confirm the presence of admixture, and date this, we used admixture LD based approaches.[7,16] This approach is based on the relationship between admixture LD, time since admixture and the difference in allelic frequency between SNPs in mixing populations. It leverages the fact that admixture LD between 2 SNPs weighted by difference in allelic frequency between two mixing populations decays exponentially as a function of time since admixture. The amplitude of the curve allows estimation of admixture proportions. We assessed multiple admixture events and identified populations most similar to ancestral mixing populations for AGVP populations, using methods described previously.[7] For these analyses, we estimated curves from a minimum distance of 0.5cM. We estimated the lower bound and upper bounds of the number of generations since admixture for each event, by assessing the rate of decay of each exponential curve. We describe these methods in further detail in **Supplementary Note 5**.

3.10 Haplotype phasing of datasets

Haplotype phasing of datasets was carried out in order to infer local ancestry among haplotypes using PCAdmix.[17] We also used phased haplotypes for long range haplotype tests for analyses of selection sweeps (as outlined in **section 4.2**), and for imputation among these populations for assessment of LD structure and fine mapping (as outlined in **section 5.3**). All datasets were phased using SHAPEIT2[18] standard parameters, and an effective population size of 17,469, as recommended for African populations. For the phasing step, all related individuals removed during quality control, and additional individuals outside of 100 subsampled individuals in each population were merged back with these datasets, in order to maximise haplotype phasing accuracy.

3.11 MASKING EURASIAN ANCESTRY IN AGVP POPULATIONS

In order to assess the effect of Eurasian ancestry on genetic variation observed among AGVP populations, we repeated PCA, population differentiation and ADMIXTURE clustering analyses after removing genomic segments of Eurasian ancestry. This was carried out by masking haplotypes of Eurasian ancestry using PCAdmix.[17] The Hapmap recombination maps were used for analysis. Masking was carried out using a fixed window length of 20 SNPs following LD pruning to an r^2 threshold of 0.8, as recommended.[17] YRI from the 1000 Genomes Project was used to represent SSA ancestry, and JPT+CHB, and CEU were included independently as reference populations in order to capture non-SSA ancestry. As PCAdmix is robust to ancestral populations being in some cases quite distant from the true ancestral populations, we believe that this choice of populations would be appropriate to capture non-SSA ancestry among the AGVP population groups. Once ancestry of windows was ascertained with PCAdmix, we only retained genomic segments that showed >90% probability of SSA ancestry, as represented by YRI. Using this strategy, even if non-SSA ancestry could not be accurately represented by European and Asian reference populations, this would still be likely to be excluded, as only genomic segments with >90% probability of YRI-like ancestry were retained. All markers between the first SNP of each SSA ancestry window, and the last SNP of the window, including LD pruned SNPs were assigned SSA ancestry if the window had >0.90 probability of YRI ancestry. We also assessed the accuracy of PCAdmix with simulations, as we describe in **Supplementary Note 6**.

3.12 ANALYSIS OF POPULATION DIFFERENTIATION AFTER EURASIAN ANCESTRY MASKING

We repeated Hudson's F_{ST} between populations, after masking of Eurasian ancestry. The proportional change in F_{ST} metrics of differentiation would indicate the level of differentiation between populations that may be explained due to gene flow between SSA and non-SSA populations. One limitation of these analyses is that masking is likely to be imperfect, especially for older admixture >50 generations ago. Using simulations, we show that in the presence of imperfect masking, we are likely to overestimate the level of differentiation between populations, so these estimates should be thought of as an upper bound (**Supplementary Note 6**).

4.0 ANALYSIS OF LOCI UNDER SELECTION

We used two methods to identify loci under selection. To examine local adaptation and differentiated loci within Africa, we carried out single locus F_{ST} tests, and in order to detect signals associated with selection sweeps, we carried out long range haplotype tests (the integrated haplotype score test). F_{ST} tests were carried out between European populations from the 1000 Genomes Project (FIN, GBR, CEU, IBS and TSI) and AGVP populations to assess loci highly differentiated between Europe and Africa. Additionally, a global F_{ST} across Africa was examined to identify loci most differentiated across the 16 African populations. Differentiation of loci were also assessed between populations exposed to certain environmental adaptive pressures, such as malaria, lassa fever, trypanosomiasis and trachoma infection. Parallel analyses were carried out using long range haplotype tests. A list of exposed and unexposed groups are presented in **SM Table 6.**

SM. Table 6: Classification of infectious disease exposure among AGV populations for Fst and IHS analyses

Disease	Highly exposed populations	Unexposed/Low exposure
Malaria	Barundi, Baganda, Banyarwanda, LWK, Somali, Oromo, Amhara, Wolof, Jola, Fula, Mandinka, Ga-Adangbe, YRI, Igbo	Sotho, Zulu, Kalenjin, Kikuyu
Lassa fever	YRI, Igbo	Barundi, Baganda, Banyarwanda, LWK, Somali, Oromo, Amhara, Wolof, Jola, Fula, Mandinka, Ga-Adangbe, Kalenjin, Kikuyu, Sotho, Zulu
Trachoma	Ethiopia (Oromo, Somali, Amhara)	Barundi, Baganda, Banyarwanda, LWK, YRI, Igbo, Wolof, Jola, Fula, Mandinka, Ga-Adangbe, Kalenjin, Kikuyu, Sotho, Zulu
Trypanosomiasis	Barundi, Baganda, Banyarwanda	LWK, YRI, Igbo, Wolof, Jola, Fula, Mandinka, Ga-Adangbe, Kalenjin, Kikuyu, Sotho, Zulu, Oromo, Amhara, Somali

Sources of data used to assign exposure included:

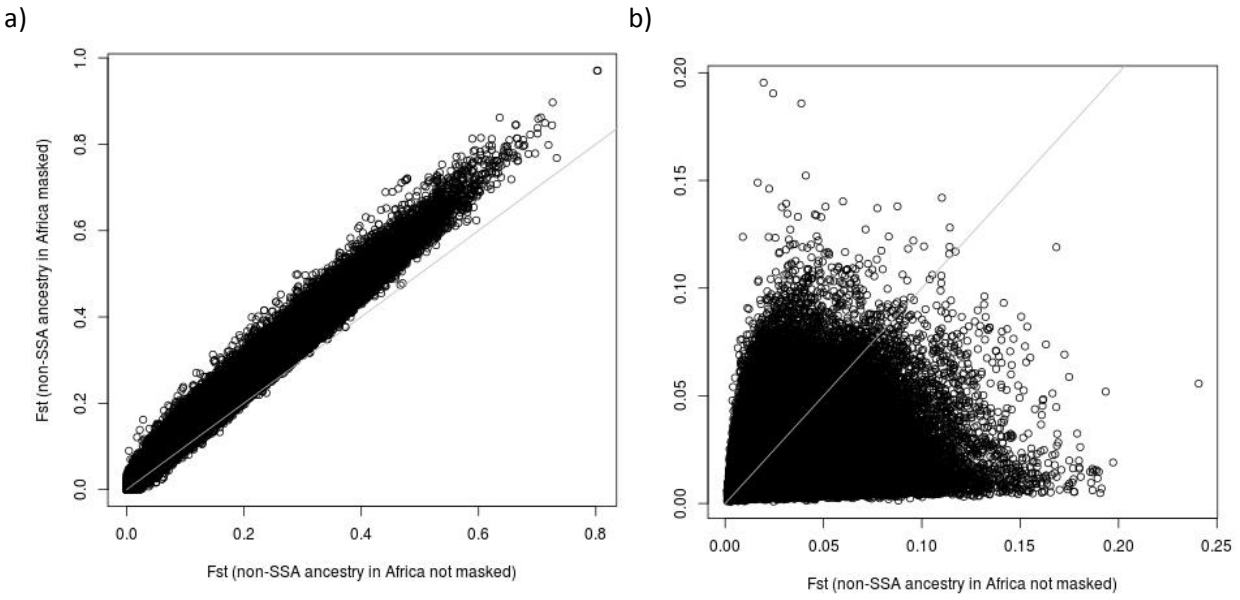
http://www.who.int/malaria/publications/world_malaria_report_2012/wmr2012_full_report.pdf
http://www.mara.org.za/pdfmaps/AfDistribution.PDF
www.cdc.gov/malaria/travelers/
http://www.trachomaatlas.org/
http://www.who.int/blindness/data_maps/en/
http://www.ij-healthgeographics.com/content/9/1/57
http://www.who.int/mediacentre/factsheets/fs259/en/
http://www.plosntds.org/article/info%3Adoi%2F10.1371%2Fjournal.pntd.0000388
http://www.who.int/csr/disease/lassafever/en/

4.1 SINGLE-LOCUS F_{ST} ANALYSIS

Locus specific F_{ST} metrics were calculated between African populations, between Africa and Europe and between groups putatively exposed/unexposed to a range of tropical diseases (**SM Table 6**) using Wright's formula [19]. We used Wright's F_{ST} metric, as this can readily be calculated from allele frequencies of variants among populations, even when single haplotypes from individuals are masked. This facilitated comparison of F_{ST} metrics before and after masking of regions of Eurasian ancestry among populations. Comparisons with Weir and Cockerham metrics on unmasked data showed very high correlation with Wright's F_{ST} (data not shown), suggesting that these metrics of differentiation were able to appropriately rank differentiated loci across the genome. The same formula was also used to calculate a single global F_{ST} across Africa, based on all the populations typed in this study. Wright's formula estimates F_{ST} from the variance of allele frequency and its mean value in each assessed population [20]. After deconvoluting each putatively admixed African genome into its SSA and non-SSA components with PCAdmix [21], the allele frequencies retrieved after masking the non-SSA chunks in each individual genome were used to calculate F_{ST}. The values retrieved for each pair and global set, before and after the deconvolution, were compared to assess the effect of Eurasian admixture on the observed differentiation (**SM Figure 4**). Furthermore, in each pair or global set, the SNPs falling in the top 0.1% of F_{ST} distribution and their genetic regions were searched for enrichment in specific biological pathways or for genes putatively involved in adaptive processes. In order to confirm that a large proportion of differentiation arose due to selection, and not random drift alone, we examined enrichment of known loci under selection among the most differentiated candidates. We carried out a chi^2 test on 1df on genes in a 50Kb flanking region of each differentiated site, assigning 46,262 as the total number of genes in the autosomal region. Known genes under selection were compiled as outlined in **section 4.3**.

As differentiation in regions with low diversity due to background selection, due to reduced effective population size and increased drift, we also assessed whether background selection was predominantly driving differentiation. We outline these analyses in **Supplementary Note 7**.

SM Figure 4: Scatterplot showing correlation between Europe-Africa and within Africa F_{ST} calculated before and after masking of Eurasian genomic ancestry



SM Figure 4a shows the correlation of Europe-Africa F_{ST} metrics before and after masking of Eurasian ancestry (r=0.99). Supp Figure 4b shows the correlation of within Africa F_{ST} metrics before and after masking of non-SSA ancestral segments (r=0.51). The moderate correlation suggests that Eurasian admixture has a substantial impact on locus specific differentiation and should be considered in selection scans among admixed populations.

4.2 Integrated Haplotype Score (iHS) Analysis

Integrated Haplotype Scores were estimated using the methods outlined by Voight et al[22] with the software WHAMM (http://coruscant.itmat.upenn.edu/whamm/index.html). Scores were calculated for each site within the AGVP dataset as a whole and standardised within allele frequency bins of 10%. Following standardisation, the proportion of absolute standardised iHS scores >2 in each 200 KB window were calculated, as outlined by Pickrell et al.[23] This method has been shown to be more powerful for identification of selection sweeps as compared to single locus scores.[22] As the use of a different window size, or overlapping sliding windows did not materially alter results in previous published work,[23] we chose to use this approach to minimise missing data, and for comparability with published literature. Following these calculations, empirical p values were calculated by binning windows by the number of SNPs in each (in bins incremented by 20 each, and excluding windows with <20 SNPs) calculating the proportion of iHS values greater than a given value, divided by the total number of windows in that bin.[23] Windows with the lowest 0.1% p values were then examined further to identify loci with the most extreme IHS scores, as these were most likely to be linked to regions under selection.

4.3 ANNOTATION OF SELECTION LOCI

The top 0.1% of F_{ST} candidates, and windows with empirical p values in the lowest 0.1% were annotated for genes using SNIPPER (http://csg.sph.umich.edu/boehnke/snipper/). For F_{ST} loci, annotation was carried out in the flanking 50 KB regions, and for IHS, the 200 KB windows with the lowest 0.1% of empirical p values were annotated. All genes in the region were annotated, and the consequence associated with each F_{ST} SNP was also annotated using VEP version 2.1. CONDEL was used for annotation of protein consequence, to identify potential non-synonymous mutations with deleterious consequence on protein. We also compiled a list of genes identified as under selection based on the recent literature[22-25] to identify known loci under selection. Enrichment for known loci was tested in 50kb regions around the most differentiated candidates using a chi^2 test.

5.0 EVALUATION OF LD STRUCTURE AND IMPUTATION ACCURACY IN AGVP

5.1 EXAMINING LD DECAY AMONG POPULATIONS

In order to compare LD decay among AGVP populations, and globally, we calculated LD decay over a distance of 300 KB in each population extracted from the global dataset, including only genetic variants with minor allele frequencies >0.05. LD was calculated from the phased haplotypes using the r^2 metric. We calculated LD on an equal number of haplotypes randomly sampled from each population, to make data comparable across populations (**Supp Figure 6**).

5.2 EVALUATION OF IMPUTATION ACCURACY IN THE CHIP

We evaluated imputation accuracy in two ways: 1. First, we evaluated the comparative accuracy of imputation into the Omni 2.5M genotype data among African and global populations using the 1000 Genomes Project reference panel; 2. Secondly, we thinned the Omni 2.5M data to the sites overlapping with the Illumina OmniExpress chip array, and compared the imputation accuracy with the full 2.5M dataset among African populations. These analyses are detailed in **Supplementary Note 10**.

The IMPUTE2 algorithm was used for phasing and imputation for both comparisons. During the process of this evaluation, the new 1000 Genomes Project reference panel (September 2013 release) that had been produced using SHAPEIT2 became available, so this was used for the imputation into individual populations in order to maximise accuracy in comparisons. Imputation was carried out in 2 MB chunks and then concatenated. Imputation accuracy was calculated by masking each locus in the genotype data sequentially and calculating the correlation between imputed data and original genotype data ('leave one out masking'). We assessed the relationship between mean imputation accuracy in each population and the minimum genetic distance from the 1000 Genomes Project reference panel. For assessment of overall imputation accuracy in each population, we calculated a weighted mean r^2. This was calculated by assigning equal weights to the mean r^2 estimates in each allele frequency bin in order to avoid bias due to differences in accuracy among populations arising from differences in allele frequency spectra.

5.3 Assessment of improvement in imputation accuracy using the AGVP reference panel

We generated a new reference panel of African populations with low coverage sequences from 3 population sets (Baganda, Zulu and Ethiopia). We assessed the improvement of imputation accuracy using a merged reference panel including the 1000 Genomes Project phase 1 v3 integrated reference panel (1000GP) and the AGVP whole genome sequence data from 320 individuals belonging to 3 distinct populations (**Supplementary Note 11**). The improvement in imputation accuracy was calculated as the difference in weighted mean r^2 (as described in **section 5.2**) when using the merged reference panel (1000GP + AGVP) in comparison with 1000GP alone. We assessed improvement in imputation accuracy as a function of distance of each population from the closest population in 1000GP panel (by F_{ST}) and the minimum distance from any population in the AGVP panel. This metric was calculated as the difference of the minimum F_{ST} from any population in the 1000GP panel and the minimum F_{ST} distance from populations in the AGVP panel. One would expect improvement to be greater for populations poorly represented in the 1000GP and well represented in AGVP; therefore, this score is expected to correlate with improvement in imputation accuracy with the AGVP reference panel. We discuss these analyses in detail in **Supplementary Note 11**.

5.4 Evaluation of more efficient chip designs in African populations

Given that current genotype array designs have been largely ascertained on European individuals, or on few African populations, we sought to evaluate a genotype array design to capture common variation across 5 different African populations (YRI, LWK, Baganda, Zulu and Ethiopia). For this evaluation, we carried out multi-population tagging using an in-house algorithm developed based on TAGster.[26] This method utilised an algorithm identical to TAGster but several fold higher in computational efficiency, allowing fast tagging across the whole genome. We describe this in more detail in **Supplementary Note 12**. As single marker tagging is not an efficient method to capture variation, we used a previously described iterative cyclical method, alternating cycles of single-marker tagging, with imputation-based tagging,[27] so as to efficiently capture sites that were tagged by haplotypes as well as single markers (see **Supplementary Note 12**).

5.5 Assessment of the fine mapping potential in AGVP populations

Although it is well known that patterns of LD are different between African and non-African populations, the differences in LD structure within Africa have not been previously characterised in detail. We sought to explore LD structure within AGVP populations, by simulating traits associated with known causal loci associated with traits identified through

GWAS and biological studies, and examining association patterns around these regions using imputed data. Following haplotype phasing, as described in **section 3.10**, imputation was carried out into the global dataset in order to examine loci not present in the genotype data. Imputation was carried out on pre-phased data with IMPUTE2 using the 1000 Genomes Project phase I integrated reference panel (March 2012 release) including multi-ethnic populations (1092 individuals), based on standard recommendations. (https://mathgen.stats.ox.ac.uk/impute/impute_v2.html#best_practices). Imputation was carried out in chunks of 2MB and then concatenated. As haplotype phasing was carried out including related individuals and individuals removed at the subsampling stage in order to maximise phasing accuracy, these individuals were removed from the dataset prior to imputation.

We chose regions that are known to be causally associated with certain traits, and have different LD structure in global populations. These included the *SORT1* locus (associated with LDL), the sickle cell variant at the *HBB* locus (associated with malaria), the *APOL1* locus (associated with chronic kidney disease and trypanosomiasis susceptibility), the *TCF7L2* locus (associated with fasting glucose), the *PRDM9* locus (associated with recombination hotspot usage) and *APOE* variants (associated with Alzheimer's disease) (**SM Table 7 and Extended Data Figure 9**). These loci known to have different LD structure globally, and the sickle cell locus (associated with malaria), is known to be differentiated across Africa. We used imputed data for analysis, after confirming that causal loci were imputed with high quality. Phenotypic traits were simulated across all populations using GCTA[28] with a heritability of 70% for each trait. A high heritability was chosen in order to simulate associations, given the small sample sizes for each population. Effect sizes were extracted from beta coefficients published in the previous literature (**SM Table 7**). One MB regions around each causal locus were extracted and association analyses was carried out on genotype dosages using GEMMA[29] mixed model analysis. Association analyses were carried out separately for European populations (CEU, GBR, FIN, IBS and TSI from the 1000 Genomes Project), Asian populations (CHB, CHS, JPT, KHV and CDX) and African populations (AGVP populations). Association signals were examined within each population to identify differences in LD structure among populations. Regional plots were visualised using LocusZoom[30] after removing the causal SNP, to examine the pattern of association and LD in different AGVP populations. We also examined LD structure around each region using HAPLOVIEW.

SM. Table 7: Parameters used for simulation of traits and biologically relevant loci

Locus	*Causal SNP*	*Importance*	*Simulated beta [ln(OR)]*	*Simulated Heritability*
SORT1[31]	rs12740374	Associated with LDL-C	-0.167	70%
HBB[32]	rs334 (sickle cell variant)	Associated with severe malaria and under selection in Africa	-2.40	70%
APOL1[15]	rs73885319	Trypanolytic factor under selection; associated with CKD in African-Americans in GWAS	1.17	70%
TCF7L2[33]	rs7903146	Associated with type 2 Diabetes	0.30	70%
APOE[34]	rs429358	Implicated in Alzheimer's disease	1.43	70%
PRDM9[35]	rs6889665	Implicated in recombination patterns and hotspot usage	1.10	70%

LDL-C: LDL cholesterol; GWAS: Genome wide association study

S. NOTE 1: CURATION OF AGVP GENOTYPE DATA AND REMOVAL OF CHIP EFFECTS

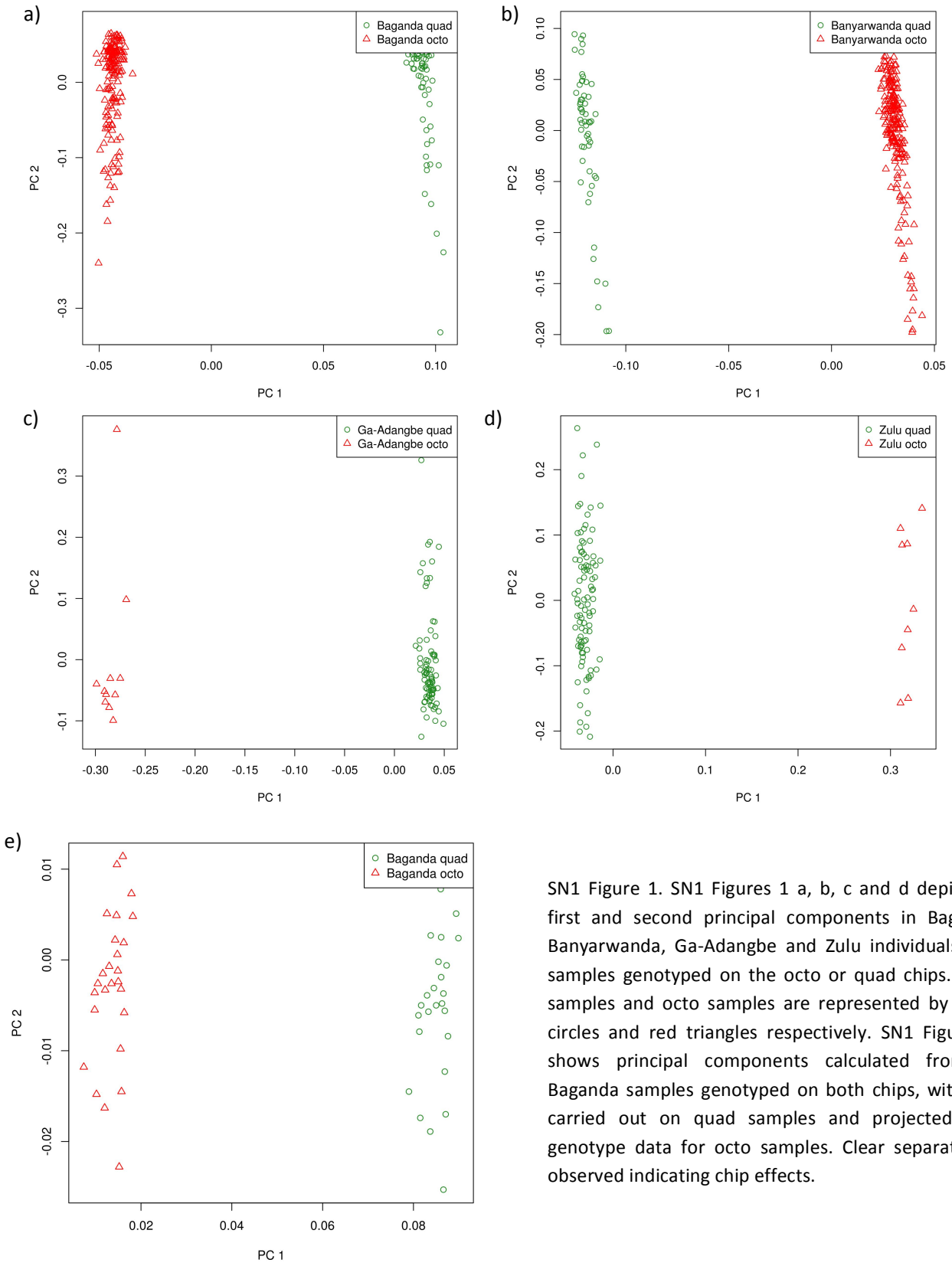
1.1 SAMPLES AND POPULATIONS

We genotyped 2,185 samples from 16 different African populations from SSA on the Illumina HumanOmni 2.5M BeadChip array and sequenced 320 individuals at 4x coverage from 7 ethno-linguistic groups (**SM Tables 1 and 2**). As the samples were genotyped at different times, 720 samples were genotyped on the Illumina HumanOmni 2.5M-quad BeadChip array (hereafter referred as quad) and 1,465 on the Illumina HumanOmni 2.5M-8 BeadChip array (hereafter referred as octo) (**SM Table 3**), which replaced the quad chip on the market during the course of the study. The octo and quad arrays were noted to be very similar, with 99.7% overlap between genotyped sites. To assess concordance of genotypes between chips, 29 samples from one population (Baganda) were genotyped on both platforms. Additionally, four populations from East, West and South Africa (Baganda, Banyarwanda, Ga-Adangbe and Zulu) were genotyped partially on the quad chip and partially on the octo chips, providing us with a further opportunity to rigorously examine and control for any chip effects. In order to supplement the African diversity panel, we also included 2.5M-quad data for Yoruba (YRI) and Luhya (LWK) individuals from the 1000 Genomes Project. Masaai (MKK) was not included, due to the small number of samples. In total, 18 population groups were studied (**SM Table 1**, **Figure 1**).

1.3 REMOVAL OF CHIP EFFECTS FROM DATASETS: AN OVERVIEW

Although concordance between the quad and octo chips for samples genotyped on both chips among Baganda samples was high (>98%), clear chip effects were observed on PCA of individual populations with samples on both chips (**SN1. Figure 1**), the global (**SN2. Figure 2**) and African datasets (**SN1 Figure 8**). As chip effects are likely to bias subsequent analyses, we sought to identify and filter out variants causing differential genotyping between chips. We identified principal components (PCs) representing chip effects for the global and African datasets, and calculated SNP loadings along these principal components, in order to remove highly weighted SNPs along these components. Correlation between SNP weights and discordancy in genotypes between Baganda samples with duplicated genotyping on both chips along PCs was high for both datasets (**SN1 Figure 3 and 10**), further substantiating the utility of this approach. We then removed these SNPs, and confirmed the absence of chip effects by repeating PCA of global, African and individual population datasets. We also confirmed the absence of chip effects in individual African populations by carrying out per population PCA for the four populations genotyped across both chips for sets of SNPs within each dataset, and inspecting statistically significant principal components for chip effects. For Baganda samples genotyped on both chips,

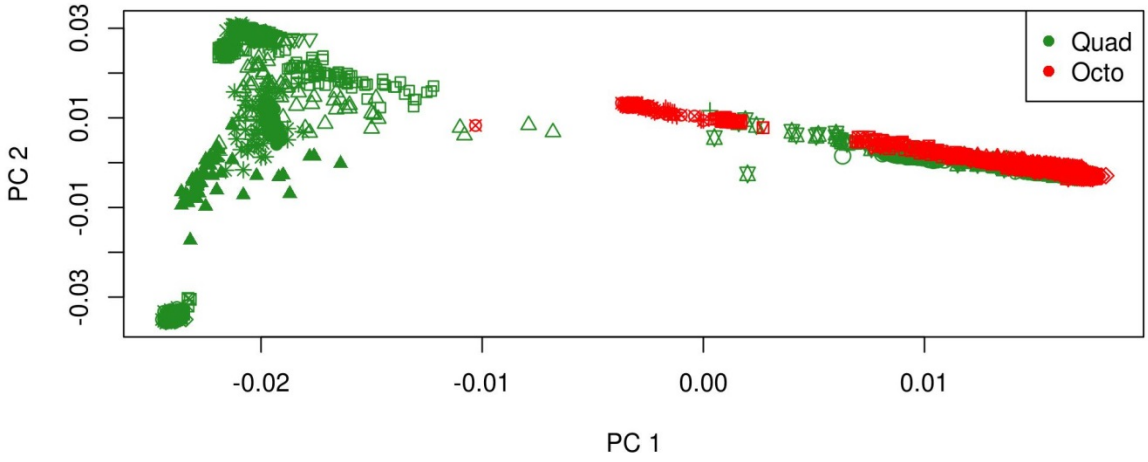
SN1. Figure 1: Chip effects apparent in African samples genotyped on quad and octo chips



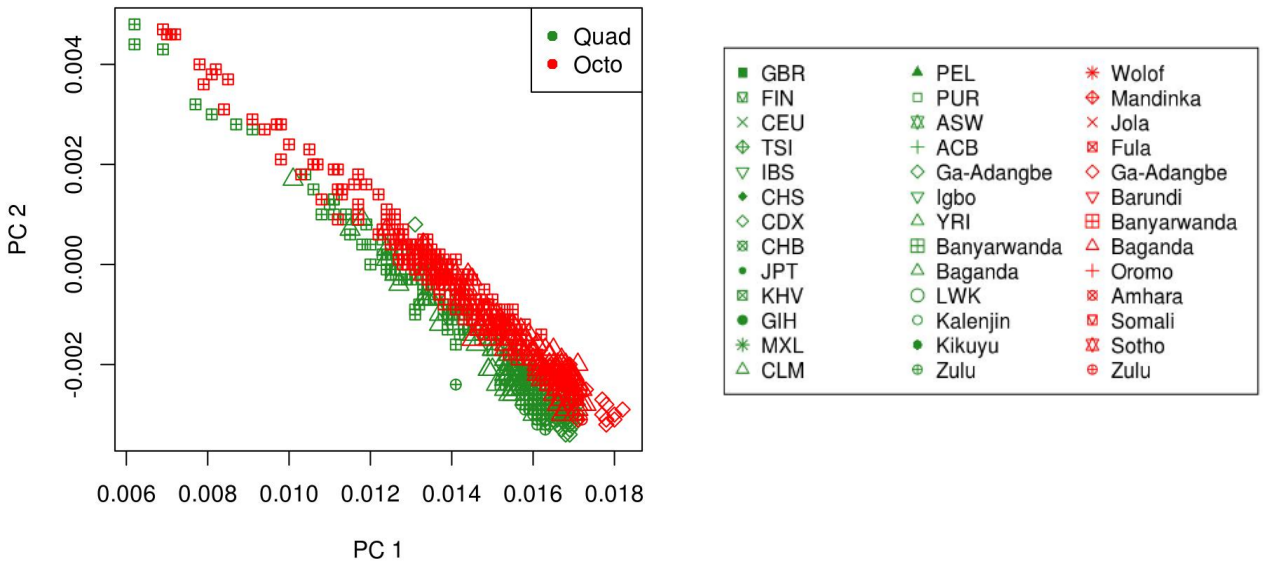
SN1 Figure 1. SN1 Figures 1 a, b, c and d depict the first and second principal components in Baganda, Banyarwanda, Ga-Adangbe and Zulu individuals with samples genotyped on the octo or quad chips. Quad samples and octo samples are represented by green circles and red triangles respectively. SN1 Figure 1e shows principal components calculated from 26 Baganda samples genotyped on both chips, with PCA carried out on quad samples and projected onto genotype data for octo samples. Clear separation is observed indicating chip effects.

SN1. Figure 2: Chip effects apparent on global dataset principal component analysis- a representation of the top 8 PCs

a)



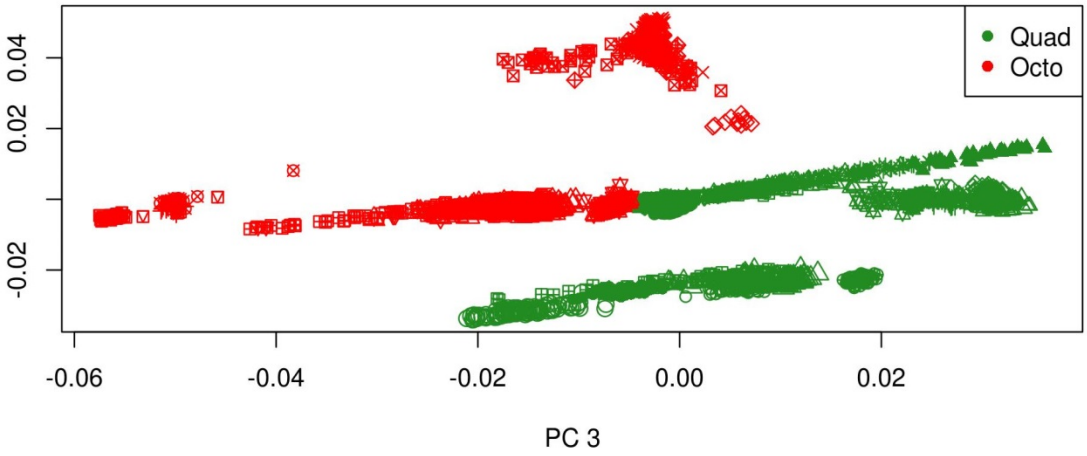
b)



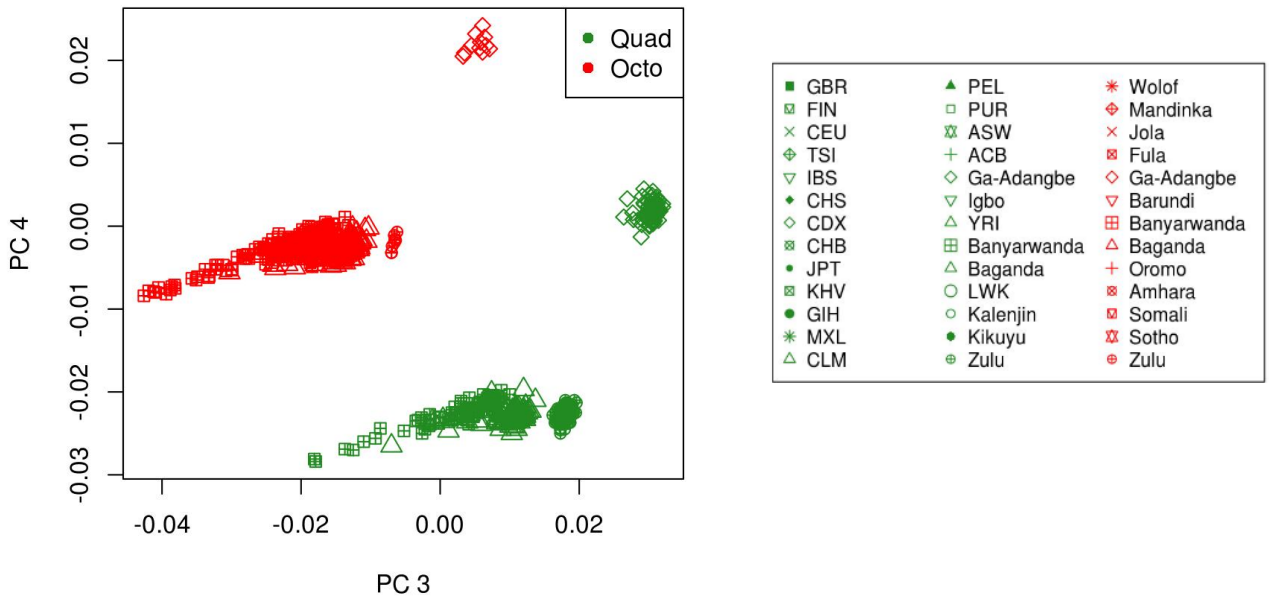
.

SN1 Figure 2a shows global samples represented along PCs 1 and 2. Samples genotyped on the quad chip are shown in green, while those genotyped on the octo platform are shown in red. SN1 Figure 2b depicts PC 1 and 2 for samples only from the four populations that were genotyped across both chips. While slight separation is seen along PC2, this does not seem to primarily represent chip effects.

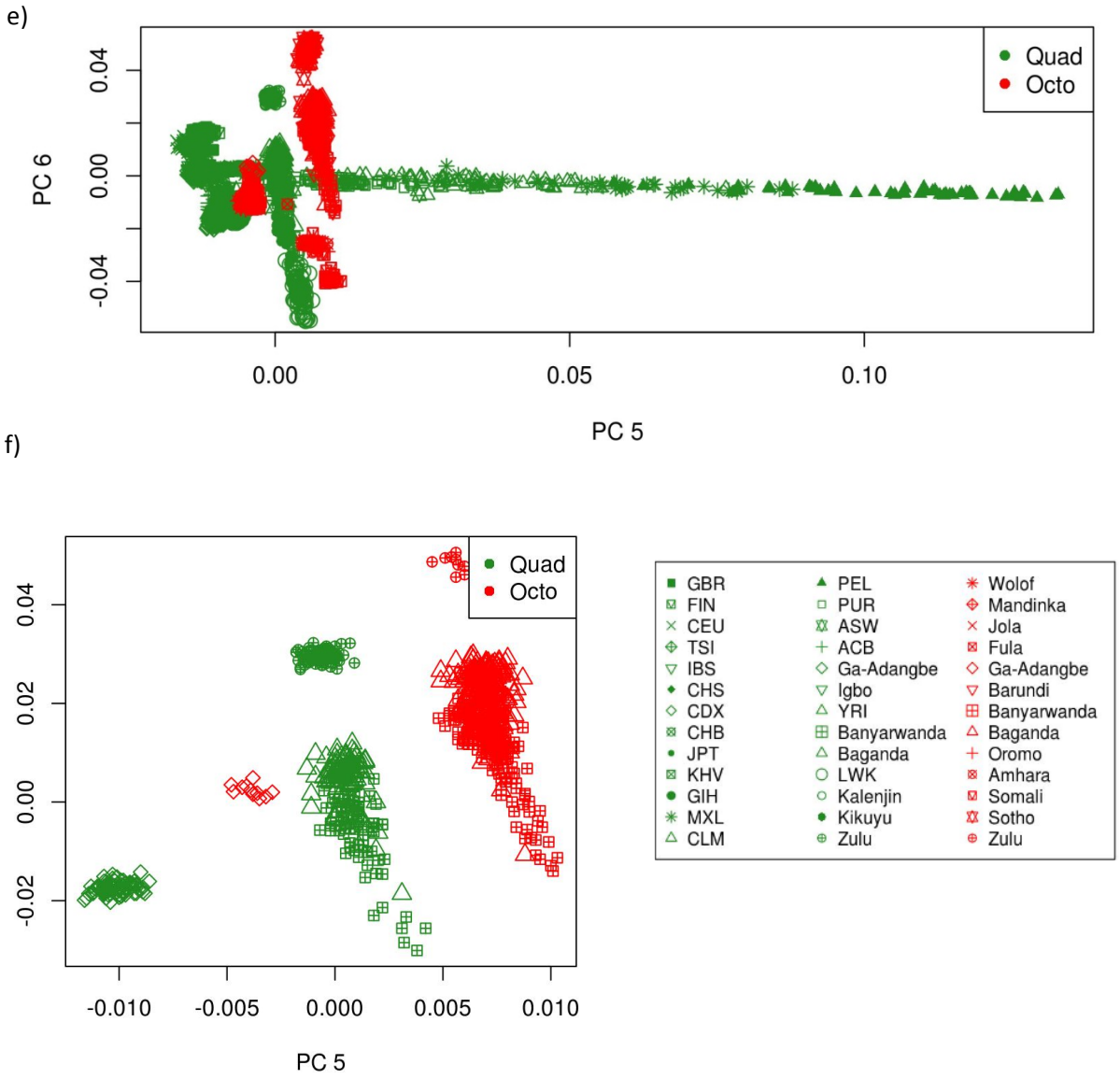
c)



d)

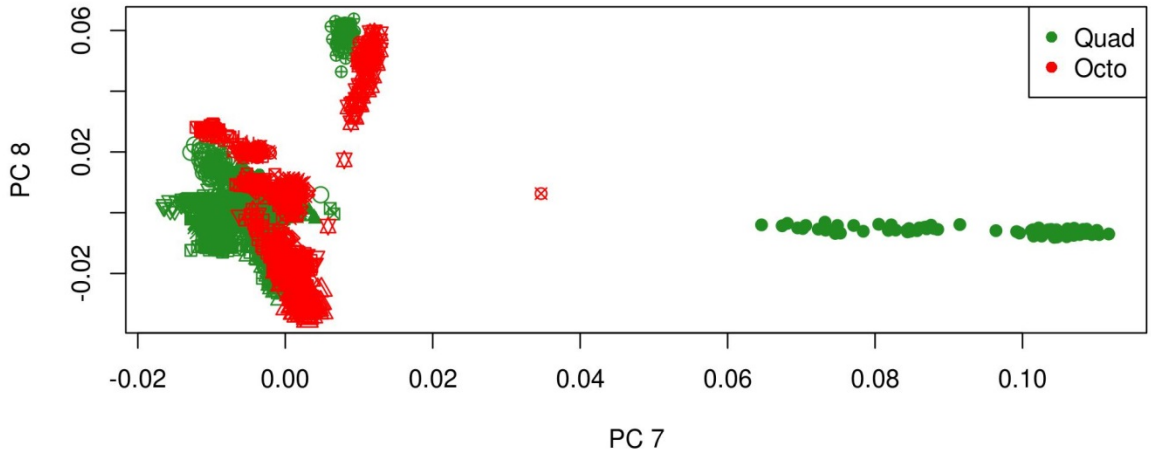


SN1 Figure 2c shows global samples represented along PCs 3 and 4. Samples genotyped on the quad chip are shown in green, while those genotyped on the octo platform are shown in red. SN1 Figure 3d depicts PCs 3 and 4 for samples only from the four populations that were genotyped across both chips. The four populations are clearly seen to be separated along PC3 (horizontally) and PC4 (vertically) by the chip they were genotyped on.

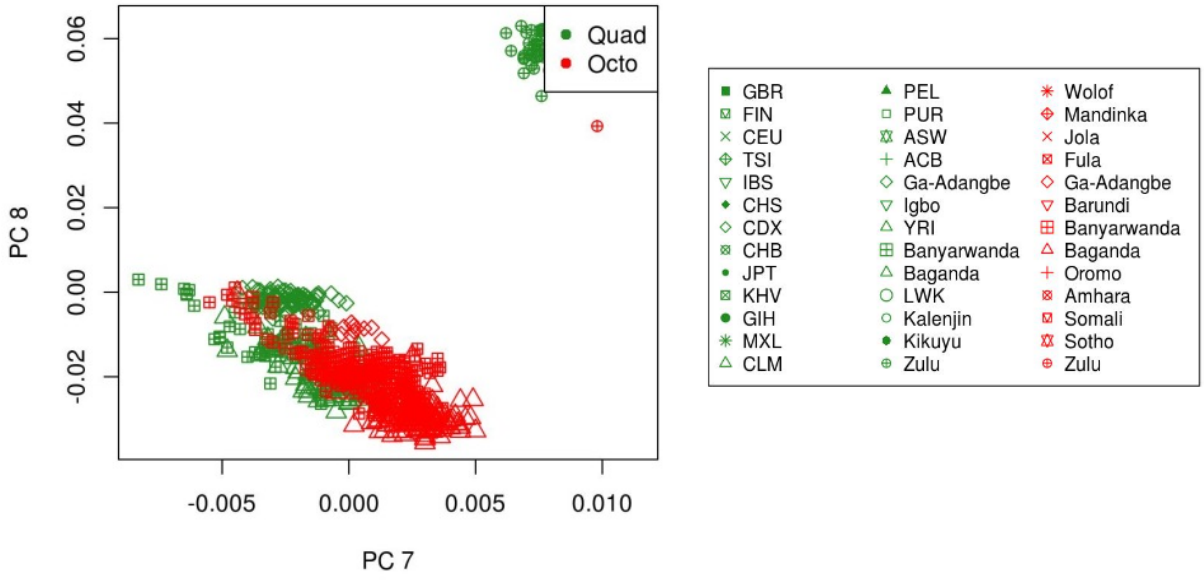


SN1 Figure 2e shows global samples represented along PCs 5 and 6. Samples genotyped on the quad chip are shown in green, while those genotyped on the octo platform are shown in red. SN1 Figure 2f depicts PCs 5 and 6 for samples only from the four populations that were genotyped across both chips. Although, the four populations are seen to be separated along PC5 (horizontally) and PC6 (vertically) by chip, these PCs do not primarily seem to represent chip effects. PC5 represents a north-south American cline, while PC6 seems to represent a cline between East and South Africa.

g)



h)



SN1 Figure 2g shows global samples represented along PCs 7 and 8. Samples genotyped on the quad chip are shown in green, while those genotyped on the octo platform are shown in red. Figure SN1 Figure 2h depicts PCs 7 and 8 for samples only from the four populations that were genotyped across both chips. Although, the four populations are seen to be separated along PC7 (horizontally) and PC8 (vertically) by chip, these PCs do not primarily seem to represent chip effects. PC7 separates GIH from other population groups, while PC8 seems to primarily separate South African populations from others.

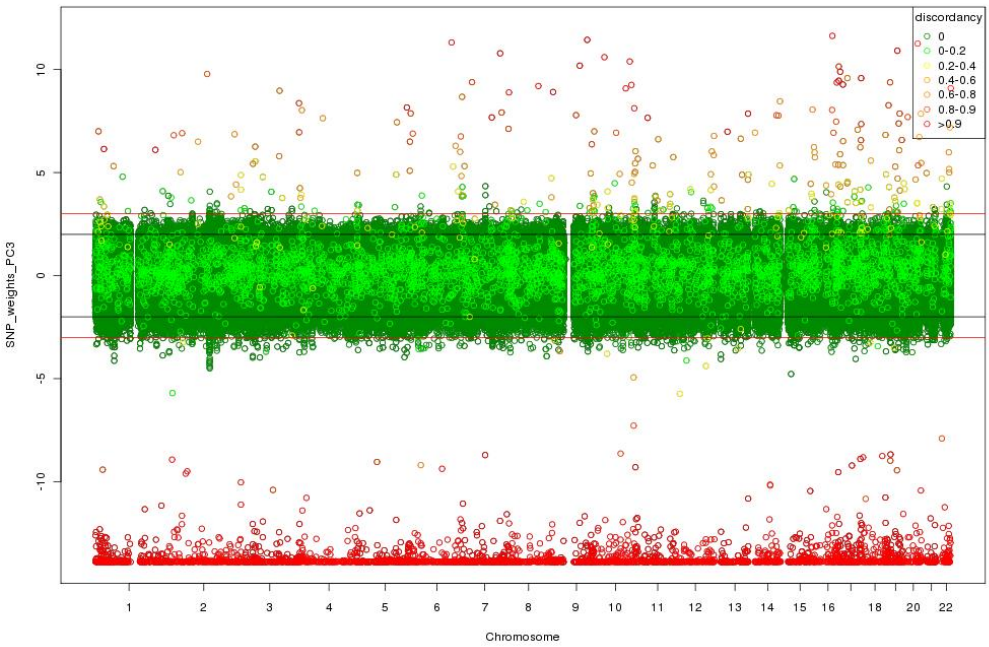
we carried out PCA on quad samples, projecting onto the octo chip, and confirmed that samples were identically projected along PCs. We also carried out clustering analysis in ADMIXTURE to identify chip related clustering of genetic data. Removal of SNPs representing chip effects among principal components also eliminated the clusters relating to chip effects in ADMIXTURE, but retained all other ancestral clusters observed in the algorithm. In order to confirm that removal of SNPs causing chip effects did not remove important ancestral effects of significance, we also compared PCs before and after removal of these variants. Removal of variants responsible for chip effects did not alter the broad interpretation of principal components except for components representing chip effects in each dataset. ADMIXTURE analysis also confirmed that all clusters remained broadly the same, except for the cluster representing chip effects when the mentioned variants were removed from each dataset. Using LD-pruned and non LD- pruned sites produced similar results in all comparisons (data not shown). SNP weights were also highly correlated between in all principal component analyses in LD pruned and non LD-pruned data (-1 for PC3 and -0.88 for PC4 in global datasets).

1.4. CURATION OF THE GLOBAL DATASET

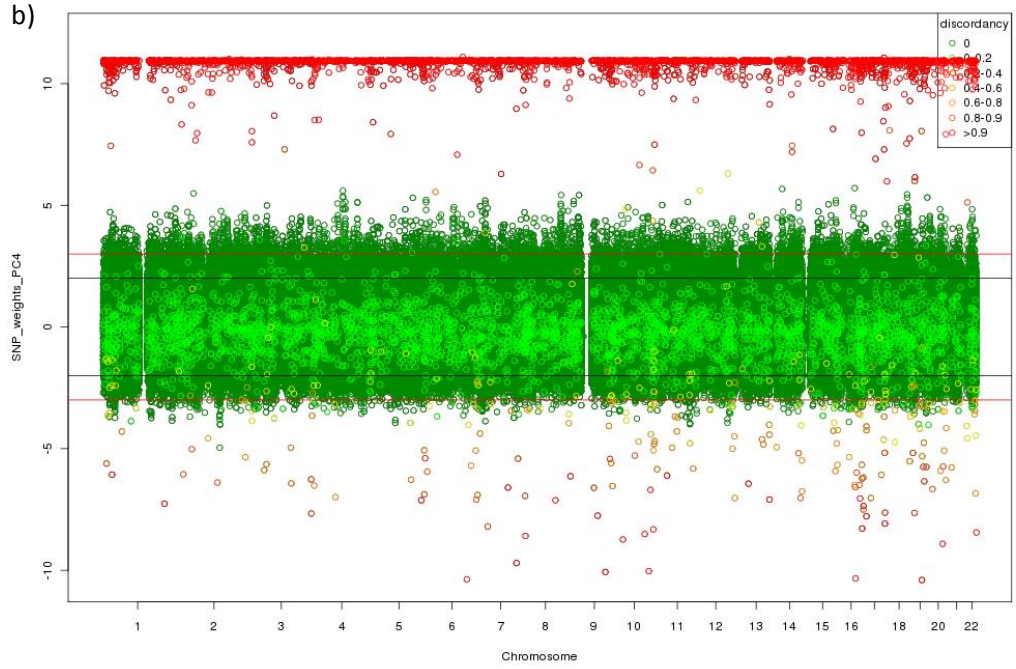
The global dataset was produced by merging genotype data from 33 different populations (**SM. Tables 4 and 5**) after completion of sample and SNP quality based filtering. PCA revealed chip effects between quad and octo genotyped data, evident as separation between chips along PCs 3 and 4 (**SN1 Figure 2**). We calculated SNP loadings along PCs 3 and 4 for the global dataset to identify SNPs causing separation of genetic data along these components. SNPs weighted highly among these components were identified as those > 3 SD from the mean. The correlation between SNP weights and genotype discordancy among duplicate Baganda samples for PCs 3 and 4 was 0.77 and 0.61, respectively, confirming that these did indeed represent chip effects (**SN1 Figure 3**). While chip separation was seen along PCs 5 and 6, the correlation between genotype discordancy among duplicate samples in Baganda, and SNP weights along these components was poor. We only removed highly weighted SNPs along PCs 3 and 4 and tested if this would eliminate separation seen along subsequent PCs. Removing these 7,432 variants eliminated separation between chips along all PCs (**SN1 Figure 4**) and clustering by chip evident on ADMIXTURE analysis (**SN1 Figure 5 and 6**). Furthermore, analysis of each population separately showed that chip effects were not apparent even when PCA was carried out in the four individual populations after filtering for the aforementioned variants (**SN1 Figure 7a-d**). PCA among Baganda quad samples with projection of components onto data from the octo chip for the same samples showed identical positioning of samples on all components examined (**SN1 Figure 7e**). Interpretation of PCA and ADMIXTURE clustering analysis was not

SN1. Figure 3: SNP weights for principal component 3 and 4 for the global dataset annotated with discordant SNPs in Baganda

a)

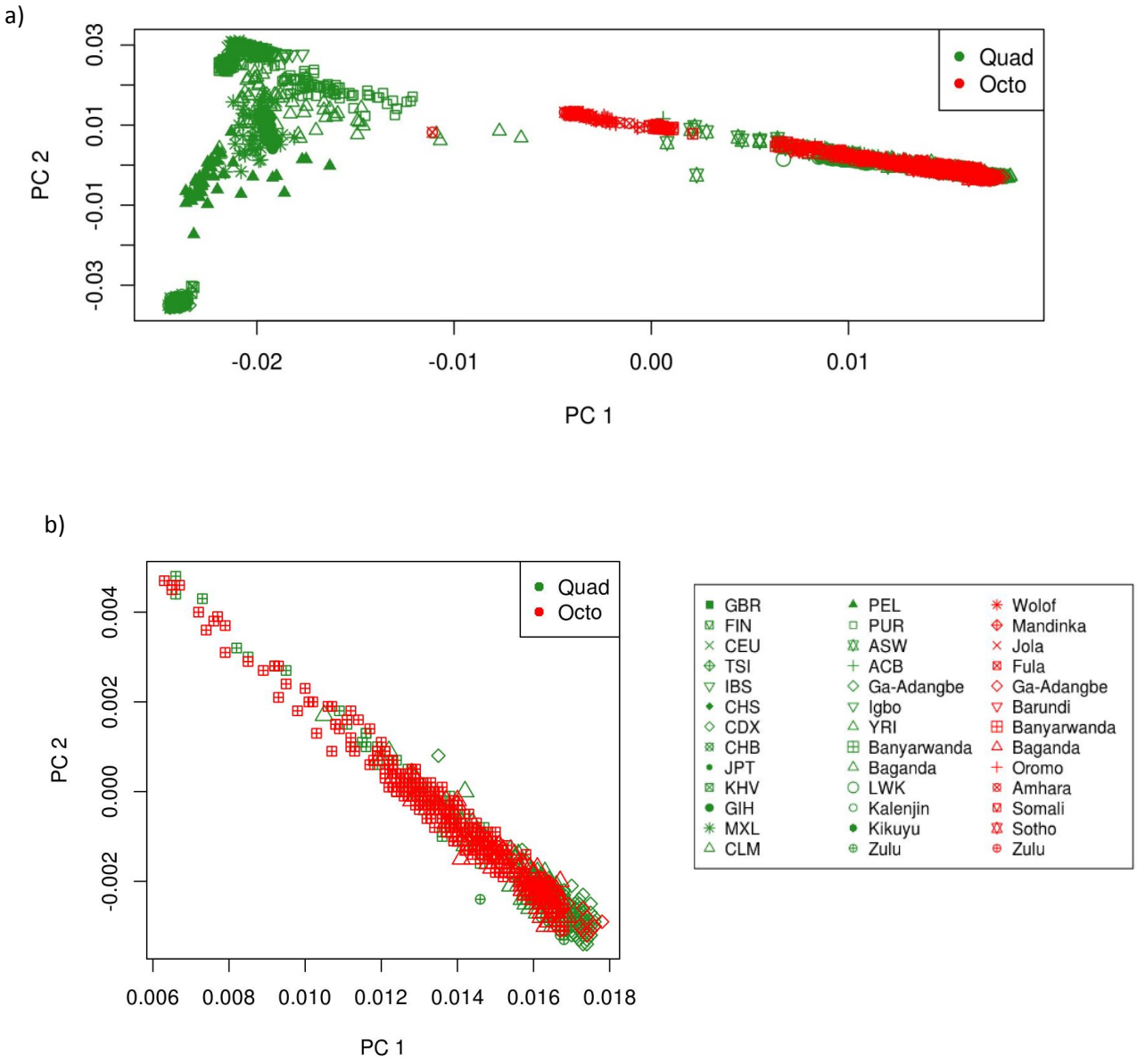


b)



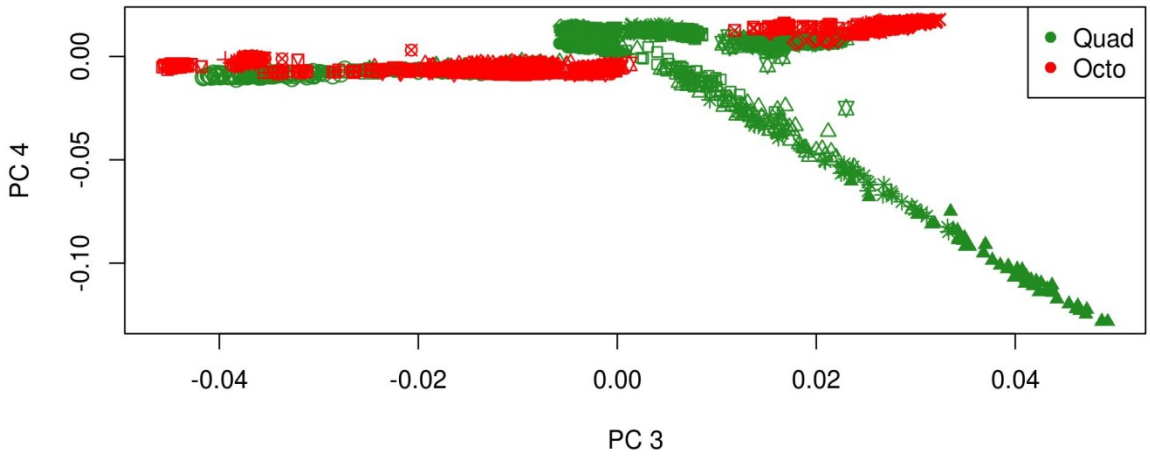
SN1 Figure 3a and 3b represent standardised SNP loadings along PCs 3 and 4 for the global dataset along chromosomes 1-22. The black and red lines represent 2 and 3 SD thresholds from the mean respectively. Sites along chromosomes are coloured by the level of discordancy in genotypes between quad and octo platforms for 26 Baganda sample duplicates genotyped on both chips. There is a strong correlation observed between SNP weights and discordancy in genotypes among the two chips (Pearson's correlation r=0.77 and 0.61 for PCs 3 and 4 respectively).

SN1 Figure 4: PCA plots of global dataset after removal of SNPs with weight >3 SD from mean along PCs 3 and 4

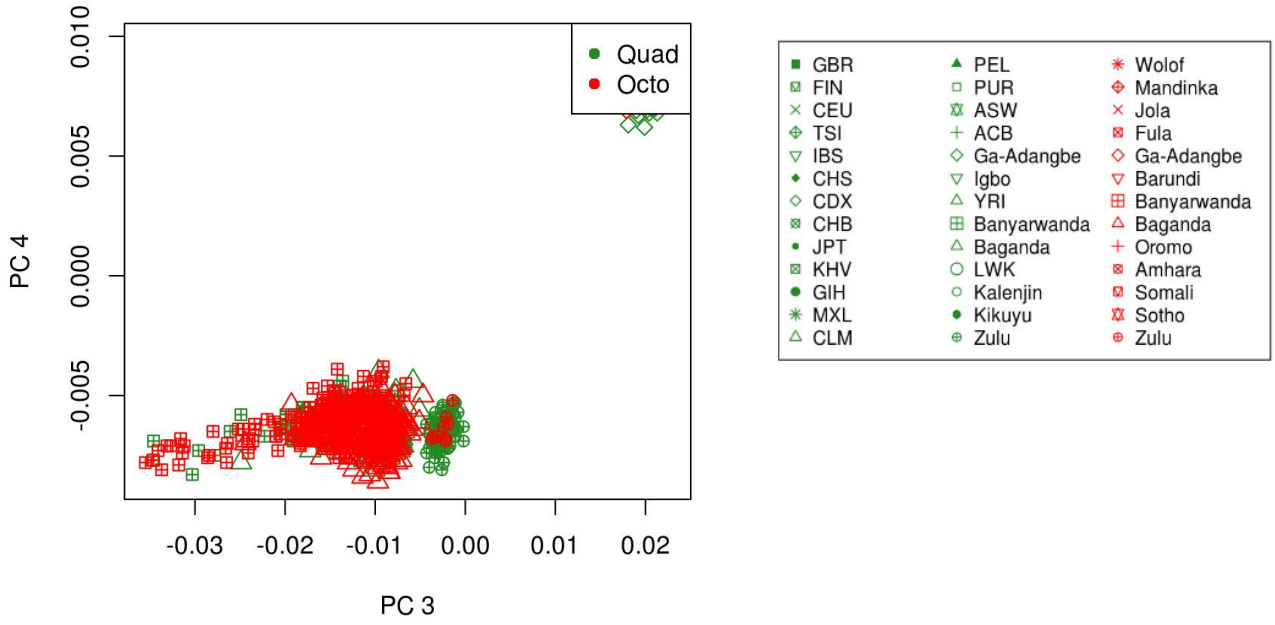


SN1. Figure 4a shows global samples represented along PCs 1 and 2 after removal of SNPs with weights > 3 SD from the mean along PCs 3 and 4. Samples genotyped on the quad chip are shown in green, while those genotyped on the octo platform are shown in red. SN1 Figure 4b depicts PCs 1 and 2 for samples only from the four populations that were genotyped across both chips. No separation is observed by chip.

c)

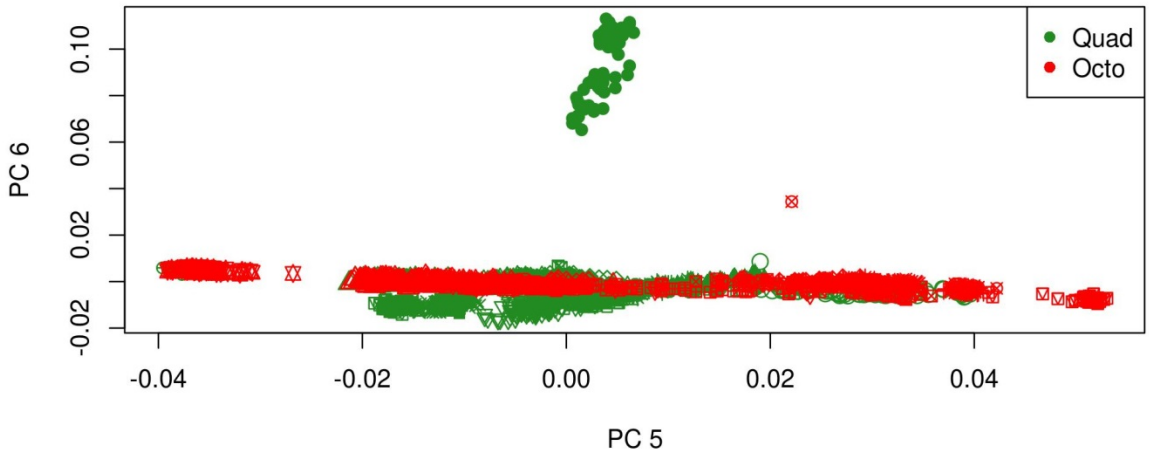


d)

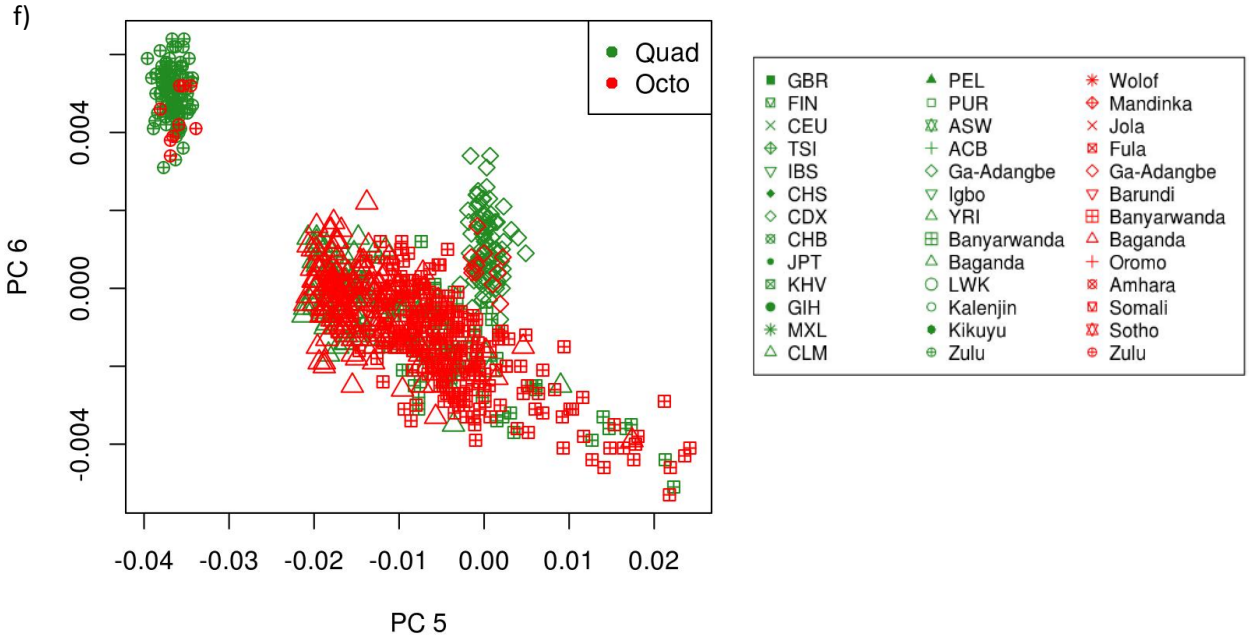


SN Figure 4c shows global samples represented along PCs 3 and 4 after removal of SNPs with weights > 3 SD from the mean along PCs 3 and 4. Samples genotyped on the quad chip are shown in green, while those genotyped on the octo platform are shown in red. SN Figure 4d depicts PCs 3 and 4 for samples only from the four populations that were genotyped across both chips. No separation is observed by chip.

e)

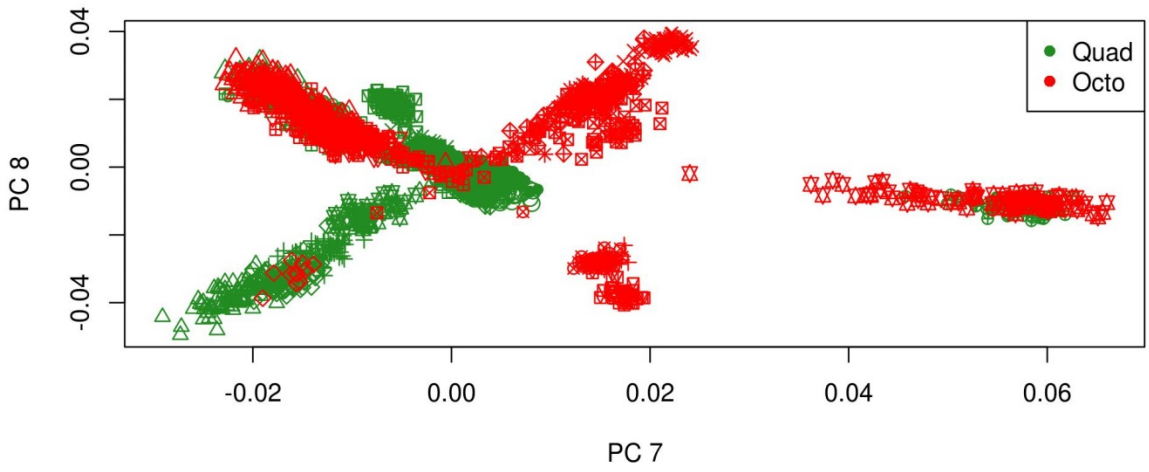


f)

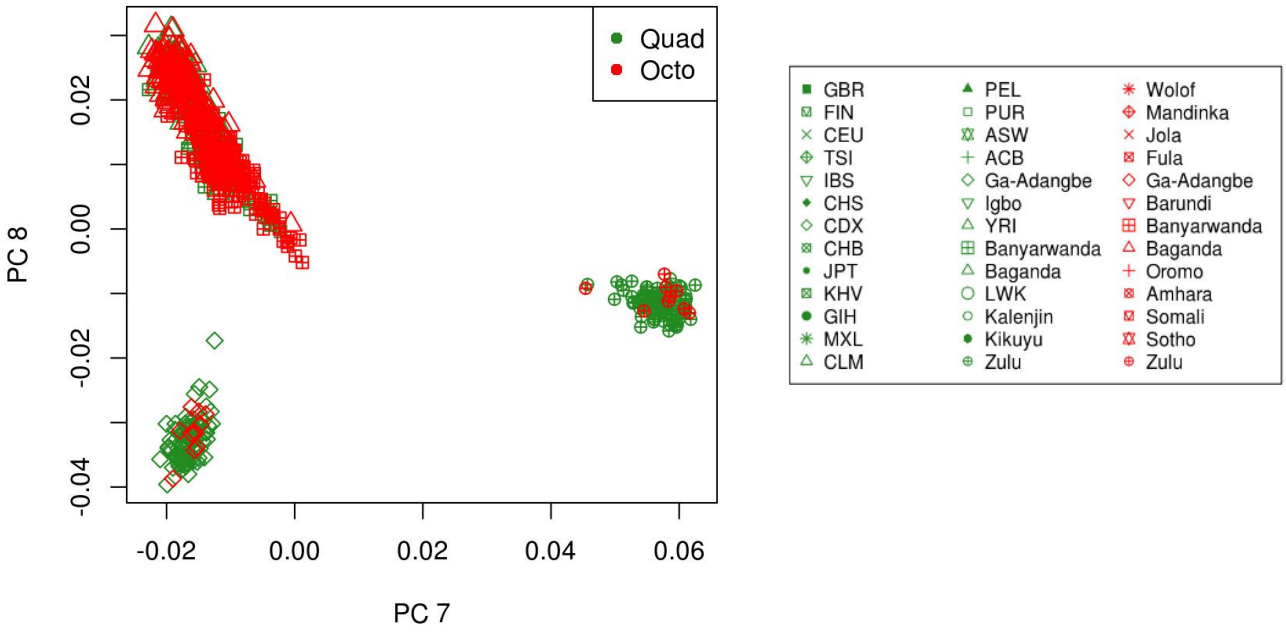


SN Figure 4e shows global samples represented along PC5s and 6 after removal of SNPs with weights > 3 SD from the mean along PCs 3 and 4. Samples genotyped on the quad chip are shown in green, while those genotyped on the octo platform are shown in red. SN Figure 4f depicts PCs 5 and 6 for samples only from the four populations that were genotyped across both chips. No separation is observed by chip.

g)

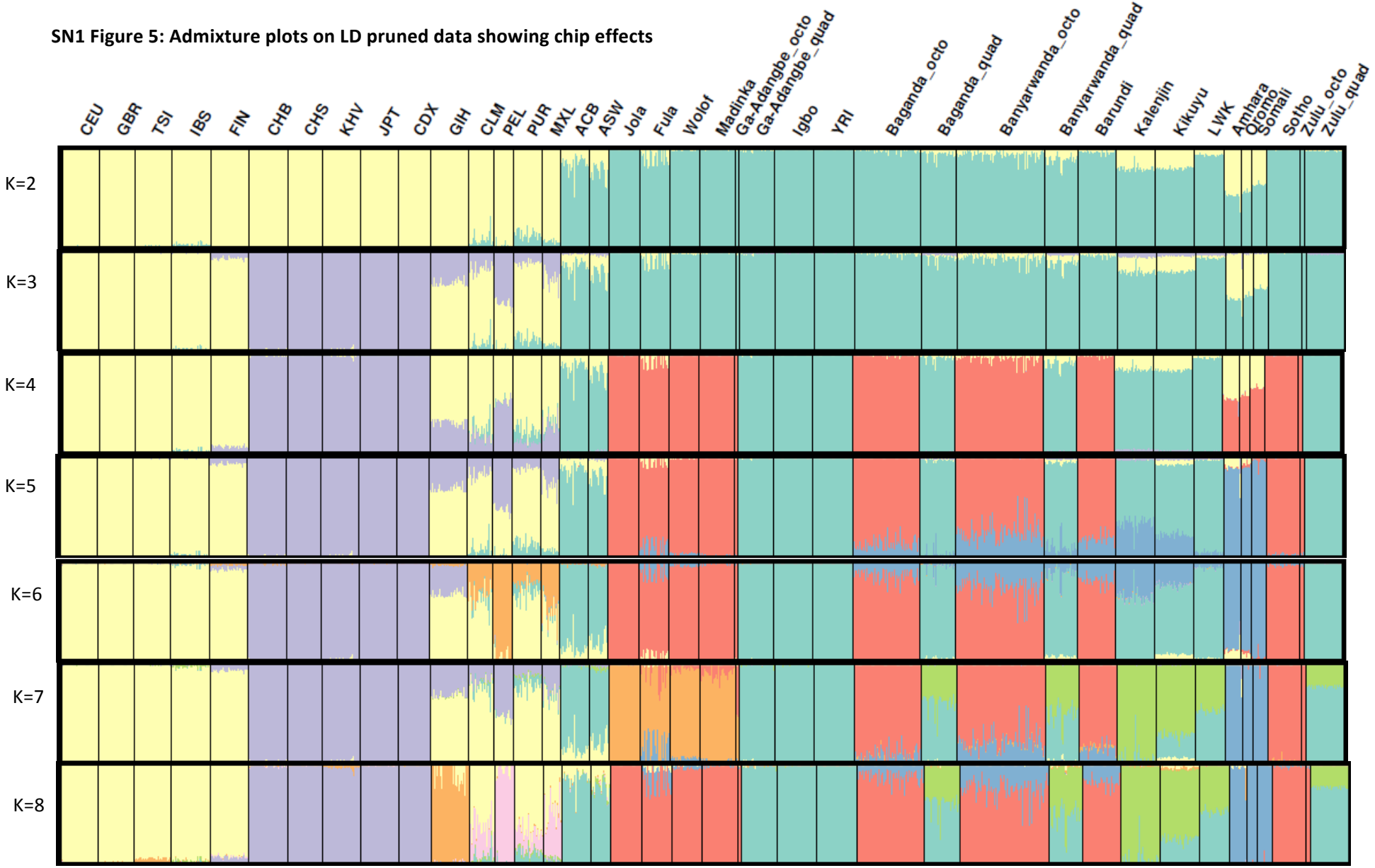


h)



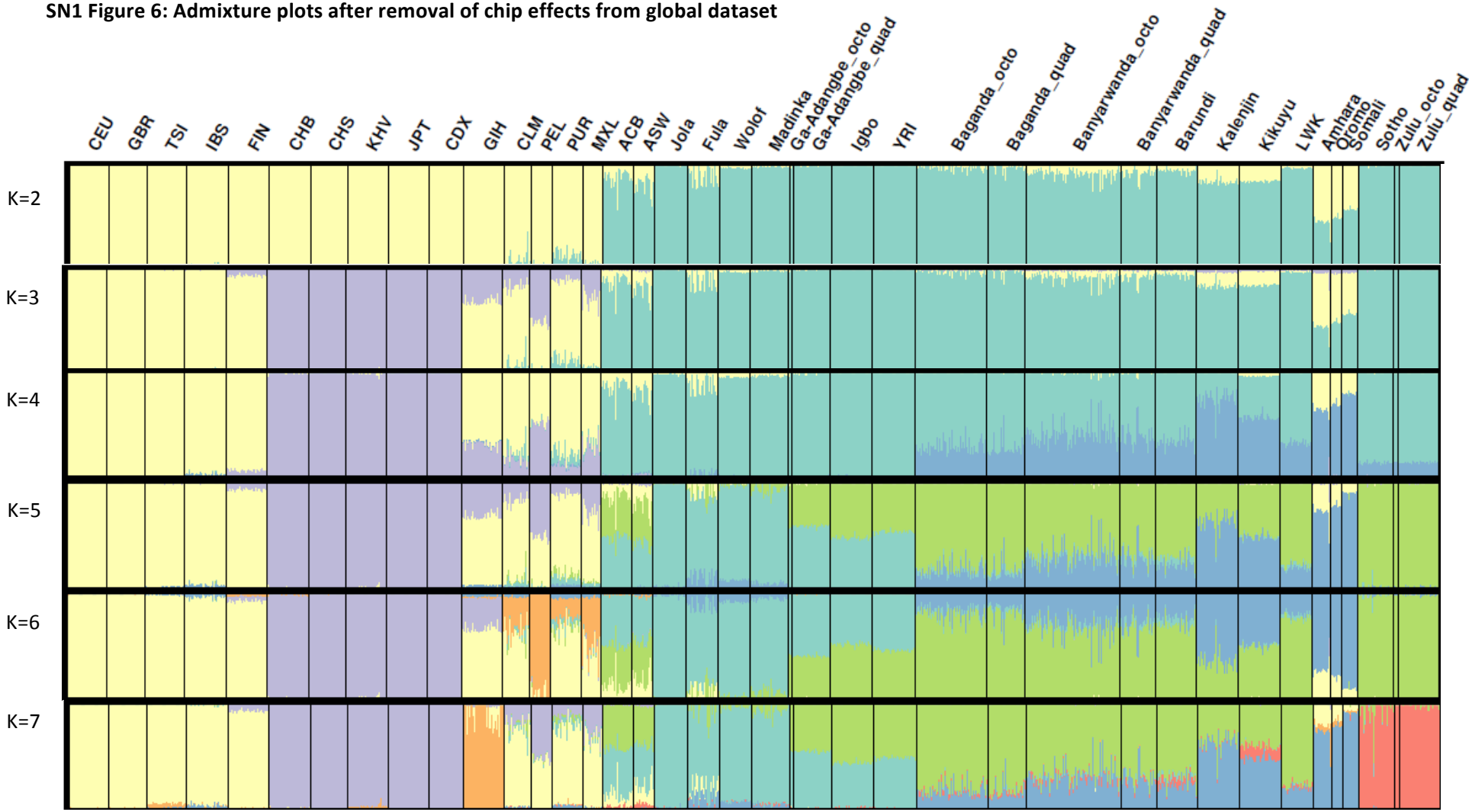
SN1 Figure 4g shows global samples represented along PCs 7 and 8 after removal of SNPs with weights > 3 SD from the mean along PCs 3 and 4. Samples genotyped on the quad chip are shown in green, while those genotyped on the octo platform are shown in red. SN1 Figure 4h depicts PCs 7 and 8 for samples only from the four populations that were genotyped across both chips. No separation is observed by chip.

SN1 Figure 5: Admixture plots on LD pruned data showing chip effects



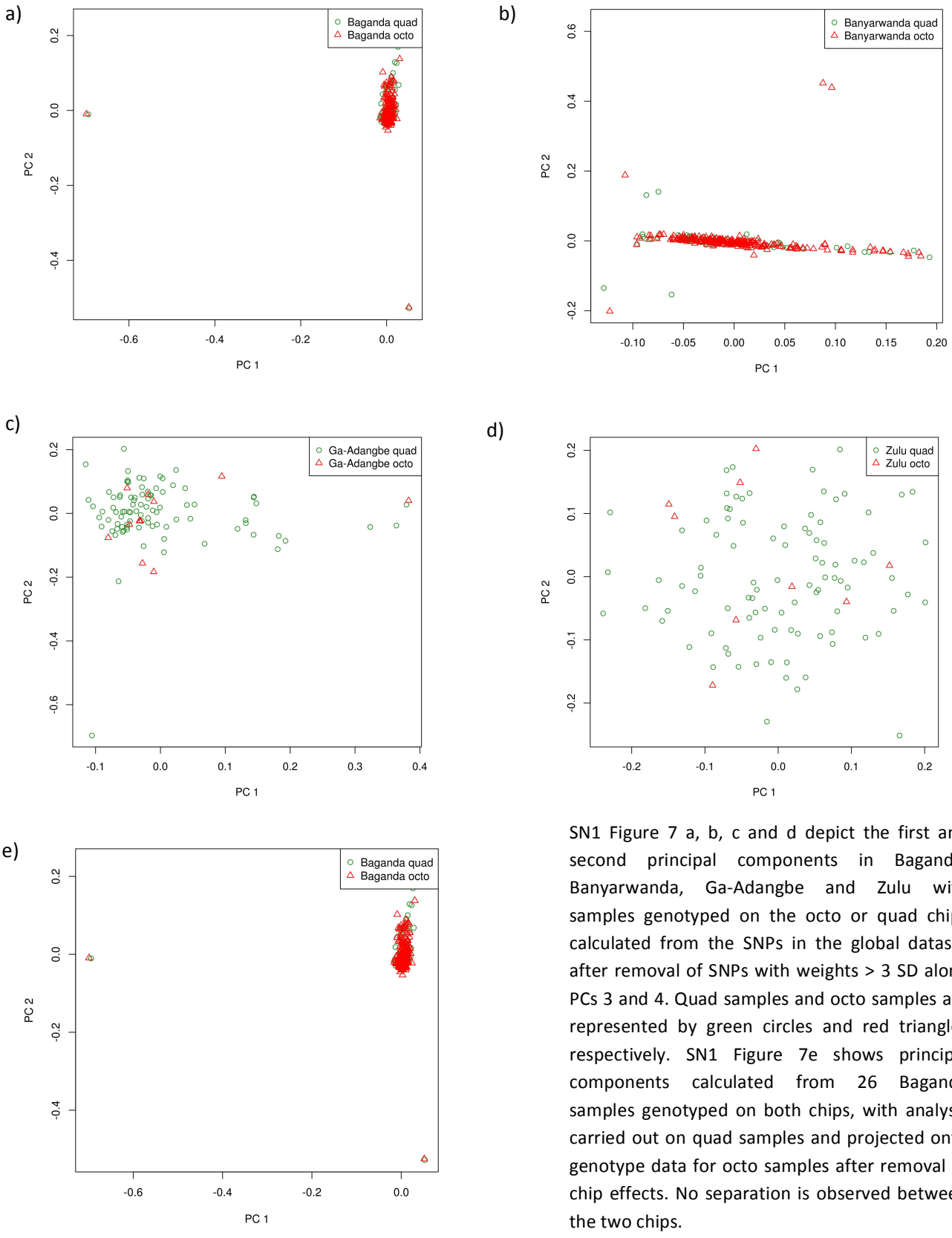
SN1 Figure 5 represents clear octo and quad chip separation on ADMIXTURE analysis at cluster K=4, suggesting the presence of chip effects.

SN1 Figure 6: Admixture plots after removal of chip effects from global dataset



SN1 Figure 6 shows ADMIXTURE clusters (K=2-7) following removal of chip effects from the global dataset. Cluster separation based on chip differences is not observed. However, other ancestral effects seem unchanged.

SN1 Figure 7: Principal component analysis within African populations after removal of SNPs with weights greater than 3 SD from mean along PCs 3 and 4 of global dataset



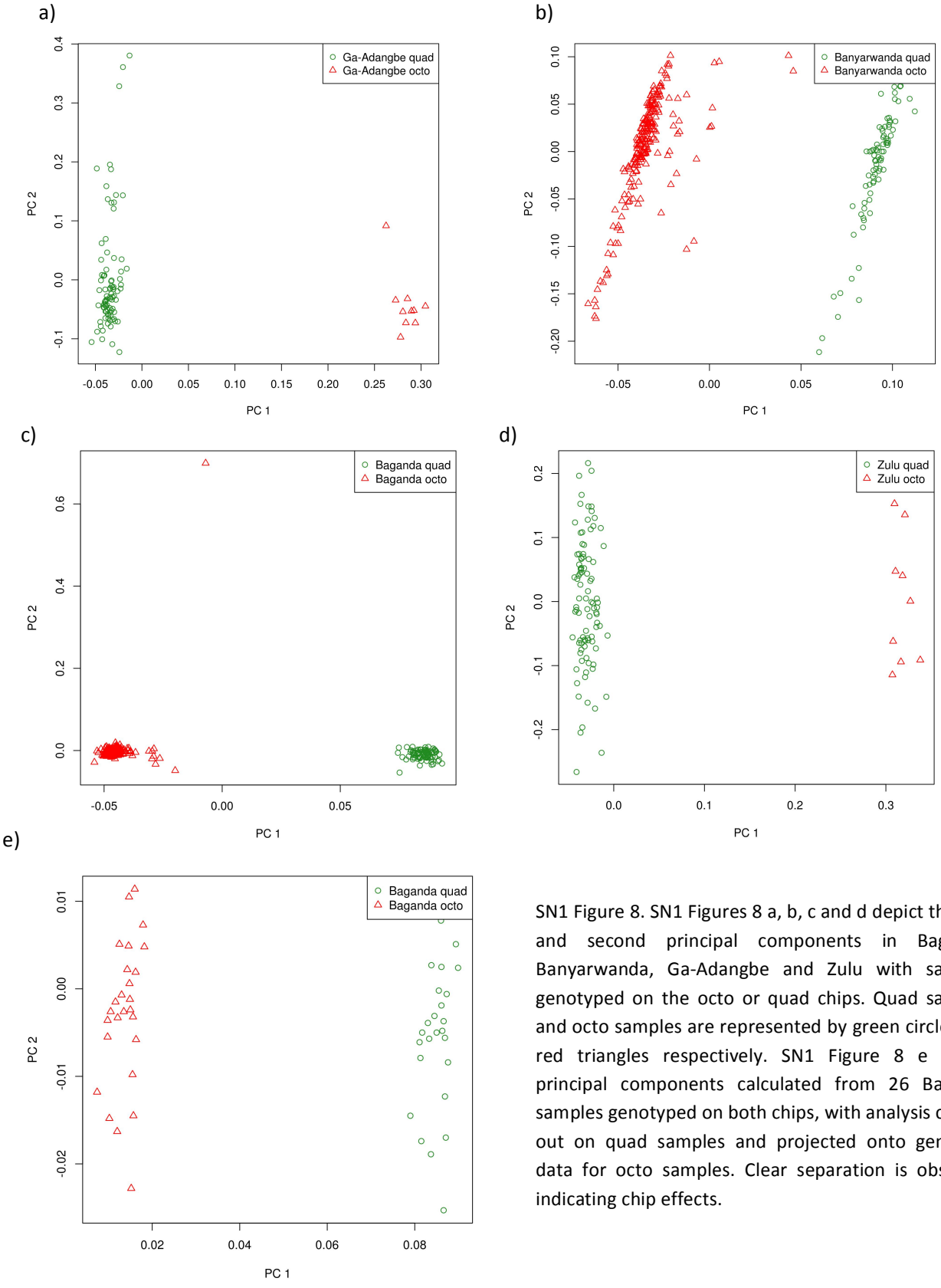
SN1 Figure 7 a, b, c and d depict the first and second principal components in Baganda, Banyarwanda, Ga-Adangbe and Zulu with samples genotyped on the octo or quad chips calculated from the SNPs in the global dataset after removal of SNPs with weights > 3 SD along PCs 3 and 4. Quad samples and octo samples are represented by green circles and red triangles respectively. SN1 Figure 7e shows principal components calculated from 26 Baganda samples genotyped on both chips, with analysis carried out on quad samples and projected onto genotype data for octo samples after removal of chip effects. No separation is observed between the two chips.

altered by removal of these variants, suggesting that these were specific to chip effects and did not represent ancestral differences among populations. Removal of SNPs greater than 2 SD from the mean of PCs 3 and 4 produced identical results, although a larger number of variants were excluded. We, therefore, decided to use a threshold of 3 SD for exclusion, in order to eliminate chip effects whilst retaining the largest number of variants.

1.5 Curation of the AGVP genotype African dataset

A similar approach was used for curation of the African populations dataset. Post-QC filtering data was combined from 16 populations, and a combined dataset including only variants that passed quality control in all 16 populations was generated, including 1,577,224 genotyped markers that passed QC in all populations (**SM Tables 4 and 5**). PCA calculation among the four populations genotyped across both chips showed clear separation between quad and octo chips for the SNPs among individual populations along the first PC (**SN1 Figure 8**). PCA on all variants across populations revealed separation by chip used for genotyping along principal components 2 and 3 (**SN1 Figure 9**). SNP loading weights were calculated for all variants along these principal components, and SNPs with weights > 3 SD from the mean were excluded from further analysis. The correlation between SNP loading weights along PCs 2 and 3, and genotype discordancy in duplicated samples among Baganda was high (0.79 and 0.72 respectively), again confirming that these components represented chip effects (**SN1 Figure 10**). Removing these variants eliminated chip separation seen on PCA of the African dataset (**SN1 Figure 11**). Furthermore, chip separation was also eliminated on per population PCA carried out only on SNPs in the African dataset (**SN1 Figure 12a-d**). Projection of PCs from Baganda quad data onto 26 duplicate samples genotyped on octo showed that samples clustered identically along all principal components analyses (**SN1 Figure 12e**). ADMIXTURE cluster analysis also showed elimination of the cluster representing different chips on removal of the aforementioned SNPs (**SN1 Figure 13 and 14**). Ancestral effects observed on ADMIXTURE analysis and PCA were not altered by removal of these SNPs, suggesting that these coded primarily for chip effects, and not ancestry. Analysis using a threshold of 2 SD from the mean for SNP weights along prinicipal components produced similar results, except with a larger proportion of variants being excluded. Although using a lower threshold resulted in a larger number of variants being excluded, the overlap of these variants with discordant variants between octo and quad chips among the 26 Baganda samples was not altered substantially, suggesting that a threshold of 3 SD appropriately identified SNPs differentiating chip genotyping.

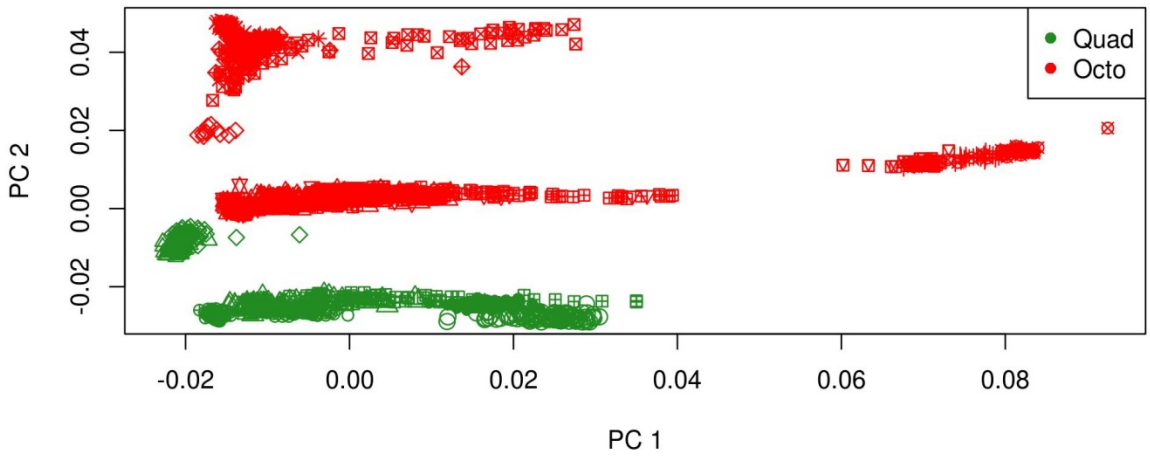
SN1 Figure 8: PC separation in individual populations for SNPs in the African dataset along PC 1



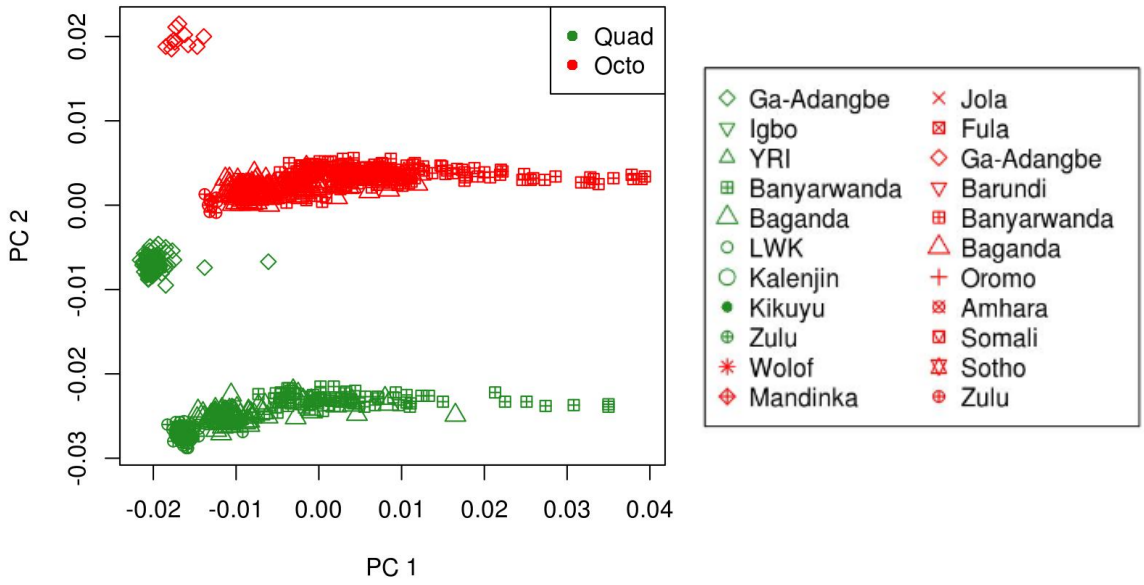
SN1 Figure 8. SN1 Figures 8 a, b, c and d depict the first and second principal components in Baganda, Banyarwanda, Ga-Adangbe and Zulu with samples genotyped on the octo or quad chips. Quad samples and octo samples are represented by green circles and red triangles respectively. SN1 Figure 8 e shows principal components calculated from 26 Baganda samples genotyped on both chips, with analysis carried out on quad samples and projected onto genotype data for octo samples. Clear separation is observed indicating chip effects.

SN1 Figure 9: PCA plots of African dataset before removal of chip effects

a)

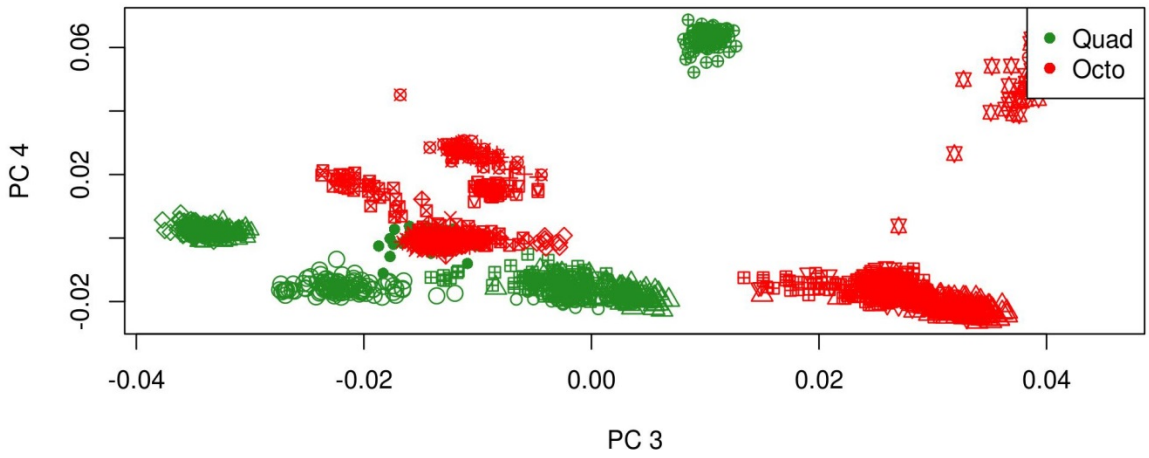


b)

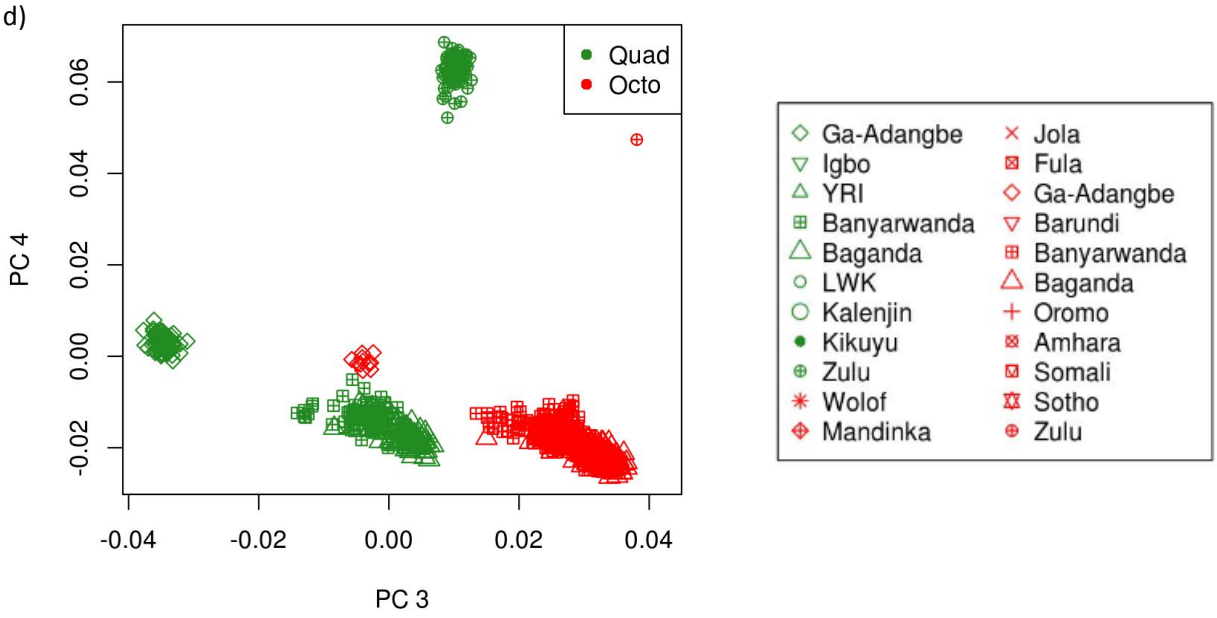


SN1. Figure 9a shows the African dataset samples represented along PCs 1 and 2. Samples genotyped on the quad chip are shown in green, while those genotyped on the octo platform are shown in red. SN1 Figure 9b depicts PCs 1 and 2 for samples only from the four populations that were genotyped across both chips. Clear separation by chip is seen along PC2.

c)

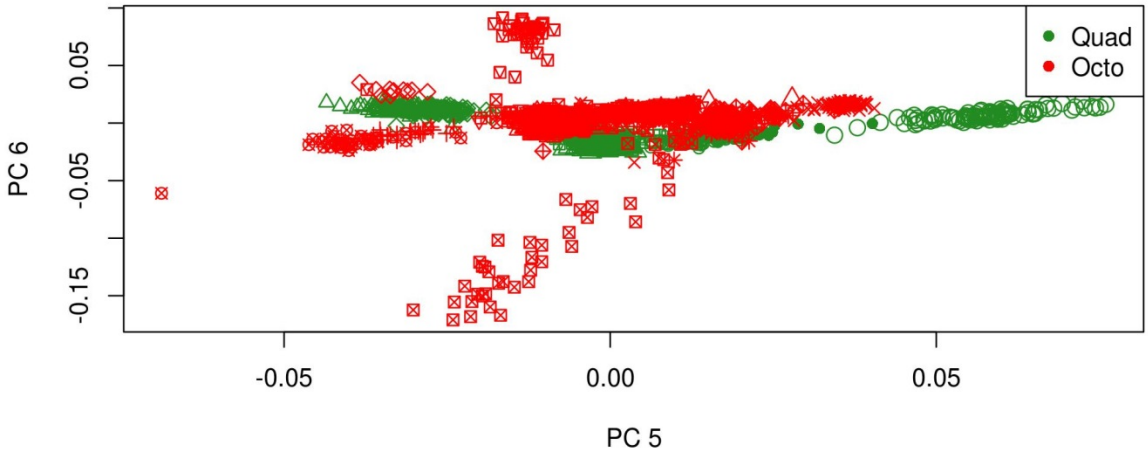


d)

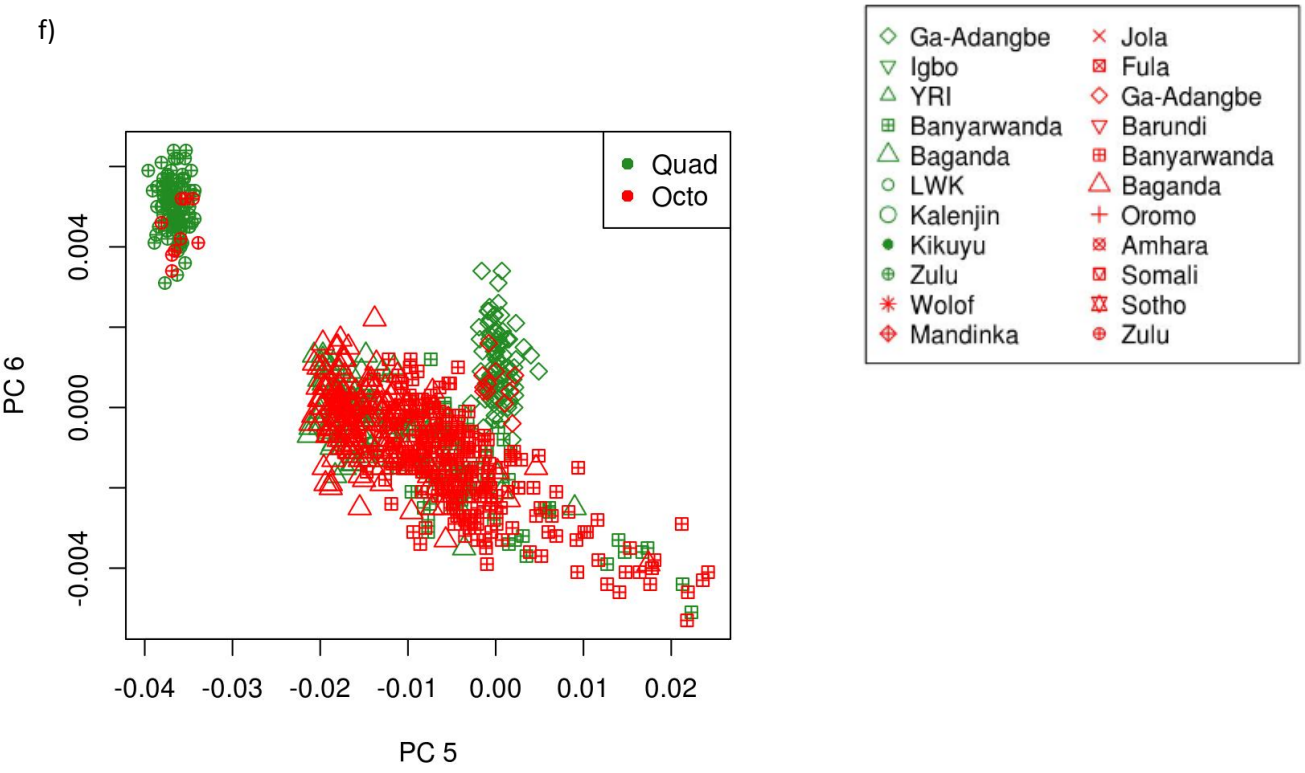


SN1 Figure 9c shows the African dataset represented along PCs 3 and 4. Samples genotyped on the quad chip are shown in green, while those genotyped on the octo platform are shown in red. SN1 Figure 9d depicts PCs 3 and 4 for samples only from the four populations that were genotyped across both chips. Clear separation by chip is seen along PC 3.

e)

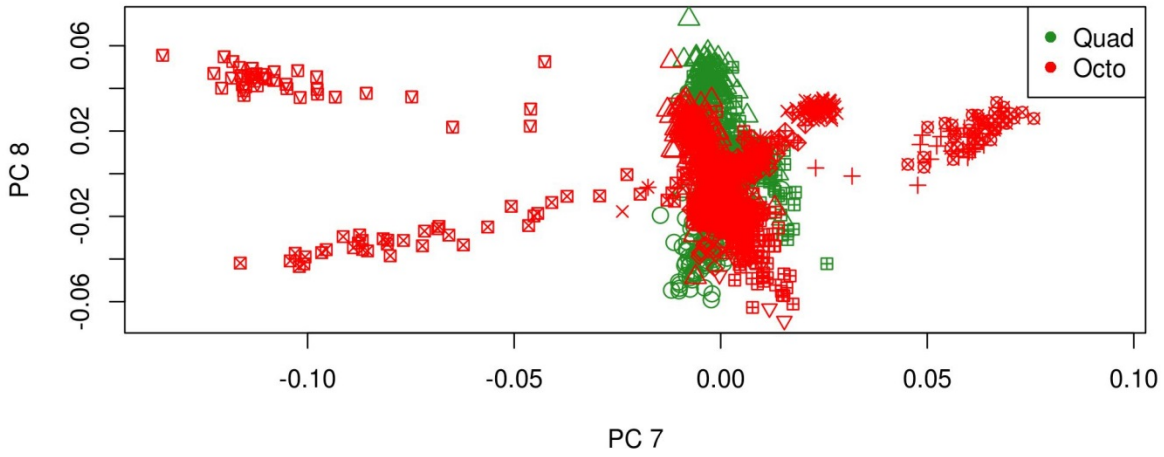


f)

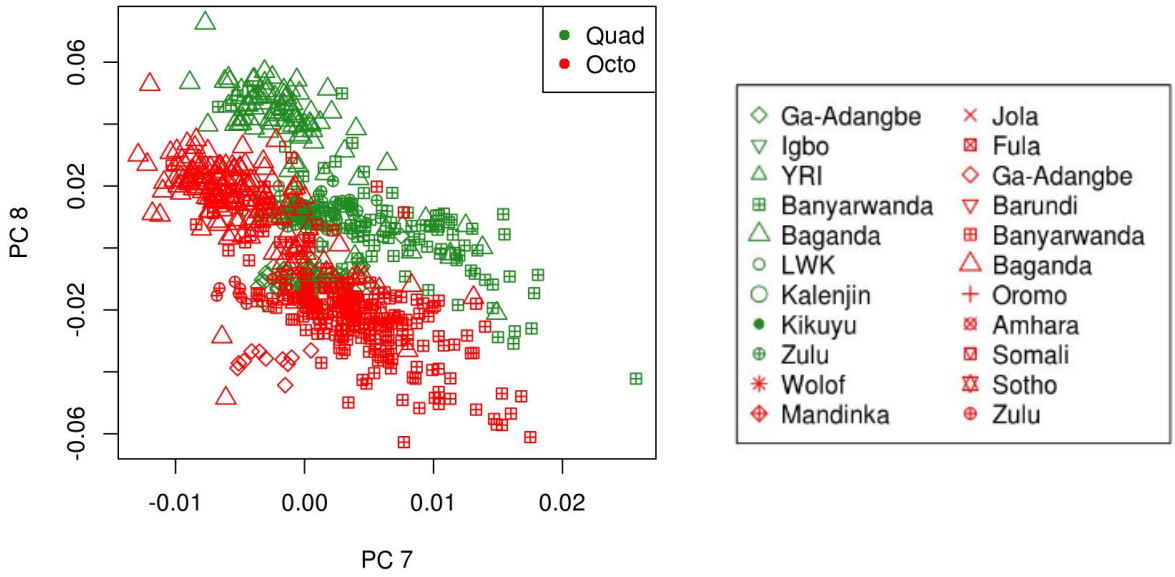


SN1. Figure 9e shows African dataset represented along PCs 5 and 6. Samples genotyped on the quad chip are shown in green, while those genotyped on the octo platform are shown in red. SN1 Figure 9f depicts PCs 5 and 6 for samples only from the four populations that were genotyped across both chips. Although some separation by chip is seen along PCs 5 and 6, these components do not seem to primarily represent chip effects.

g)



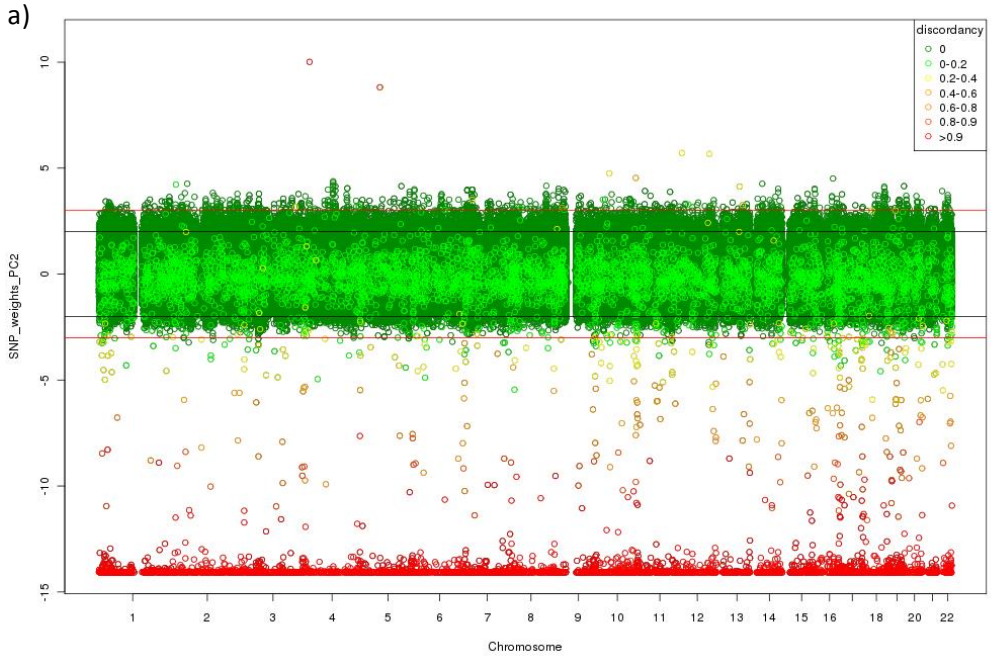
h)



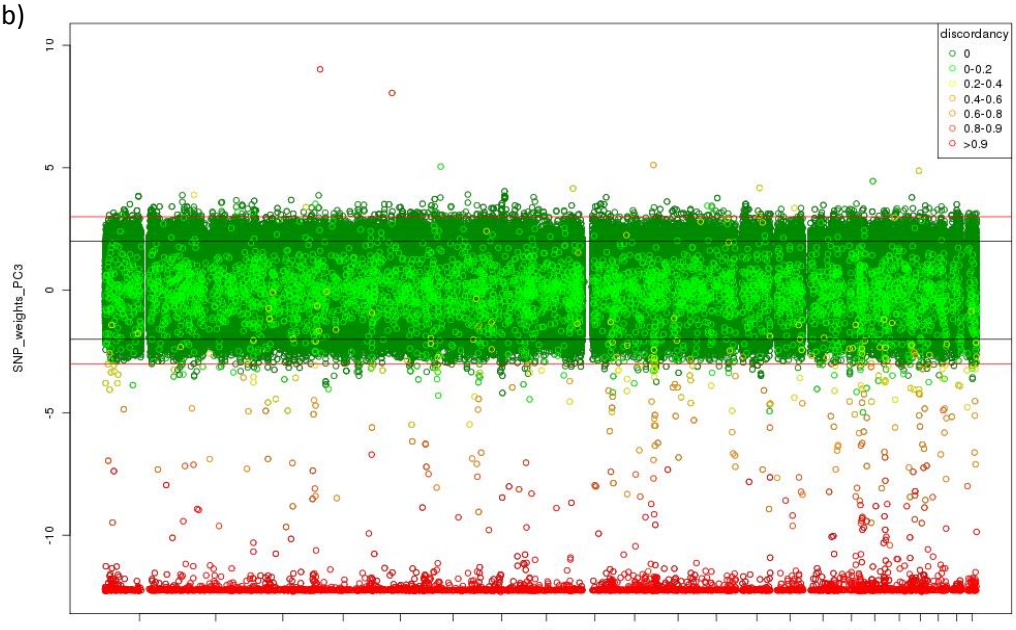
SN1. Figure 9g shows African dataset represented along PCs 7 and 8. Samples genotyped on the quad chip are shown in green, while those genotyped on the octo platform are shown in red. SN1. Figure 9h depicts PCs 7 and 8 for samples only from the four populations that were genotyped across both chips. Although some separation by chip is seen along PCs 7 and 8, these components do not seem to primarily represent chip effects.

SN1. Figure 10: SNP weights along PCs 2 and 3 for the African dataset annotated for discordancy between Baganda octo and quad genotypes for duplicate samples

a)



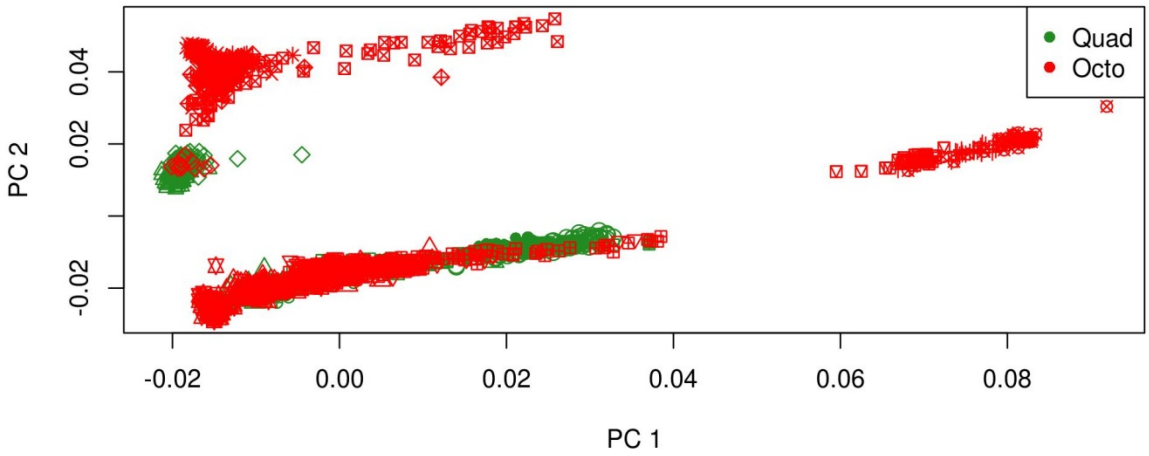
b)



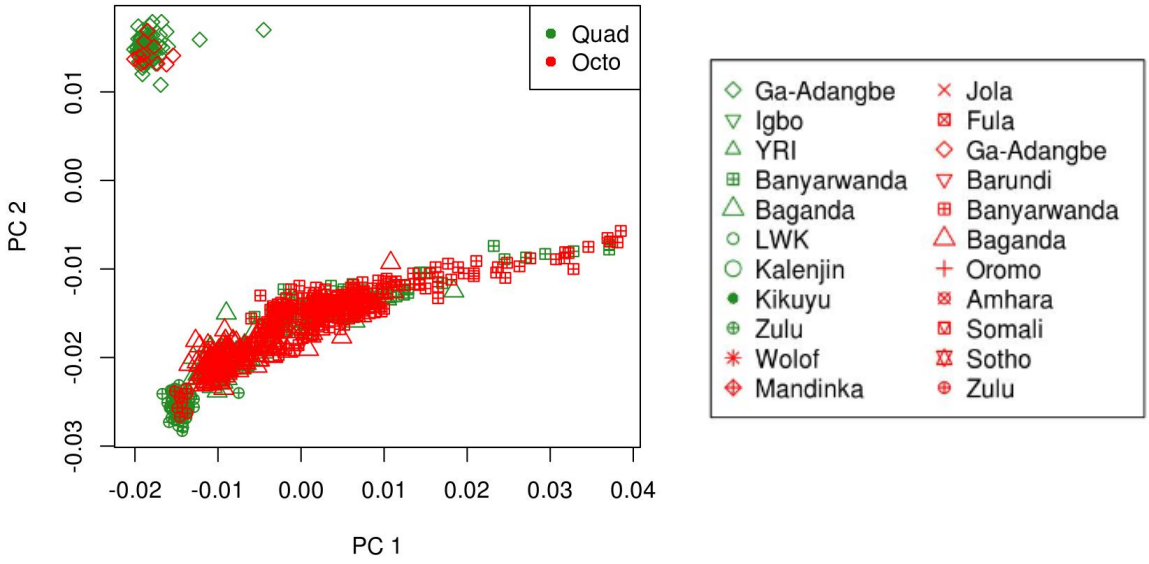
SN1 Figure 10a and b represent standardised SNP loadings along PCs 2 and 3 for the African dataset along chromosomes 1-22. The black and red lines represent 2 and 3 SD thresholds from the mean respectively. Sites along chromosomes are coloured by the level of discordancy in genotypes between quad and octo platforms for 26 samples duplicates genotyped on both chips. There is a strong correlation observed between SNP weights along PCs 2 and 3 and discordancy in genotypes among the two chips (Pearson's correlation r=0.79 and 0.72 for PCs 2 and 3, respectively).

SN1. Figure 11: PCA of African curated dataset after removal of chip effects

a)

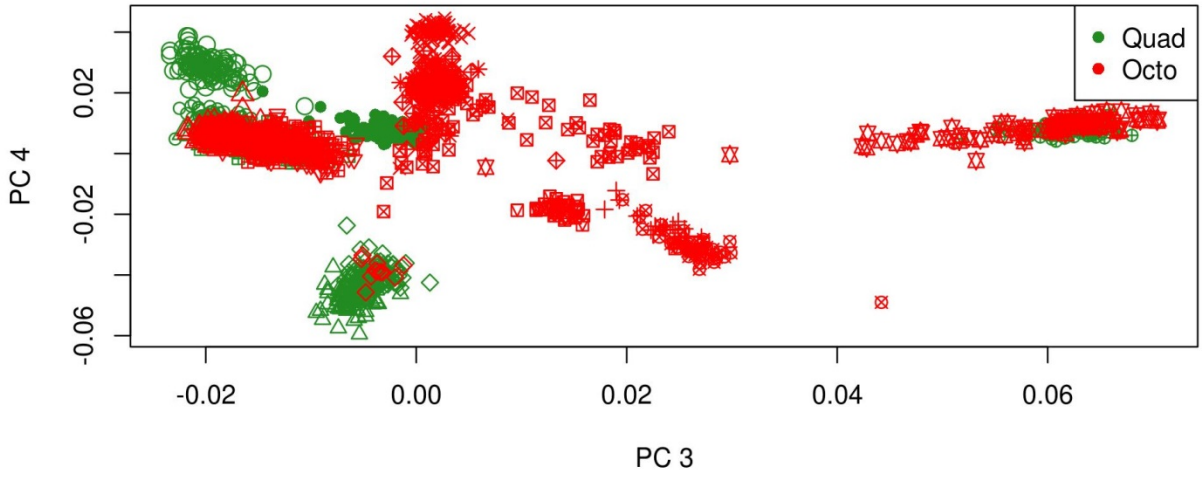


b)

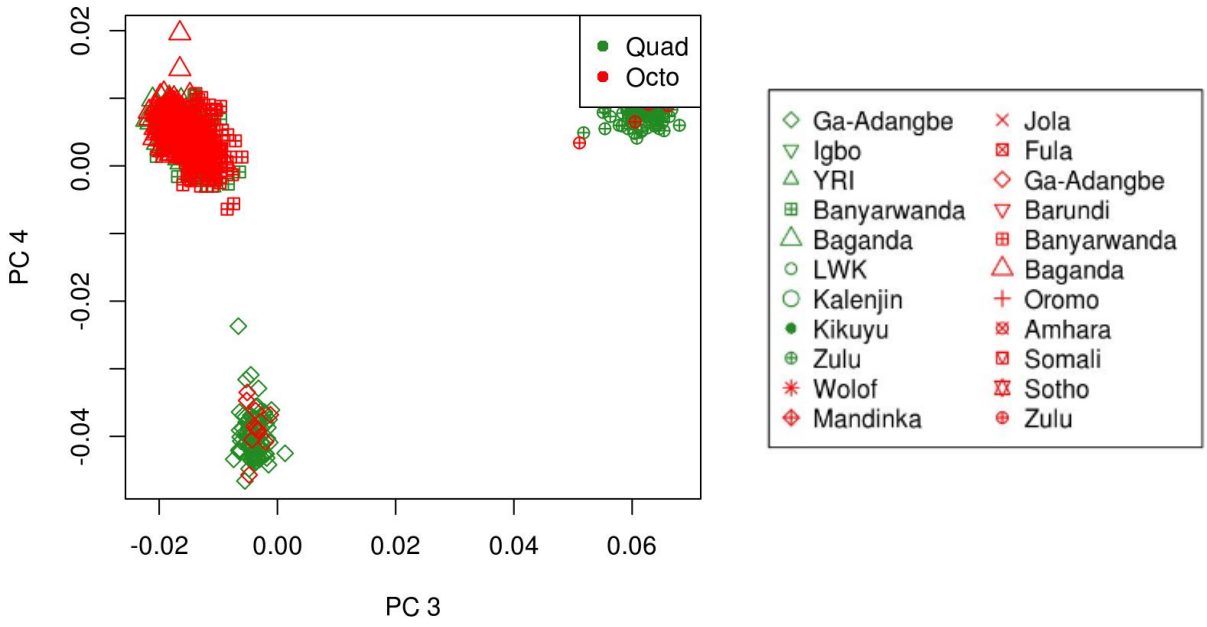


SN1. Figure 11a shows the African dataset samples represented along PCs 1 and 2 after removal of SNPs with weights > 3 SD from the mean along PCs 2 and 3. Samples genotyped on the quad chip are shown in green, while those genotyped on the octo platform are shown in red. SN1. Figure 11b depicts PCs 1 and 2 for samples only from the four populations that were genotyped across both chips. No separation is observed by chip.

c)

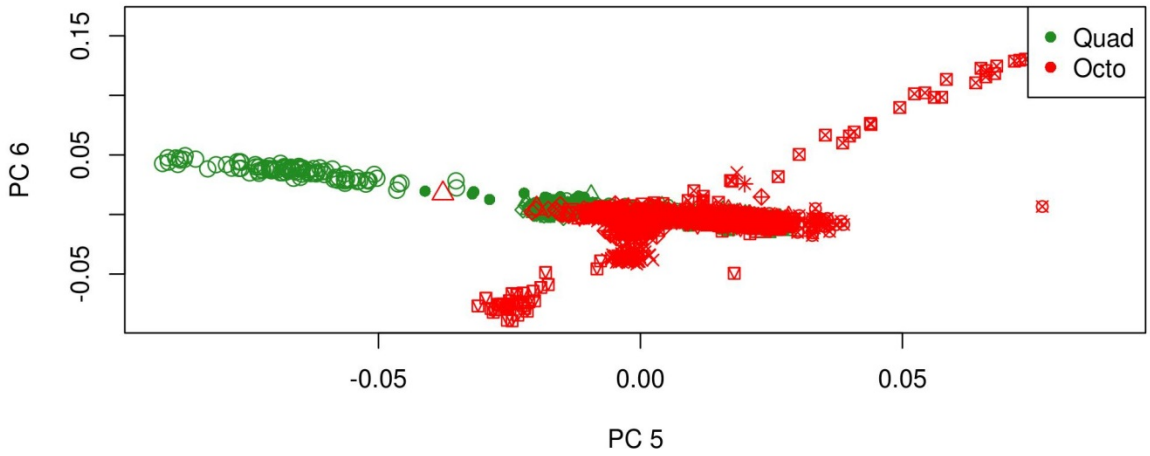


d)

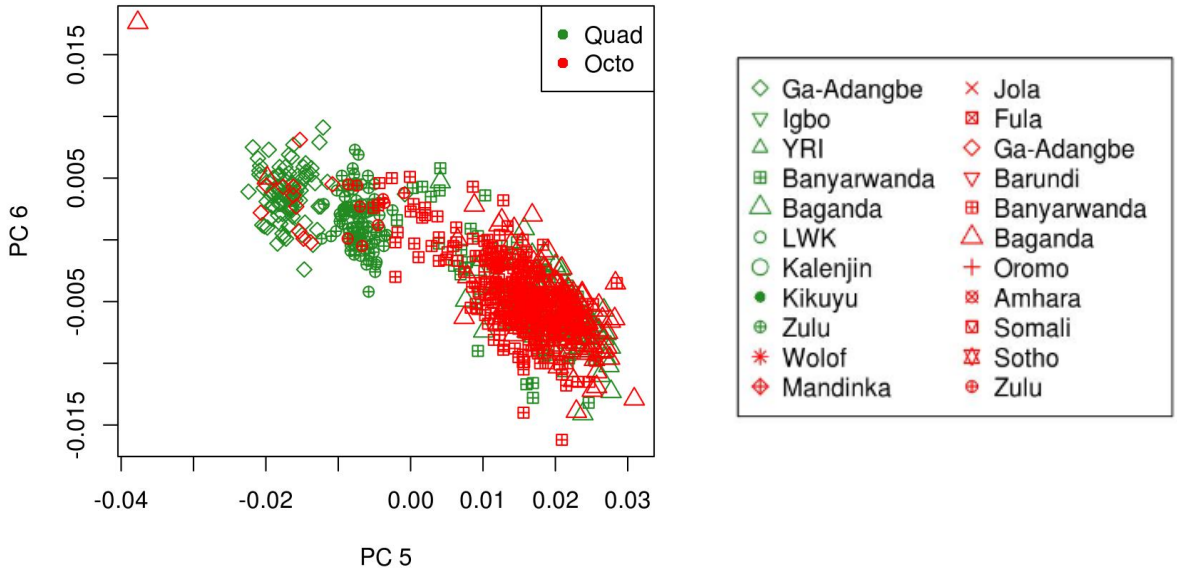


SN1. Figure 11c shows the AGVP African dataset samples represented along PCs 3 and 4 after removal of SNPs with weights > 3 SD from the mean along PCs 2 and 3. Samples genotyped on the quad chip are shown in green, while those genotyped on the octo platform are shown in red. SN1. Figure 11d depicts PCs 3 and 4 for samples only from the four populations that were genotyped across both chips. No separation is observed by chip.

e)

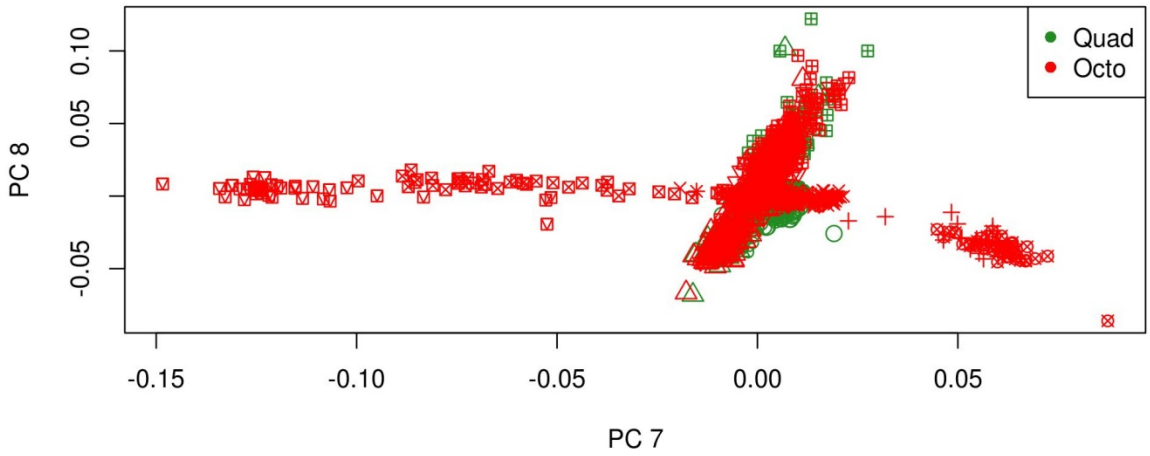


f)

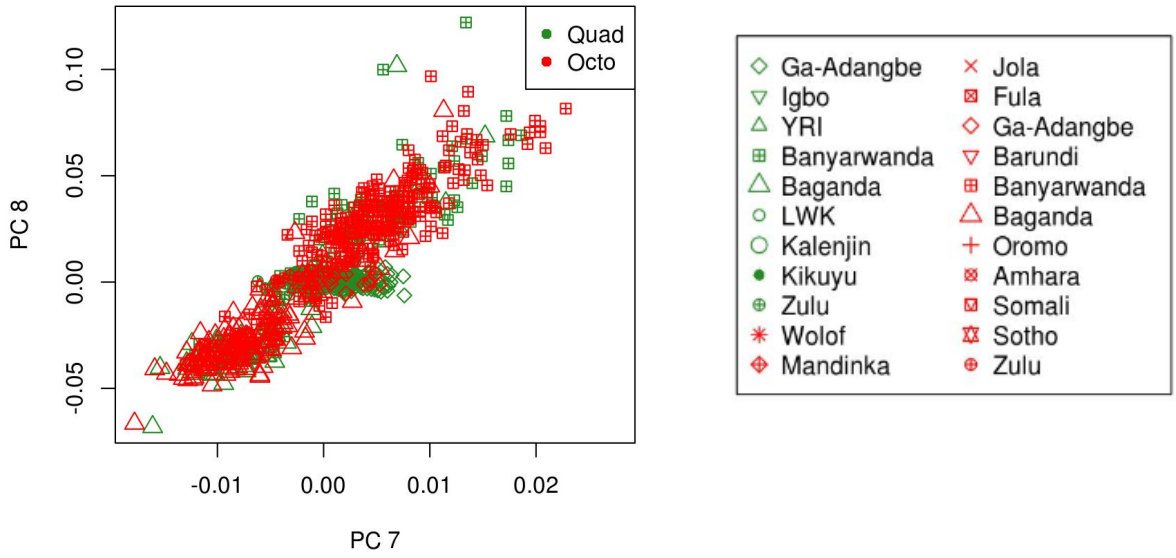


SN1. Figure 11e shows the African dataset samples represented along PCs 5 and 6 after removal of SNPs with weights > 3 SD from the mean along PCs 2 and 3. Samples genotyped on the quad chip are shown in green, while those genotyped on the octo platform are shown in red. SN1. Figure 11f depicts PCs 5 and 6 for samples only from the four populations that were genotyped across both chips. No separation is observed by chip.

g)



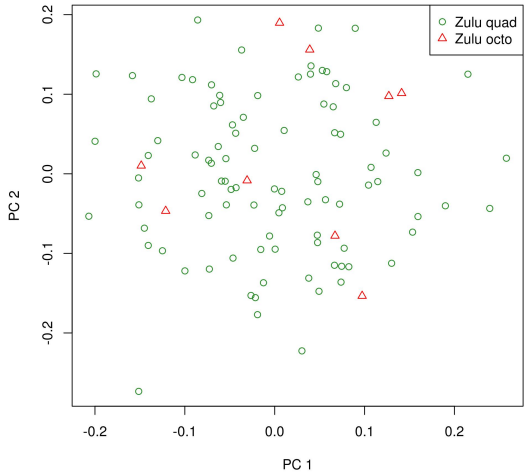
h)



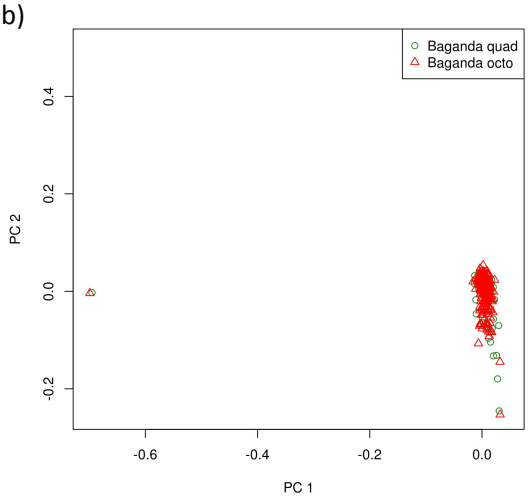
SN1. Figure 11g shows the AGVP African dataset samples represented along PCs 7 and 8 after removal of SNPs with weights > 3 SD from the mean along PCs 2 and 3. Samples genotyped on the quad chip are shown in green, while those genotyped on the octo platform are shown in red. SN1. Figure 11h depicts PCs 7 and 8 for samples only from the four populations that were genotyped across both chips. No separation is observed by chip.

SN1. Figure 12: PCA plots of African populations with SNP weights > 3 SD from the mean removed along PCs 2 and 3 of the African dataset

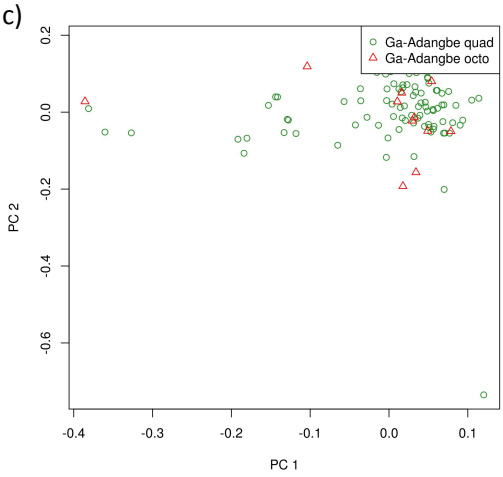
a)



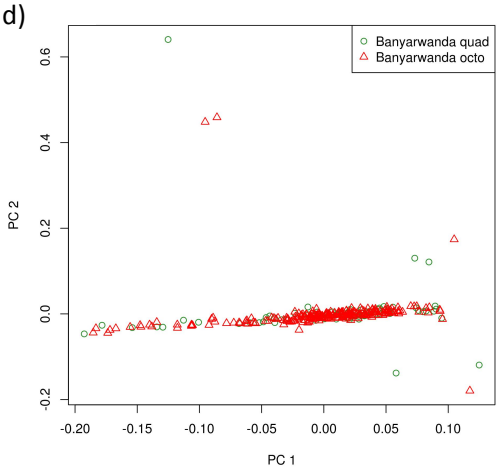
b)



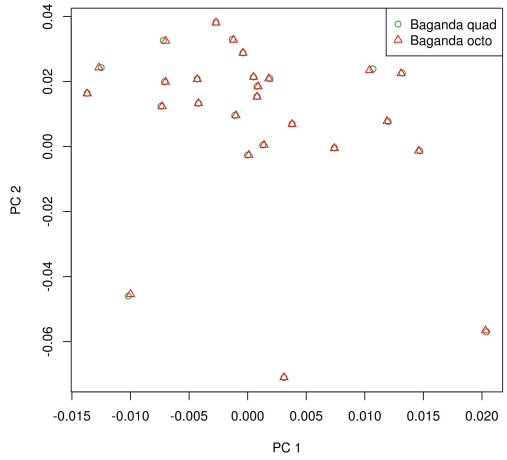
c)



d)

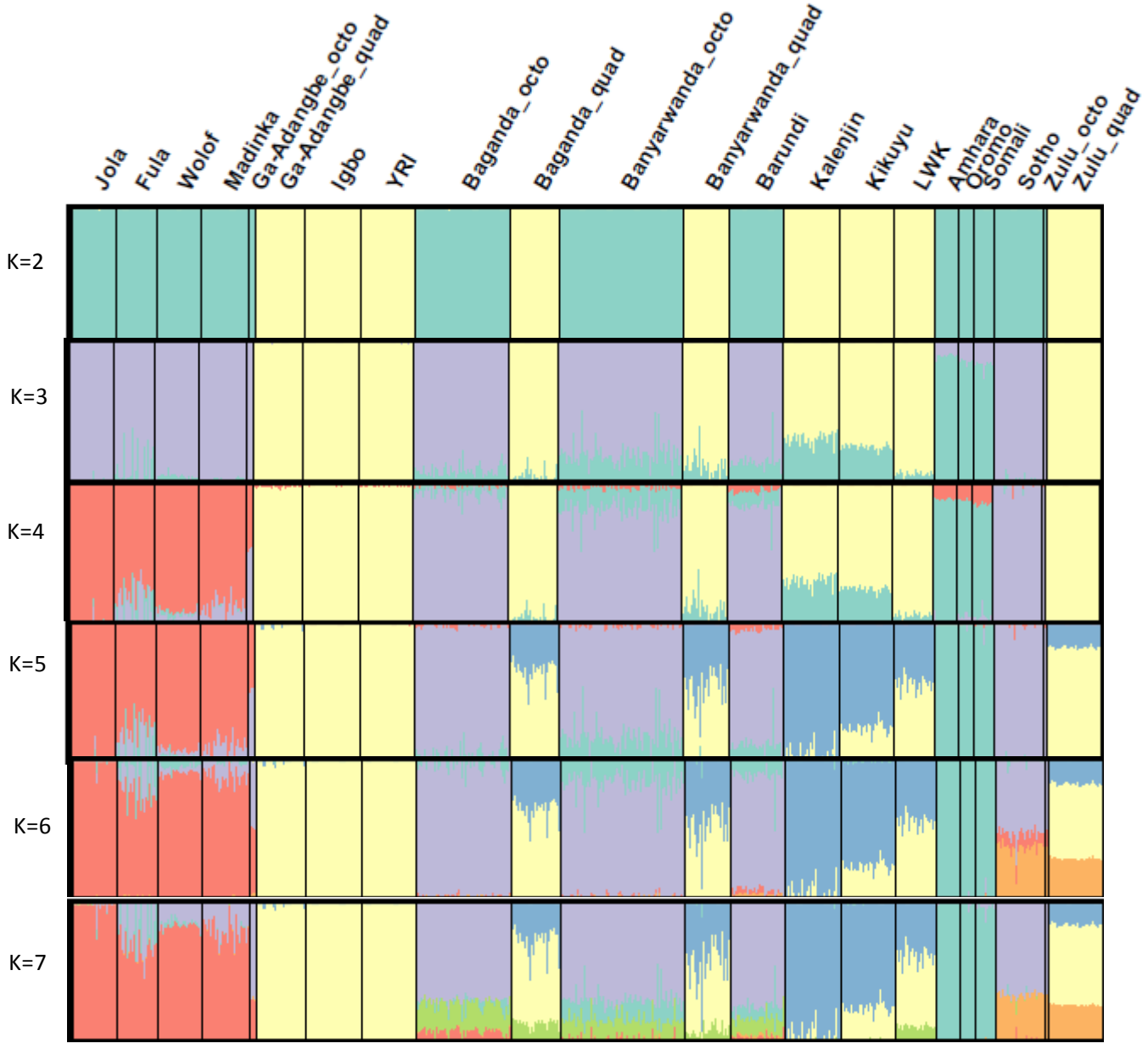


e)



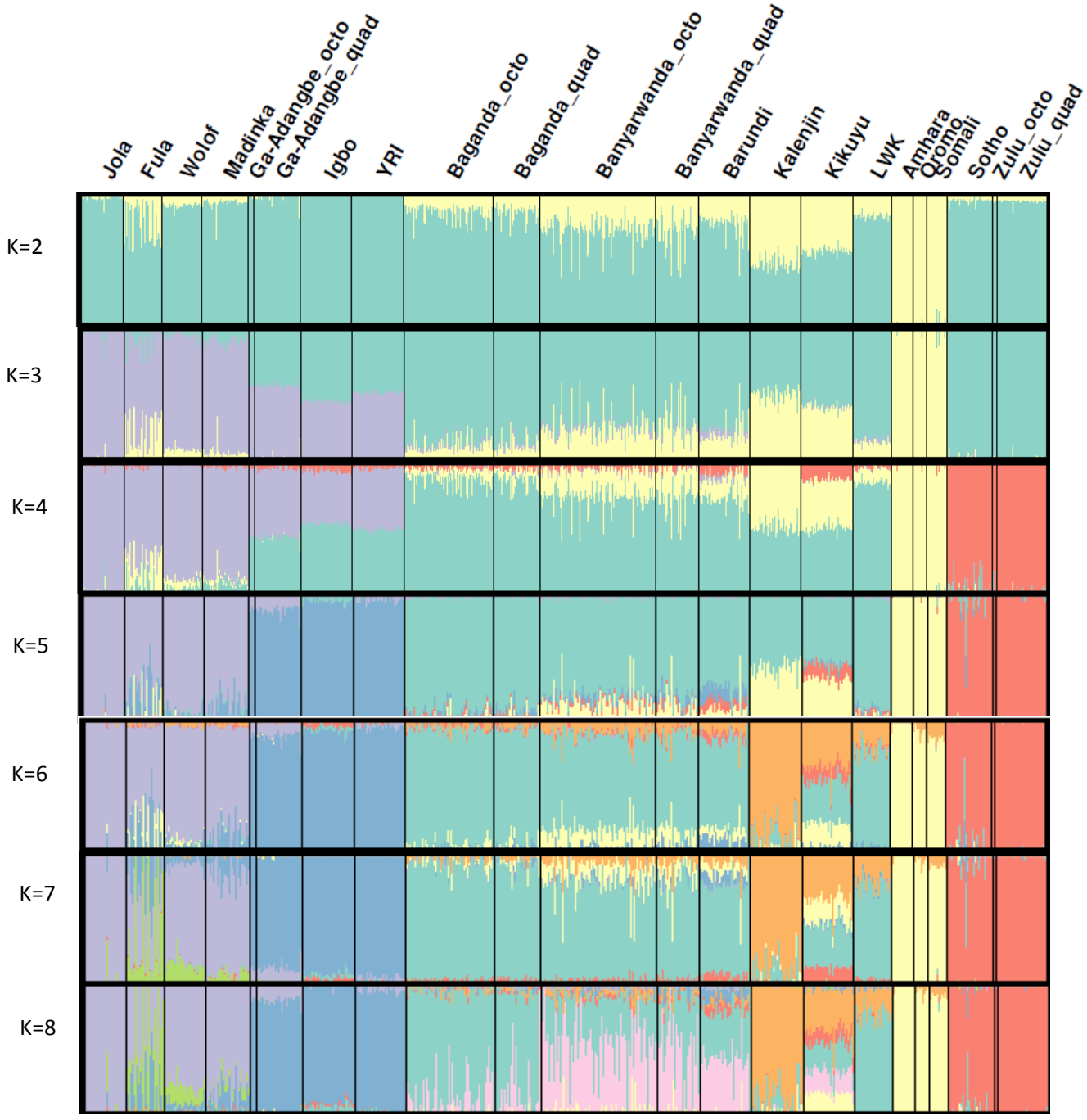
SN1 Figure 12. Figures 12 a, b, c and d depict the first and second PCs in Baganda, Banyarwanda, Ga-Adangbe and Zulu with samples genotyped on the octo or quad chips calculated from the SNPs in the African dataset after removal of SNPs with weights > 3 SD along PCs 2 and 3. Samples genotyped on the quad and octo chips are represented by green circles and red triangles respectively. SN1 Figure 12e shows principal components calculated from 26 Baganda samples genotyped on both chips, with analysis carried out on quad samples and projected onto genotype data for octo samples after removal of chip effects. No separation is observed between the two chips.

SN1 Figure 13: Admixture clustering analysis of African dataset before removal of chip effects from African dataset



SN1. Figure 13 shows ADMIXTURE clusters (K=2-7) before removal of chip effects from the African dataset. Cluster separation based on chip differences is observed from K=2 onwards.

SN1. Figure 14: Admixture clustering analysis of the African dataset after removal of chip effects



SN1 Figure 14 shows ADMIXTURE clusters (K=2-8) following removal of chip effects from the African dataset. Cluster separation based on chip differences is not observed. However, other ancestral effects seem unchanged.

1.6 Removal of outliers and related individuals

Following removal of chip effects, the four populations with samples across both chips were extracted from the global dataset and re-examined for related individuals, as previous QC had been carried out per population per chip. Pairwise IBD was calculated between all individuals in each population using the set of SNPs from the curated global dataset after LD pruning of data to an r^2 threshold of 0.5 using PLINK. Individuals with IBD>0.05 were iteratively removed until all individuals in each population were unrelated. PCA outliers were identified by inspecting the top ten PCs of the curated global dataset. Related individuals and PCA outliers were removed from the global dataset, and reciprocally also removed from the African data and individual population data.

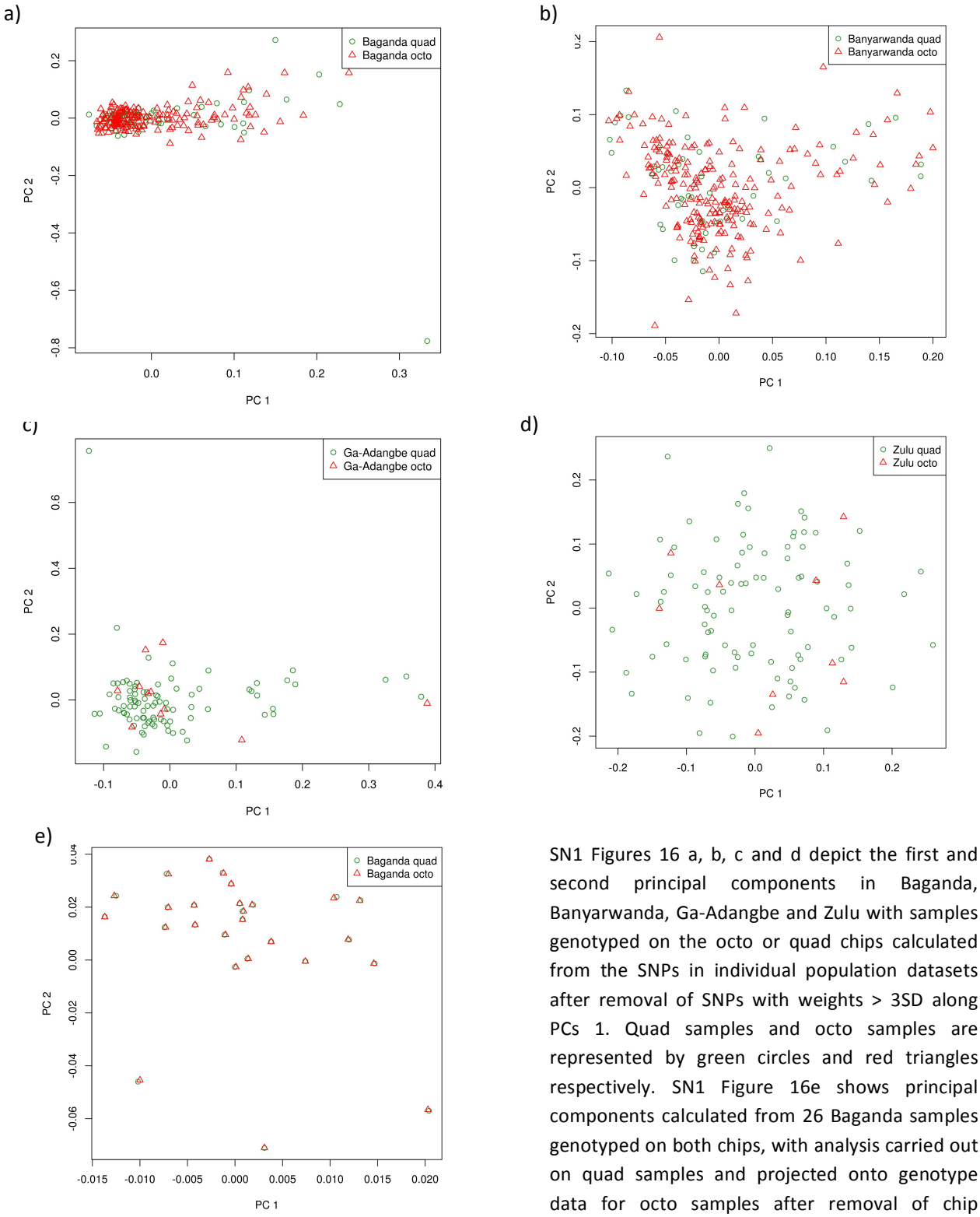
1.7 Curation of individual population data

Following QC carried out per population per chip, data from the four populations with data across both chips were curated as described above with removal of highly weighted SNPs along PC 1 representing chip based separation, after removal of PCA outliers and related individuals as identified in the global dataset (**SN1 Figures 1 and 15**). The number of SNPs retained in each population following QC and removal of chip effects can be found in **SM Table 4 and 5**.

1.8 Subsampling of global and African datasets

In order to account for differences in sample numbers in different populations, for subsequent analyses, both the global and African curated datasets were subsampled to approximately 100 individuals for each population randomly. For populations with less than 100 individuals, all individuals were retained (**SM Table 3**).

SN1 Figure 15: PCA plots of African populations with PC > 3SD removed along PC 1 in individual population data



SN1 Figures 16 a, b, c and d depict the first and second principal components in Baganda, Banyarwanda, Ga-Adangbe and Zulu with samples genotyped on the octo or quad chips calculated from the SNPs in individual population datasets after removal of SNPs with weights > 3SD along PCs 1. Quad samples and octo samples are represented by green circles and red triangles respectively. SN1 Figure 16e shows principal components calculated from 26 Baganda samples genotyped on both chips, with analysis carried out on quad samples and projected onto genotype data for octo samples after removal of chip effects. No separation is observed between the two chips.

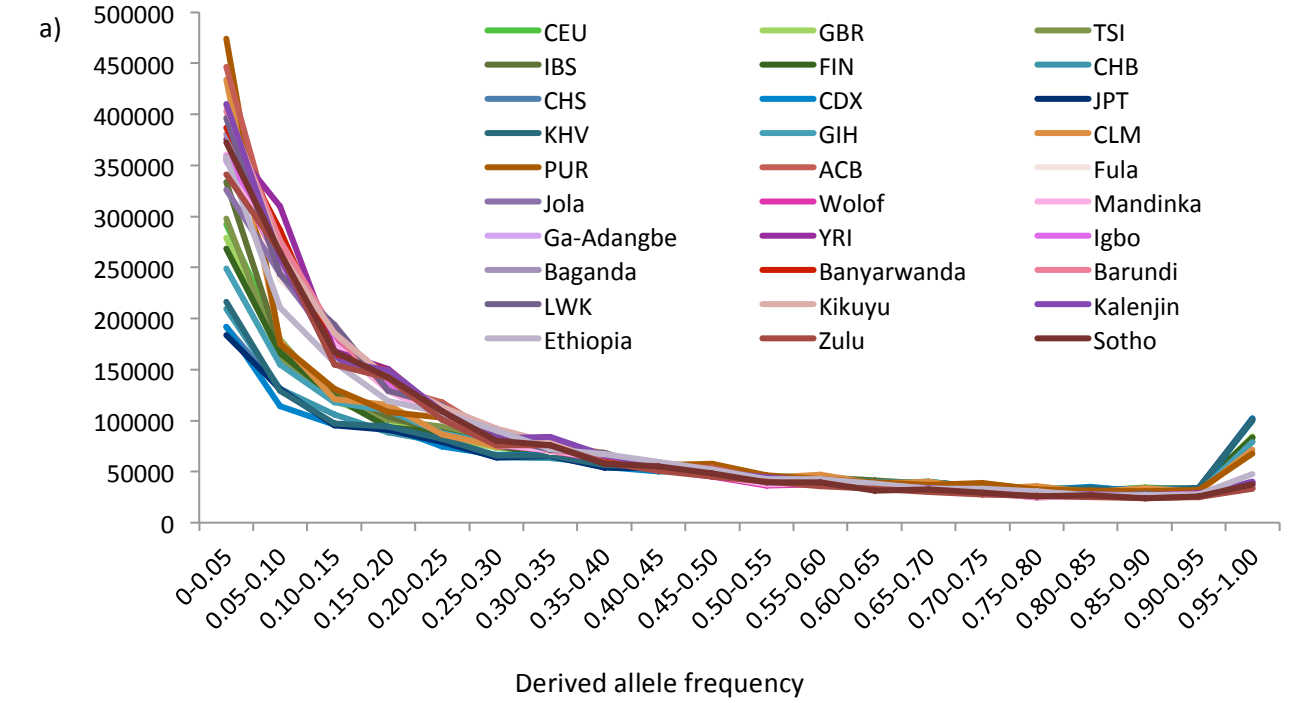
S. NOTE 2: ASSESSING ASCERTAINMENT BIAS ON THE OMNI2.5M ARRAY

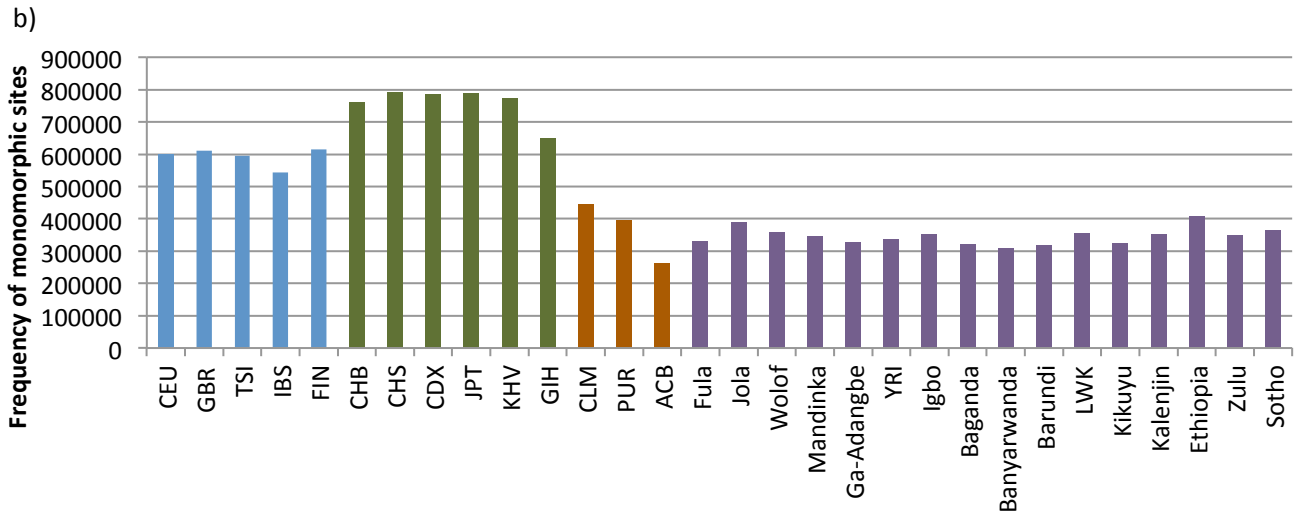
Although genotype array based studies have been extensively used in the study of population genetics, it is well recognised that ascertainment bias can distort various statistics relative to different populations.[36] This bias arises from the ascertainment of SNPs on the chip being dependent on sequence data from a few individuals from population sets that do not fully represent the population being genotyped. Therefore, such ascertainment can bias site frequency spectra (SFS), and statistics dependent on this (e.g. F_{ST}) differentially among populations depending on how well those populations are represented by ascertained markers on the array.[36] Relative to this, whole genome sequencing data can be considered less biased; however, it has been shown that even low coverage sequence calls are poor at capturing rare variation, leading to bias in the SFS towards common variation.[9] In order to assess ascertainment bias on the Illlumina Omni 2.5M genotype array, we used a variety of approaches: 1. We compared the SFS determined in different populations on the genotype array with the whole genome sequencing based SFS; 2. We compared F_{ST} statistics obtained by genotype data with those calculated from WGS data; and 3. We assessed the impact of different ascertainment schemes on *f3* statistics and inferences from these to examine if ascertainment would render our results unreliable. For calculation of the SFS for low coverage sequence data, we used the direct SFS maximum likelihood approach for inference in order to avoid biases against rare variant that can occur during genotype calling of such data.[9]

SN2.1 COMPARISON OF THE SFS OF SEQUENCE DATA AND GENOTYPE ARRAY DATA

In order to directly assess ascertainment bias on the Illumina Omni 2.5M, we estimated the SFS on the sequence data and for the genotype array. First we compared the allele frequency spectra among African populations on the Omni 2.5M genotype array; we found an excess of rare variation in all African populations relative to European and Asian populations (**SN2 Fig 1a**). The frequency of monomorphic variants was noted to be highest among Asian populations, followed by European populations. African populations had the lowest proportion of monomorphic variants, with a similar distribution among populations (**SN2 Fig 1b**). This argues against substantial ascertainment in favour of European populations.

SN2 Fig 1: Derived allele frequency spectra in different African populations

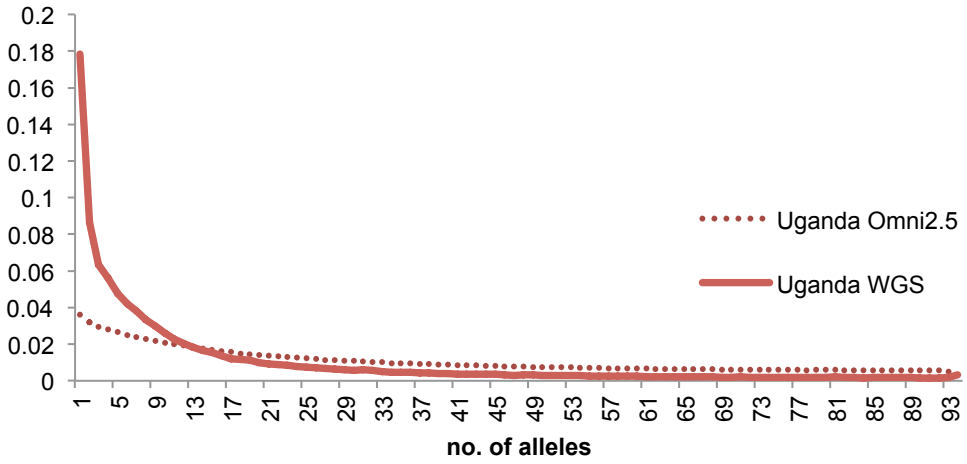




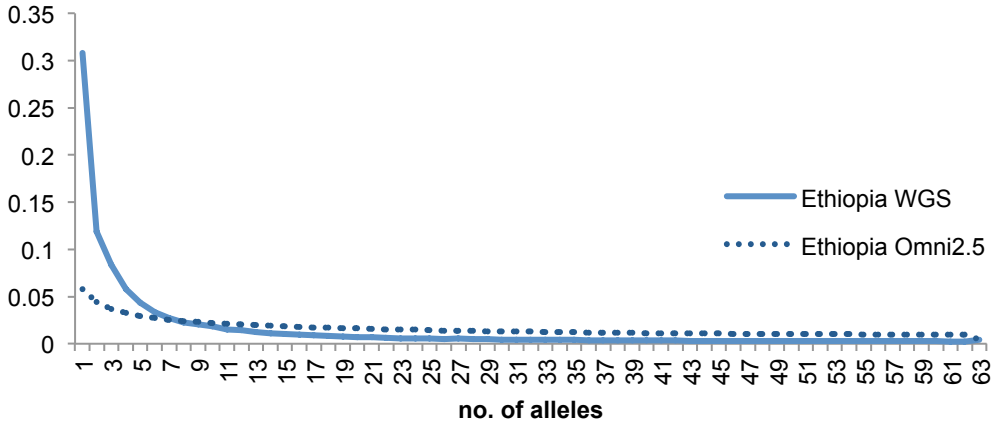
SN2 Fig 1 represents the allele frequency spectra of African populations in a global context. SN2 Fig 1a shows the derived allele frequency spectra for African populations and populations from the 1000 Genomes Project. African populations show a higher proportion of rare variants compared to European and Asian population groups. SN2 Fig 1b shows the count of monomorphic sites across the autosomal regions on the 2.5M Illumina Omni chip array for all populations. An excess of monomorphic sites is observed in European and Asian populations, with broadly similar counts of monomorphic sites across African populations.

SN2 Fig 2: Site frequency spectra for WGS and genotype array data

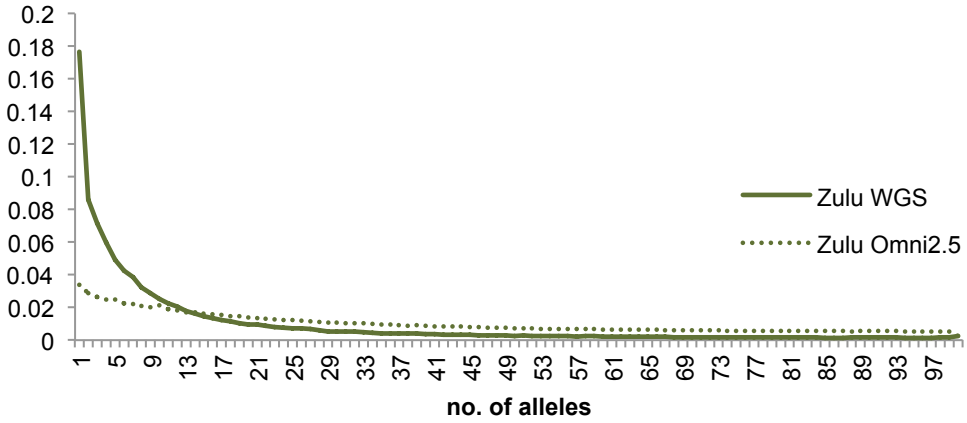
a)



b)



c)



SN2 Fig 2 depicts the site frequency spectra derived directly for 4x WGS and genotype data for Uganda, Ethiopia and Zulu. We see evidence of clear ascertainment bias, biased against detection of rare variation on the genotype array.

Next, we compared the SFS estimated from low coverage sequencing data with that estimated from genotype array data for all populations. Expectedly, we find evidence for biased SFS in genotype array data, with poor capture of rare variation among all populations (**SN2 Fig 2**).

SN2.2 Comparison of F_{ST} statistics

As F_{ST} metrics are dependent on the SFS of a given population, these are likely to be affected by ascertainment bias on genotype arrays. We assessed ascertainment bias on the Omni 2.5M chip array by comparing F_{ST} metrics calculated using genotypes on the array, with F_{ST} estimates obtained from whole genome sequence data for the same individuals. F_{ST} estimates were obtained using the Hudson's method with EIGENSOFT version 4. We found that F_{ST} estimates particularly in relation to Ethiopian populations were underestimated on the genotype array, while difference among these populations was overestimated (**SN2 Table 1**).[36] Estimates for F_{ST} among the Bantu populations (Zulu and Baganda) were very similar, which is consistent with Bantu populations in the 1000 Genomes Project used for ascertainment of markers on the chip. This suggests that the ascertainment scheme on the genotype array may distort differentiation with respect to Ethiopian populations.

SN2 Table 1: Comparison of F_{ST} estimates between sequencing and genotype data

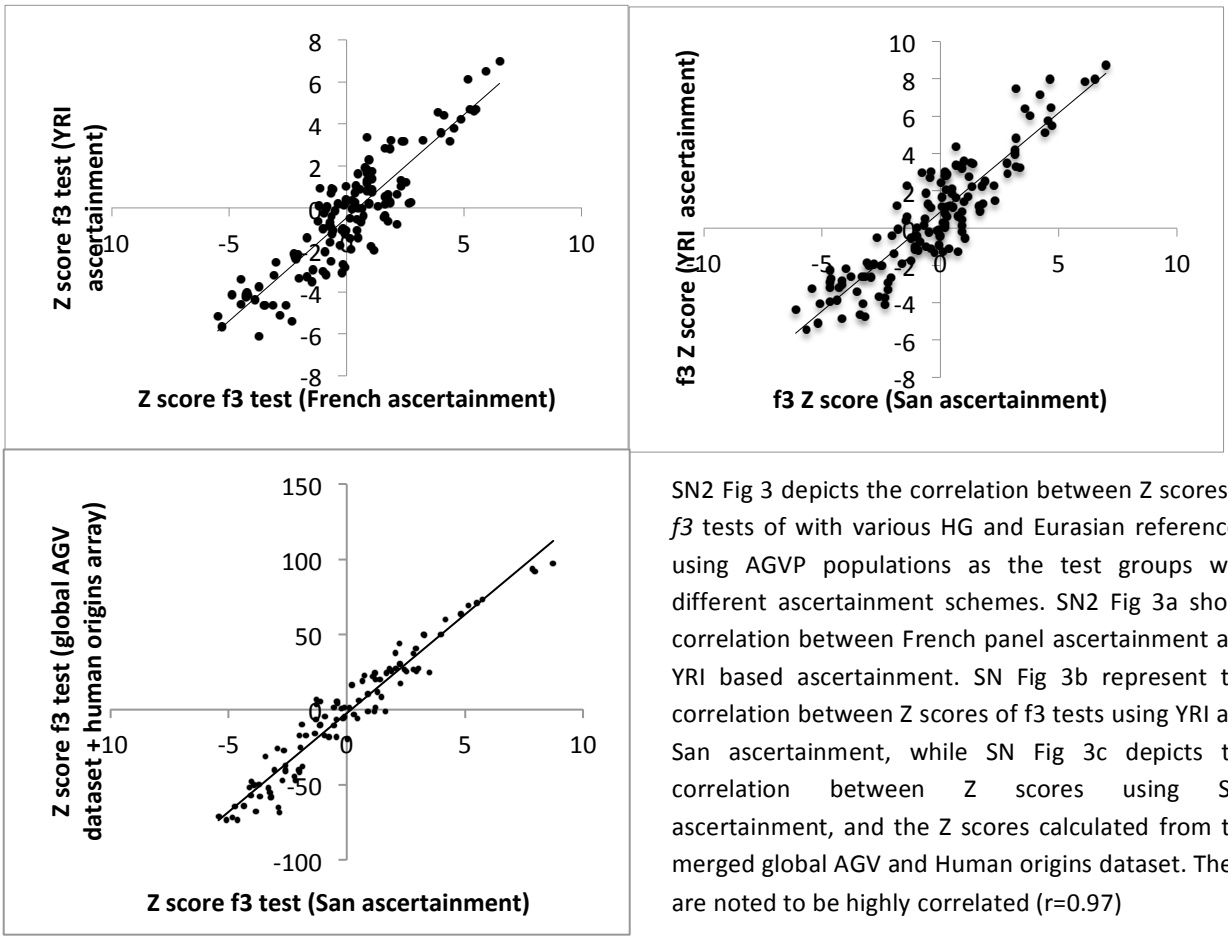
			F_{ST} from low coverage whole genome sequence		
Population	**Baganda**	**Zulu**	**Oromo**	**Amhara**	**Somali**
Baganda	0	0.008	0.035	0.039	0.035
Zulu	0.009	0	0.041	0.045	0.041
Oromo	0.025	0.032	0	0.000	0.008
Amhara	0.025	0.032	0.002	0	0.009
Somali	0.037	0.044	0.018	0.017	0
F_{ST} from Illumina Omni 2.5M genotype array					

SN2.3 Influence of ascertainment bias on *F3* statistics

We assessed the influence of such ascertainment bias on the *f3* test, a formal test for admixture, to assess whether this ascertainment bias may have impact our inferences on population admixture in Africa. *F3* statistics are considered formal tests for admixture. These have been reported to be robust to complex demographic history among populations, as well as ascertainment bias.[12] However, we sought to confirm this in our dataset, to ensure our inferences were reliable and not influenced by a biased scheme of ascertainment on the Omni 2.5M genotype array. In order to assess this, we compared *f3* tests calculated using 3 different

ascertainment schemes (French, YRI and San), using ascertainment schemes defined on the Human origins array.[12] As we needed to merge the Human origins dataset with the African AGVP dataset to examine admixture, there was substantial loss of data, due to non-overlapping markers. In order to avoid this, we imputed both datasets up to the 1000 Genomes sequence reference panel, and then carried out very stringent filtering to include only very high quality sites (info score>0.90). We then merged the data for overlapping sites, and examined *f3* tests using these data with different ascertainment schemes. Although the number of markers overlapping between the datasets was still quite low (185, 187 and 287 markers in the French, Yoruba and San ascertainment panel respectively), we were able to explore ascertainment effects on statistics. We noted that <10% of all markers were overlapping between the different ascertainment panels, suggesting that any correlation among f3 statistics was unlikely to result from this overlap. We compared Z scores obtained when using various reference populations, including Ju/'hoan North, Mbuti Pygmy, Biaka Pygmy, Hadza, French, Basque, TSI, IBS and YRI, and AGVP populations as test populations (**SN2 Fig 3**). On comparing different ascertainment

SN2 Figure 3: Correlation of Z scores from f3 statistics using different ascertainment schemes



SN2 Fig 3 depicts the correlation between Z scores of *f3* tests of with various HG and Eurasian references, using AGVP populations as the test groups with different ascertainment schemes. SN2 Fig 3a shows correlation between French panel ascertainment and YRI based ascertainment. SN Fig 3b represent the correlation between Z scores of f3 tests using YRI and San ascertainment, while SN Fig 3c depicts the correlation between Z scores using San ascertainment, and the Z scores calculated from the merged global AGV and Human origins dataset. These are noted to be highly correlated (r=0.97)

panels, we found that *f3* statistics (Z scores) were highly correlated between French and YRI ascertainment (r=0.89), YRI and San ascertainment (0.91), and French and San ascertainment (r=0.92) (**SN2 Fig 3a and 3b**). Additionally, very high correlation was observed between Z scores calculated from the San ascertainment panel, and the Z scores we calculated from the merged global AGV and human origins dataset (r^2=0.97) (**Supplementary Table 3 and SN2 Figure 3c**). This suggests that our *f3* statistics and inferences regarding population admixture are robust to ascertainment, as has been previously suggested.[12]

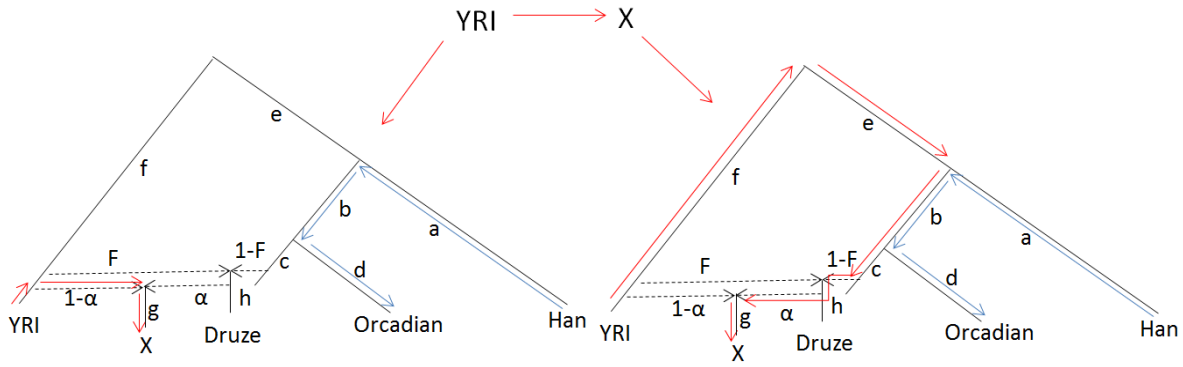
In summary, we note some ascertainment bias, specifically with regards to Afro-Asiatic populations on the Omni 2.5M genotype array. However, our analyses suggest that this ascertainment bias is unlikely to influence key findings and conclusions drawn, due to test statistics used being robust to these biases.

S. NOTE 3: ESTIMATING EURASIAN ANCESTRY AMONG AGVP POPULATIONS

We found evidence of extensive Eurasian admixture among AGVP populations based on ADMIXTURE analysis and *f3* tests (**Figure 1 and Supplementary Tables 2 and 3**). This is in keeping with previous reports of widespread non-SSA ancestry in South and East Africa.[7] In order to estimate non-SSA ancestry among AGVP populations, we applied the *f4* ratio test calculating the ratio between *f4*(Han,Orcadian;YRI,X) and *f4*(Han,Orcadian;YRI,Druze), similar to that described by Pickrell et al.[7] To reduce bias in estimation relating to non-Eurasian ancestry among Druze, and non-SSA ancestry among YRI, we used a variety of methods, as described below. Using these methods, we find evidence of extensive Eurasian ancestry ranging from <1% to 50% among different populations in East, West and South Africa.

SN3 Fig. 1: Topology of *f4* test for estimation of Eurasian ancestry

$$f4(Han,Orcadian;YRI,X)=(1-F-\alpha)b \qquad f4\ ratio=\frac{f4(Han,Orcadian;YRI,X)}{f4(Han,Orcadian;YRI,Druze)}=\frac{(1-F-\alpha)b}{(1-F)b}=\frac{1-F-\alpha}{1-F}$$



SN 3 Fig 1. The figure depicts the topology f4(Han, Orcadian; YRI, Druze) utilised to estimate the proportion of Eurasian ancestry in X. We account for the presence of non-zero African-like ancestry in Druze, and consider that only 1-F proportion of Druze represents Eurasian gene flow. Therefore the *f4* ratio shown above can be represented as 1-F-α/1-F, where F is approximately 0.05.[7]

To reduce bias, we used the following two methods:

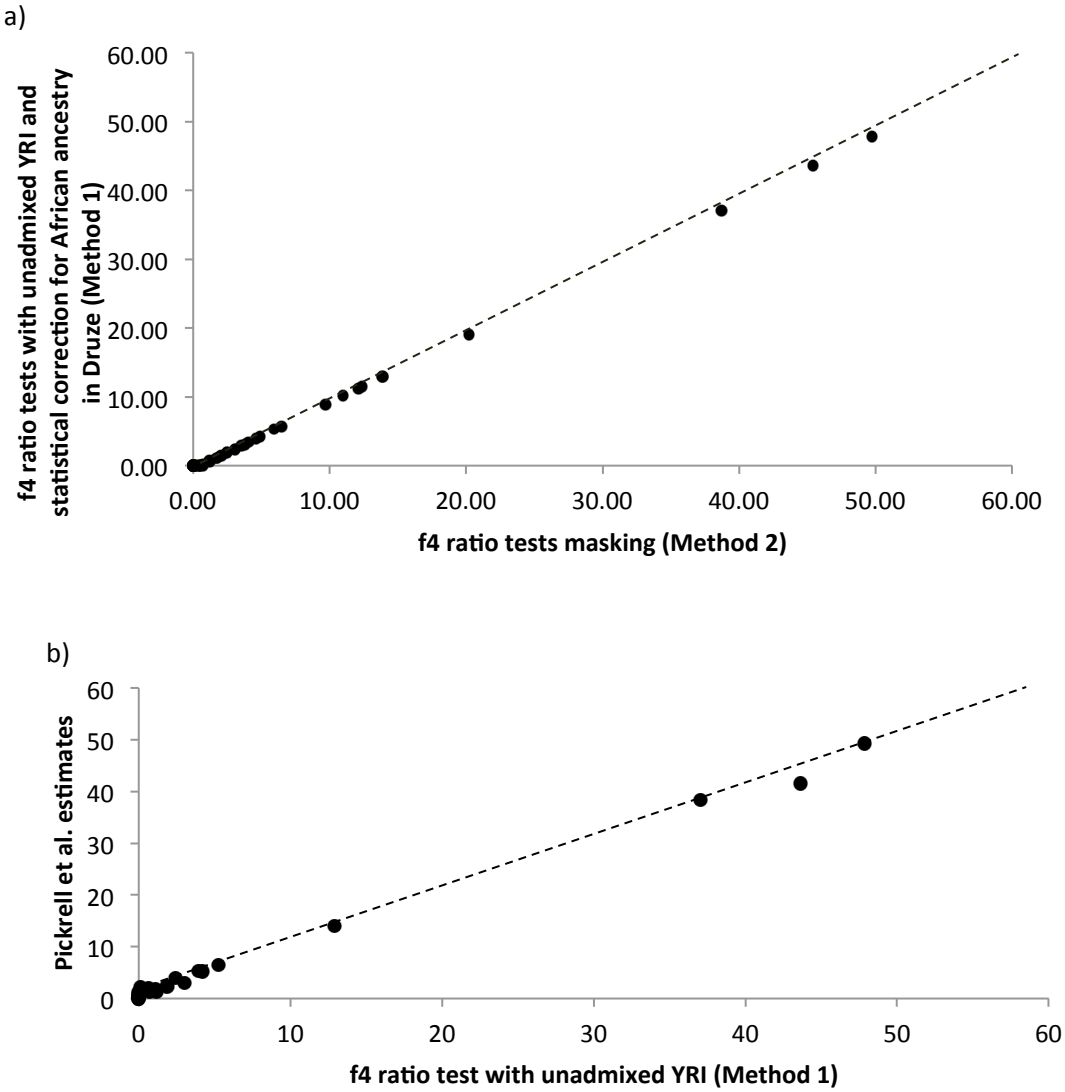
1. In order to reduce bias due to Eurasian ancestry in YRI, we only included 13 YRI individuals with <0.0025% Eurasian ancestry estimated with ADMIXTURE, cluster K=18. We refer to these as 'pure YRI'. Additionally, to account for non-zero African ancestry among Druze, we estimated that the *f4* ratio would provide an estimate dependent on both the fraction of Eurasian admixture in test population X (α), as well as the fraction of African ancestry among Druze (F), as represented in **SN3 Fig**

1. We used as estimate of F=0.05 as in Pickrell et al.[7], in order to calculate α, the proportion on Eurasian ancestry among test populations.

2. In order to further validate the approach used above, we used a masking approach, where we masked Eurasian ancestry in YRI, and African ancestry among Druze. Masking of haplotypes was carried out with PCAdmix, using multiple reference populations, including 13 pure YRI individuals, 180 TSI, French, Basque and Han individuals from HGDP and the 1000 Genomes Project, 10 Mbuti Pygmy and 5 Ju/'hoan North individuals with <0.0025% non-hunter-gatherer ancestry on ADMIXTURE analysis (referred to as 'pure Mbuti Pygmy' and 'pure Ju/'hoan North' subsequently). Any segment of the genome with >0.10 probability of non-African ancestry in YRI was masked, and any segment of the genome in Druze with >0.10 of YRI-like ancestry was masked.

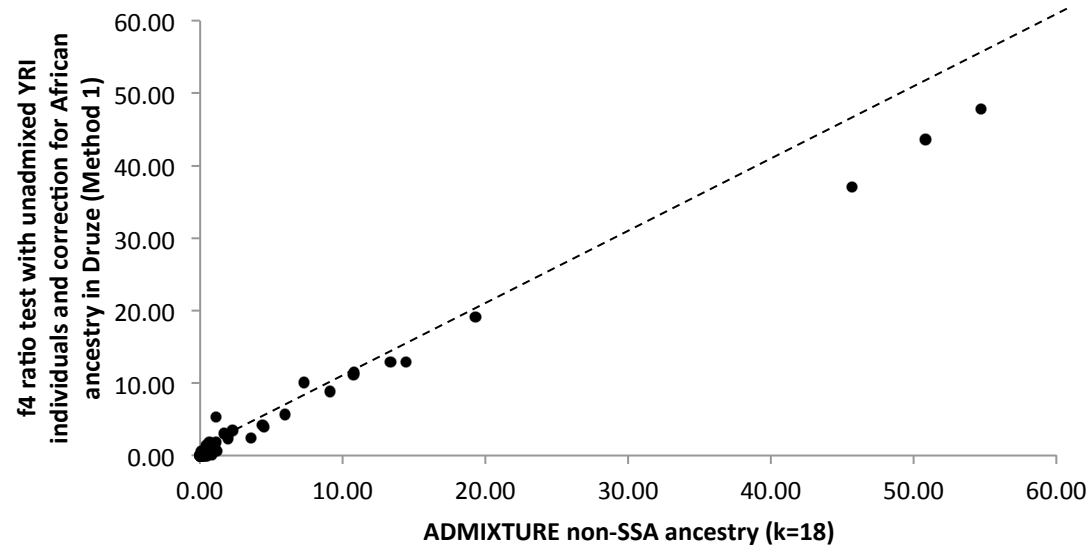
Both methods produced highly correlated results (r^2=0.99) (**SN3 Fig 2a**). Additionally, our *f4* ratio estimates calculated with Method 1 were found to be very strongly correlated with those produced by Pickrell et al[7] (**SN3 Fig 2b)** and ADMIXTURE k=18 (r2=0.99 and r2=1 respectively) **(SN3 Fig 2c)**. We note that *f4* ratio tests produced slightly lower estimates of Eurasian ancestry for Ethiopian populations as compared to ADMIXTURE estimates. We further evaluated these estimates by using Dinka as a reference (rather than YRI), as this population may be more representative of African ancestry in Ethiopian populations.[7] Using Dinka instead of YRI, did not alter estimates for the Ethiopian populations (0.44, 0.48 and 0.38 for Oromo, Amhara and Somali, respectively), suggesting that estimates are robust to the reference population used. We also noted that not accounting for Eurasian ancestry in YRI, and using admixed YRI individuals led to very slight over-estimation of Eurasian ancestry among populations (**SN3 Fig 2d**)**.** We present the *f4* ratio estimates obtained using different methods in **Supplementary Table 4**.

SN3 Fig 2: A comparison of f4 ratio estimates of non-SSA ancestry in AGVP populations using different methods

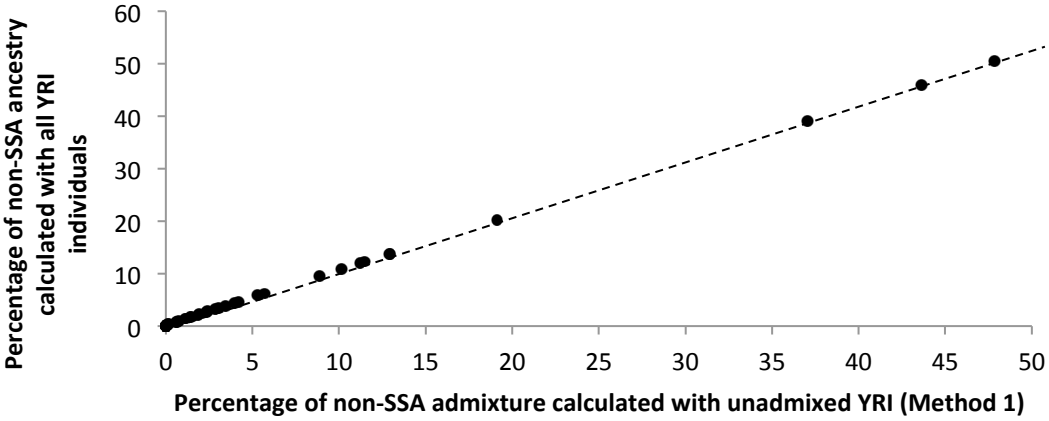


SN3 Fig 2 represents the correlation between different methods used to assess proportional Eurasian ancestry among AGVP populations. SN3 Fig 1a represents the correlation between Method 1 (using unadmixed individuals from YRI and statistically correcting for African ancestry in Druze) and Method 2 (masking Eurasian ancestry in YRI and African ancestry in Druze) estimates. Correlation is noted to be high. The dashed line is the line of perfect equality between the two methods. SN3 Fig 2b represents the correlation between estimates of Eurasian ancestry obtained using Method 1 and estimates published by Pickrell et al.[7] in the same populations.

c)



d)

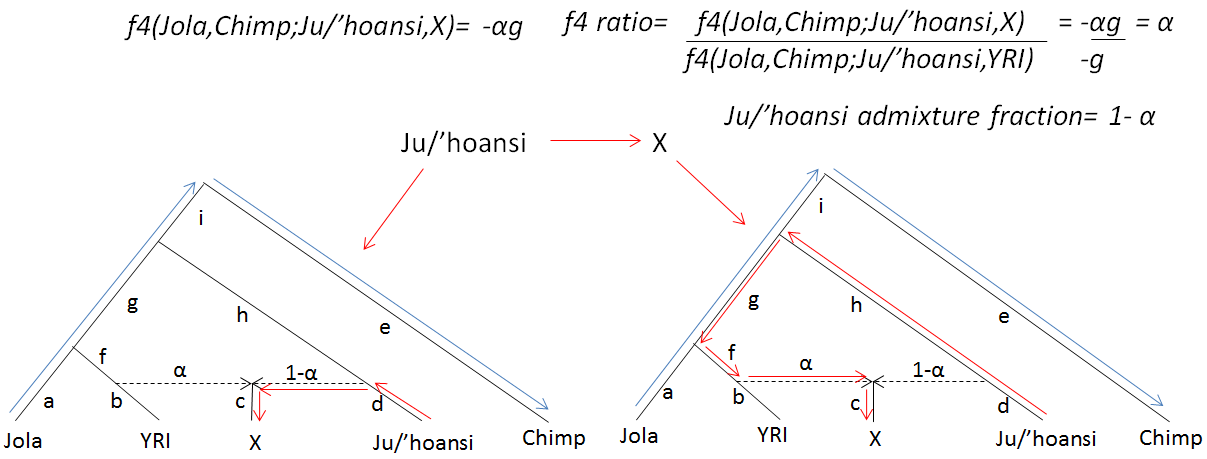


SN3 Fig 2 represents the correlation between different methods used for determination of Eurasian ancestry among AGVP populations. SN3 Fig 2c represents good correlation between estimates obtained using Method 1 and ADMIXTURE estimates from the best fitting cluster. The dashed line represents the line of perfect equality. SN3 Fig 2d represents the correlation between estimates obtained using all YRI individuals as a reference, and using only those with <0.0025% Eurasian ancestry on ADMIXTURE analysis. Both estimates are highly correlated, with very slight over-estimation of estimates when all YRI individuals, including admixed individuals are included in analysis.

S. NOTE 4: ESTIMATING HUNTER-GATHERER (HG) ANCESTRY IN SSA

We found evidence for HG admixture among several SSA populations on PCA, ADMIXTURE and *f3* test analyses (**Figure 1, Supplementary Tables 2 and 3**). In order to estimate HG proportions in SSA, we applied a variety of methods utilising *f4* ratio tests, and accounting for the multiple sources of ancestry among these populations. In order to determine the true proportion of HG ancestry among SSA populations the *f4* ratio test (Chimp, Jola; HG, X)/ (Chimp, Jola; HG, YRI) can be applied. An example of such a topology is represented in **SN4 Fig 1**. For populations where the source of ancestry is likely to be from Mbuti Pygmy-like ancestral populations, we can estimate the ratio *f4*(Chimp, Jola; Mbuti Pygmy, X)/f4(Chimp, Jola; Mbuti Pygmy, YRI). Using these approaches we find evidence for HG ancestry in West, East and South Africa, ranging from <1%-24% among the AGVP populations.

SN4 Figure 1: An example topology for estimating HG admixture among test populations



SN4 Fig 1. To calculate the proportion of HG ancestry a given population X, we first calculate *f4*(Jola, Chimp; YRI, X). Different paths can be traced based on the topology of populations specified, as shown. The drift parameters for overlap between paths (small letters a-i along the graph) are considered for these calculations. If the path from Jola to the outgroup population (Chimp) overlaps with the path from YRI to X, the *f4* parameter equates to the drift parameter of the overlapping component of the graph weighted by ancestral proportions (-αg); as this overlap is in the opposite direction, this parameter is negative. This attribute is used to calculate proportional admixture, as *f4*(Jola, Chimp; YRI, X)=α(Jola, Chimp; YRI, HG). Here, we have used Ju/'hoansi as a possible reference HG population, but others can be similarly used with this topology. For a more detailed discussion on *f4* ratio tests, please refer to Patterson et al.[12]

However, the proportional estimates will only be accurate if Jola, Ju/'hoan North/Mbuti Pygmy and YRI individuals are unadmixed, and if the admixed population X has only two sources of ancestry- HG-related, and YRI-related. It is clear that these assumptions are violated, as it is known that Ju/'hoan North has non-zero Bantu and Eurasian ancestry.[7] Similarly Jola and YRI possibly have a non-zero proportion of non-SSA ancestry, as noted for YRI in our subsequent analyses (**see Supplementary Note 5**). Additionally, most SSA populations have multiple sources of admixture, including Eurasian and HG admixture, so cannot be thought of in terms of a simple two-population mixture model. In order to account for the admixture in the reference populations used in the *f4* tests, we used two approaches:

1. In one analytic approach, we only included YRI/Jola individuals with <0.0025% non-SSA and HG ancestry, and Ju/'hoan North individuals with only <0.0025% non-Khoe-San ancestry. We refer to these individuals as 'YRI pure', 'Jola pure' and 'Ju|'hoan_North pure'. Additionally, in order to account for multiple sources of admixture in the test population, we further corrected the *f4* test (Chimp, Jola pure; HG pure, X) for non-SSA ancestry among the test populations (X). In order to correct for the *f4* test (Chimp, Jolapure; Juhoan_North, X), we used the methodology described by Reich et al.[37] We briefly outline our methods below:

$$f4(Chimp, Jola\ pure; Ju|'hoan\ pure, X)_{corr}$$
$$= \frac{f4(Chimp, Jola\ pure; Ju|'hoan\ North\ pure, X) - e\ f4(Chimp, Jola\ pure; Ju|'hoan\ North\ pure, French)}{1 - e}$$

where *e* is the proportion of Eurasian ancestry in X as determined by ADMIXTURE, k=18. We therefore calculate the *f4* ratio for Khoe-San like admixture as below:

$$1 - \frac{f4(Chimp, Jola\ pure; Ju|'hoan\ North\ pure, X)_{corr}}{f4(Chimp, Jola\ pure; Ju|'hoan\ North\ pure, YRI\ pure)}$$

Similarly the proportion of Mbuti-Pygmy like admixture can be calculated as:

$$1 - \frac{f4(Chimp, Jola\ pure; Mbuti\ Pygmy\ pure, X)_{corr}}{f4(Chimp, Jola\ pure; Mbuti\ Pygmy\ pure, YRI\ pure)}$$

However, this proportion, although corrected for the proportion of European ancestry among test populations, assumes the test population is a mixture of only two populations. Thus the proportion of HG ancestry must be corrected for the proportion of European ancestry among

these populations by multiplying by a factor of (1-e), e being the proportion of Eurasian ancestry in the test population, as estimated by ADMIXTURE (k=18).

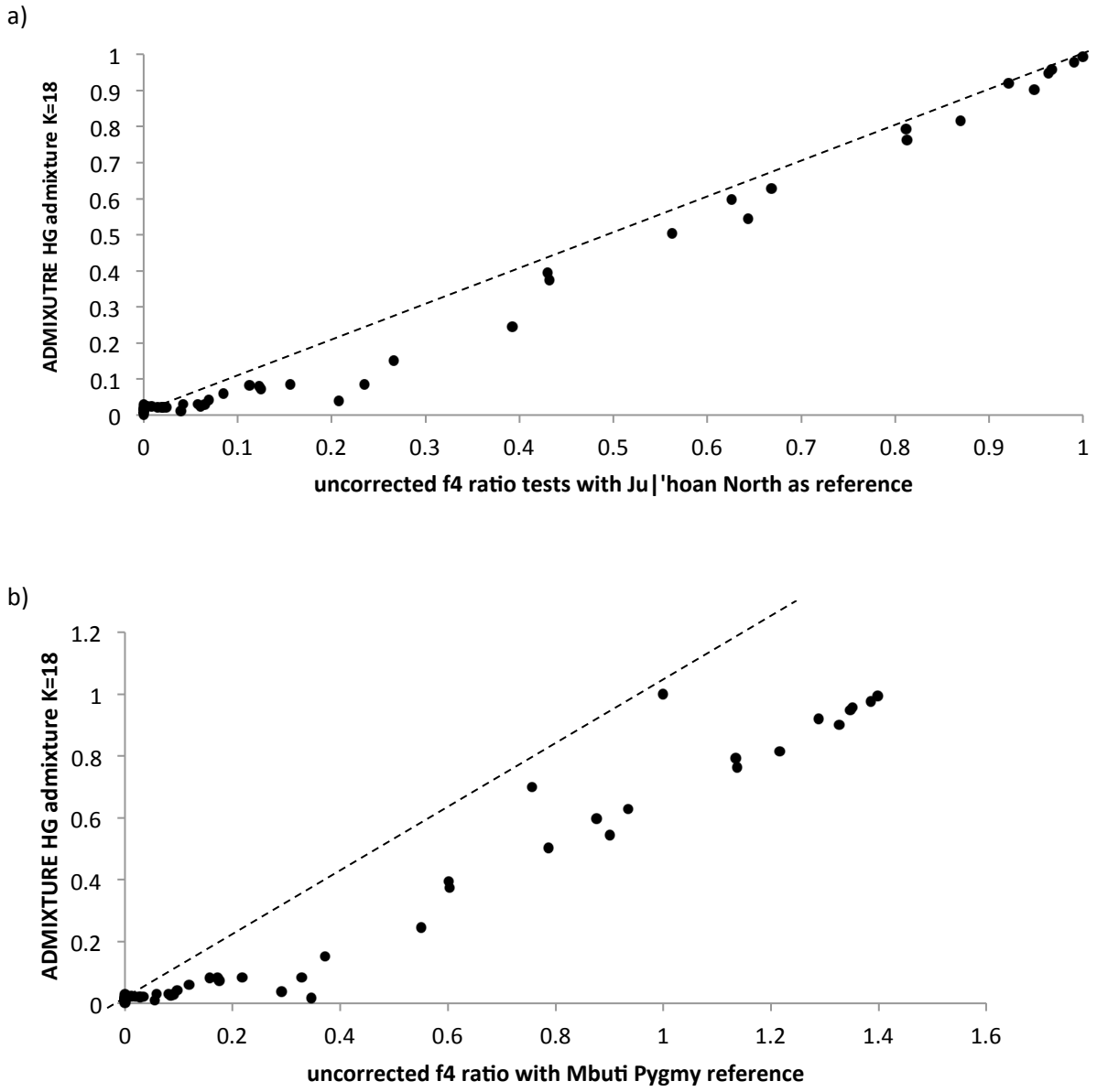
2. In the second analytic approach, we estimated the ratio of Khoe-San or Mbuti-Pygmy like admixture by masking Eurasian ancestry among YRI, Jola and test populations, and non-HG ancestry among Ju/'hoan North and Mbuti Pygmy. Masking was carried out in the same way as described in **Supplementary Note 3**. For YRI, Jola, and test populations, any Eurasian ancestry with a probability of >0.10 was masked, and for Ju|'hoan North and Mbuti Pygmy, any genomic window with a probability of >0.10 Eurasian and YRI-like ancestry combined was masked. This approach produces estimates of HG like ancestry among test populations, assuming two way admixture between a YRI-like and HG-like population. We, therefore, corrected the final proportion by accounting for the proportion of non-SSA ancestry in each population as follows:

$$Proportion\ of\ HG\ ancestry = \left(1 - \frac{f4(Chimp, Jola_{masked}; Ju|'hoan\ \ North_{masked}, X_{masked})}{f4(Chimp, Jola_{masked}; Ju|'hoan\ \ North_{masked}, YRI_{masked})}\right)(1-e)$$

The proportion of Mbuti Pygmy like ancestry was calculated similarly by replacing the Ju|'hoan North masked individuals with Mbuti Pygmy masked individuals.

We first carried out *f4* ratio testing without any correction for admixture, to assess bias due to the issues outlined above. Using uncorrected estimates with either a Ju/'hoan North or an Mbuti Pygmy reference seems to overestimate HG ancestry as compared with ADMIXTURE estimates (**SN4 Fig 2a and b**), with greater over-estimation using Mbuti Pygmy as a reference population (**SN4 Fig 2b**). We discuss the biases arising from using reference populations different from the true mixing HG population later in this section.

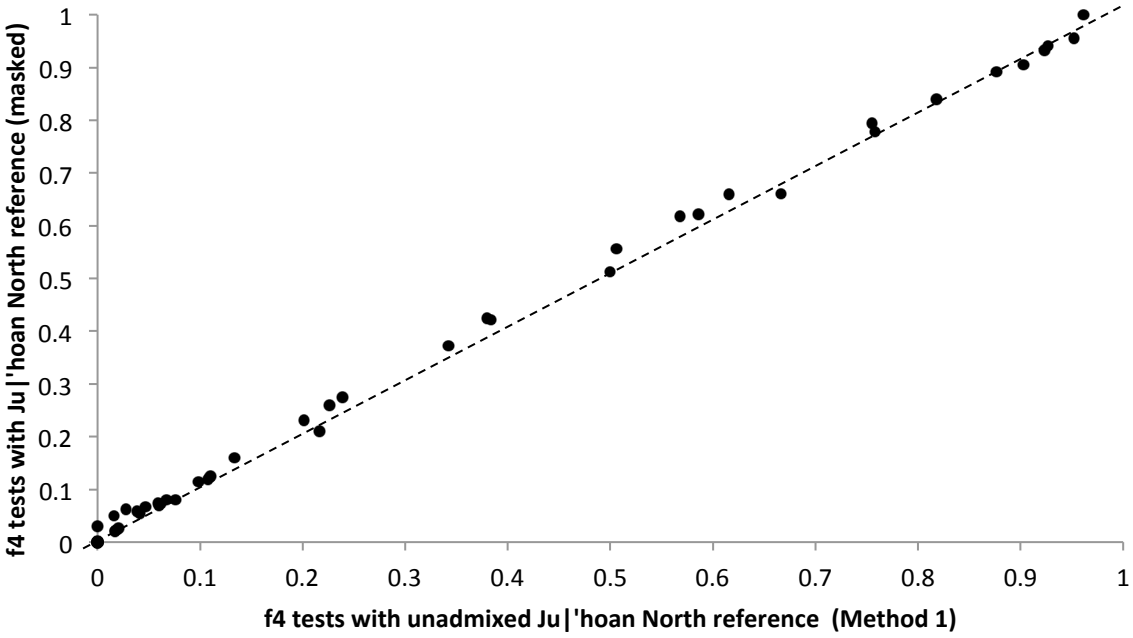
SN4 Fig 2: Comparison of uncorrected f4 ratio estimates of HG proportional ancestry with ADMIXTURE estimates.



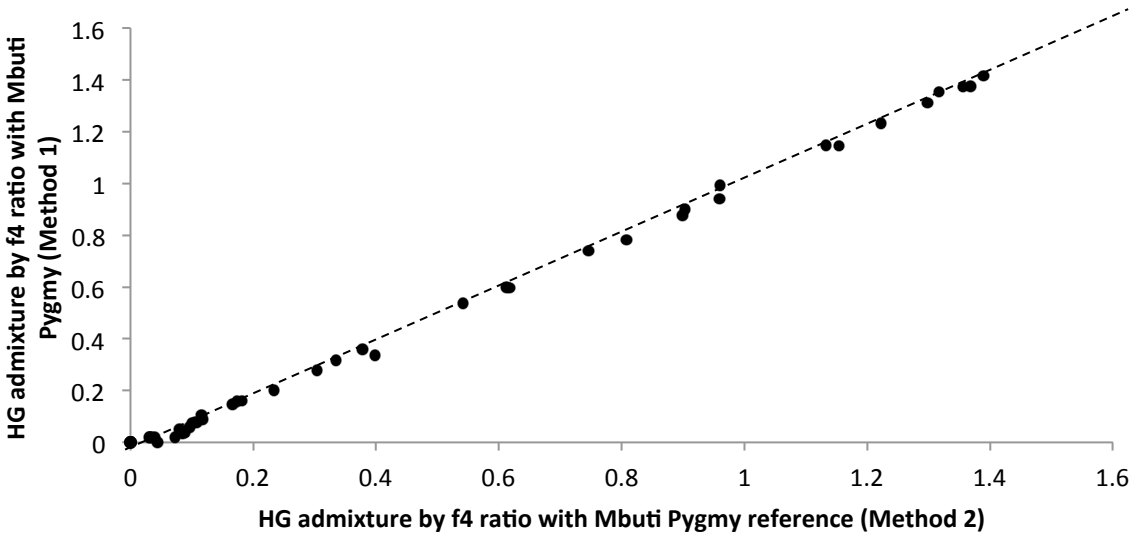
SN4 Fig 2 represents the correlation between estimates of HG ancestry among test populations using uncorrected f4 ratio estimates and ADMIXTURE analysis (k=18). SN4 Fig 2a depicts overestimation of HG ancestry using Ju/'hoan North as a reference population, when f4 statistics are left uncorrected and admixed reference populations are used in analysis. SN4 Fig 2b shows marked inflation in estimated proportions when using Mbuti Pygmy references relative to ADMIXTURE estimates. This is likely to arise from Mbuti Pygmy being a poor reference population for HG ancestry among several population groups, as we discuss later in this section.

SN4 Fig 3: Comparison of admixture estimates obtained using Method 1 and 2 with Ju|'hoan North and Mbuti Pygmy reference populations

a)



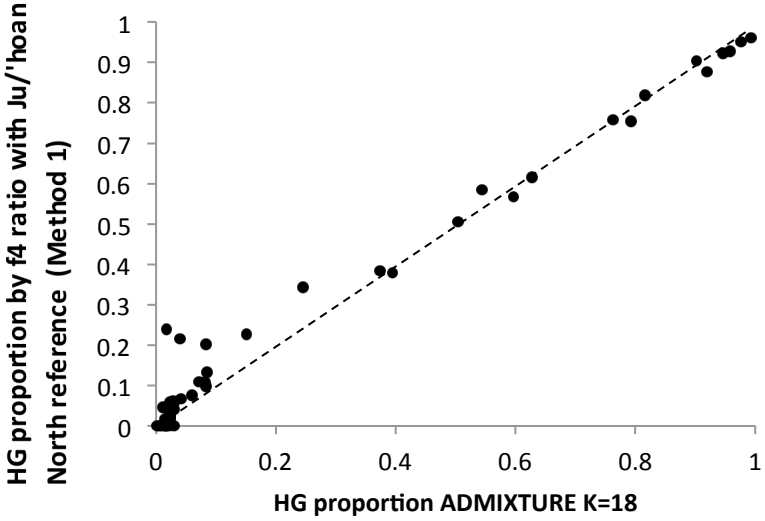
b)



SN4 Fig 3 represents the correlation between Method 1 and Method 2 for estimation of HG ancestry among test populations using different HG reference populations. The dashed line represents the line of equality. Both methods produce highly consistent estimates of HG ancestry.

Corrected estimates using Ju/'hoan North and Mbuti Pygmy as a reference seemed to be highly correlated between Method 1 and Method 2 (r2=1 for both). Both also correlated well with ADMIXTURE estimates for HG ancestry (**SN3 Fig 4**), except for South African Bantu populations such as Zulu and Sotho, where estimates from ADMIXTURE seem much lower than those estimated by the f4 ratio tests. We think this is likely to be due to inaccurate estimation of HG ancestry among Southern Bantu populations by ADMIXTURE in the cluster examined. We find marked variation in HG ancestry among different ADMIXTURE clusters, suggesting that these estimates may be unstable (**Figure 1**). Additionally, ADMIXTURE analysis in a merged dataset including global AGVP populations and Khoe-San populations from Brenna et al, produced much greater estimates of HG admixture among these populations (**Extended Data Figure 6**), further suggesting that these estimates may be unstable between different datasets, and clusters examined. We confirm this by calculating ALDER lower bound estimates for HG ancestry among Zulu and Sotho; lower bound estimates using Ju/'hoan North as a reference are 13% and 15% among Zulu and Sotho, respectively, which are both higher than the estimates produced by ADMIXTURE. This suggests that our *f4* ratio tests provide more accurate estimates of HG ancestry among these populations.

SN4 Fig 4: correlation between HG proportions using f4 ratio test with Ju|'hoan North as a reference and ADMIXTURE estimates (K=18)

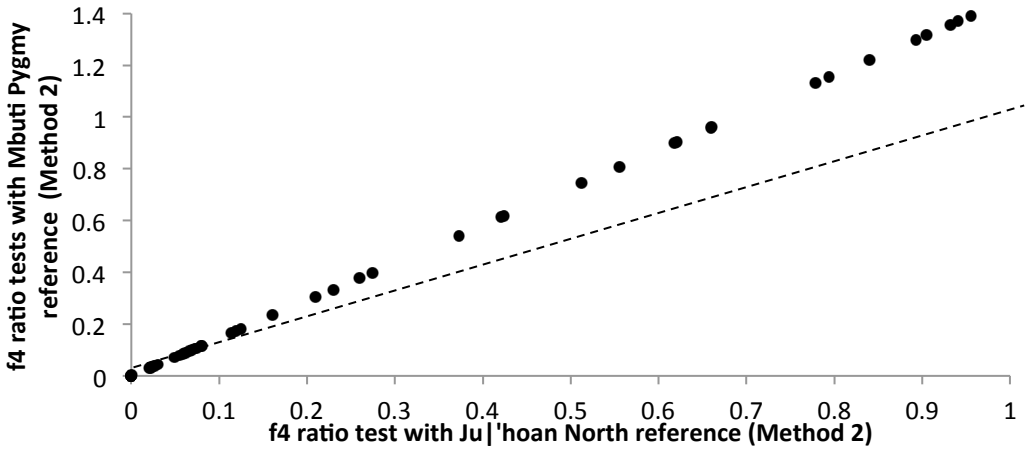


SN4 Fig 4 represents to correlation between HG proportions calculated using Method 1 with Ju/'hoan North as a reference and the proportions obtained from ADMIXTURE cluster K=18. The correlation appears to be high, except for South African Bantu populations, where the f4 ratio provides higher estimates than ADMIXTURE (discussed further in text).

As the true source of HG ancestry is unclear in many African populations, we have carried out analyses using two different reference populations- the Ju|'hoan North and Mbuti Pygmy. Using these reference populations may result in different biases depending on the nature of the true source of HG ancestry. We note that estimates of HG ancestry using Mbuti Pygmy reference seem systematically higher than with Ju|'hoan North, and in many cases estimates exceed one (**SN4 Fig 5**).

We can show that if Mbuti Pygmy represents the true source of HG ancestry but Ju/'hoan North is used as a reference population instead, the estimated admixture fraction will be underestimated by a fixed quantity that depends on the relative drift between the separation of Pygmy and YRI from the root population, compared to drift between the separation of Ju/'hoan North and YRI from the drift population (g/(g+h) in **SN4 Fig 6**). Similarly, we can show that if the true ancestral mixing population is ancestrally similar to Ju/'hoan North, and Mbuti Pygmy is used as a reference, we are likely to overestimate the proportion of admixture by the inverse of the quantity described above (**SN4 Fig 6**). Irrespective of the true reference population, a ratio of these two values obtained by using the Ju|'hoan North reference and the Mbuti Pygmy reference would provide us with the ratio of the relative drift from where Mbuti Pygmy and YRI arise from the root population, and where Ju|'hoan North and YRI arise from the root (**SN4 Fig 5**). We find this is indeed true, with this ratio being equal to 0.68 for all non-zero f4 ratio tests, suggesting that the Pygmy populations may have diverged from the root much closer to Khoe-San populations as compared to Bantu populations, as studies have previously suggested.[38] This also explains why using Mbuti Pygmy references produces values of admixture above 1, exclusively in South African populations, where the true sources of HG ancestry are likely to be Khoe-San.

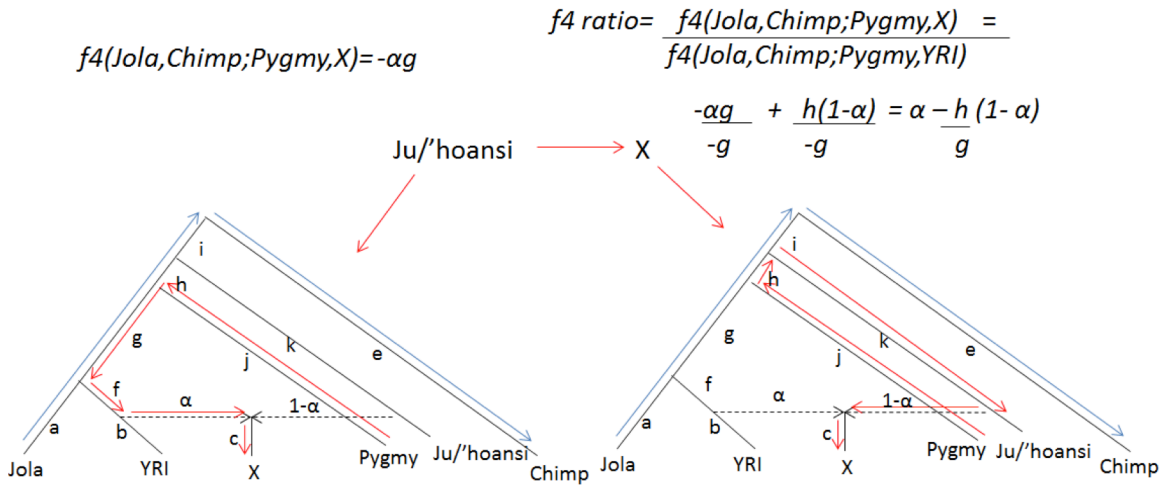
SN4 Figure 5: Comparison between estimates of HG ancestry obtained using different HG references



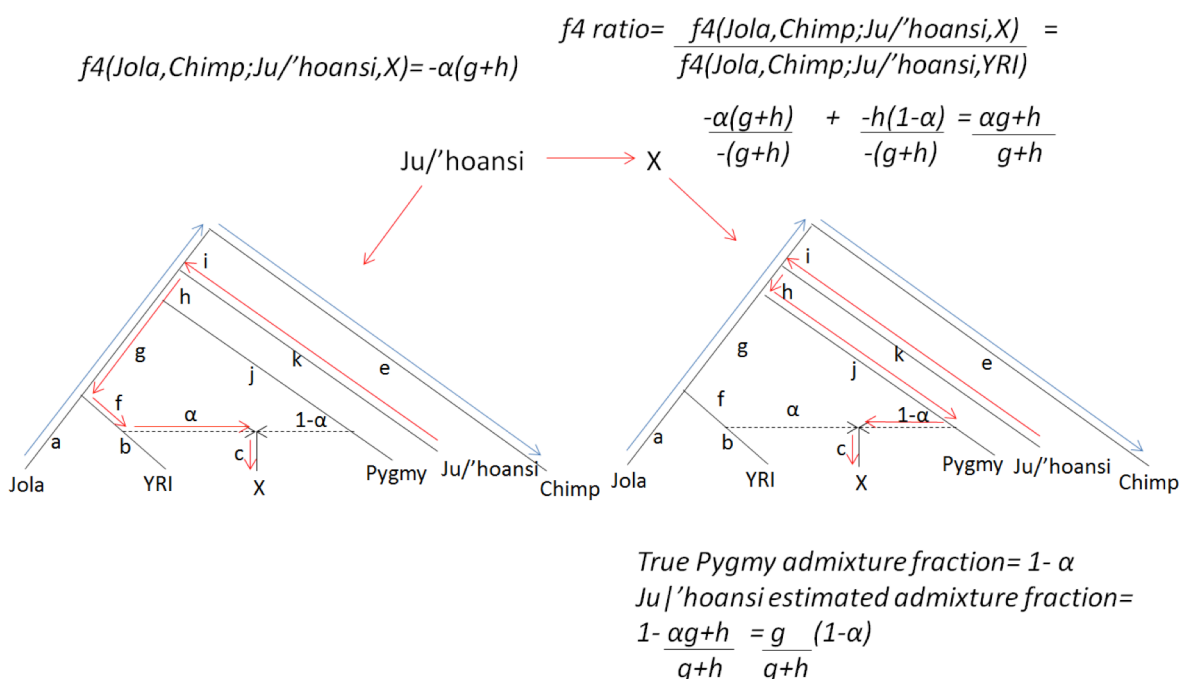
SN4 Fig 5 represents the comparison between estimates of HG ancestry among African populations when using Ju/'hoan North and Mbuti Pygmy individuals as reference populations in f4 ratio tests. We find the ratio of these is constant, and represents the demographic history of these populations (discussed further in text)

SN4 Fig 6. Biases in estimation of proportion of admixture depending on source of HG ancestry

a)



b)



SN4 Fig 6 depicts the statistical bias in estimation of HG ancestry using the f4 ratio test (Jola, Chimp; HG, X)/f4(Jola, Chimp; HG, YRI), by misclassifying the source of HG ancestry. SN4 Fig6a represent the bias in estimation when the true source of ancestry is Pygmy and Ju/'hoan North is used as the reference population. We find that true Pygmy ancestry is underestimated by the quantity g/g+h, where g and h represent drift parameters for the topography specified. SN4 Fig 6b shows the bias in HG ancestry estimates when the true mixing population is Khoe-San, and Mbuti Pygmy is specified as a reference population. Here, the fraction of ancestry is overestimated by (g+h)/g. In any scenario, the ratio of the two ancestral proportions can provide an estimate of g/(g+h).

It must be noted that these estimates are sensitive to any violations of the simple topology presented in **SN4 Figure 6**, and admixture in the reference populations, or inaccurate masking or estimation of proportional non-SSA admixture in the test population can influence these tests. Although the two methods we have used produce highly correlated results, we think it is unlikely that 'unadmixed' individuals or segments of the genome that remain after masking are truly unadmixed. These estimates should therefore be thought of as approximations of the true underlying admixture proportions.

We present estimates of ancestry as estimated using Ju/'hoan North and Mbuti Pygmy ancestry in **Supplementary Table 5**. All negative estimates have been presented as 0% admixture, and estimates above 1 using the Mbuti Pygmy reference have been capped at 1. As the true proportion of hunter-gatherer ancestry depends on the relationship of the reference population with the true ancestral population contributing hunter-gatherer ancestry to each population, we investigate this issue in some detail in the subsequent section.

S. NOTE 5: EXPLORING ADMIXTURE AMONG AFRICAN POPULATIONS IN SSA

We find strong evidence for HG and non-SSA ancestry among several AGVP populations based on PCA analysis, ADMIXTURE analysis and *f3* tests. Although, we find convincing evidence of HG ancestry in SSA, it is unclear what the sources of this ancestry might be in different parts of Africa. Here, to provide new insights into African pre-history, we try to better characterise aspects of this admixture and answer the following questions: 1. What are the most likely sources of HG ancestry in different regions of Africa?, 2. When can this Eurasian and African HG gene flow be dated to? and 3. Is this admixture limited to single events, or multiple events in admixed populations?

In order to identify the possible sources of admixture among different populations in Africa, we applied admixture linkage disequilibrium based approaches. These approaches utilise the concept that the decay of linkage disequilibrium between markers in an admixed population is exponential when weighted by the product of the difference in allele frequencies of markers between two reference admixing populations. In this case, the amplitude of this decaying curve is proportional to the amount of reference admixture in the admixed population, and the rate of decay of the exponential curve is a function of the date of admixture. Leveraging this, we can compare different reference populations, and identify most likely mixing populations, or populations closest to the ancestral mixing population as those that produce the greatest amplitude on these tests. Additionally, this allows us to date different streams of gene flow into SSA. Given the complex history of migration into and within Africa, it is likely that multiple events of Eurasian and HG admixture have occurred in some regions. We sought to explore this by assessing if the admixture LD decay was better fit by a sum of multiple exponential curves in some populations compared to a single exponential curve. We used methods described by Pickrell et al[7] to assess this (MALDER).

Here, we discuss novel findings of extensive HG and Eurasian admixture across East, West and South Africa, with evidence for complex admixture in many regions. We find previously undetected HG admixture dating to ~300 generations ago in West Africa, as well as evidence of Eurasian admixture among YRI dating to the same period consistent with previous findings of Neanderthal ancestry among YRI.[39] We suggest the presence of HG populations predating Bantu settlements in this region. We find novel evidence for recent admixture among several other populations in West Africa, including the Fula, Ga-Adangbe and Jola. Additionally, we substantiate findings of complex Eurasian admixture dating to ~100 generations ago in east Africa. We also show the first evidence for complex HG admixture among South African Bantu

populations, which corresponds to the period of reciprocal Bantu gene flow into co-located Khoe-San populations.[7] We discuss our findings in detail subsequently.

SN5.1 Assessing the probability of admixture events using MALDER

In addition to identifying multiple admixture events and most likely source populations using MALDER, we assessed the probability of each Eurasian and HG-like admixture event by using a process similar to that described by Pickrell et al (**SN5 Table 1 and SN5 Fig 1**).[7] For populations with the highest amplitude of admixture LD produced by the combination of Eurasian and African reference populations, we compared the highest amplitude with the highest amplitude produced when both reference populations had <1% Eurasian admixture (defined based on ADMIXTURE K=18 and *f4* ratio tests). We calculated this as follows:

$$Z_{EUR} = \frac{Amp_{max} - Amp_{max_{EUR}<1\%}}{\sqrt{SE_{max}^2 + SE_{maxEUR<1\%}^2}}$$

Z_{EUR} represents the statistical difference between the highest amplitude and the highest amplitude when both populations have <1% Eurasian ancestry. Similarly, we estimated the probability of HG admixture when the highest amplitude included either a Khoe-San, Hadza or Pygmy population, as follows:

$$Z_{HG} = \frac{Amp_{max} - Amp_{max_{HG}<1\%}}{\sqrt{SE_{max}^2 + SE_{maxHG<1\%}^2}}$$

where the proportion of HG admixture was estimated from ADMIXTURE analysis as the sum of Khoe-San and Pygmy like ancestry (K=18). For admixture events that clearly included a HG population, when Z_{HG} was above 2, we also calculated the probability that the hunter-gatherer event was due to a specific population. We describe these analyses subsequently. Overall, a Z score of above 2.0 was used describe a high probability event in both cases, similar to the approach used by Pickrell et al.[7]

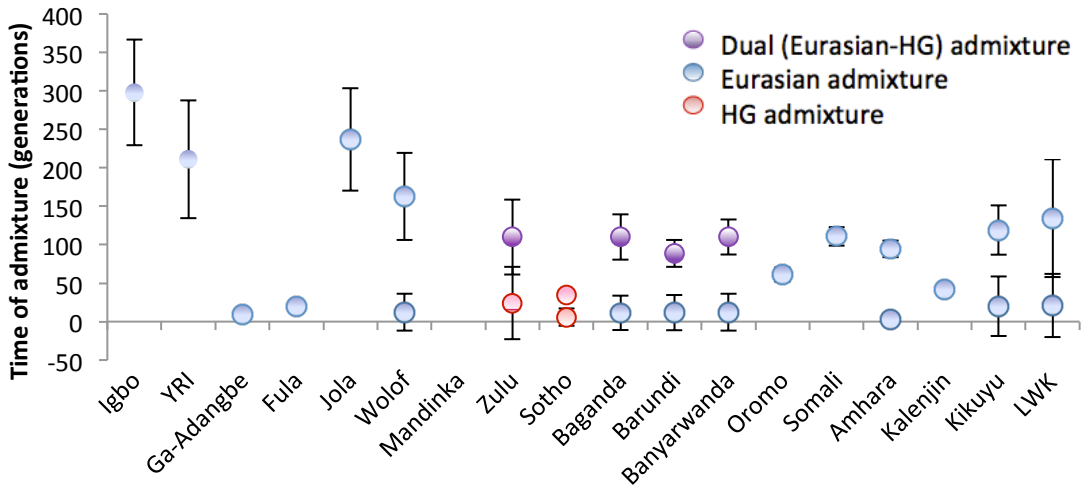
For some populations, admixture events with the highest amplitude included both a hunter-gatherer (Khoe-San, Mbuti Pygmy or Biaka Pygmy) and Eurasian population. For these we calculated the separate probability of HG and Eurasian admixture using the Z scores described above. The source of admixture was considered to be the population with a Z score of above 2. If both Z_{HG} and Z_{EUR} were >2, this was considered a dual admixture event where a HG-like

SN5 Table 1: MALDER results for AGVP populations

MALDER results	ADM1	Z	TIME1	ADM2	Z	TIME2
Igbo	Kgalagadi;Sardinian	4.27	298+/-35			
YRI	French;Himba	3.35	211+/-39			
Ga-Adangbe	Juhoan_North;Orcadian	11.7	9.2+/-1.5			
Fula	Jola;Sardinian	11	19.6+/-3			
Jola	BiakaPygmy;Sardinian	4.16	237+/-34			
Wolof	Sardinian;Wambo	5.97	163+/-29	Sardinian;Taa_East	6.04	12.2+/-2.5
Mandinka	NA					
Mandenka	BiakaPygmy;Sardinian	3.615	196+/-38	Adygei;Taa_North	7.17	8.3+/-1.9
Zulu	Juhoan_South;Orcadian	10.42	110+/-25	Juhoan_South;Yoruba	9.54	24+/-2.1
Sotho	Druze;Juhoan_South	37.86	35+/-2.3	Basque;Taa_West	5.13	5.8+/-1
Baganda	MbutiPygmy;Sardinian	11.94	110+/-15.1	Sardinian;Wambo	7.35	11+/-1.7
Barundi	MbutiPygmy;Sardinian	16.28	89+/-9	Sardinian;Wambo	0.01	12+/-1.3
Banyarwanda	MbutiPygmy;Sardinian	20.63	110+/-12	Sardinian;Wambo	8.52	12+/-1.5
Oromo	Sardinian;Taa_West	16.9	61+/-4.7			
Somali	Dinka;Sardinian	17.33	111+/-6.2			
Amhara	Dinka;Sardinian	19.09	95+/-5.5	MbutiPygmy;Tuscan	7.07	2.9+/-0.86
Kalenjin	Sardinian;Wambo	24.32	42+/-2.5			
Kikuyu	MbutiPygmy;Sardinian	14.29	119+/-16	Sardinian;Wambo	7.39	19.8+/-2.2
LWK	MbutiPygmy;Sardinian	6.04	134+/-39	BantuSouthAfrica; Sardinian	4.32	20.9+/-3.6

SN5 Table 1 depicts the results for MALDER analysis of AGVP populations using different reference populations. ADM1 represents the best representative populations of the first admixture event with the Z score indicating the statistical deviation of the amplitude from zero. ADM2 represents the same for the second admixture event, among populations where two events were found to fit better than a single event.

SN5 Figure 1: Malder results for AGVP population using genotype array data



SN5 Fig 1 represents the time and sources of admixture with confidence intervals for different AGVP populations. Circular markers with a line drawn around them represent high probability events, while those with no line around them represent low probability events, as described in the text.

ancestral population or an ancestral population with HG ancestry mixed with a Eurasian population or a population with Eurasian ancestry.

SN5.2 ADMIXTURE IN WEST AFRICA

PCA, ADMIXTURE analysis, *f3* and admixture linkage disequilibrium based tests all suggest extensive Eurasian and HG gene flow in several populations in West Africa. This ancient Eurasian and HG admixture among West African populations may have critical relevance as it provides new insights into the prehistory of Africa prior to the Bantu expansion. We therefore examined 1. Ancient HG admixture in Igbo; 2, Ancient Eurasian admixture in YRI; and 3. Recent complex admixture among populations in the Gambia.

SN5.2.1 Ancient HG admixture in Igbo

Our analyses confirm HG ancestry among Igbo on *f3* tests (**Supplementary Tables 2 and 3**), with consistently negative *f3* scores, even when using very different reference HG populations. This finding is of historical relevance as it may provide important insights into the pre-history of Africa in this region. In order to examine possible sources of ancestry in Igbo, we first examined the relative amplitudes carrying out ALDER analysis with the one-reference test using different population sets (**SN5 Fig 2**). Although, we found that amplitudes were highest for Eurasian population references followed by Khoe-San populations, the standard errors around the amplitudes were large with marked overlap between different populations, making it difficult to unravel the sources of ancestry among Igbo (**SN5 Fig 2**). We carried out testing with multiple reference populations to assess different possible combinations of admixing populations using MALDER, as outlined previously[7]. We did not see clear evidence for HG admixture among Igbo on MALDER analysis, in contrast with our results from *f3* tests. We hypothesised that this may be due to lack of resolution due to the low number of markers examined. The highest amplitude for LD decay was produced by a combination of Kgalagadi and Sardinian reference populations, dating to ~300 generations ago. Kgalagadi is a south African bantu population with ~34% HG admixture; this finding may reflect several events, including admixing of ancestral Eurasian, Bantu and/or HG populations. However, the Eurasian admixture event was low confidence (Z_{EUR}<2).

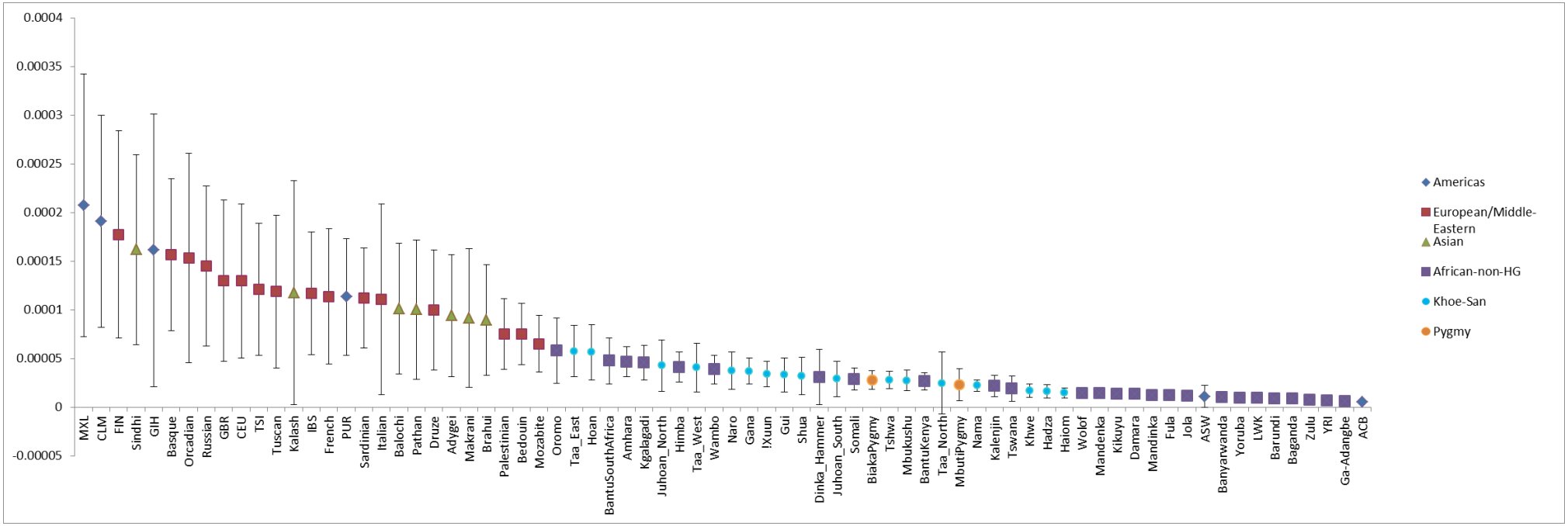
Given the ambiguity in ancestral mixing populations in Igbo, and to improve resolution to assess admixture, we carried out imputation into each dataset (the global AGVP dataset and the human origins array data) separately and then combined high confidence sites obtained from both.

Imputation was carried out into the global AGV dataset (~1.4M markers) and into the Human origins dataset (~600K markers), using a merged reference panel including the 1000 Genomes phase I version 3 integrated sequence data, and low coverage sequencing data from 3 AGVP populations, as has been described in **Supplementary Methods section 5.3** and **Supplementary Note 11)**. In order to avoid bias due to poor imputation, we only included markers that were imputed with an info score of >0.90, suggesting very high certainty in genotypes. We examined 4.7M variants of high quality combined across all populations to assess sources and dating of admixture in SSA. As our reference panel for imputation did not include any HG populations, we would expect the approach we have used with the imputed data to bias our results against finding HG admixture events, rather than towards, as it is more likely that poorly imputed HG haplotypes would drop out from analysis given the stringent filters applied.

Using this approach, we noted markedly improved resolution of admixture events in Igbo (**SN5 Table 2 and Figure 2**), with clear evidence of hunter-gatherer admixture ~300 generations ago, and ~150 generations ago, consistent with our *f3* tests. The greatest amplitude was observed for admixture between French and Ju|'hoan North for the event ~300 generations ago, suggesting possible admixture from an ancestral population most similar to current Khoe-San populations in South Africa (**SN5 Table 2**).

To explore the source of this HG admixture in more detail, we assessed if the amplitude for admixture with Ju|'hoan North was statistically different from the highest amplitude arising from Mbuti Pygmy and Biaka Pygmy as one of the reference populations. We found the amplitude to be statistically significantly different among these groups (Z=3.57 and 3.34, respectively). This analysis suggests that modern day Khoe-San populations more closely represent HG ancestry in Igbo than other modern rainforest HG populations, including Mbuti and Biaka Pygmies in our data. Collectively, ADMIXTURE analysis, *f3* tests, *f4* ratio tests with sensitivity analyses and admixture LD based approaches all support the presence of HG admixture in Igbo.

SN5 Fig2: relative amplitude of exponential curve for Igbo using different reference populations in a one population reference test



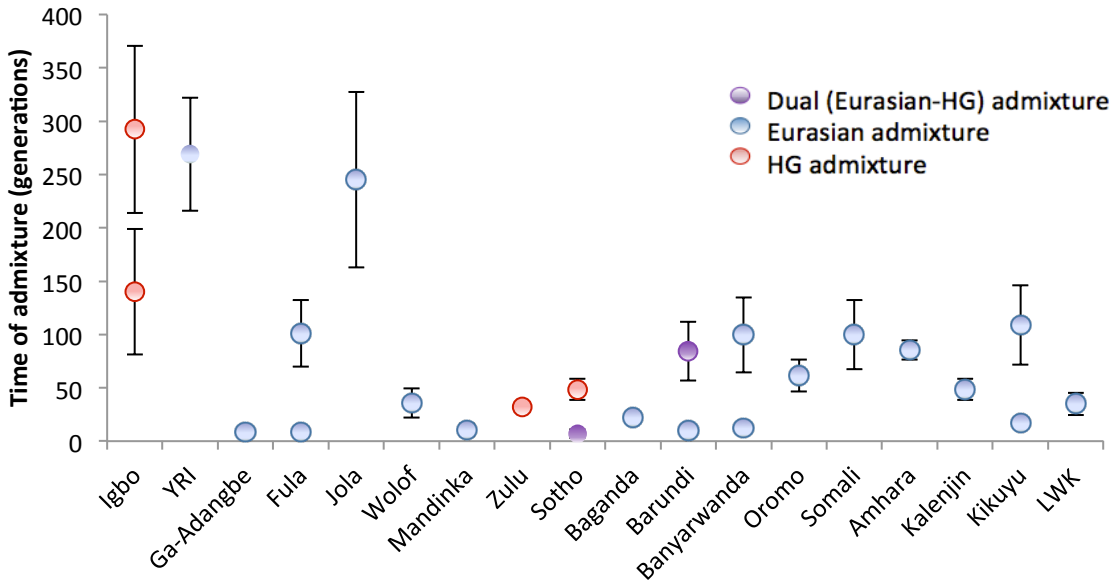
SN5 Fig2 shows the relative amplitudes for admixture in Igbo, when using different reference populations. The error bars represent a single standard error, showing that the amplitude is not significantly different from 0 for many reference populations, with the error bars being wide, making it difficult to identify the best source of ancestry in this population.

SN5 Table 2: MALDER analyses of imputed data

population	ADM1	TIME1	Z1	ADM2	TIME2	Z2
Igbo	French;Juhoan_North	292+/-40	7.33	MbutiPygmy;Pathan	140 +/-30	4.69
YRI	Mbukushu;Oroqen	269+/-27	9.87			
Ga-Adangbe	Orcadian;Tswana	8.5+/-1	7.19			
Baganda	Sardinian;Wambo	22/-2	9.56			
Banyarwanda	Basque;Wambo	100 +/- 18	5.42	Sardinian;Wambo	12+/2	5.6
Barundi	Basque;MbutiPygmy	84+/-14	5.87	Sardinian;Wambo	10+/-2.5	4.01
LWK	Kgalagadi;Sardinian	35+/-5.3	6.57			
Oromo	Basque;MbutiPygmy	62+/-7.6	8.04			
Somali	Dinka_Hammer;Sardinian	100+/-17	6.09			
Amhara	Dinka_Hammer;Sardinian	85+/-4.7	18.24			
Kikuyu	Sardinian;Wambo	109+/-19	5.64	Sardinian;Wambo	17+/-2.3	7.51
Kalenjin	Sardinian;Tswana	48+/-5	9.72			
Sotho	Juhoan_North;Yoruba	37+/-4.	9.24	Basque;Juhoan_North	6.9+/-2	3.37
Jola	Basque;Wambo	245+/-41	5.88			
Fula	Basque;Jola	101+/-16	6.26	Sardinian;Wambo	8.7+/-1.9	4.53
Wolof	French;Kgalagadi	36+/-7.3	4.91			
Mandinka	BiakaPygmy;Sardinian	10.5+/-3.1	3.37			
MbutiPygmy	Juhoan_North;Yoruba	27+/-6.8	3.96			

SN5 Table 2 depicts the results for MALDER analysis of AGVP populations using imputed data to improve statistical resolution. ADM1 represents the best representative populations of the first admixture event with the Z score indicating the statistical deviation of the amplitude from zero. ADM2 represents the same for the second admixture event, among populations where two events were found to fit better than a single event.

SN5 Fig 3: complex HG and Eurasian admixture in SSA on MALDER analysis of imputed data



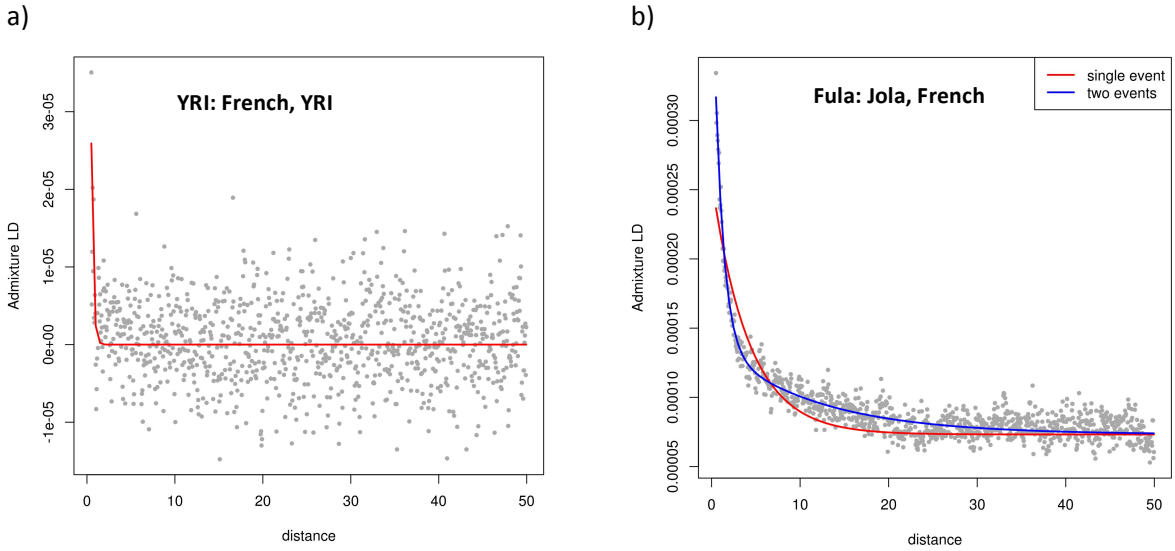
SN5 Fig 3 represents the time and sources of admixture with confidence intervals for different AGVP populations. Circular markers with a line drawn around them represent high probability events, while those with no line around them represent low probability events, as described in the text.

Although our analyses suggest that a modern day Khoe-San population may be the closest representative to the admixing HG ancestry in Igbo, there is no historical evidence that a Khoe-San like population existed in this region within the last 10,000 years. Moreover, there is no evidence of click consonants in languages in this region. However, the historical evidence is incomplete. Archaeological evidence from burials at Gobero in South Western Chad suggest the presence of distinct HG populations in this region during the mid-Holocene period when conditions were humid and foraging conditions were optimal for survival.[40] These graves suggest the presence of an HG population with distinct morphological features, including tall stature,[40] not typical of present Pygmy populations in central Africa. Archaeological findings of distinct rock art are also consistent with the presence of such populations in central and Western Africa prior to bantu settlements. This archaeological evidence dates to the Neolithic period. Collectively, this genetic and archaeological evidence suggests that HG populations may have inhabited Western parts of Africa as well as Eastern and Southern Africa 10,000 years ago. Our findings provide new insights into the ancient HG admixture in Africa. A critical next step would be to identify and deep sequence multiple populations across the region to help refine the source and complexity of this admixture. Additionally, given the extensive migration and replacement of ancient populations across Africa, it is likely that present populations do not represent ancient hunting and foraging people inhabiting these parts of Africa. This highlights the need to explore archaeological sites in these regions, as the study of ancient DNA could provide important insights into ancient populations located in this region during different periods, as has been discussed previously[41].

SN5.2.2 Ancient Eurasian admixture in YRI

In systematic analyses using *f3* tests, we did not find evidence of Eurasian admixture in YRI, which is also consistent with our *f4* ratio tests for Eurasian ancestry in YRI. However, linkage disequilibrium tests, suggested the presence of ancient Eurasian admixture in West Africa, specific to this ethno-linguistic group. Importantly, this finding of ancient admixture in YRI populations is consistent with findings suggestive of Neanderthal ancestry in YRI as a result of Eurasian admixture in YRI from the back to Africa migration.[39] The lack of negativity on *f3* tests may be due to the very low levels of Eurasian ancestry in this population in addition to the ancient nature of admixture resulting in prominent drift since the admixture event, rendering the *f3* statistic positive. *f3* and linkage disequilibrium based tests have been shown to be distinct but complementary tests, and may be differently powered to detect specific admixture events.[16] Specifically, we found evidence of Eurasian gene flow in YRI dating to ~300 generations

SN5 Fig 4: MALDER analysis of admixture events in West Africa



SN5 Fig 4 depicts the decay of admixture LD in YRI and Fula with French and Bantu African reference populations. The decay of admixture LD in YRI is consistent with very old admixture, while in Fula, this seems relatively recent. In Fula, the fit of a sum of exponential curves representing two discrete events appears to be better than that modelling a single event.

ago, although this was low confidence (Z_{EUR}<2) in both our genotype data and imputed analysis. We believe these findings may well reflect relatively old Eurasian ancestry in West Africa dating to between 6500-10,000 years ago, consistent with the non-zero Neanderthal ancestry in YRI reported recently that dated to a similar time period.[39] This admixture predates the Bantu expansion, and may have possibly occurred during the period the Sahara desert became lush,[42] allowing Eurasian migration across this region. It is interesting to note that we do not see evidence for such an event in other Bantu populations in East and South Africa, suggesting that this may not have been the origin of the Bantu expansion, and Eurasian admixture from this

period may have been limited to this region. Alternatively, it is also possible that we may not have had the power to detect very small quantities of relatively old admixture in other population groups, where multiple admixture events with large proportions of Eurasian ancestry have occurred. One important future question will be to conduct much deeper sampling and sequencing of populations across the region to provide more reliable estimates of admixture and dating to help reconstruct the bantu expansion, and the structure of African populations predating this, as discussed in SN 5.2.1.

SN5.2.3 Recent Eurasian admixture in West Africa

We also see more recent evidence of admixture among the Ghanian (Ga-Adangbe) and the Gambian populations (Fula, Jola, Wolof and Mandinka) (**SN5 Tables 2 and 3, and SN5 Figures 3 and 4**). This is consistent with the large variation in proportion of individual non-SSA ancestry observed among these populations on ADMIXTURE. These findings are also substantiated by highly negative *f3* statistics for Fula and Wolof. We also found evidence for complex Eurasian admixture in Fula, with an older event ~100 generations ago, and a more recent event ~10 generations ago, consistent with the nomadic nature of these populations and extensive migration of these groups during this period. Modelling two admixture events among these populations substantially improved the fit of the model to data points as shown in **SN4 Fig 4b**.

SN5.2.4 Summary of findings

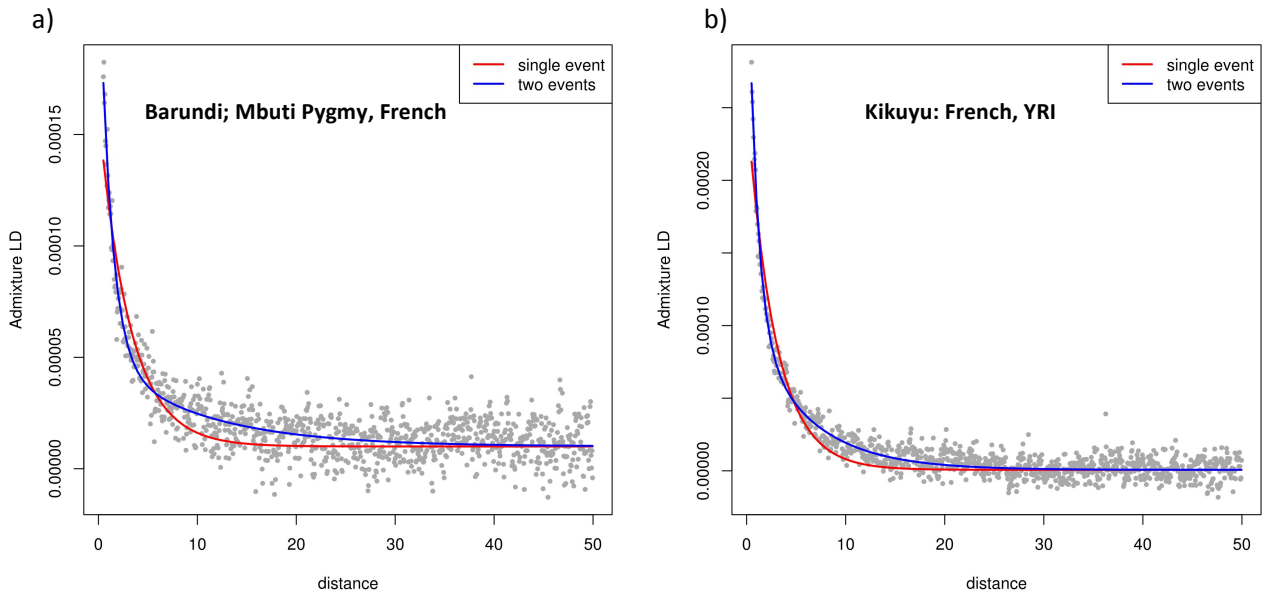
In summary, we provide novel insights into African pre-history and admixture. Specifically, we identify a signal of ancient HG admixture in W. Africa dating to ~300 generations ago, and hypothesise that this signal may arise from distinct HG populations that existed in this region during the early Holocene period. We also report ancient Eurasian admixture localised to YRI in West Africa, corroborating evidence for ancient admixture identified previously through signals of Neanderthal ancestry in YRI. Additionally, we show novel evidence for complex recent Eurasian admixture in Ghanian and Gambian populations. Although we find convincing evidence for ancient HG admixture in West Africa, we cannot reliably identify the most likely modern day representatives of the ancestral populations. This vital question will require further exploration in more diverse population sets, and with exploration of archaeological sites for ancient DNA.

SN5.3 Eurasian and HG Admixture in East Africa

Collectively, on PCA, ADMIXTURE, *f3* tests and MALDER, we find strong evidence for extensive and complex Eurasian and HG admixture in East Africa (**Supplementary Table 2, SN5 Table 2 and SN5 Fig 3**). Our findings of Eurasian ancestry in this region are very consistent with those of Pickrell et al., and suggest that several waves of gene flow may have occurred, with older events ~100 generations ago, and very recent admixture in some populations 5-20 generations ago (**SN5 Fig 5**). The proportion of admixture is variable among different groups, approaching 50% in Ethiopian population groups and varies between 2-14% in other east African populations.

We also find strong novel evidence of HG admixture in East Africa, particularly in the Ugandan populations on MALDER and *f3* tests. Our analyses suggest dual admixture events that may have occurred between a West Eurasian-like population and a population with hunter-gatherer-like ancestry, most similar to present Mbuti rHG among the populations examined. Admixture appears to have occurred ~100 generations ago, which is consistent with when the Bantu expansion is likely to have spread to this region. It is likely that hunter-gatherer ancestry was assimilated during the expansion by these populations between 3000-4500 years ago around the same time that West Eurasian ancestry entered this region. Our imputation analyses additionally suggest admixture from populations similar to present day South African Bantu populations with Khoe-San populations. This is consistent with a hypothesis of extensive migration and gene flow between East and South Africa, corroborated by findings of East African Eurasian ancestry in South African Khoe-San populations previously reported by Pickrell et al. This is also consistent with the clear clines we see between Ugandan and South African Bantu population on PCA, and shared South African clusters on ADMIXTURE analysis. We do not see clear evidence for hunter-gatherer ancestry among other East African populations on *f3* tests and MALDER, suggesting that such historical admixture may be population specific and have occurred in geographically and historically associated HG and Bantu populations. While previous reports examining admixture among Bantu and Pygmy populations have focused largely on gene flow from Bantu groups into Pygmies, this is the first characterisation of possible Pygmy ancestry among Bantu populations in East Africa. While our analyses suggests that Mbuti Pygmies are the closest modern day representatives of the mixing populations in this region, it is possible that denser sampling of other HG groups in the regions, including Twa Pygmies co-located with these population groups may reveal other more proximal sources of this ancestry, suggesting the need for more diverse sampling of HG populations in East Africa.

SN5 Fig 5: MALDER analysis of admixture events in East Africa

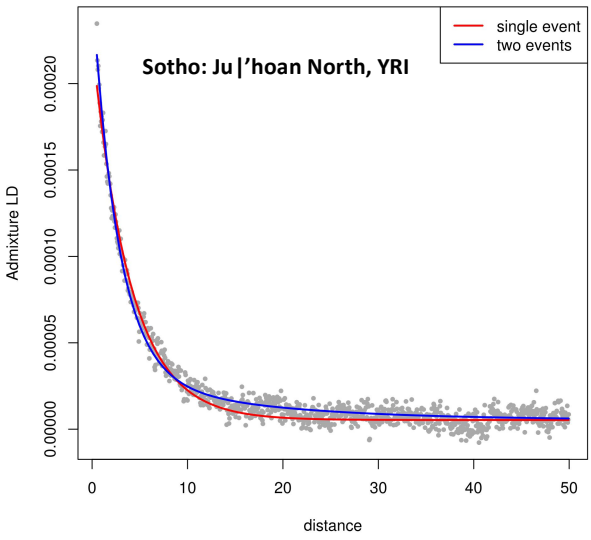


SN5 Fig 5 a) and b) represent the decay of admixture over distance for Barundi and Kikuyu, using specific references. Here we show that the decay suggests a better fit with a two admixture event model as compared to a single admixture event in both cases.

SN5.4 Hunter gatherer admixture in South Africa

Although previous studies have reported the presence of HG admixture among South African Bantu populations, the nature and extent of this admixture has not been fully characterised. Here, we show strong evidence of hunter-gatherer admixture in South Africa (*f3* tests, *f4* ratio tests and MALDER analyses), most likely from Khoe-San populations that are geographically co-located. This is consistent with the assimilation of click consonants into the Zulu language. We also see evidence suggestive of complex admixture among Sotho (**SN5 Fig 6**), which is consistent with these populations groups being geographically proximal to present

SN5 Fig 6: MALDER analysis of admixture events in South Africa



SN5 Fig 6 represents the decay of admixture LD by distance for Sotho using Khoe-San and YRI references, showing a better fit for a two-event admixture model

Khoe-San populations in the region. Admixture appears to be relatively recent, having occurred over the past 1500 years, which is consis10tent with the dating of reciprocal bantu admixture among the Khoe-San popualtions, suggesting bilateral gene flow between these groups over this period. Although modelling two events appears to improve the fit of the exponential curve, it is possible that there has been continuous gene flow, or gene flow at multiple time points. (**SN5 Fig 6**) These findings are important, as they suggest that extensive tracts of Khoe-San ancestry occur among South African haplotypes. This has relevance to the efficiency of large scale genetic studies among these populations, and suggests the need for representing these HG haplotypes in current reference panel in order to improve imputation accuracy among these populations. We discuss this further in **Supplementary Notes 10 and 11**.

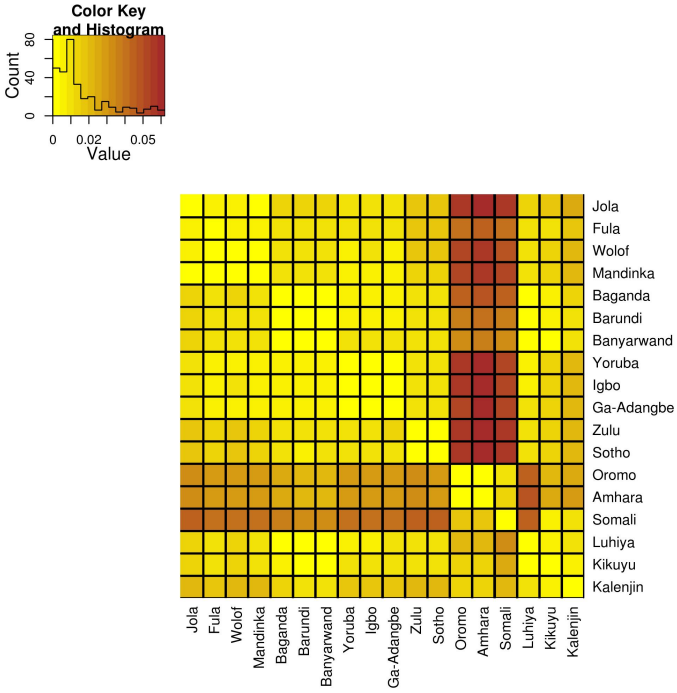
S. NOTE 6: EVALUATION OF MASKING OF EURASIAN ANCESTRY IN PCADMIX

We provide the first comprehensive assessment of selection and differentiation in SSA in the presence of complex admixture using the AGVP panel. For our selection analyses, and specific admixture analyses, in order to reduce bias, we have carried out masking of regions of the genome of Eurasian ancestry in several African populations.

Masking Eurasian ancestry markedly reduced population differentiation among populations as measured by a decline in mean pairwise F_{ST} from 0.021 to 0.015 (**SN6 Figure 1**), and the proportion of variation captured by PC 1 from 21% to 8%, with much of the decrease in F_{ST} attributable to reduced differentiation between Afro-Asiatic and other populations; this suggests that Eurasian ancestry has a substantial impact on differentiation among populations across SSA and should be considered in examining population structure.

For our selection analysis, on examining the most differentiated sites across the genome among African populations, we hypothesised that Eurasian admixture could lead to spurious differentiation between populations at sites, where differentiation had not been spurred by positive or negative selective forces. Of course, positive or negative selection also can lead to increased or reduced probability of specific local ancestry at loci- however, we do not consider such type of selection in these analyses, as we lack the resolution to examine this, particularly in populations with low levels of non-SSA ancestry.

SN6 Figure 1: F_{ST} differentiation among populations before and after masking

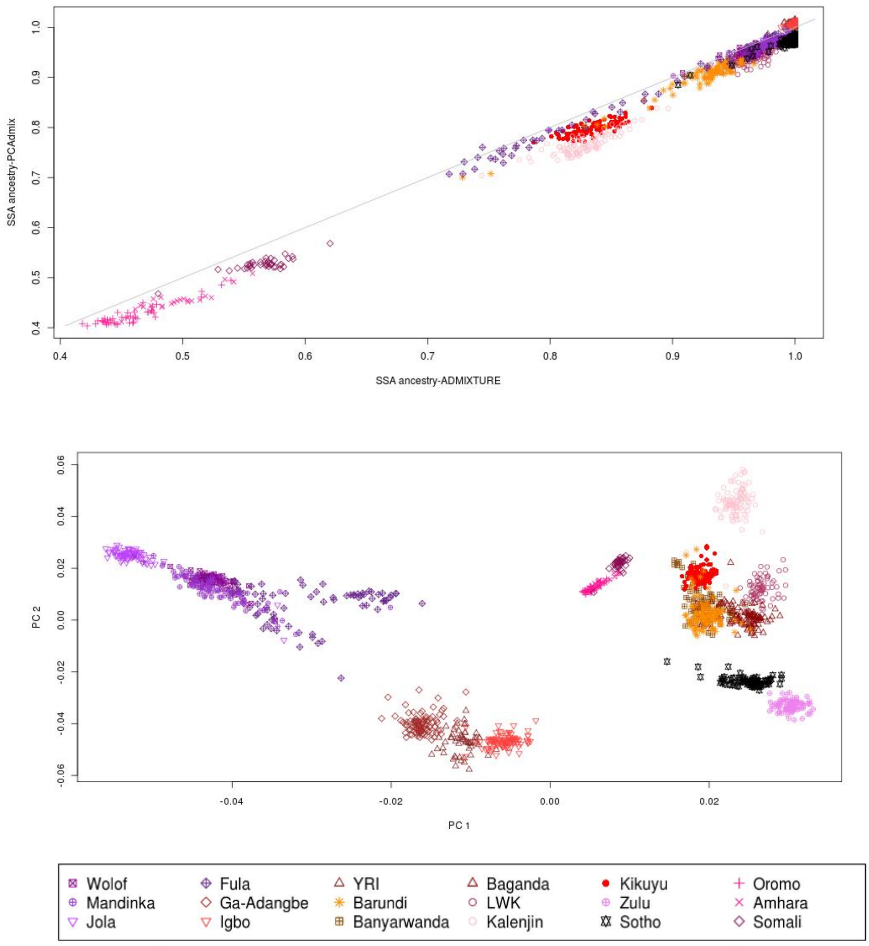


SN6 Figure 1 represents a heat map of the F_{ST} differentiation among AGVP populations before and after masking of Eurasian ancestry among these. The greatest differentiation is observed between Ethiopian (Oromo, Amhara and Somali) and other populations groups (top right triangle). This differentitation is noted to be substantially diminished after masking of Eurasian ancestral segments among these populations (lower left triangle), suggesting that a large proportion of differentiation among these populations can be explained by variable Eurasian admixture.

Although most algorithms for ancestry deconvolution are accurate for populations such as the African-Americans, where most of the admixture is thought to have occurred in the recent past, ancestry assignment can be less accurate when admixture has occurred >50 generations in the past. In order to

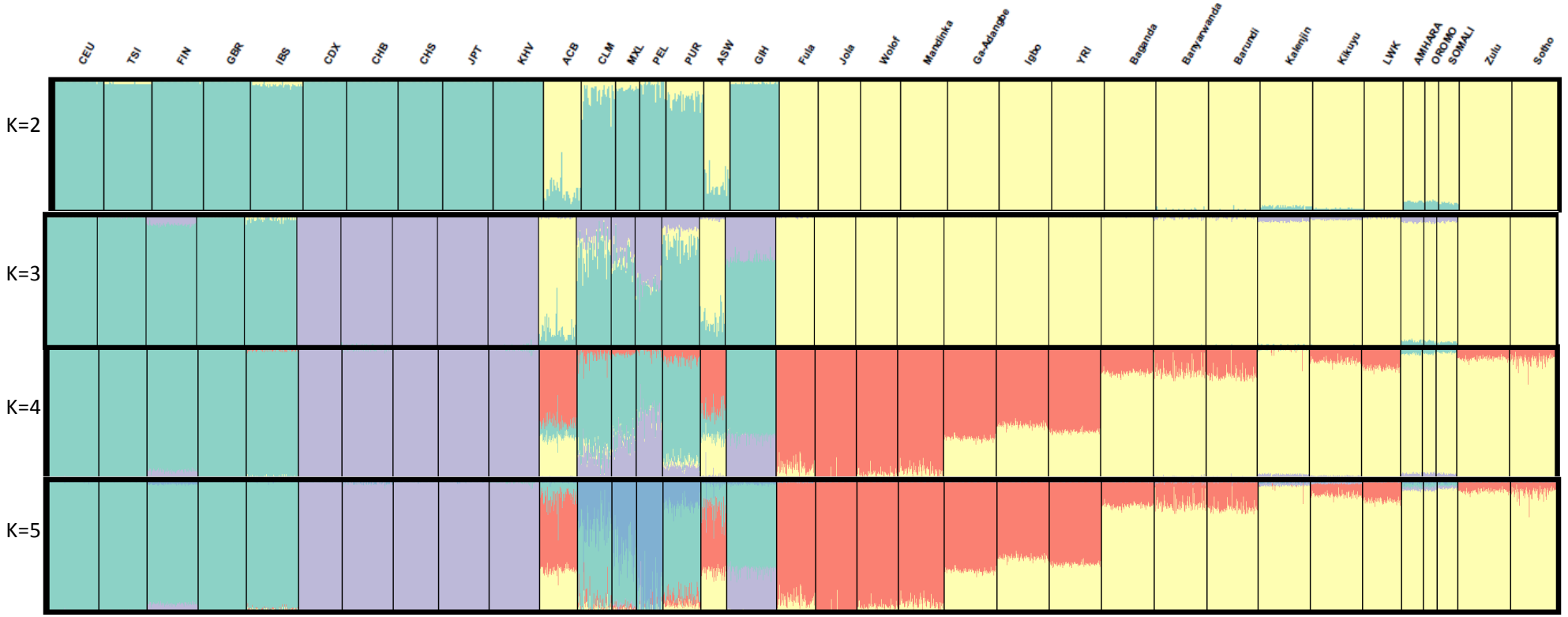
assess accuracy of masking, we first examined consistency in proportional ancestry identified using PCAdmix and ADMIXTURE. We found high consistency between proportion of SSA ancestry using ADMIXTURE (K=3 clusters) compared with PCAdmix estimates (**SN6 Fig 2**). In order to confirm that we had successfully masked Eurasian ancestry among AGV populations, we repeated PCA and ADMIXTURE clustering analyses on the non-SSA ancestry masked data (**SN6 Figs 2 and 3**). We find almost no evidence of residual Eurasian ancestry in these analyses, suggesting masking is likely to be reasonably accurate.

SN6 Fig 2. Correlation between proportional individual SSA ancestry assigned by PCAdmix and ADMIXTURE analysis and PCA among African populations after Eurasian ancestry masking



SN6 Fig 2a shows the correlation between SSA ancestry as identified by YRI-like ancestry in PCAdmix, and ADMIXTURE (K=3). A strong correlation is seen for individual proportional ancestry for all populations. SN6 Fig 2b represents PCs 1 and 2 for the African dataset after masking of Eurasian ancestry using PCAdmix. The clines seen in most African populations including Ethiopians are not observed, suggesting that Eurasian ancestry has been appropriately masked.

SN6 Fig 3.: ADMIXTURE clustering analysis after masking of non-SSA ancestry among African populations



SN6 Fig 3 represents admixture clustering analysis of AGVP samples in a global context after masking for Eurasian ancestral segments in these data. ADMIXTURE clustering suggests almost complete masking of Eurasian ancestry among the African populations.

In order to formally assess the accuracy of ancestry masking for older admixture, we carried out simulations of admixed populations. For this, we simulated a population of 100 haplotypes (50 diploid individuals), with varying proportions of CEU and YRI ancestry. The global population set with ~1.6M SNPs was used to generate this. YRI and CEU haplotypes were generated by phasing all related individuals in order to improve accuracy. Following this, related individuals were removed and individuals were generated with means of 40% CEU ancestry (SD=0.05), using the methods outlined previously.[43] Briefly, we first drew randomly from a beta distribution with the assigned mean and SD, and then drew random haplotypes from the two reference populations. For the first marker, the probability from the beta distribution was used to assign ancestry. Following this, the probability of recombination between two markers was estimated as 1-e-λg, where λ was the time of point admixture in generation time, and g was the morgan distance between the two markers. Based on this probability, whenever recombination occurred, we chose ancestry based on the probability of each ancestral type. Following this, we ran PCAdmix using default parameters, using different reference populations, as shown in **SN6 Table 1**. We used lambda (50, 100 and 300) values to assess the accuracy of PCAdmix over different times of admixture. As it is unrealistic that CEU and YRI would necessarily be accurate representatives of the true ancestral population being examined, we also examined relatively divergent reference populations, to see if PCAdmix was sensitive to misclassification of ancestral populations. To this extent, we used FIN and GIH in two separate analyses as surrogates for CEU (F_{ST} 0.006 and 0.03 with CEU, respectively) and Zulu as a surrogate for YRI (F_{ST}=0.01), and vice versa. For stringency, we only considered sites with >0.90 posterior probability of YRI-like ancestry as having African ancestry. The chosen procedure is likely to elevate sites that are true African ancestry being assigned non-African and being masked, yet providing conservative results in analyses.

We find that masking produces relatively accurate results, and predictably, the greatest proportion of misclassification of Eurasian ancestry classified as African (6%) in the scenario where there is 40% of non-SSA admixture dating 300 generations ago. Changing the window size did not alter results. In most cases, however, misclassification of non-African ancestry was below 5%. In fact, accuracy was very high (95%) for admixture up to 50 generations ago. While using surrogate populations that were relatively diverged from the true reference population affected accuracy, this difference was modest.

Given that most populations have non-SSA ancestry that dates only to approximately 100 generations ago, this masking should give us reasonably good accuracy. Also, populations with older admixture such as YRI, have <1% Eurasian ancestry; therefore, inaccuracy in masking in

unlikely to produce large biases, even if 33% of this ancestry is left unmasked, as suggested by our simulations.

Although our primary concern is accurate masking of non-SSA ancestry, over-masking of African ancestry can also lead to bias. As our masking is based on probability of African ancestry in a region, we are more likely to mask regions that are poorly differentiated between African and non-African populations, as these are unlikely to produce clear delineation of probabilities, even when these regions are of African ancestry. This would lead to over-representation of regions that are differentiated between African and non-African populations, and under-representation of regions of true African ancestry that are poorly differentiated from European populations. This can potentially bias all F_{ST}s calculated using masked data upwards when examining differentiation between African and non-African. While this is a likely bias, sites that lie in the extreme 0.1% distribution of these F_{ST}s are still likely to represent the most differentiated sites between these population groups and will be unaffected by this bias. However, effects of such over-masking when studying differentiation among African populations can potentially be more complex. The extent of overmasking appears to depend on the true level of non-SSA admixture in populations, as well as the dating of admixture, suggesting that such effects may be differential among different African populations, leading to biased results. We hypothesise that in spite of this bias, masking would overall produce estimates closer to the estimates for the level of differentiation between ancestrally pure populations, compared to not masking at all. We further explore this in a simulation. We simulate two populations with 50 individuals (100 haplotypes) on chromosome 10 each derived from the Zulu and Jola. Zulu and Jola are considered the original ancestrally pure individuals. We simulate differentially admixed populations from these: an admixed population with admixture between Zulu (60%) and CEU (40%) 100 generations ago, and a population with admixture between Jola (95%) and CEU (5 %) 300 generations ago. We examine the locuswise differentiation between these populations. We carry out masking of non-SSA admixture among these two simulated populations and compare the results from calculation of locuswise F_{ST} with those from unmasked admixed population sets. To account for the differentiation between populations used for masking, and true ancestral mixing populations, we used FIN and YRI populations to represent the mixing populations. We find that results from masked analyses are a much better approximation of differentiation between the true ancestral populations compared to analyses of admixed populations that do not take non-SSA admixture into account at all. The overall F_{ST} across the region from masked analysis (0.014) was closer to the true estimate (0.012) than from unmasked analysis (0.015), suggesting that although masking might not provide completely unbiased results, these are closer to true estimates than not considering

admixture at all. The correlation between 'true' locuswise F_{ST}s and those from the masked data were much higher than the correlation with unmasked simulated data (r2=0.57 and 0.34 respectively). Among the top 1% of F_{ST} distribution, the masked F_{ST} algorithm captured 31.3% of the F_{ST}s in the tail of the top 1% of the true distribution based on the unadmixed populations, as compared to only 14.7% captured by the F_{ST}s calculated without any masking. This suggests that although masking does not allow us to completely remove bias and assess the true distribution of F_{ST}s between unadmixed populations, this is more accurate than not taking admixture into account at all. This would also argue for the development of better methods for masking, to be able to identify sites differentiation truly due to positive/negative selection rather than differential admixture alone. Our simulations would suggest that previous selection scans based on differentiation in Africa would need to consider biases arising from the presence of admixture.

SN6 Table 1: Accuracy of masking in simulated admixed African populations

True ancestral	Input reference	Window size	Proportion (SD) of European admixture simulated	Mean proportion actually in simulated samples	Time of admixture (generations)	Correlation between simulated proportional ancestry and PCAdmix ancestry	Accuracy (sites matching true assignment)	Non-African ancestry misclassified as African (% of all sites)- False negatives	African ancestry misclassified as non-African (% of all sites)- False positives	Proportion of CEU ancestry misclassified (%)	Proportion of African ancestry misclassified
CEU, YRI	CEU, YRI, JPT+CHB	20	0.40(0.05)	0.40(0.08)	50	0.98	95%	0.8%	4.1%	2.0%	7.0%
CEU, YRI	CEU, YRI, JPT+CHB	20	0.40 (0.05)	0.41(0.07)	100	0.98	92.5%	1.6%	5.9%	4%	10%
CEU, Zulu	FIN, YRI, JPT+CHB	20	0.40(0.05)	0.41 (0.08)	100	0.97	90.7%	2.1%	7.2%	5.3%	12.4%
CEU, YRI	CEU, YRI, JPT+CHB	20	0.40(0.05)	0.41(0.06)	300	0.94	85.0%	4.8%	10.1%	12%	17%
CEU, YRI	CEU, YRI, JPT+CHB	10	0.40(0.05)	0.41(0.06)	300	0.94	84%	4.9%	10.7%	12%	18.4%
CEU, YRI	FIN, Zulu JPT+CHB	20	0.40(0.05)	0.41(0.06)	300	0.94	83.7%	5.7%	10.5%	14.3%	18.1%
CEU, YRI	GIH, Zulu JPT+CHB	20	0.40(0.05)	0.41(0.06)	300	0.93	82.7%	6.5%	10.8%	16.1%	18.5%
CEU, YRI	CEU, YRI JPT+CHB	20	0.05(0.01)	0.05 (0.02)	300	0.74	94.7%	1.6%	3.7%	33.3%	3.9%
CEU, Jola	FIN, YRI JPT+CHB	20	0.05 (0.01)	0.05 (0.02)	300	0.72	92.7%	1.5%	5.8%	31.3%	6.1%

S. NOTE 7: ASSESSMENT OF BACKGROUND SELECTION IN DIFFERENTIATED REGIONS

One of the methods we used to examine loci under selection was to study differentiation at individual loci using Wright's F_{ST}. We then examined the sites in the tail distribution of the highest 0.1% of F_{ST}s and hypothesised that these might be under selection. Differentiation can arise due to positive selection in response to selection pressures that are differential or localised to certain regions. Alternatively, background selection, leading to a reduction in effective population size, and therefore augmenting drift forces, can also lead to differentiation among sites. We would expect that if background selection were the predominant force, genomic diversity in surrounding regions would be reduced. In order to formally assess this, we examined the expected fraction of neutral diversity at each site across the genome to assess if sites with lower diversity were enriched among tail end of the F_{ST} distribution. We did this by examining the B score, a score generated by Vicker et al., which indicates the expected fraction of neutral diversity that is present at each site along the genome, with values close to 0 representing near complete removal of diversity as a result of selection and values near 1 indicating little effect. We systematically assessed differences in B scores among our 0.1% highly differentiated sites and random sites from the genome. We only included unlinked highly differentiated sites in our analysis, and random sites were selected from each segment of the genome (using 1MB segments), eliminating linkage disequilibrium.

We compared the distribution of B scores among our top hits with a set of sites selected randomly from each 1MB segment within the genome. We then generated an empirical null distribution of p values by randomly sampling one site each from 1MB segments of the genome (not including sites among the top 0.1% of the F_{ST} distribution). A number equal to the number of variants in the top 0.1% of the distribution were chosen and compared to the remaining variants in 1000 permutations and p values

SN7 Table 1: Assessment of differences in B score distribution among highly differentiated and other sites among different dataset

Comparison	Masking	MAF matching	P value	Z
AGV vs Europe	unmasked	No	0.001	3.10
	unmasked	Yes	<0.0001	3.45
Within AGV	unmasked	No	0.48	0.69
	unmasked	Yes	0.16	1.43
	masked	No	0.40	-0.80
	masked	Yes	0.80	-0.28
Malaria	unmasked	No	0.52	0.65
	unmasked	Yes	0.11	1.60
	masked	No	0.40	0.88
	masked	Yes	0.38	0.86
Lassa fever	unmasked	No	0.01	2.61
	unmasked	Yes	0.002	2.88
	masked	No	0.06	1.82
	masked	Yes	0.11	1.64

SN7 Table 1 depicts the evidence for differences in distributions of B scores between highly differentiated and remaining sites. The Z score represent the direction of difference, with positive scores suggesting higher B scores among differentiated sites.

were calculated with the Wilcoxon rank sum test. We calculated an empirical p value as the ratio of the number of p values in the null distribution below the p value obtained by comparing the top Fst distribution to the rest of the genome and the total number of p values generated. We also carried out sensitivity analyses by conditioning on minor allele frequencies, and found this made no material difference to results. We do not find any evidence for reduced genomic diversity among our differentiated sites, making it unlikely that this differentiation is due to drift arising in the vicinity of genes under negative selection. By contrast, we identify increased diversity in some of our selection scans (**SN7 Table 1**), providing strong evidence against background selection as an explanation for our selection signals.

S. NOTE 8: LONG RANGE HAPLOTYPES UNDER SELECTION IDENTIFIED BY IHS

Integrated Haplotype Score (iHS) analysis within African populations identified several signals previously reported as under selection (**SN8 Table 1**). In addition, several novel loci including the *ARPC3* and *IL18* involved in the innate immune response and cytokine signalling were identified. Other gene regions showing evidence for selection sweeps included genes involved in auditory perception and pathways (*OR5B21*, *LOXHD1*, *CNGA3* and *FOX1*), embryogenic, spermatogenic development and normal fertility (*LBH*, *DHX30*, *GGN*, *CATSPERG*, *VPS13B*, *ZEB1*, *GNRH1*, *SPAG4* and *DDX4*), neurological development and signalling (*SIM1*, *HTR7*), variation in body height (*MAP3K3*, *LBH*), response to hypoxia (*HMOX2*), and hair and periodontal tissue structure, possibly involved in dietary adaptation (the *KRT* and *KRTAP* clusters).

Among the top 0.1% of highly differentiated loci between Europe and Africa, 25 also showed evidence for selective sweeps (**SN8 Table 2**). Among the most differentiated candidates in African populations, 16 also showed evidence for selective sweeps (**SN8 Tables 3 and 4**). Such loci are likely to represent selection sweeps that have occurred independently across Africa, leading to allelic differentiation at hitchhiking loci within these long-range haplotypes. Several of these loci were noted to be important in development. The *P2RX1* gene was one such locus (**SN 8 Tables 3 and 4**); this gene belonging to the *P2X* family of G-protein-coupled receptors is thought to be essential for male reproductive function in mice in addition to being important for HIV viral entry into macrophages,[44] and thymic selection among humans,[45] suggesting a possibly vital role in regulation of the host immune response.

SN8 Table 1: Top 0.1% iHS signals within Africa

Chr	Window (bp)	p value	Genes
11	90018141-90218141	0.0000	
11	90418141-90618141	0.0000	
11	91018141-91218141	0.0000	
11	91618141-91818141	0.0000	
11	91818141-92018141	0.0000	
11	92018141-92218141	0.0000	*FAT3*
1	206120733-206320733	0.0000	*AVPR1B,FAM72A,C1orf186,CTSE*
12	108401242-108601242	0.0000	*WSCD2*
1	222920733-223120733	0.0000	*FAM177B,DISP1*
12	62201242-62401242	0.0000	*FAM19A2*
13	30523899-30723899	0.0000	
13	95723899-95923899	0.0000	*ABCC4*
14	62997969-63197969	0.0000	*KCNH5*
15	58029561-58229561	0.0000	
19	38696458-38896458	0.0000	*SPRED3,DPF1,SPINT2,PPP1R14A,FAM98C,PSMD8,GGN,KCNK6,SIPA1L3,C19orf33 ,CATSPERG,YIF1B*
19	51896458-52096458	0.0000	*SIGLEC12,SIGLEC10,FLJ30403,ZNF175,LOC100129083,CEACAM18,SIGLEC8,SIGLEC6*
2	223863495-224063495	0.0000	*KCNE4*
3	120713534-120913534	0.0000	*STXBP5L*
3	65913534-66113534	0.0000	*MAGI1*
6	167986117-168186117	0.0000	*C6orf123*
8	13398327-13598327	0.0000	*C8orf48*
8	5398327-5598327	0.0000	
12	83001242-83201242	0.0004	*TMTC2*
1	223120733-223320733	0.0005	*TLR5,DISP1*
1	39720733-39920733	0.0005	*MACF1,KIAA0754*
9	24809754-25009754	0.0006	
14	65997969-66197969	0.0007	*FUT8*
14	98597969-98797969	0.0008	
4	3750683-3950683	0.0008	*ADRA2C*
17	3456962-3656962	0.0009	*TRPV1,TRPV3,EMC6,P2RX5-TAX1BP3,ITGAE,SHPK,P2RX5,CTNS,GSG2,TAX1BP3*
4	118950683-119150683	0.0010	*NDST3*
9	114809754-115009754	0.0012	*SUSD1,MIR3134,PTBP3*
3	114713534-114913534	0.0013	*ZBTB20*
18	43957275-44157275	0.0013	*LOXHD1,RNF165*
16	22915657-23115657	0.0014	*USP31,HS3ST2*
12	79601242-79801242	0.0015	*SYT1*
4	99150683-99350683	0.0015	*RAP1GDS1*
7	20260033-20460033	0.0016	*ITGB8*
1	222320733-222520733	0.0017	
10	118162102-118362102	0.0018	*PNLIP,PNLIPRP1,PNLIPRP3*
11	58218141-58418141	0.0019	*ZFP91,OR5B21,LPXN,ZFP91-CNTF,CNTF*
1	235120733-235320733	0.0020	*RBM34,TOMM20,SNORA14B*

8	100198327-100398327	0.0020	*VPS13B*
1	153720733-153920733	0.0021	*SLC27A3,CRTC2,DENND4B,GATAD2B,INTS3*
3	110713534-110913534	0.0022	*PVRL3*
2	179063495-179263495	0.0023	*OSBPL6,MIR548N*
6	130586117-130786117	0.0024	*TMEM200A*
6	40386117-40586117	0.0024	*LRFN2*
1	71520733-71720733	0.0025	*ZRANB2,MIR186*
2	12263495-12463495	0.0025	
17	4056962-4256962	0.0025	*CYB5D2,UBE2G1,ANKFY1*
1	242120733-242320733	0.0027	*PLD5,MAP1LC3C*
11	110618141-110818141	0.0028	
12	77801242-78001242	0.0029	
9	24409754-24609754	0.0030	
6	79586117-79786117	0.0030	*PHIP,IRAK1BP1*
11	110418141-110618141	0.0030	*ARHGAP20*
1	223520733-223720733	0.0032	*SUSD4,CAPN8,C1orf65*
15	55029561-55229561	0.0032	
3	120513534-120713534	0.0033	*STXBP5L*
1	223320733-223520733	0.0034	*SUSD4*
11	61018141-61218141	0.0034	*DAK,SDHAF2,TMEM138,DDB1,CPSF7,VWCE,TMEM216,CYB561A3,PGA5*
14	32197969-32397969	0.0035	*NUBPL*
22	46273468-46473468	0.0037	*LOC730668,WNT7B,LOC150381,C22orf26*
18	6157275-6357275	0.0038	*L3MBTL4*
4	13150683-13350683	0.0038	*HSP90AB2P*
2	98863495-99063495	0.0039	*INPP4A,CNGA3,VWA3B*
12	132801242-133001242	0.0040	*GALNT9,LOC100130238*
3	5513534-5713534	0.0040	
8	68598327-68798327	0.0041	*CPA6*
7	17860033-18060033	0.0042	*SNX13*
8	23998327-24198327	0.0042	*ADAM28*
6	100786117-100986117	0.0043	*ASCC3,SIM1*
9	24209754-24409754	0.0044	
3	44513534-44713534	0.0045	*ZKSCAN7,ZNF660,ZNF197,ZNF35,ZNF445*
7	40860033-41060033	0.0047	*C7orf10*
17	45256962-45456962	0.0047	*MYL4,CDC27,EFCAB13,ITGB3*
3	10713534-10913534	0.0048	*SLC6A11*
10	102162102-102362102	0.0049	*HIF1AN,WNT8B,NDUFB8,SEC31B*
19	8896458-9096458	0.0049	*MUC16,MBD3L1,ZNF558*
9	8409754-8609754	0.0049	*PTPRD*
20	47287112-47487112	0.0051	*PREX1*
12	110801242-111001242	0.0052	*PPTC7,ARPC3,FAM216A,ANAPC7,RAD9B,GPN3,VPS29*
10	92362102-92562102	0.0052	*HTR7*
2	107663495-107863495	0.0053	
4	106350683-106550683	0.0053	*EEF1A1P9,ARHGEF38,PPA2*
3	2113534-2313534	0.0054	*CNTN4*
2	122063495-122263495	0.0054	*CLASP1*

17	61656962-61856962	0.0055	*TACO1,DDX42,STRADA,MAP3K3,CCDC47,LIMD2,DCAF7*
2	217463495-217663495	0.0056	*IGFBP2,IGFBP5*
10	60562102-60762102	0.0057	*BICC1*
11	111818141-112018141	0.0059	*PIH1D2,DIXDC1,TIMM8B,C11orf57,DLAT,IL18,SDHD*
13	33123899-33323899	0.0059	*PDS5B*
17	39056962-39256962	0.0059	*KRT39,KRTAP1-1,KRTAP1-3,KRTAP1-5,KRT23,KRTAP3-2,KRTAP3-3,KRTAP3-1,KRTAP4-7,KRTAP2-1,KRTAP2-4,KRTAP2-2,KRTAP4-8,KRT40*
16	47315657-47515657	0.0060	*ITFG1,PHKB*
8	67198327-67398327	0.0062	*ADHFE1,RRS1*
12	124001242-124201242	0.0064	*MIR3908,TCTN2,DDX55,ATP6V0A2,EIF2B1,TMED2,RILPL1,GTF2H3*
10	22562102-22762102	0.0064	*COMMD3,SPAG6,COMMD3-BMI1,BMI1,LOC100499489*
18	4557275-4757275	0.0064	
22	35873468-36073468	0.0064	*RASD2,MB,APOL6*
12	28201242-28401242	0.0066	
12	49801242-50001242	0.0067	*KCNH3,FAM186B,SPATS2,MCRS1*
7	154060033-154260033	0.0067	*DPP6*
8	25198327-25398327	0.0068	*DOCK5,GNRH1,KCTD9,CDCA2*
6	115586117-115786117	0.0069	
10	124162102-124362102	0.0070	*PLEKHA1,DMBT1,HTRA1,ARMS2,MIR3941*
13	47723899-47923899	0.0070	
7	43460033-43660033	0.0071	*STK17A,HECW1*
12	59201242-59401242	0.0072	*LRIG3*
5	4090162-4290162	0.0072	
10	74162102-74362102	0.0074	*MIR1256,MICU1*
5	169490162-169690162	0.0074	*FOXI1,DOCK2,LCP2,C5orf58*
10	31562102-31762102	0.0074	*ZEB1*
1	12120733-12320733	0.0076	
16	4515657-4715657	0.0076	*UBALD1,CDIP1,NMRAL1,HMOX2,MGRN1*
9	5009754-5209754	0.0079	*INSL6,JAK2*
3	47713534-47913534	0.0079	*SMARCC1,DHX30,MIR1226,MAP4*
1	71320733-71520733	0.0080	*PTGER3*
1	97520733-97720733	0.0080	*DPYD*
4	146950683-147150683	0.0080	*LSM6*
2	190463495-190663495	0.0081	*ASNSD1,ORMDL1,PMS1,ANKAR,OSGEPL1*
2	30463495-30663495	0.0081	*LBH*
12	79401242-79601242	0.0082	*SYT1*
5	166290162-166490162	0.0082	
3	110513534-110713534	0.0084	
8	118398327-118598327	0.0085	*MED30*
17	3656962-3856962	0.0085	*ATP2A3,C17orf85,CAMKK1,ITGAE,P2RX1*
7	140860033-141060033	0.0088	
8	5598327-5798327	0.0088	

Sites highlighted in red depict selection signals identified in previous studies

SN8 Table 2: Highly differentiated signals between Europe and African populations that showed overlap with extreme iHS scores

chr:pos	Genes within 50kb	A1	A2	AF_CEU	AF_YRI	AF_CHB	AF_Ethiopia	AF_Zulu	AF_LWK	AF_Jola	AF_Kalenjin	F_{ST}
1:206302435	CTSE,C1orf186	G	A	0.01	0.88	0.00	0.43	0.85	0.74	0.82	0.69	0.59
11:61122878	DDB1,CPSF7,TMEM216,TMEM138,CYB561A3,DAK	G	A	0.01	0.67	0.03	0.67	0.81	0.80	0.84	0.85	0.58
11:61122341	DDB1,CPSF7,TMEM216,TMEM138,CYB561A3,DAK	T	C	0.01	0.67	0.03	0.67	0.81	0.80	0.84	0.85	0.57
10:118165879	PNLIPRP3,CCDC172	T	C	0.94	0.10	0.53	0.59	0.09	0.20	0.14	0.21	0.55
4:3881341	NA	T	G	0.02	0.78	0.34	0.50	0.72	0.81	0.82	0.71	0.54
6:130608255	NA	A	G	0.03	0.78	0.00	0.42	0.79	0.78	0.79	0.59	0.53
10:118173471	PNLIPRP3,CCDC172	A	G	0.92	0.10	0.53	0.58	0.09	0.20	0.14	0.21	0.53
10:118187399	PNLIPRP3,CCDC172	G	T	0.92	0.10	0.43	0.64	0.08	0.18	0.14	0.22	0.53
10:118193020	PNLIPRP3	C	T	0.92	0.11	0.43	0.65	0.08	0.18	0.14	0.22	0.52
17:3503921	CTNS,TRPV1,TRPV3,SHPK	C	T	0.00	0.79	0.07	0.40	0.79	0.70	0.79	0.58	0.52
3:110882326	PVRL3	T	G	0.85	0.03	0.29	0.34	0.04	0.04	0.01	0.13	0.52
14:32328698	NUBPL	A	G	0.02	0.79	0.00	0.43	0.72	0.75	0.82	0.59	0.51
10:31712158	ZEB1	A	C	0.06	0.83	0.24	0.59	0.79	0.81	0.88	0.76	0.51
10:22568695	COMMD3,BMI1,COMMD3	A	G	0.01	0.69	0.00	0.46	0.72	0.82	0.74	0.72	0.51
10:31670562	ZEB1	G	A	0.06	0.83	0.24	0.59	0.79	0.80	0.87	0.76	0.50
9:8561542	PTPRD	A	G	0.81	0.05	0.69	0.48	0.06	0.09	0.04	0.15	0.50
2:12384394	NA	C	T	0.00	0.75	0.00	0.34	0.73	0.70	0.68	0.52	0.49
9:24883179	NA	C	A	0.88	0.06	0.79	0.49	0.03	0.09	0.13	0.19	0.49
14:32328830	NUBPL	T	C	0.01	0.78	0.00	0.42	0.59	0.73	0.76	0.58	0.48
10:31592407	ZEB1	A	G	0.06	0.82	0.24	0.59	0.70	0.74	0.89	0.74	0.48
9:8554938	PTPRD	T	C	0.85	0.11	0.70	0.42	0.07	0.17	0.06	0.13	0.48
10:22666657	BMI1,SPAG6,COMMD3	G	T	0.03	0.69	0.00	0.44	0.70	0.82	0.75	0.73	0.48
3:110848243	PVRL3	G	A	0.95	0.20	0.73	0.48	0.25	0.20	0.14	0.37	0.47
11:61078988	DDB1,VWCE,CYB561A3,DAK	T	C	0.00	0.59	0.03	0.61	0.70	0.72	0.68	0.75	0.47
10:22673699	SPAG6	T	C	0.03	0.69	0.00	0.44	0.68	0.80	0.74	0.72	0.46

Allele frequencies are presented with respect to the allele in the A1 column. Estimates of Europe-Africa differentiation for each locus are presented in the F_{ST} column.

SN8 Table 3: Top differentiated signals (unmasked F_{ST}) among African populations that were also candidates for selection sweeps on iHS analysis

Chr	Pos	Genes within 50kb	AGVP F_{ST} (unmasked)	AGVP F_{ST} (masked)	A1	A2	CEU	YRI	CHB	Ethiopia	Zulu	LWK	Kalenjin
16	47333345	*ITFG1*	0.15	NA	G	A	0.33	0.00	0.64	0.29	0.00	0.00	0.05
16	47486817	*PHKB,ITFG1*	0.14	NA	T	C	0.34	0.00	0.67	0.29	0.00	0.01	0.06
4	99309404	*RAP1GDS1*	0.14	0.01	A	G	0.46	0.01	0.10	0.33	0.01	0.02	0.14
8	118478102	NA	0.12	0.01	T	C	0.55	0.00	0.18	0.31	0.01	0.03	0.05
14	63098431	NA	0.12	0.07	G	A	0.89	0.21	0.51	0.64	0.10	0.20	0.49
8	5408130	NA	0.12	0.02	T	G	0.28	0.00	0.19	0.26	0.00	0.01	0.11
10	118193020	*PNLIPRP3*	0.12	0.03	C	T	0.92	0.11	0.43	0.65	0.08	0.18	0.22
17	3821988	*ATP2A3,CAMKK1,P2RX1*	0.12	0.06	C	A	0.58	0.04	0.89	0.50	0.07	0.05	0.19
15	58219859	*ALDH1A2*	0.12	0.01	A	C	0.30	0.00	0.51	0.23	0.01	0.02	0.05
10	118187399	*PNLIPRP3,CCDC172*	0.12	0.02	G	T	0.92	0.10	0.43	0.64	0.08	0.18	0.22
1	12217140	*TNFRSF1B,TNFRSF8*	0.12	0.06	T	C	0.46	0.00	0.28	0.30	0.02	0.05	0.10
3	10891237	*SLC6A11*	0.11	0.01	T	C	0.21	0.01	0.26	0.27	0.02	0.03	0.07
3	110882326	*PVRL3*	0.11	NA	T	G	0.85	0.03	0.29	0.34	0.04	0.04	0.13
8	25364331	*CDCA2,KCTD9*	0.11	0.05	A	G	0.34	0.02	0.12	0.33	0.00	0.02	0.06
12	124147831	*DDX55,ATP6V0A2,EIF2B1,TCTN2,GTF2H3*	0.11	0.01	A	G	0.53	0.02	0.83	0.29	0.02	0.03	0.09
9	8561542	*PTPRD*	0.11	0.01	A	G	0.81	0.05	0.69	0.48	0.06	0.09	0.15

Allele frequencies are presented with respect to the allele in the A1 column.

AGVP F_{ST} masked and unmasked represent the F_{ST} calculated across all 16 African populations when non-SSA ancestry was masked, and when this was left unmasked, respectively.

SN8 Table 4. Top differentiated signals (masked F_{ST}) among African populations that were also candidates for selection sweeps in iHS analysis

Chr	Pos	Genes within 50kb	AGVP F_{ST} (masked)	AGVP F_{ST} (unmasked)	A1	A2	CEU	CHB	YRI	Ethiopia	Zulu	LWK	Kalenjin
17	3815031	*ATP2A3,CAMKK1,P2RX1*	0.09	0.10	C	T	0.53	0.81	0.08	0.52	0.11	0.20	0.33
12	77903833	NA	0.08	0.05	G	A	0.03	0.00	0.00	0.12	0.00	0.00	0.04
12	62377692	*FAM19A2*	0.08	0.02	T	C	0.00	0.00	0.04	0.14	0.06	0.04	0.05
2	223937249	KCNE4	0.08	0.03	T	G	0.00	0.00	0.03	0.11	0.01	0.03	0.09
11	91929956	NA	0.08	0.07	C	T	0.00	0.00	0.26	0.01	0.34	0.18	0.07
1	12223451	*TNFRSF8,TNFRSF1B*	0.08	0.10	A	G	0.38	0.31	0.02	0.37	0.03	0.09	0.18
11	91932735	NA	0.08	0.07	T	C	0.00	0.00	0.26	0.01	0.34	0.18	0.07
11	91933439	NA	0.08	0.07	A	G	0.00	0.00	0.26	0.01	0.34	0.18	0.07
9	24522869	NA	0.08	0.02	C	T	0.01	0.00	0.06	0.14	0.06	0.06	0.13
15	55179901	NA	0.08	0.05	A	G	0.05	0.16	0.03	0.18	0.02	0.09	0.20
15	55183281	NA	0.08	0.05	T	C	0.05	0.16	0.03	0.18	0.02	0.09	0.21
11	61148456	*DDB1,CPSF7,TMEM138,TMEM216,SDHAF2,CYB561A3,DAK*	0.08	0.05	G	A	0.01	0.00	0.25	0.42	0.24	0.38	0.60
2	223931853	*KCNE4*	0.08	0.03	C	T	0.00	0.00	0.03	0.12	0.02	0.03	0.12
11	91690669	NA	0.07	0.04	C	T	0.24	0.16	0.03	0.23	0.06	0.03	0.08
11	61074128	*DDB1,VWCE,CYB561A3,DAK*	0.07	0.04	T	C	0.00	0.00	0.08	0.17	0.10	0.11	0.21
11	61207310	*CPSF7,TMEM216,SDHAF2,PPP1R32*	0.07	0.05	T	G	0.15	0.03	0.20	0.43	0.25	0.37	0.53

Allele frequencies are presented with respect to the allele in the A1 column.

AGVP F_{ST} masked and unmasked represent the F_{ST} calculated across all 16 African populations when non-SSA ancestry was masked, and when this was left unmasked, respectively.

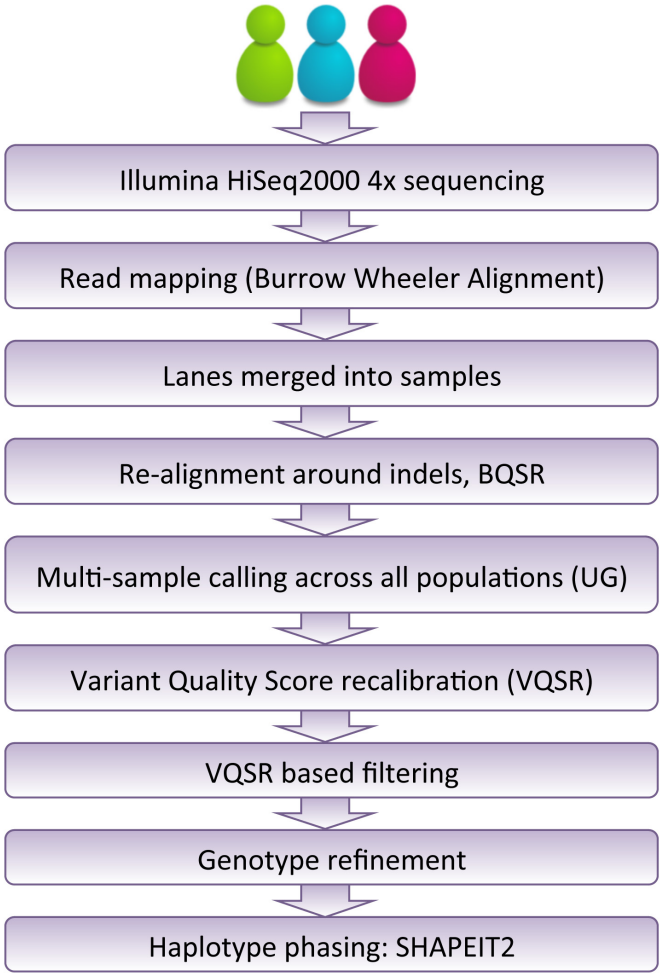
S. NOTE 9: CURATION OF AGVP SEQUENCE DATA AND REFERENCE PANEL

The AGVP provides a whole genome sequence resource of African populations available to researchers. In order to provide a valid resource, and examine the impact of calling variants within and across multiple diverse populations, we evaluated various strategies for calling variants in diverse African populations. Here, we describe these methods, and the workflow for the curation of the AGVP sequence data.

SN9.1 READ MAPPING AND BAM IMPROVEMENT

Following generation of raw reads, duplicate reads were marked, and mapping was carried out using BWA to the human reference genome (GRCh37). Lanes were merged into samples, and sample level bam improvement was carried out using GATK. This consisted of re-alignment of reads around indels in addition to recalibration of base qualities using the GATK BQSR function. Two samples from the Complete Genomics and the Platinum Genomes highly curated set were included for validation of the data processing pipeline (NA12878 and NA19240) (http://www.illumina.com /platinumgenomes/). PCR-free reads were used for these validation samples, to avoid PCR artefacts. These validation samples were downsampled to 4x coverage, and processed through the same pipeline, to provide a comparator against high coverage 30x data. The Platinum Genomes sample (NA12878) represents a sample from a 12 person pedigree that has gone through extensive curation and validation of variants. This was considered the gold standard for evaluation. The accuracy of called data from a 4x sample would provide a

SN 9 Fig 1: Workflow for variant calling



SN9 Fig 1 depicts the workflow for variant calling across the AGVP low coverage sequence from 3 populations. Following read mapping and bam improvement, calling was carried out across all populations using Unified Genotyper (UG). This was followed by VQSR based filtering and genotype refinement with Beagle.

guide to the accuracy of the workflow applied. For comparisons of ultra-low coverage sequencing data with higher coverage data (**see Supplementary Note 12**), we also generated and called 8x coverage data among Ethiopian samples. We used a similar approach for calling 8x sequence data; these data were called across the 120 Ethiopian samples alone, with the same workflow as represented in **SN9 Fig 1**.

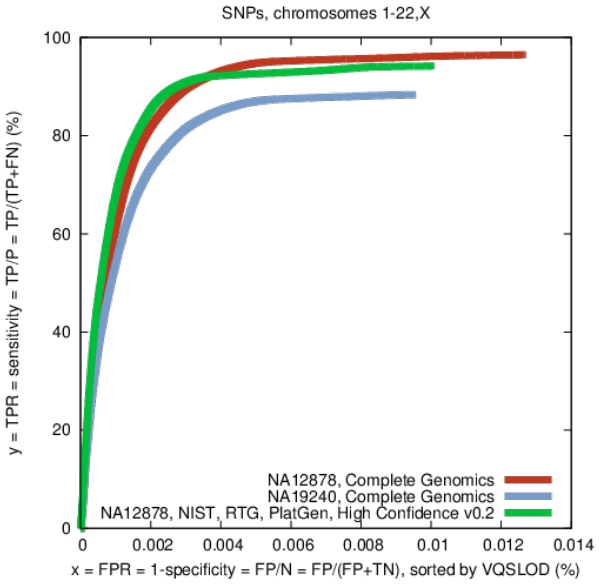
SN9.2 Quality control of sequence data

In order to ensure the quality of the large quantity of BAMs produced for the project, an automatic quality control system was employed to reduce the number of data files that required manual intervention. This system was derived from the one designed for the UK10K project (http://www.uk10k.org) and used a series of empirically derived thresholds to assess summary metrics calculated from the input BAMs. These thresholds included: percentage of reads mapped; percentage of duplicate reads marked; various statistics measuring INDEL distribution against read cycle and an insert size overlap percentage. Any lane that fell below the "fail" threshold for any of the metrics were excluded; any lane that fell below the "warn" threshold on a metric would be manually examined; and any lane that did not fall below either of these thresholds for any of the metrics was given a status of "pass" and allowed to proceed into the later stages of the pipeline.

SN9.3 Data processing workflow

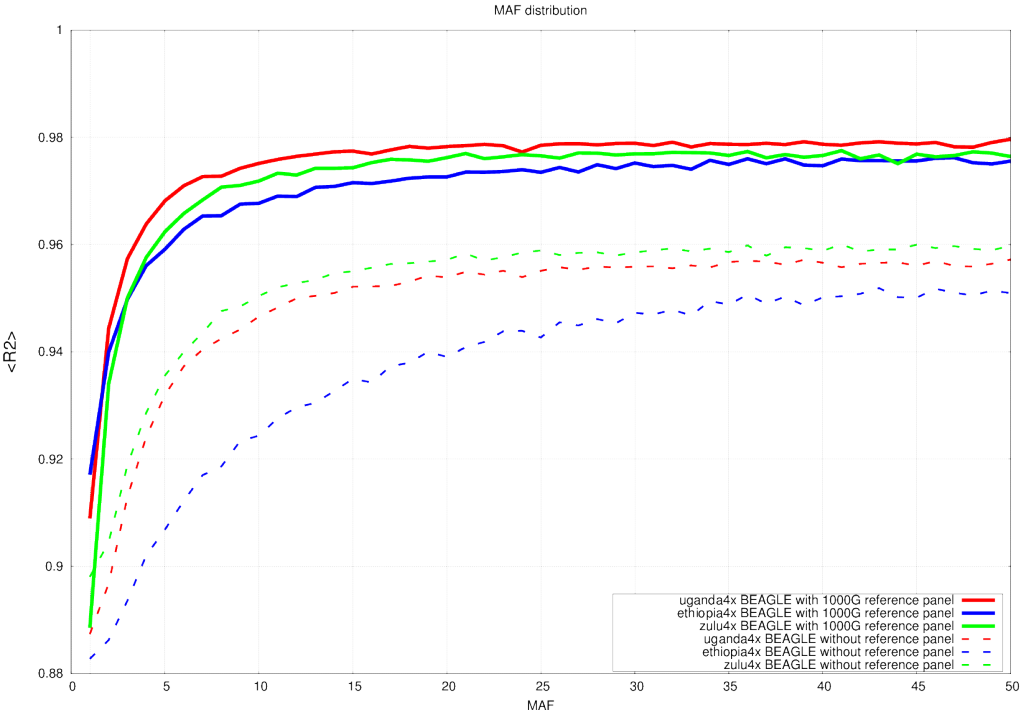
The workflow for data processing is represented in **SN9 Fig 1**. Multi-sample variant calling was carried out across all populations with the Unified Genotyper using the -emit_variant_only option. Annotations (DP, QUAL, FS, HaplotypeScore, MQRankSum, ReadposRankSum) for filtering were calculated across all samples, to input these into models for calculation of Variant Quality Scores for recalibration. Following this, Variant Quality Score Recalibration (VQSR) was carried out using the HapMap and 1000 Genomes Omni 2.5 M data as truth and training datasets. For validation, we examined the downsampled validation samples which had been called along with all 320 samples for sensitivity and specificity relative to the curated high coverage gold standard data within high confidence regions available for these samples, and identified the tranche sensitivity threshold corresponding to the highest point on the ROC curve (**SN9 Fig 2**). We chose a threshold of 99% tranche sensitivity, corresponding to a sensitivity of 88% and specificity of 76%. The correlation of the final curated dataset in comparison to Omni 2.5M genotypes was noted to be high (**SN9 Fig 3**), with mean correlation >0.95 for all three populations.

SN9 Fig 2: ROC curve of low coverage whole genome sequencing data for NA12878 and NA19240 compared to 30x coverage sequencing for different VQSLOD filtering scores.



SN9 Fig 2 shows the ROC curves for the sensitivity and specificity of 4x sequence data from Complete genomics and Platinum genomes samples with respect to high coverage curated 30x genomes from these individuals. The green curve represents the sensitivity and specificity of 4x sequence data for NA12878 in comparison with gold standard data within high confidence intervals. The red and blue lines represent these parameters calculated across the genome for NA12879 and NA19240. Each point on the curve corresponds to the the sensitivity and specificity obtained at a particular VQSLOD threshold used for filtering. We find the highest sensitivity and specificity for the real time genomics sample NA12878 calculated for high confidence sites (green line) corresponds to a tranche-sensitivity threshold of 99%

SN9 Fig 3: Correlation between curated 4x WGS data and Omni 2.5M genotypes



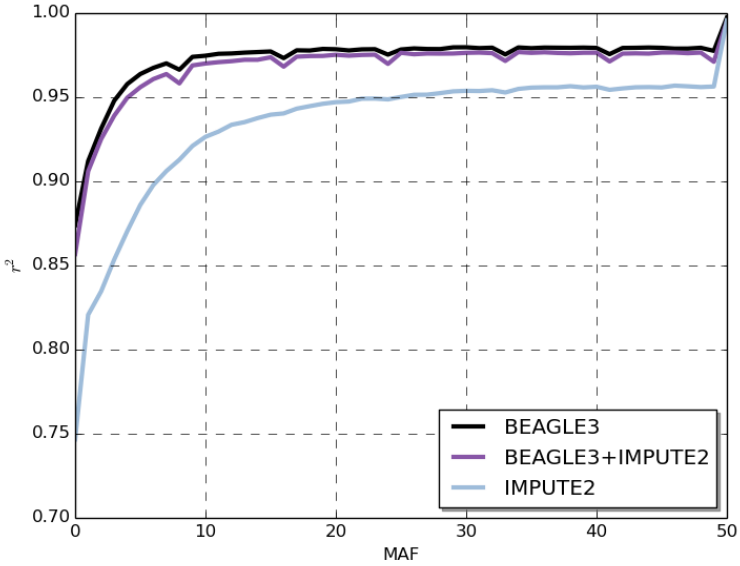
SN9 Fig 3 represents the high correlation between WGS 4x curated data (following refinement with the 1000 Genomes reference panel) and the Omni 2.5M genotypes for the same individuals

We anticipated that the false positive rate would drop following imputation based genotype refinement. Following this we carried out genotype refinement, as we describe subsequently.

SN9.4 Evaluation of methods for imputation based refinement

In order to obtain a highly curated set of variants, we evaluated methods for imputation based genotype refinement, to use the most effective approach. It has been previously suggested that sequential imputation approaches can provide modestly more accurate results compared to single step imputation for imputation based refinement.[46] We, therefore, compared using a single algorithm such as Beagle with the 1000 Genomes Project phase I reference panel, to using Beagle and IMPUTE2 sequentially. We assessed accuracy by correlation with genotypes on the Omni 2.5M genotype array. This comparison was carried out for the Bagandan population alone. For this, calling of the 4x data from the Ugandan population was carried out separately across 100 Bagandan samples, using the same workflow as described in SN9 Fig 1, using a tranche sensitivity of 96% for filtering. Beagle v3 used alone performed better than sequential imputation with a high correlation with genotypes (**SN9 Fig 4**), suggesting that Beagle alone provides better refinement of genotype likelihoods. This has important implications for curation of large-scale genome sequence data, as computational burden can be substantially reduced by omitting an additional imputation step from data processing. The final sensitivity and specificity of the improved dataset was noted to be 97% and 91% respectively for the Platinum genomes sample.

SN9 Fig 4: Comparison of Beagle with IMPUTE2 and sequential imputation with both for genotype refinement of data



SN9 Fig 4 represents the correlation between 4x Bagandan sequence data and Omni 2.5M genotypes, when using different algorithms for genotype refinement (Beagle v3, Beagle v3 followed by IMPUTE2, IMPUTE2 alone). Using Beagle alone for genotype refinement appears to produce the most accurate genotype likelihoods. Addition of IMPUTE2 does not provide any added advantage.

SN9.5 GENERATION OF A MERGED REFERENCE PANEL FOR IMPUTATION

Given the poor representation of many African populations in existing sequencing reference panels, we sought to improve this by integrating the AGVP panel with the 1000G reference panel. For this, we generated a merged reference panel for imputation. First, we phased the data generated from the 3 populations with SHAPEIT2, using an approach that assigned discrete genotypes by identifying the genotype with the greatest genotype likelihood. We generated a merged reference panel using the process outlined in IMPUTE2 with the **-merge_ref_panels_output_ref** command (http://mathgen.stats.ox.ac.uk/impute/merging_reference_panels.html). Briefly, this method generates a merged reference panels for imputation, by reciprocally imputing missing sites between the two panels. The final merged reference panel included whole genome sequence data from 1,412 individuals and 40,163,067 variants, including SNPs and structural variants. We describe our evaluation of this reference panel subsequently in **Supplementary Note 11**.

SN 9.6 DESCRIPTION OF CURATED WHOLE GENOME SEQUENCE DATA

Based on the AGVP WGS data, we identified a total of 20.5, 20.3M and 20.5M SNPs among Ethiopian, Zulu and Bagandan individuals, respectively (**Extended Data Figure 1**). We found a substantial proportion of unshared variants among these populations (11%-23%). The greatest proportion of unshared variants was in the Ethiopian populations, which is consistent with their greater differentiation. In all 3 populations, we identified a large proportion of novel variants relative to the 1000 Genomes phase 1 data (16-24%), the majority of which were unshared and rare. Again, this was greatest in the Ethiopian populations (**Extended Data Figure 1**). These findings recapitulate the need for large-scale sequence data from populations across Africa, including genetically divergent populations.

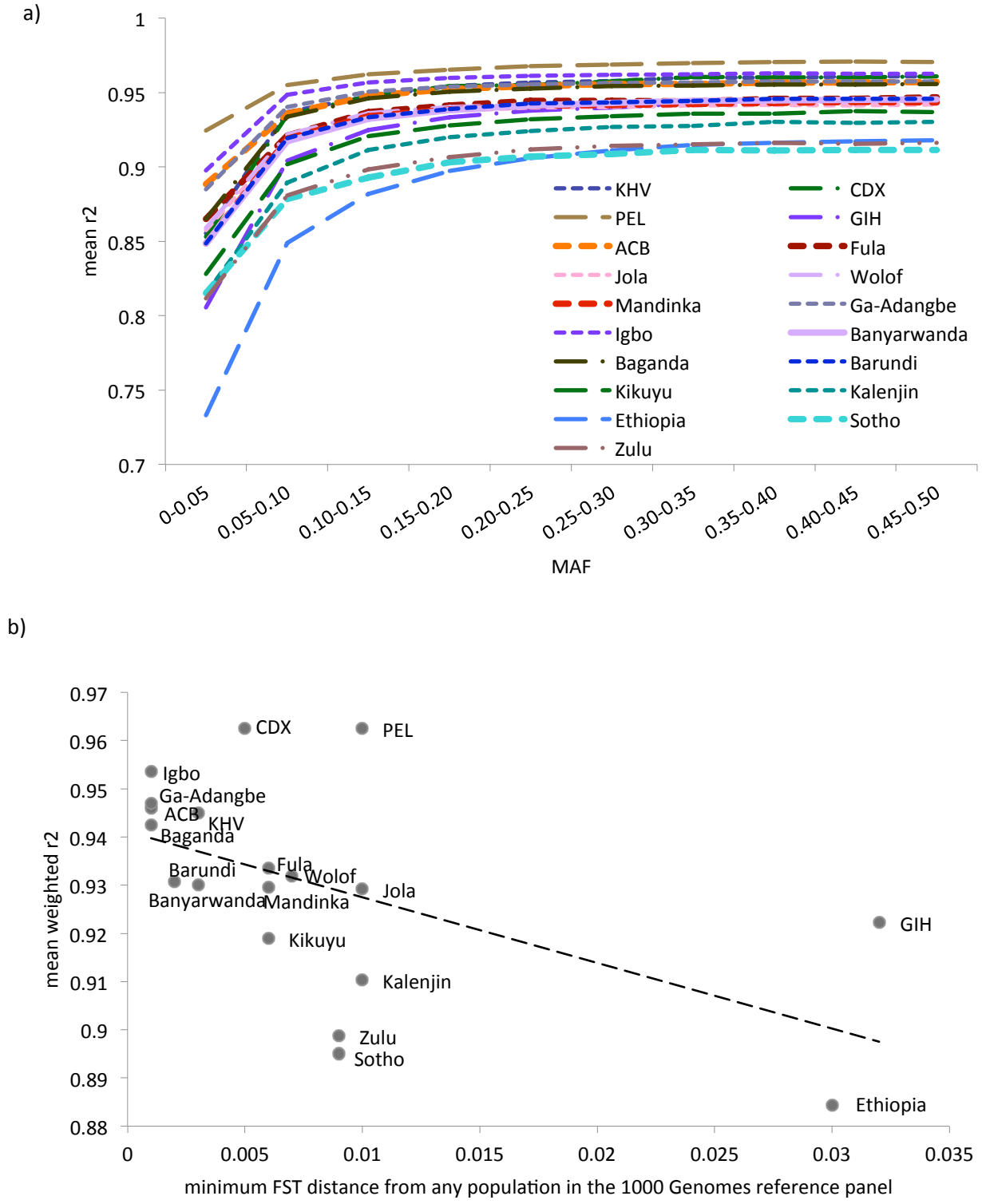
S. NOTE 10: IMPUTATION INTO GENOTYPE ARRAYS IN AFRICAN POPULATIONS

Conventionally, imputation into genotype data from African populations is thought to be poor due to greater LD decay with distance resulting in shorter tracts of LD compared with European and Asian populations.[47] However, the accuracy of imputation into African populations using existing WGS reference panels, and dense chip arrays remains unevaluated. We sought to evaluate imputation accuracy across 16 populations in the AGVP using the 1000 Genomes phase 1 version 3 integrated reference panel for imputation. Imputation was carried out into the Illumina Omni 2.5 M genotype array for each population separately. Pre-phasing was not carried out prior to imputation to avoid any loss of accuracy.

For assessment of imputation accuracy, we used the r^2 metric outputted by IMPUTE2. This metric is generated by 'leave one out masking', in which each marker is sequentially left out of the genotyped dataset and imputed, as if it were missing. This parameter, therefore, represents the correlation between imputed and genotyped data for markers on the array. We calculated mean r^2 within minor allele frequency bins of 5%. In order to generate a single estimate comparable across all populations, we calculated a mean r^2 across all genotyped sites for each population. However, as r^2 is dependent on minor allele frequency, and frequency spectra for different populations may be different, we weighted this estimate equally across all allele frequency bins, so as to avoid effects due to ascertainment bias.

We observed reasonably accurate imputation across all AGVP populations (**SN10 Fig 1**). Weighted mean r^2 varied between 0.88-0.95 among different populations, with the lowest accuracy for Ethiopian populations, and the greatest for Igbo. This variation in accuracy is likely due to the poor representation of Afro-Asiatic linguistic groups in the 1000 Genomes sequencing reference panel, and relatively good representation of West African Bantu population groups like Igbo by YRI. Next, we confirmed that the imputation accuracy in each population was directly correlated with the minimum genetic distance of the population from any population in the reference panel (r=-0.55) (**SN10 Fig 1b**).

SN10 Fig 1. Comparative imputation accuracy into the Illumina Omni 2.5M genotype array for different African populations



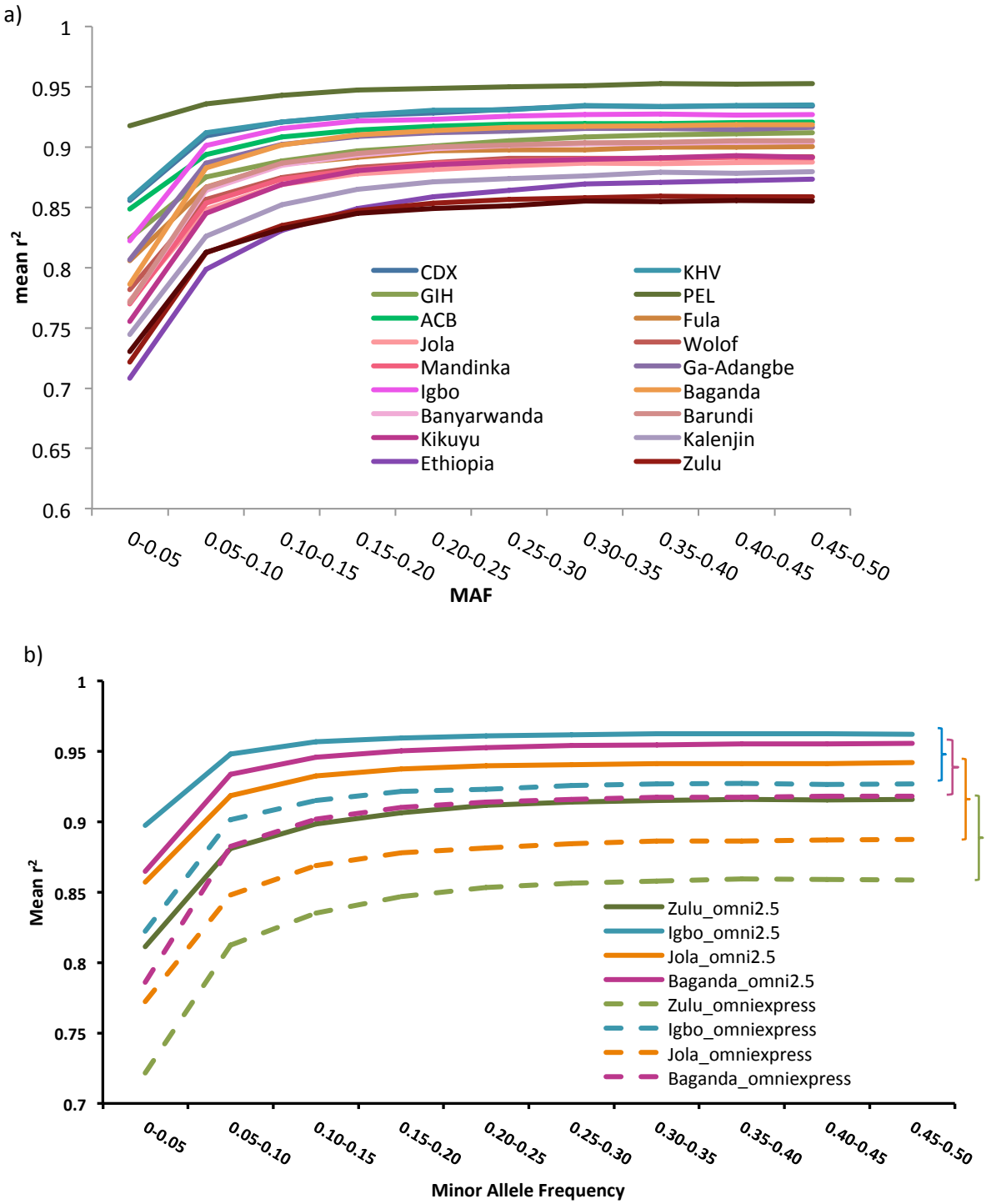
SN10 Fig 1a shows the imputation accuracy into the Omni 2.5M genotype array for different populations using the 1000 Genomes Project reference panel. SN9 Fig 1b depicts the correlation between the mean weighted r^2 across the array and the minimum genetic distance of each population from the reference panel.

SN10 Fig 1b suggests that imputation accuracy is determined by genetic distance from a given reference panel. However, it is also important to consider the influence of admixture here. For example, we would expect that Ethiopia would perform poorly as it is not well represented in the reference panel, with a high minimum F_{ST} from any population in the reference; however, at least a proportion of this differentiation from the closest population in the panel could be attributed to high levels of Eurasian admixture in this population group. It is unlikely that this Eurasian admixture would reduce imputation accuracy, as these haplotypes are likely to be well represented in existing reference panels. We would therefore expect that accuracy in Ethiopian populations would be greater than expected by the level of differentiation from the reference panel. Similarly, we would expect differentiation to be over-estimated due to admixture in other admixed groups such as GIH, CDX, PEL and ACB, making these outliers. On exclusion of these groups, we observe that the negative correlation between imputation accuracy and genetic distance from the reference panel becomes much stronger (r=-0.79).

SN9.1 IMPACT OF REDUCTION IN CHIP DENSITY ON IMPUTATION ACCURACY

Imputation accuracy is a function of the reference panel used for imputation, as well as the genotype array into which imputation is carried out. Use of dense genotype arrays can potentially improve imputation accuracy by providing better delineation of haplotypes. We assessed the impact of altering the marker density on the genotype array on imputation accuracy. We thinned the Omni 2.5M genotype array to include only those markers on the Omni express genotype array. A total of ~600,000 markers were noted to be overlapping, and included in the analysis. Expectedly, we found a decline in imputation accuracy across all AGVP populations, with mean r^2 varying between 0.85-0.93 among African populations. This decline was observed across the whole allele frequency spectrum; however, we note this decline is not uniform, and was greater in some populations relative to others (**SN10 Fig 2**). Specifically, we find that minimum genetic distance from any population in the reference panel was weakly correlated (r=0.14) with the level of decline in mean r^2 seen among different populations. (**SN10 Fig 3**) However, here too, we find that Ethiopia represents an outlier, as much of the differentiation from the 1000 Genomes reference panel can be explained by substantial Eurasian admixture among Ethiopian populations, which would not affect imputation proportionately, as these haplotypes are likely to be well represented in the reference panel. Excluding Ethiopia markedly improved the correlation observed to 0.89 (**SN10 Fig 3b**). This suggests that decline in imputation accuracy with loss of marker density is greater for populations poorly represented in reference panels, arguing the need for dense and efficient genotype arrays in combination with better sequencing reference panels for these populations.

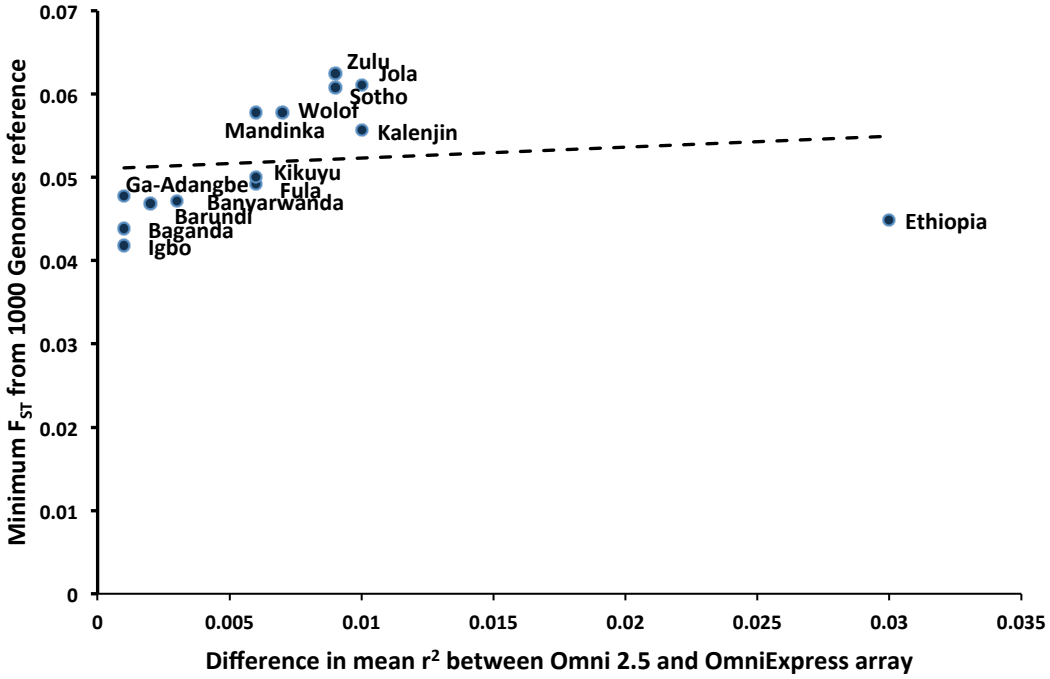
SN10 Fig 2 : Loss of imputation accuracy in different African populations after thinning of data



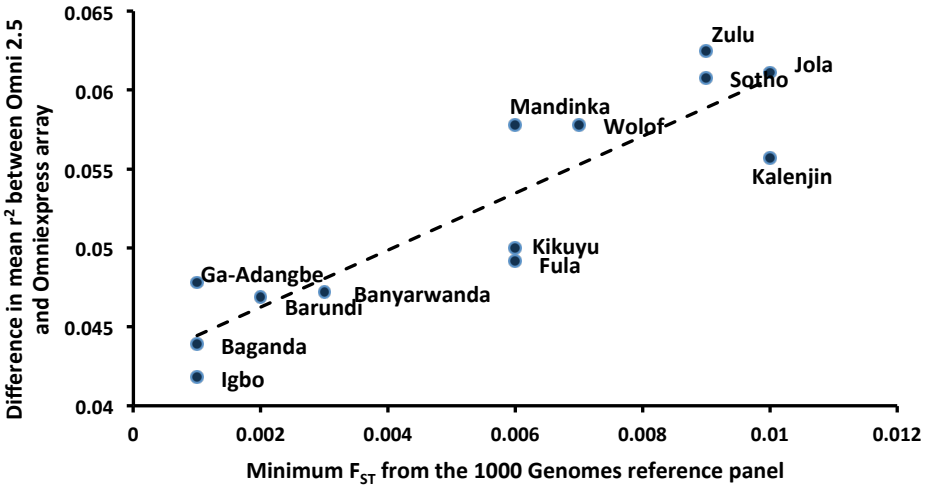
SN10 Fig 2 shows the reduction in imputation accuracy on thinning the Illumina Omni 2.5 M chip data to Illumina OmniExpress (~650,000 sites) when using the 1000 Genomes Project phase I integrated dataset as a reference panel. There is clear reduction in accuracy observed in all populations, with the reduction in accuracy being greater for certain populations (Zulu and Jola) compared to others (Igbo and Baganda). The reduction in accuracy appears to be related to the genetic distance of a given population to the reference panel, arguing that denser chip arrays may be required for more accurate imputation in populations poorly represented in reference panels.

SN10 Fig3: Correlation between loss of imputation accuracy and distance from the 1000 Genomes reference panel among African populations

a)



b)



SN10 Fig 3 shows the correlation between loss of imputation accuracy when imputing into the Illumina Omni 2.5M chip compared to the Illumina OmniExpress chip array, and the minimum genetic distance from any population within the 1000 Genomes Project reference panel. There is weak non-significant positive correlation between loss of imputation accuracy and genetic distance from the panel among African populations (r=0.14, p=0.64). Ethiopia appears to be an outlier; this is likely to be due to a large proportion of genetic distance from the 1000 Genomes panel being due to substantial Eurasian admixture, which would not affect the imputation accuracy proportionately, as these haplotypes are likely to be well represented in the panel. On excluding Ethiopians, the correlation became much stronger (r=0.89, p=0.00003).

S. NOTE 11: IMPUTATION USING THE AGVP REFERENCE PANEL

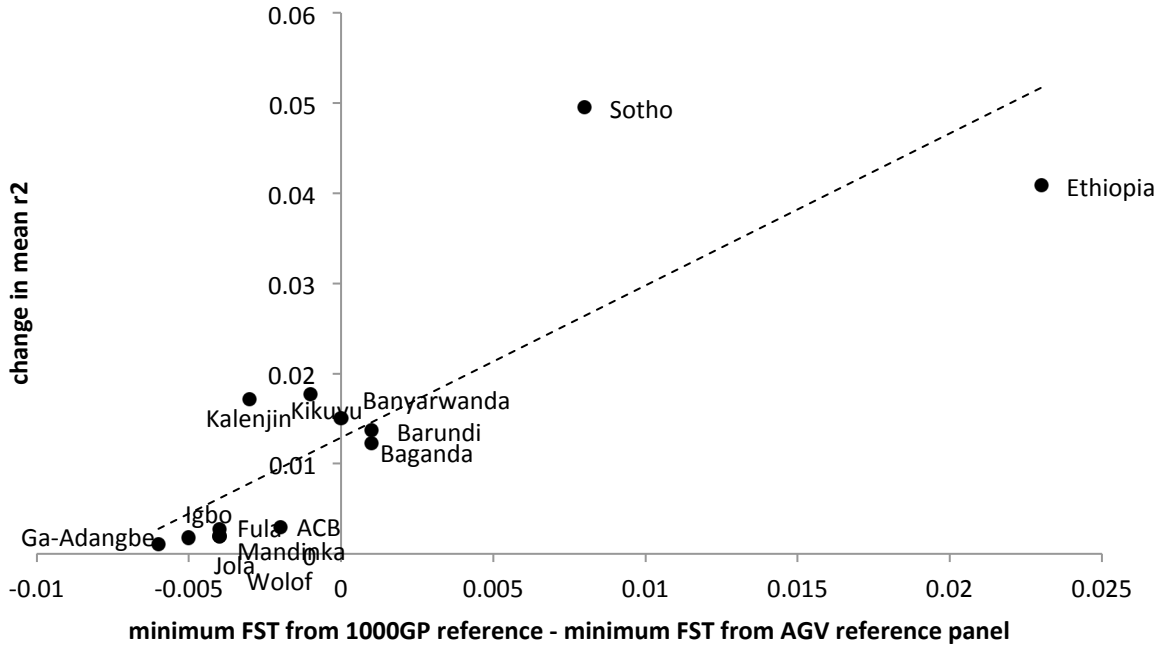
We show in Supplementary Note 10 that imputation accuracy is highly correlated with representation of haplotypes in a given reference panel. Existing reference panels have focused largely on European and Asian population groups, with very limited representation of African populations. The most commonly used reference panel currently is the 1000 Genomes phase I whole genome sequence reference panel, that includes only ~185 individuals from two African populations (YRI, and LWK), both belonging to the same broad ethno-linguistic group (Bantu speaking), with limited differentiation (F_{ST}=0.007). We also note that subsequent phases of the 1000 Genomes project include very similar population sets from West Africa, with only Bantu speaking ethno-linguistic groups being represented. In order to extend this reference panel, we included WGS data from 320 individuals belonging to 7 different ethno-linguistic groups (Baganda, Zulu, Oromo, Somali, Amhara, Gumuz and Wolayta). Several of these groups represent Afro-Asiatic populations that have not previously been represented in any reference panel. Additionally, we included Baganda and Zulu individuals as these have non-trivial HG (Mbuti Pygmy and Khoe-San) ancestry, and represent distinct regions in Africa. The generation of this reference panel has been described previously (**Supplementary Note 9**).

Following generation of the merged reference panel (with the 1000 Genomes Project sequencing panel), we carried out imputation into all AGVP populations except those included in the reference panel. We also included 18 Ethiopian individuals and 74 Bagandan individuals who were included in the genotyped set, but did not have corresponding sequence data in the reference panel. With this new reference panel, we observed variable improvement in imputation accuracy across the AGVP populations. There was very slight improvement in imputation for populations from West Africa that were well represented by the existing 1000 Genomes reference panel. By contrast, we observed marked improvement in imputation accuracy among Ethiopian and South African populations, across the whole range of the allele frequency spectrum. We hypothesise that this improvement is likely to be due to poor representation of haplotypes within these populations in current reference panels. There are no populations representing an Afro-Asiatic ethno-linguistic group in the 1000 Genomes panel, including in later phases of the project. Sotho is a South African Bantu population with ~20% HG admixture from Khoe-San populations, which are not represented in existing panels. Inclusion of such divergent haplotypes from the Zulu population in our reference panel is likely to have contributed to the marked improvement in imputation accuracy seen in Sotho. Notably, we only saw modest improvement in imputation Baganda and other Ugandan populations, in spite of inclusion of 94 Bagandan individuals in the reference set. This is likely to be due to

relatively good representation of these haplotypes by LWK individuals in the 1000 Genomes reference panel. We also evaluated whether a collective score indicating relative distances from the two reference panels (AGV and 1000 Genomes Project) was a good determinant of imputation accuracy (**SN11 Fig 1**). We found high correlation between these two parameters (r2=0.70), providing further evidence that representation of more divergent haplotypes in sequencing panels is essential to improve accurate capture of variation in populations across Africa.

These findings collectively argue against the need for population-specific sequencing to obtain high imputation accuracy among several populations, as we find that several populations can be well represented by a single population group, particularly for Bantu populations that do not show much differentiation. Instead, a more efficient strategy is likely to be to sequence diverse population sets to capture divergent haplotypes, e.g. HG haplotypes that are likely to be widespread in several population groups. In addition to providing a novel resource to researchers for use in imputation, we additionally provide important insights to inform the design of future large-scale sequencing efforts to develop resources for researchers.

SN11 Fig 1: Improvement in imputation accuracy with AGVP panel as a function of representation in reference panels



SN11 Fig 1 represents the correlation between the improvement in imputation accuracy with the merged reference panel as a function of the distance from the AGVP and 1000 Genomes reference panel. A combined metric of relative differentiation between the two panels appears to be the most important determinant of imputation accuracy.

S. NOTE 12: AN EVALUATION OF ULTRA-LOW COVERAGE SEQUENCING AND GENOTYPE ARRAY DESIGNS IN AFRICA

SN12.1 RATIONALE FOR EVALUATION

Several studies have evaluated extremely low coverage sequencing in relation to low coverage sequencing and genotype array designs, to develop potentially cost-effective and efficient way to capture common variation in European populations for genome wide association studies.[46,48] While these designs have been found to be cost-effective and reasonably sensitive in relation to low coverage sequencing, the utility of such designs, and current genotype arrays have not been assessed in African populations. Although costs of whole genome sequence are declining, current costs still restrict sequencing of samples for GWAS on a large scale; however, with further drops in costs, genotype arrays are becoming very cost-effective and practical for large-scale GWAS in Africa. In **Supplementary Note 10**, we discussed the imputation accuracy using the Illumina Omni 2.5 M array in African populations, highlighting the utility of existing chips in Africa. Here, we extend on these findings to discuss the utility of this chip array for capture of common variation as compared to low coverage sequencing data. We also assess ULC designs in relation to 4x and 8x sequencing designs, and assess the efficacy of various designs for a fixed budget study. Examining this has broader implications for developing large-scale frameworks for GWAS in Africa.

SN12.2 CURATION OF DATA

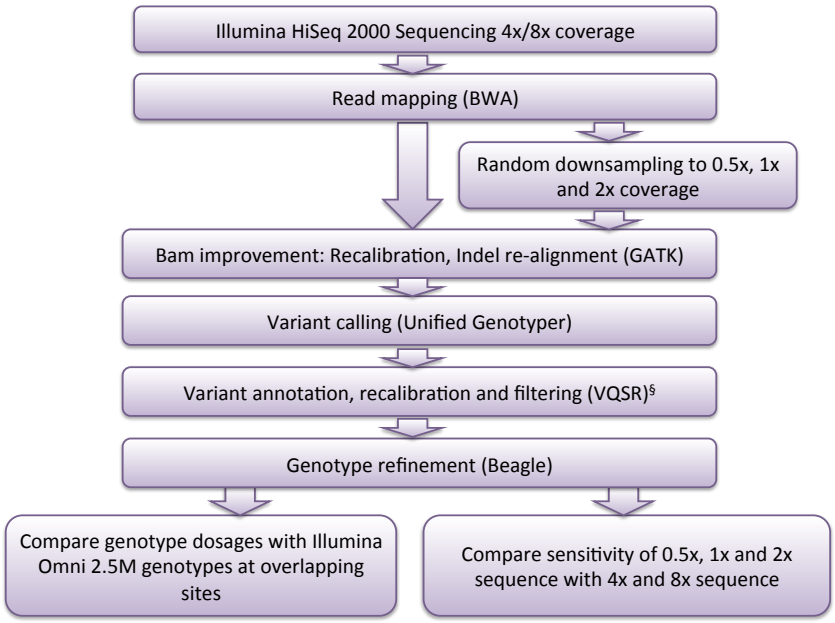
In order to generate data simulating ultra low coverage (ULC) sequence data, we randomly downsampled low coverage whole genome sequence data from 3 populations – Zulu (4x), Baganda (4x) and Ethiopia (4x and 8x) to 0.5x. 1x and 2x average depth using GATK. Downsampling was carried out by proportionately downsampling each sample, so that the overall distribution of coverage depths across samples was maintained. Downsampled data, like 4x coverage sequence data, were called across all 320 individuals from 3 populations, as described in **Supplementary Note 9**. Filtering was carried out using VQSR applying the same threshold as with 4x data (**Supplementary Note 9**). This was followed by genotype refinement across all samples with Beagle. As 8x coverage data was available for Ethiopian samples alone, this dataset of 120 individuals was curated separately, while 4x data was called with other populations as described previously.

SN12.3 Comparing ULC with low coverage sequencing designs

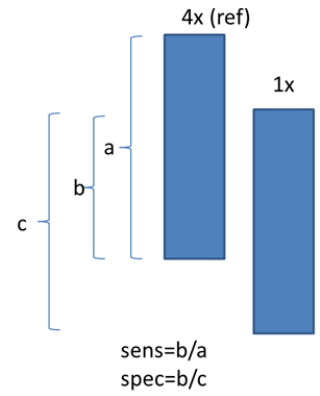
In order to evaluate the utility of these designs, we relied on two approaches (**SN12 Fig 1**):

1. First, the curated set of variants generated in each population were compared to higher coverage data. For Baganda, and Zulu, we compared these to variants called from 4x data, while for Ethiopia, we were able to compare to 8x data, as higher coverage sequencing data were available for this population. We used two metrics to compare downsampled and imputed chip data to 4x/8x low coverage sequencing designs, as has been done before:

SN12 Fig 1. Workflow for evaluation of ULC designs



SN12 Fig 1 outlines the workflow for curation of low coverage and ULC WGS for comparison. All three populations were called together for 4x, 2x, 1x, and 0.5x coverage. For 8x coverage, 120 Ethiopian samples were called separately using the same workflow. We compared ULC sequencing data with low coverage sequencing data regarding sensitivity of capture of variatnts. To further assess accuracy of genotype capture, we compared genotype calls with those from the Illumina Omni 2.5 genotype array for overlapping sample sets



Overall sensitivity= $= \frac{\sum_{i=1}^{n} sens_i}{n}$

Overall specificity= $= \frac{\sum_{i=1}^{n} spec_i}{n}$

SN12 Fig 2: Sensitivity and Specificity of ULC designs as compared to 4x and 8x sequencing designs.

If the blue boxes represent the pool of variants in the reference sample (4x/8x), and the test sample (0.5x/1x/2x) respectively, then sensitivity is the proportion of variants which are accurately captured by the test sample relative to the reference sample. This metric is then averaged across all samples. By accurate capture, we imply that variants should not only be present, but genotypes should be accurately captured as captured as homozygotes or heterozygotes relative to the reference sample. Specificity, in contrast, represents the proportion of the test sample that is captured accurately by variation in the reference sample, averaged over all samples.

a. Sensitivity: this represents the proportion of genotypes in a reference sample of 4x/8x low coverage whole genome sequence that are accurately captured by the design

specified (extremely low coverage/imputed Omni 2.5 M array), averaged across all samples (**SN12 Figure 2**).

b. Specificity: this represents the proportion of genotypes on the test sample that are concordant with the reference sample (4x/8x) (**SN12 Fig 2**).

Here, sensitivity can be seen as a composite of such metrics to quantify identification of genetic variation across the genome, as well as the accuracy in determining the correct genotypes. Specificity can be considered a metric for the false positive rate, as (1- Specificity), which represents the proportion of variation in a test sample that is not concordant with the reference sample.

2. As genotype data were available for a subset of sequence samples (**SM Figure 2**), we also assessed accuracy of each design by examining the correlation between refined data from ULC designs with genotype data on the 2.5M Omni chip array. We assessed correlation between genotype dosages from ULC data after refinement, and the chip at overlapping sites. This correlation metric is useful as this can be directly translated to reflect the power a given genome wide association study may have for a fixed budget, as we discuss later.[46]

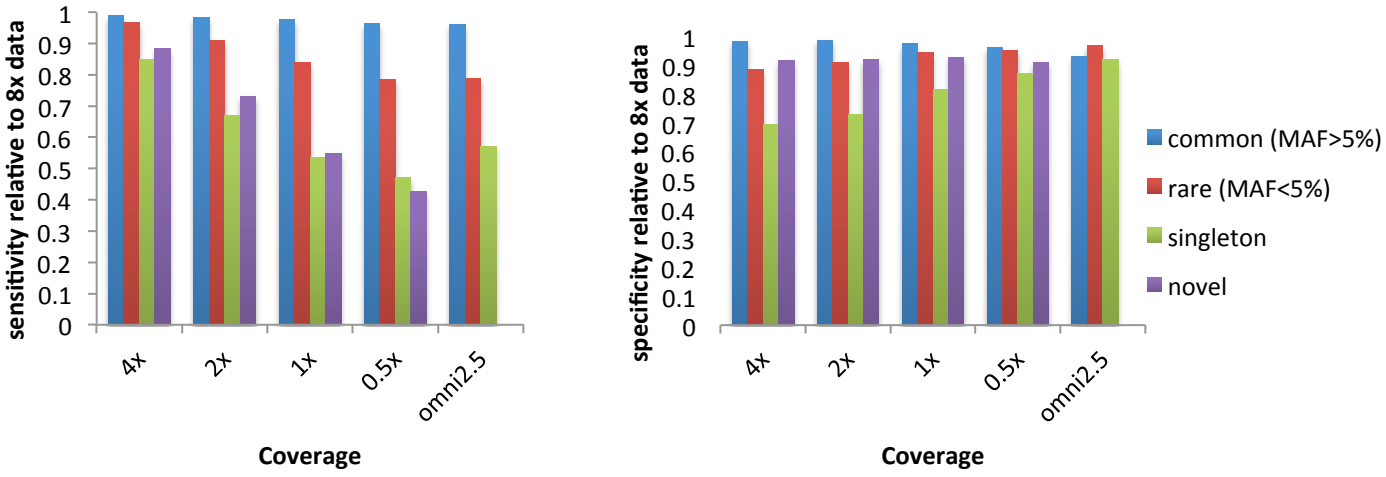
SN12.4 COMPARING GENOTYPE ARRAY DESIGNS WITH WHOLE GENOME SEQUENCING DESIGNS

Given the limited infrastructure for large-scale whole genome sequencing and curation of data in Africa, it would be helpful to evaluate the utility of genotype arrays in capturing genetic variation relative to whole genome sequence using existing imputation panels. We sought to assess this for the Illumina Omni 2.5M array. In **Supplementary Note 10**, we described the imputation accuracy of this array with existing reference panels. We calculate this metric by assessing the correlation between the chip genotypes, and imputed dosages at each site on the chip using 'leave one out masking', as outlined in IMPUTE2 (https://mathgen.stats.ox.ac.uk/impute/impute_v2.html). In this method, each site is sequentially masked, and imputed as if it were missing, allowing direct comparison between genotypes on the array and imputed genotypes, providing a metric for imputation accuracy at each site. In order to evaluate the sensitivity of genotype array and imputation designs with low coverage sequence designs, we calculated the sensitivity and specificity of the Omni 2.5M genotypes imputed with the 1000 Genomes reference panel relative to 4x and 8x sequencing in the populations studied.

The sensitivity of all ULC designs was >95% for accurate capture of common variation with respect to 8x for Ethiopians and 4x sequence data for all populations (**SN12 Fig 3** and **4**). For variants with a MAF <5%, the sensitivity varied between 73% to 81% with respect to higher coverage (4x/8x) designs. As expected, sensitivity for capture of singletons and novel variation was relatively poor (**SN12 Fig 3 and 4**). On comparison of Omni2.5M genotype array data imputed up to the 1000 Genomes panel, we found relatively high sensitivity for capture of common and rare variation, comparable to the 0.5x ULC design (**SN12 Fig 3 and 4**). The primary disadvantage of such a design would be the inability to capture novel variation, which may become less important as imputation panels become larger and more diverse. All designs were noted to have high specificity relative to higher coverage sequencing, suggesting a relatively low false positive rate, or errors correlated between ULC and higher coverage designs (**SN12 Fig 3 and 4**).

Our findings suggest that ULC and existing dense genotyping designs perform very well for capture of common variation in Africa, even when existing imputation panels are used for genotype refinement or imputation. This could potentially open up new possibilities for cost-effective large-scale genetic studies in the region. We expect our results to be further improved by the development of larger and more diverse reference panels for genotype refinement of sequence data, and imputation into genotype arrays. We subsequently assess the cost-effectiveness of these designs.

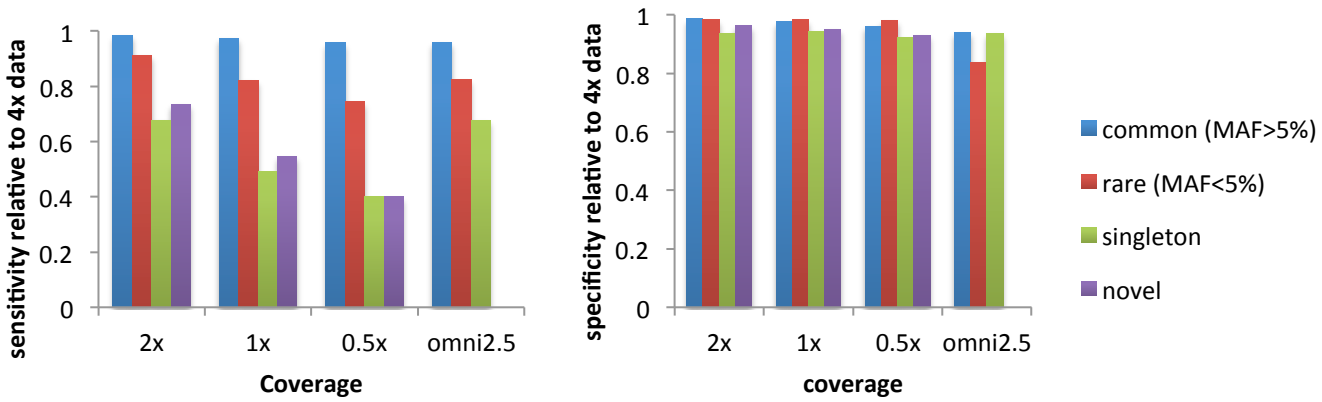
SN12 Fig 3: Sensitivity and specificity of different designs relative to 8x coverage data



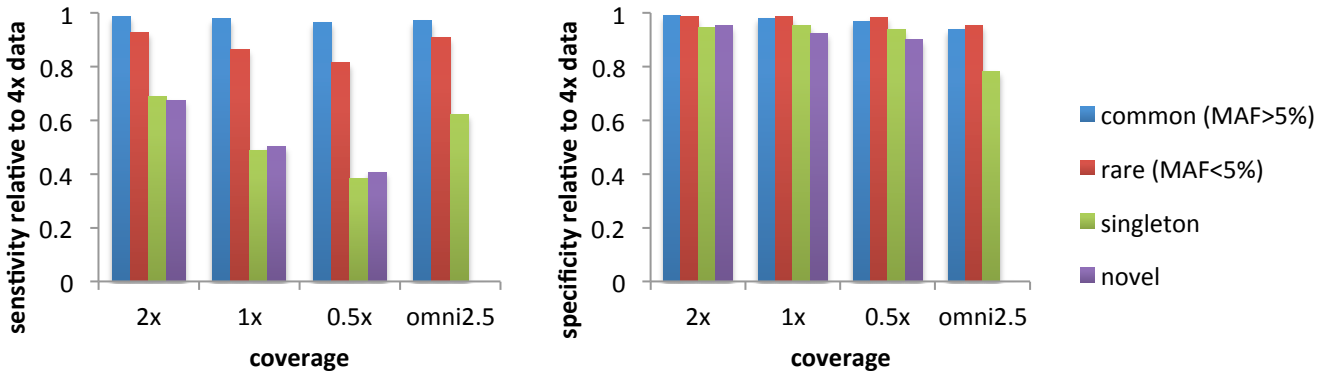
SN12 Fig 3 represents the sensitivity and specificity of different designs (2x, 1x, 0.5x ULC sequencing, Omni 2.5M genotyping with imputation) with respect to 4x WGS among Ethiopian individuals.

SN12 Fig 4: Sensitivity and specificity of ULC and imputed Omni 2.5M chip array relative to 4x data among different populations

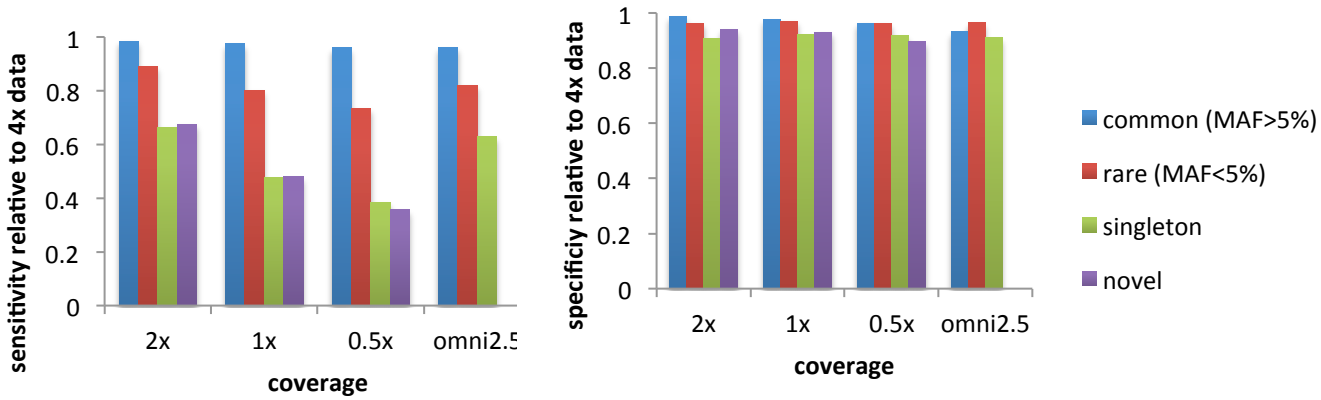
a) Zulu



b) Uganda



c) Ethiopia



SN12 Fig 4 shows the sensitivity and specificity (as defined in the text) for different ULC sequencing designs, and imputed data from the omni 2.5M genotype array relative to 4x WGS data from different populations.

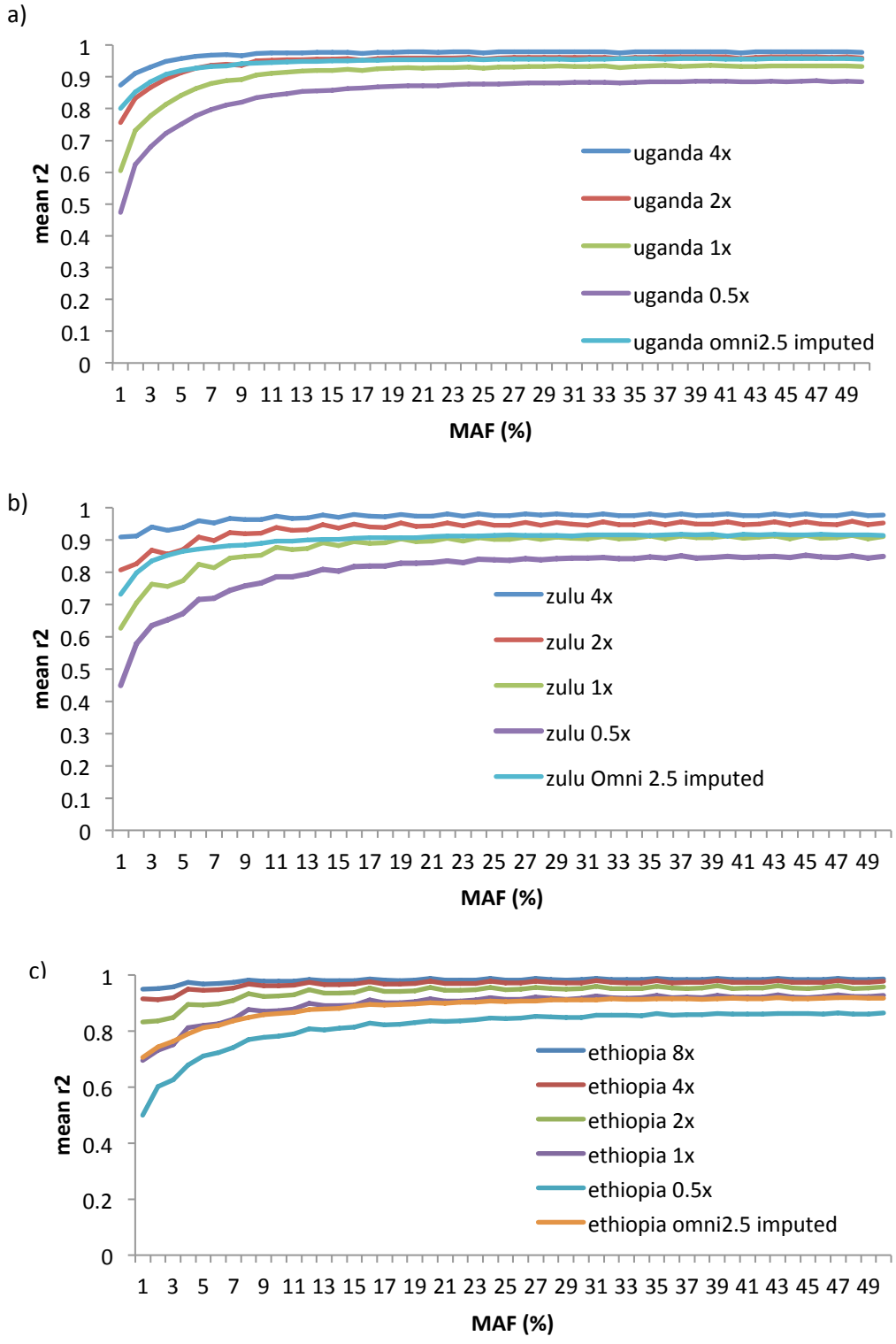
SN12.5 Assessing cost-effectiveness of different study designs

In order to compare cost-effectiveness among different designs, we assumed a fixed budget GWAS study. Effective sample size was calculated as the product of the number of samples that could be sequenced and called at a given read depth (N) with a fixed budget, and the accuracy of genotype calling of sequence data (measured as the r^2 between sequence and chip genotype data). We included several scenarios for evaluation of the number of samples that could be sequenced and processed for a given budget. For the Omni 2.5M genotype array, the r^2 metric obtained by imputation was used for calculation of effective sample size, as this represents the accuracy for variants imputed onto the chip in a given allele frequency bin. We acknowledge that the metrics for rare variation obtained from imputation onto the Omni 2.5M genotype array may be affected by SNP ascertainment, and should therefore be viewed as upper limits. However, ascertainment is unlikely to substantially influence the correlations we obtain for common variant.

We find high correlations between 4x and 2x data and Omni 2.5M genotypes among all populations (**SN12 Fig 5**). While, correlations are lower for lower coverage designs, for rare and common variation for all populations (**SN12 Fig 5**), we find that the effective sample size for these designs, as well as the Omni 2.5M genotype array is still much greater, as the higher sample size for a given fixed budget more than compensates for the inaccuracy in identifying true genotypes; hence leading to a greater effective sample size for a study, and therefore greater efficiency for examination of common and rare variation (**SN12 Fig 6**).

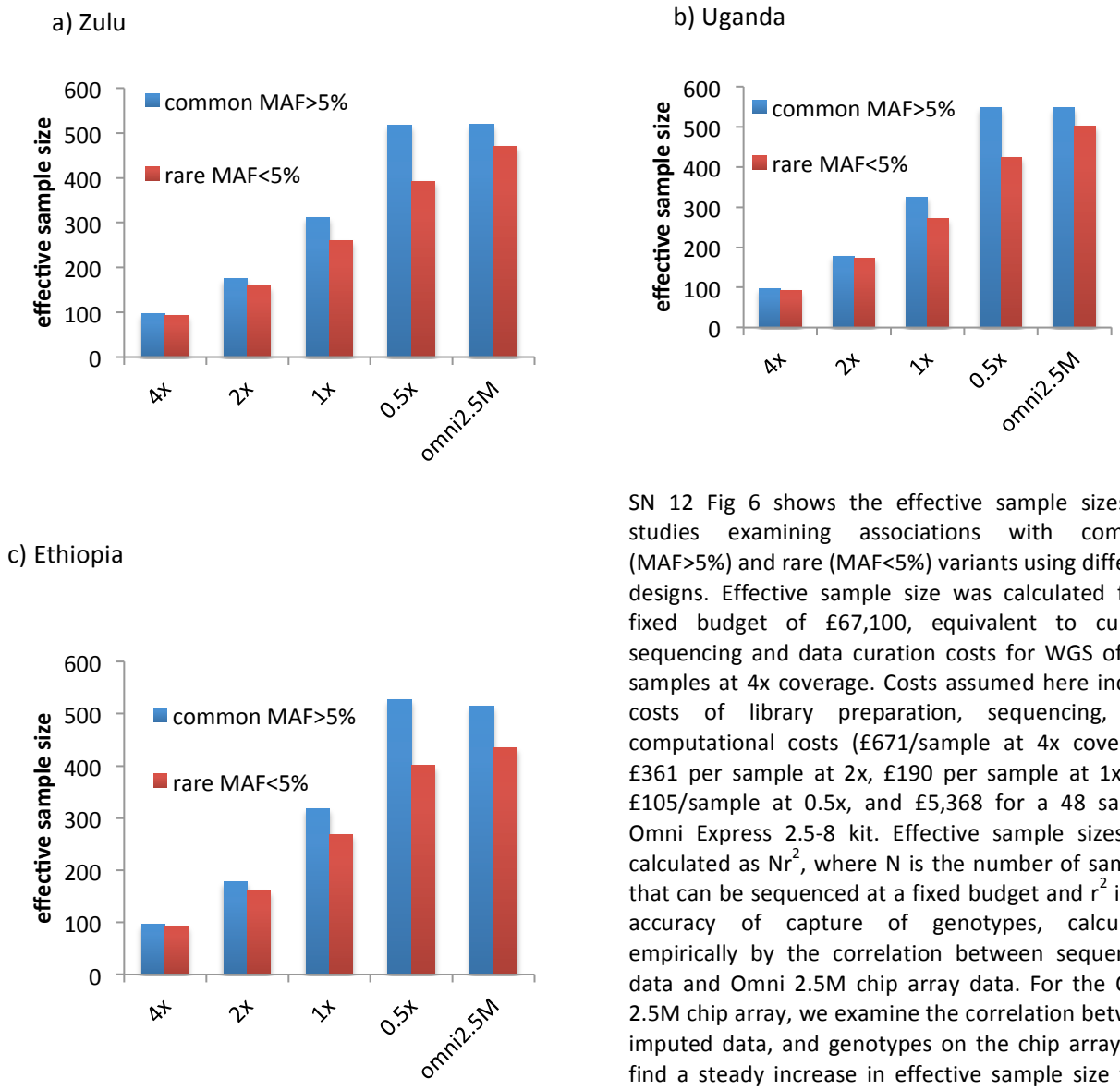
Our findings in African populations are consistent with findings of higher efficiency for GWAS of common variants with ULC sequencing among European populations. Furthermore, we present the first evidence to suggest that given computational costs and infrastructure needed for analysis of sequencing data, dense and cost-effective genotype arrays may provide useful and scalable alternatives for the study of genetic variation associated with disease in Africa. We discuss this in the subsequent section.

SN12 Fig 5 correlation between different designs and omni 2.5 M genotypes



SN 12 Fig 5 a), b), and c) represent the correlations between different designs and overlapping genotype data from the Omni 2.5M array among Uganda, Zulu and Ethiopia respectively. The line labelled "Omni 2.5M imputed" for each design represents the correlation between data imputed from the Omni 2.5M array using the 1000 Genomes Project phase 1 v3 integrated panel, and the original data genotyped on the chip array

SN 12 Fig 6: Effective sample sizes for different designs calculated from correlations in different African populations



SN 12 Fig 6 shows the effective sample sizes for studies examining associations with common (MAF>5%) and rare (MAF<5%) variants using different designs. Effective sample size was calculated for a fixed budget of £67,100, equivalent to current sequencing and data curation costs for WGS of 100 samples at 4x coverage. Costs assumed here include costs of library preparation, sequencing, and computational costs (£671/sample at 4x coverage, £361 per sample at 2x, £190 per sample at 1x and £105/sample at 0.5x, and £5,368 for a 48 sample Omni Express 2.5-8 kit. Effective sample sizes are calculated as Nr^2, where N is the number of samples that can be sequenced at a fixed budget and r^2 is the accuracy of capture of genotypes, calculated empirically by the correlation between sequencing data and Omni 2.5M chip array data. For the Omni 2.5M chip array, we examine the correlation between imputed data, and genotypes on the chip array. We find a steady increase in effective sample size for a fixed budget, with lower sequencing coverage, with Omni 2.5M designs performance broadly equivalent to 0.5x sequencing.

S. NOTE 13: EVALUATION OF A CHIP DESIGN SPECIFIC TO AFRICA

With next generation genotyping, and the decline in genotyping costs, it has now been possible to carry out large-scale dense genotyping across individuals for genome wide association studies. While these chip designs capture variation in European populations very well, their utility in capturing variation in more diverse African populations is unclear. In **Supplementary Note 12**, we show that existing chip designs with imputation using large-scale sequencing panels can capture common variation well relative to low coverage sequencing designs. We also show marked improvement in imputation accuracy, even for common variation, with addition of a novel AGVP reference panel. We hypothesise that development of more efficient chip designs specific to African populations would complement the development of novel reference panels, and improve efficiency of variant capture by genotyping arrays even further. While existing chip designs have included African populations in the ascertainment set, European populations have dominated development of these arrays, and African populations in current reference panels are not representative of more differentiated population groups within Africa. Here, for the first time, we describe the development of such an array, ascertained on relatively large samples from distinct population groups across Africa. We describe an array that captures common genetic variation (MAF>5%) across 5 geographically distinct populations in Africa, including the YRI, Zulu, Baganda, LWK and Ethiopia. In order to capture variation across all populations, we developed an in-house algorithm to maximise efficiency of such a chip by applying cycles of tagging and imputation, as has been described before.[27]

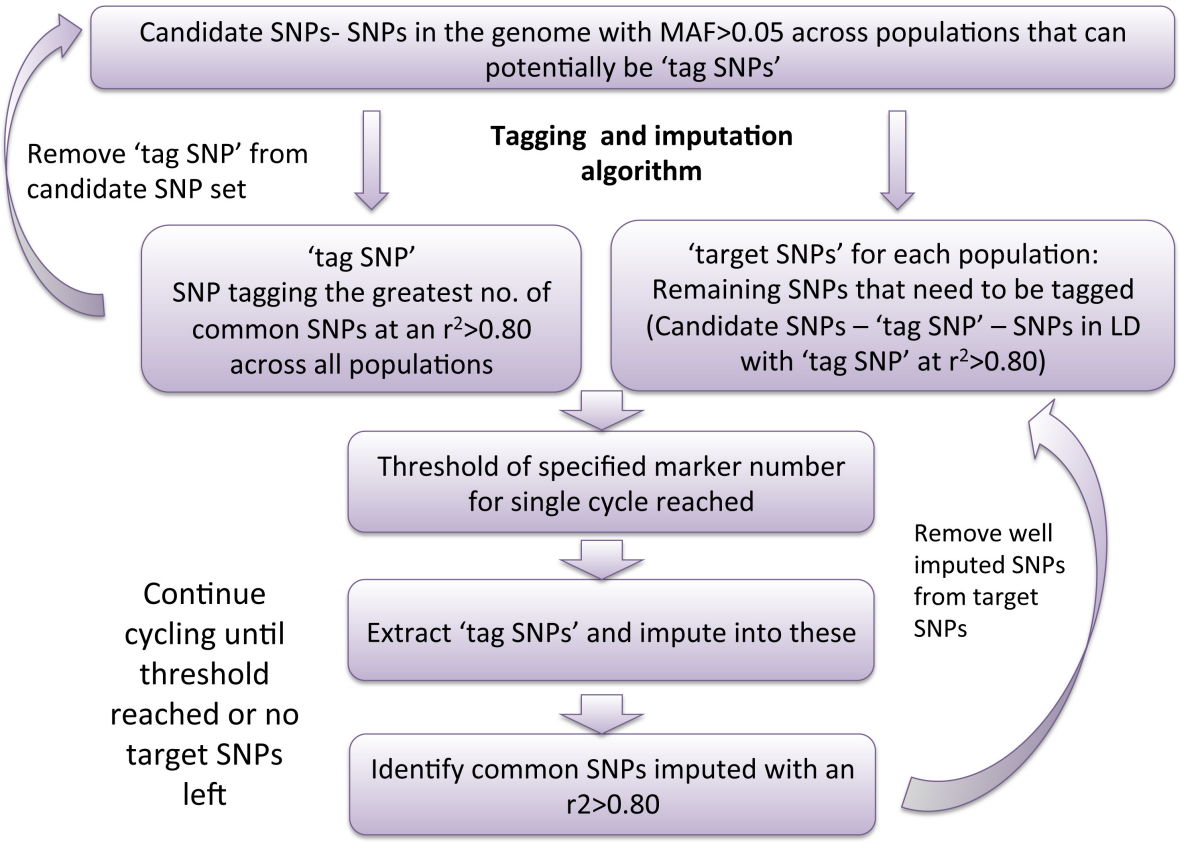
We first developed a multi-population tagging algorithm based on the algorithm TAGster for WGS data. [26] The methods we used for tagging were identical to those used by TAGster; however, by using seeking and indexing approaches we were able to optimise the computational efficiency of the algorithm by an order of magnitude (unpublished data, Carstensen et al.). We briefly outline the tagging algorithm as follows (**SN13 Fig 1**):

1. Calculate LD (r^2) between each SNP and all other SNPs in the flanking 250 KB region across all 5 populations. MAF thresholds are imposed at this stage, and only pairs of SNPs where both exceed the MAF threshold are included.
2. For each SNP not already in the tagging set, a count of SNPs in the target set that are in LD> exceeding a given threshold r_t^2 with it is generated across the genome.
3. The most informative SNP (the SNP with most target SNPs in LD with it) is chosen as the tagging SNP and added to the set of tagging SNPs.

4. This tagging SNP and SNPs in LD with it are now removed from the set of target SNPs. However, SNPs in LD with the tagging SNP can still be picked up as tagging SNPs themselves if they independently tag the maximum no. of SNPs in any iteration.

Steps 2-3 are repeated until either a specified number of SNPs or all target SNPs (chosen as SNPs above a specific MAF threshold) are tagged across all population sets, or until a specific number of SNPs is reached, as specified.

SN13 Fig 1: Hybrid algorithm for tagging and imputation applied for chip design



Here, we have referred to 'tagging SNPs' as those that tag other SNPs above a specified LD threshold, and target SNPs as the SNPs we hope to capture at a given threshold of LD by tagging.

As single marker tagging can be inefficient, we sought to improve efficiency by introducing cycles of tagging and imputation, as described previously.[27] We carried out tagging across 505 individuals from 5 populations (Zulu, Baganda, Ethiopia, YRI and LWK) at an LD threshold of 0.8 to capture markers at an MAF >0.05, until a given threshold of markers was reached. Following this, we assessed coverage across the genome, by carrying out an imputation step. For imputation, we used a combination of the 1000 Genomes and AGV reference panel. However, as individuals from all five populations were also represented in the reference panel, we randomly sampled half the individuals from the five populations, and carried out imputation into these individuals, retaining the remaining individuals in the reference panel. In this way, we were able to efficiently identify multi-marker tagged variation in all five populations. We then removed all variants captured at an r2>0.8 from the target set. This hybrid approach would be likely to improve the efficiency of markers captured, as imputation is likely to capture a greater number of markers at a given LD threshold in comparison with single-marker tagging. In this way, we are able to remove additional markers that have been captured by imputation from the target set in each cycle. We carry out several such cycles, until 1M tagging SNPs were identified across the 5 populations. We compared results from the hybrid method with just pairwise tagging, in addition to comparing 5 cycles and 10 cycles for imputation. Coverage was calculated as the proportion of markers in different allele frequency bins that were captured by the tagging set in each population at an r2>0.8.

Our coverage for common variation >5% across all populations using 1M tagging variants was noted to be high (93%), (**Extended Data Figure 10**) suggesting that a 1M density array could potentially capture substantial proportions of variation across Africa. Using the hybrid approach improved the efficiency of tagging (**Extended Data Figure 10**), and using 10 cycles further improved the efficiency over using fewer cycles (data not shown).

To our knowledge this is the first comprehensive evaluation of a chip design for capture of common variation across 505 individuals from 9 ethno-linguistic groups representative of many population groups within Africa. We show for the first time that a 1M chip array design is theoretically feasible for sensitive capture of common variation, and potentially scalable to large-scale GWAS in Africa. These findings have important implications for the design of large-scale genomic studies of health and disease in Africa.

References

1 Consortium, H. A. Research capacity. Enabling the genomic revolution in Africa. *Science* **344**, 1346-1348, doi:10.1126/science.1251546 (2014).

2 Genomes Project, C. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65, doi:10.1038/nature11632 (2012).

3 Teo, Y. Y. *et al.* A genotype calling algorithm for the Illumina BeadArray platform. *Bioinformatics (Oxford, England)* **23**, 2741-2746 (2007).

4 Schlebusch, C. M. *et al.* Genomic variation in seven Khoe-San groups reveals adaptation and complex African history. *Science* **338**, 374-379, doi:10.1126/science.1227721 (2012).

5 Henn, B. M. *et al.* Genomic ancestry of North Africans supports back-to-Africa migrations. *PLoS genetics* **8**, e1002397, doi:10.1371/journal.pgen.1002397 (2012).

6 Henn, B. M. *et al.* Hunter-gatherer genomic diversity suggests a southern African origin for modern humans. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 5154-5162, doi:10.1073/pnas.1017511108 (2011).

7 Pickrell, J., Patterson N, Loh P, Lipson M, Berger B, Stoneking M, Pakendorf B, Reich D. Ancient west Eurasian ancestry in southern and eastern Africa. *unpublished* (2013).

8 Nielsen, R., Korneliussen, T., Albrechtsen, A., Li, Y. & Wang, J. SNP calling, genotype calling, and sample allele frequency estimation from New-Generation Sequencing data. *PloS one* **7**, e37558, doi:10.1371/journal.pone.0037558 (2012).

9 Han, E., Sinsheimer, J. S. & Novembre, J. Characterizing bias in population genetic inferences from low-coverage sequencing data. *Molecular biology and evolution* **31**, 723-735, doi:10.1093/molbev/mst229 (2014).

10 Price, A. L. *et al.* Long-range LD can confound genome scans in admixed populations. *American journal of human genetics* **83**, 132-135; author reply 135-139, doi:10.1016/j.ajhg.2008.06.005 (2008).

11 Bhatia, G., Patterson, N., Sankararaman, S. & Price, A. L. Estimating and interpreting FST: The impact of rare variants. *Genome research* **23**, 1514-1521, doi:10.1101/gr.154831.113 (2013).

12 Patterson, N. *et al.* Ancient admixture in human history. *Genetics* **192**, 1065-1093, doi:10.1534/genetics.112.145037 (2012).

13 Thathy, V., Moulds, J. M., Guyah, B., Otieno, W. & Stoute, J. A. Complement receptor 1 polymorphisms associated with resistance to severe malaria in Kenya. *Malaria journal* **4**, 54, doi:10.1186/1475-2875-4-54 (2005).

14 Kosoy, R. *et al.* Evidence for malaria selection of a CR1 haplotype in Sardinia. *Genes and immunity* **12**, 582-588, doi:10.1038/gene.2011.33 (2011).

15 Genovese, G. *et al.* Association of trypanolytic ApoL1 variants with kidney disease in African Americans. *Science* **329**, 841-845, doi:10.1126/science.1193032 (2010).

16 Loh, P. R. *et al.* Inferring admixture histories of human populations using linkage disequilibrium. *Genetics* **193**, 1233-1254, doi:10.1534/genetics.112.147330 (2013).

17 Brisbin, A. *et al.* PCAdmix: principal components-based assignment of ancestry along each chromosome in individuals with admixed ancestry from two or more populations. *Human biology* **84**, 343-364, doi:10.3378/027.084.0401 (2012).

18 Delaneau, O., Zagury, J. F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nature methods* **10**, 5-6, doi:10.1038/nmeth.2307 (2013).

19 Wright, S. Evolution in Mendelian Populations. *Genetics* **16**, 97-159 (1931).

20 Holsinger, K. E. & Weir, B. S. Genetics in geographically structured populations: defining, estimating and interpreting F(ST). *Nature reviews. Genetics* **10**, 639-650, doi:10.1038/nrg2611 (2009).

21 Brisbin, A. *et al.* PCAdmix: Principal Components-Based Assignment of Ancestry Along Each Chromosome in Individuals with Admixed Ancestry from Two or More Populations. *Hum. Biol.* **84**, 343-364, doi:10.3378/027.084.0401 (2012).
22 Voight, B. F., Kudaravalli, S., Wen, X. & Pritchard, J. K. A map of recent positive selection in the human genome. *PLoS biology* **4**, e72, doi:10.1371/journal.pbio.0040072 (2006).
23 Pickrell, J. K. *et al.* Signals of recent positive selection in a worldwide sample of human populations. *Genome research* **19**, 826-837, doi:10.1101/gr.087577.108 (2009).
24 Sabeti, P. C. *et al.* Positive natural selection in the human lineage. *Science* **312**, 1614-1620, doi:10.1126/science.1124309 (2006).
25 Granka, J. M. *et al.* Limited evidence for classic selective sweeps in African populations. *Genetics* **192**, 1049-1064, doi:10.1534/genetics.112.144071 (2012).
26 Xu, Z., Kaplan, N. L. & Taylor, J. A. TAGster: efficient selection of LD tag SNPs in single or multiple populations. *Bioinformatics* **23**, 3254-3255, doi:10.1093/bioinformatics/btm426 (2007).
27 Hoffmann, T. J. *et al.* Design and coverage of high throughput genotyping arrays optimized for individuals of East Asian, African American, and Latino race/ethnicity using imputation and a novel hybrid SNP selection algorithm. *Genomics* **98**, 422-430, doi:10.1016/j.ygeno.2011.08.007 (2011).
28 Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *American journal of human genetics* **88**, 76-82, doi:10.1016/j.ajhg.2010.11.011 (2011).
29 Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nature genetics* **44**, 821-824, doi:10.1038/ng.2310 (2012).
30 Pruim, R. J. *et al.* LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* **26**, 2336-2337, doi:10.1093/bioinformatics/btq419 (2010).
31 Global Lipids Genetics, C. *et al.* Discovery and refinement of loci associated with lipid levels. *Nature genetics* **45**, 1274-1283, doi:10.1038/ng.2797 (2013).
32 Jallow, M. *et al.* Genome-wide and fine-resolution association analysis of malaria in West Africa. *Nature genetics* **41**, 657-665, doi:10.1038/ng.388 (2009).
33 Palmer, N. D. *et al.* Resequencing and analysis of variation in the TCF7L2 gene in African Americans suggests that SNP rs7903146 is the causal diabetes susceptibility variant. *Diabetes* **60**, 662-668, doi:10.2337/db10-0134 (2011).
34 Bertram, L., McQueen, M. B., Mullin, K., Blacker, D. & Tanzi, R. E. Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database. *Nature genetics* **39**, 17-23, doi:10.1038/ng1934 (2007).
35 Hinch, A. G. *et al.* The landscape of recombination in African Americans. *Nature* **476**, 170-175, doi:10.1038/nature10336 (2011).
36 Albrechtsen, A., Nielsen, F. C. & Nielsen, R. Ascertainment biases in SNP chips affect measures of population divergence. *Molecular biology and evolution* **27**, 2534-2547, doi:10.1093/molbev/msq148 (2010).
37 Reich, D. *et al.* Reconstructing Native American population history. *Nature* **488**, 370-374, doi:10.1038/nature11258 (2012).
38 Patin, E. *et al.* Inferring the demographic history of African farmers and pygmy hunter-gatherers using a multilocus resequencing data set. *PLoS genetics* **5**, e1000448, doi:10.1371/journal.pgen.1000448 (2009).
39 Prufer, K. *et al.* The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**, 43-49, doi:10.1038/nature12886 (2014).
40 Sereno, P. C. *et al.* Lakeside cemeteries in the Sahara: 5000 years of holocene population and environmental change. *PloS one* **3**, e2995, doi:10.1371/journal.pone.0002995 (2008).

41 Pickrell, J. K. & Reich, D. Toward a new history and geography of human genes informed by ancient DNA. *Trends in genetics : TIG* **30**, 377-389, doi:10.1016/j.tig.2014.07.007 (2014).
42 Kuper, R. & Kropelin, S. Climate-controlled Holocene occupation in the Sahara: motor of Africa's evolution. *Science* **313**, 803-807, doi:10.1126/science.1130989 (2006).
43 Price, A. L. *et al.* Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS genetics* **5**, e1000519, doi:10.1371/journal.pgen.1000519 (2009).
44 Hazleton, J. E., Berman, J. W. & Eugenin, E. A. Purinergic receptors are required for HIV-1 infection of primary human macrophages. *Journal of immunology* **188**, 4488-4495, doi:10.4049/jimmunol.1102482 (2012).
45 Chvatchko, Y. *et al.* The involvement of an ATP-gated ion channel, P(2X1), in thymocyte apoptosis. *Immunity* **5**, 275-283 (1996).
46 Pasaniuc, B. *et al.* Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. *Nature genetics* **44**, 631-635, doi:10.1038/ng.2283 (2012).
47 Huang, L. *et al.* Haplotype variation and genotype imputation in African populations. *Genetic epidemiology* **35**, 766-780, doi:10.1002/gepi.20626 (2011).
48 Flannick, J. *et al.* Efficiency and power as a function of sequence coverage, SNP array density, and imputation. *PLoS computational biology* **8**, e1002604, doi:10.1371/journal.pcbi.1002604 (2012).