SUPPLEMENTARY INFORMATION

doi:10.1038/nature13997

TABLE OF CONTENTS

SUPPLEMENTARY METHODS 2
1.0 CURATION OF AGVP GENOTYPE DATA 2
2.0 CURATION OF AGVP SEQUENCE DATA 9
3.0 ANALYSIS OF POPULATION STRUCTURE AND GENE FLOW 10
4.0 ANALYSIS OF LOCI UNDER SELECTION 17
5.0 EVALUATION OF LD STRUCTURE AND IMPUTATION ACCURACY IN AGVP 21
S. NOTE 1: CURATION OF AGVP GENOTYPE DATA AND REMOVAL OF CHIP EFFECTS 25
S. NOTE 2: ASSESSING ASCERTAINMENT BIAS ON THE OMNI2.5M ARRAY 56
S. NOTE 3: ESTIMATING EURASIAN ANCESTRY AMONG AGVP POPULATIONS 62
S. NOTE 4: ESTIMATING HUNTER-GATHERER (HG) ANCESTRY IN SSA 66
S. NOTE 5: EXPLORING ADMIXTURE AMONG AFRICAN POPULATIONS IN SSA 75
S. NOTE 6: EVALUATION OF MASKING OF EURASIAN ANCESTRY IN PCADMIX 88
S. NOTE 7: ASSESSMENT OF BACKGROUND SELECTION IN DIFFERENTIATED REGIONS 95
S. NOTE 8: LONG RANGE HAPLOTYPES UNDER SELECTION IDENTIFIED BY IHS 97
S. NOTE 9: CURATION OF AGVP SEQUENCE DATA AND REFERENCE PANEL 104
S. NOTE 10: IMPUTATION INTO GENOTYPE ARRAYS IN AFRICAN POPULATIONS 109
S. NOTE 11: IMPUTATION USING THE AGVP REFERENCE PANEL 114
S. NOTE 12: AN EVALUATION OF ULTRA-LOW COVERAGE SEQUENCING AND GENOTYPE ARRAY DESIGNS IN AFRICA 116
S. NOTE 13: EVALUATION OF A CHIP DESIGN SPECIFIC TO AFRICA 124

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

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

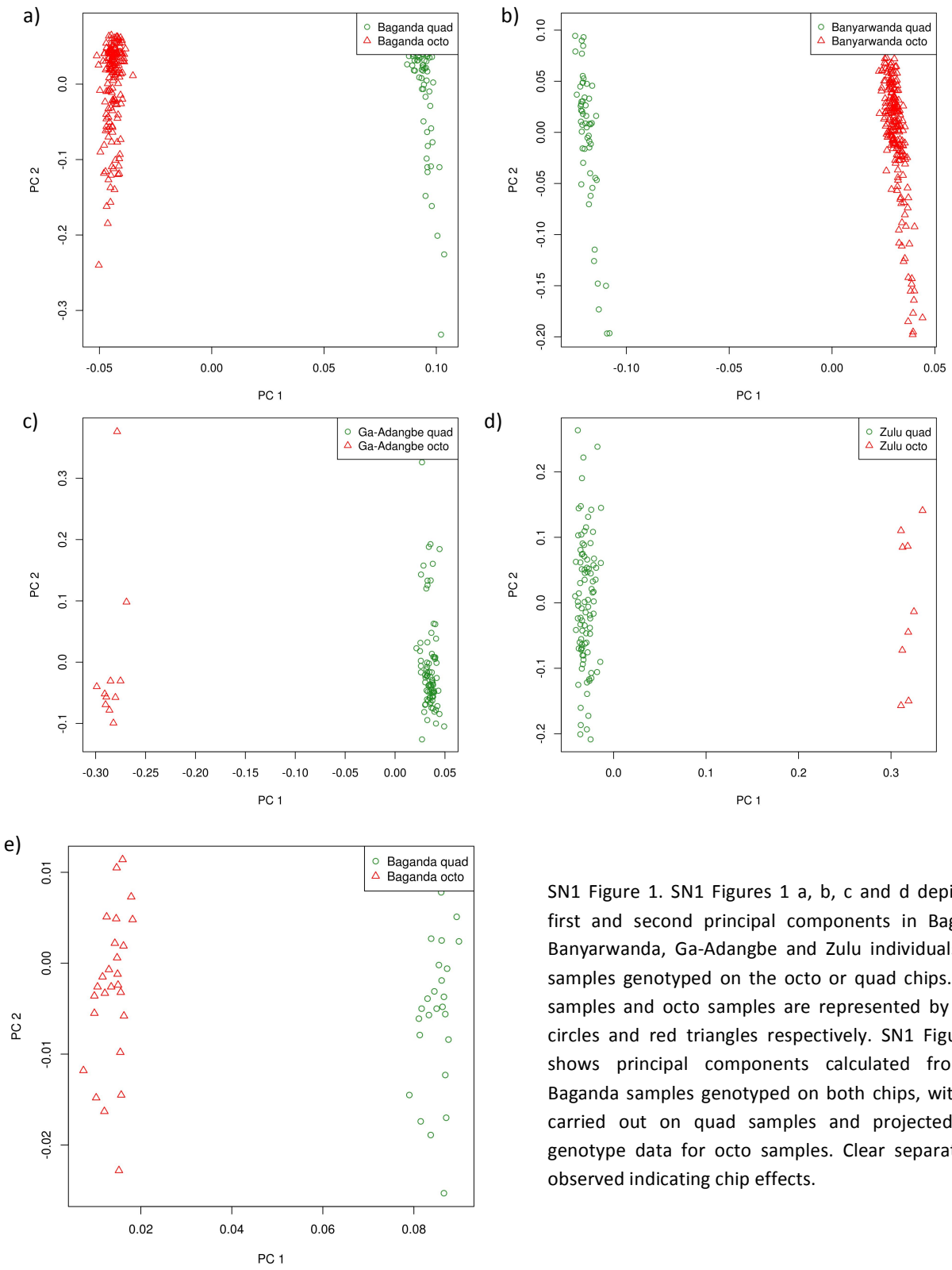
1.1 SAMPLES AND POPULATIONS

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,

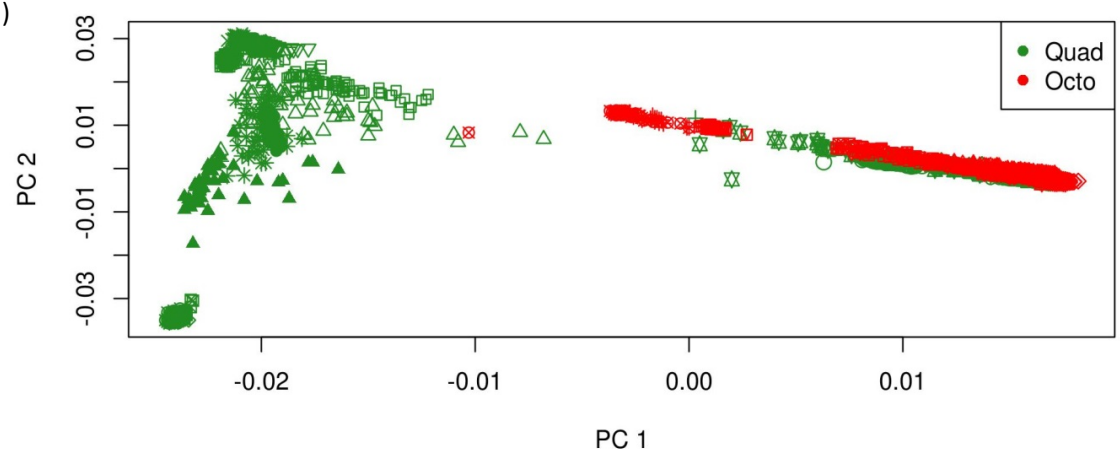
SN1. Figure 1: Chip effects apparent in African samples genotyped on quad and octo chips



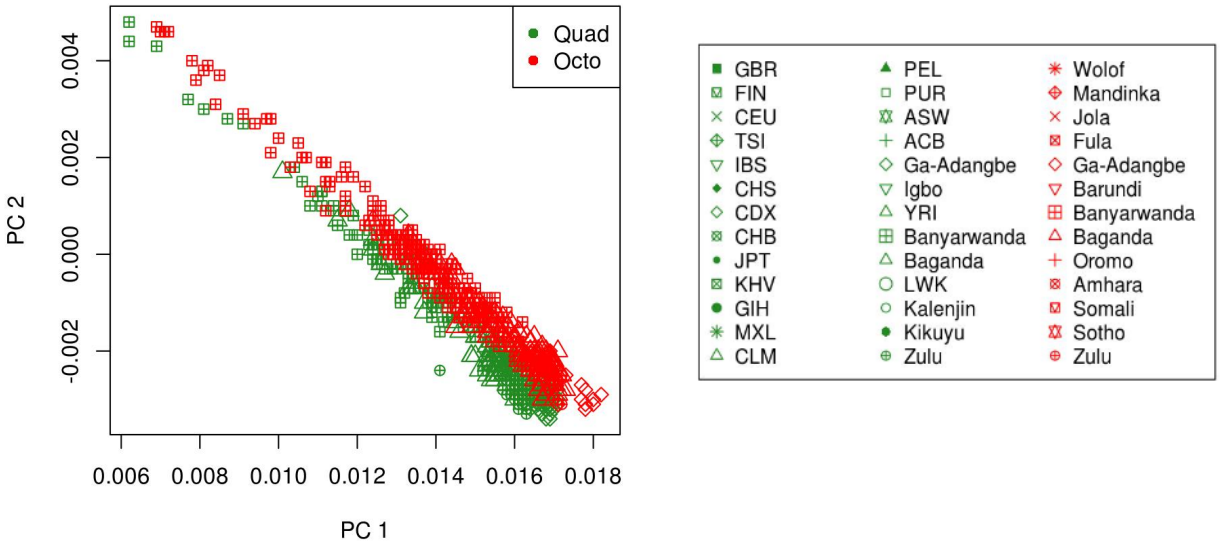
SN1 Figure 1. SN1 Figures 1 a, b, c and d depict the first and second principal components in Baganda, Banyarwanda, Ga-Adangbe and Zulu individuals with samples genotyped on the octo or quad chips. Quad samples and octo samples are represented by green circles and red triangles respectively. SN1 Figure 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)



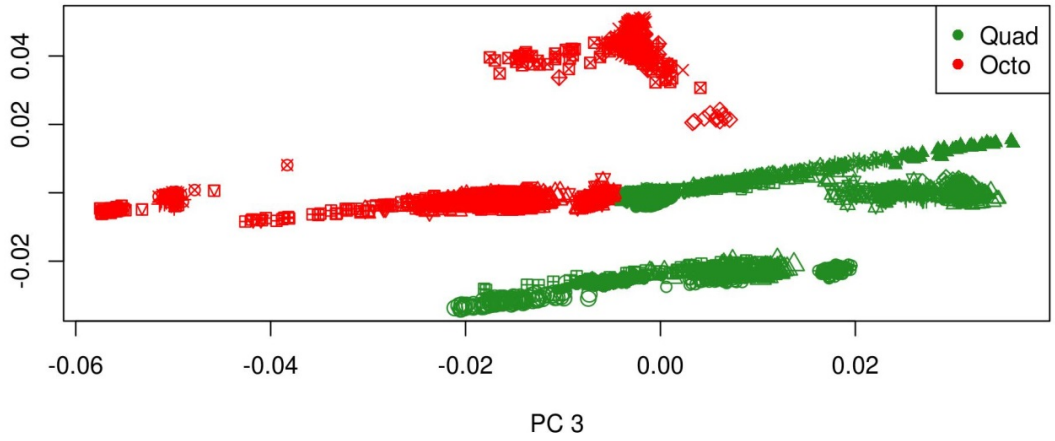
b)



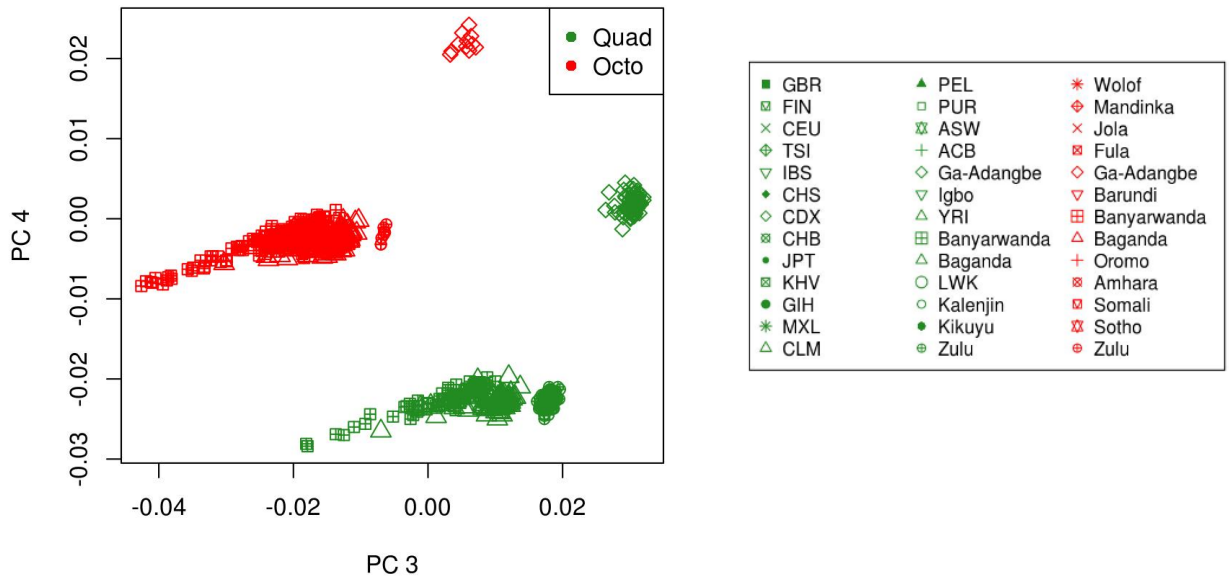
.

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.

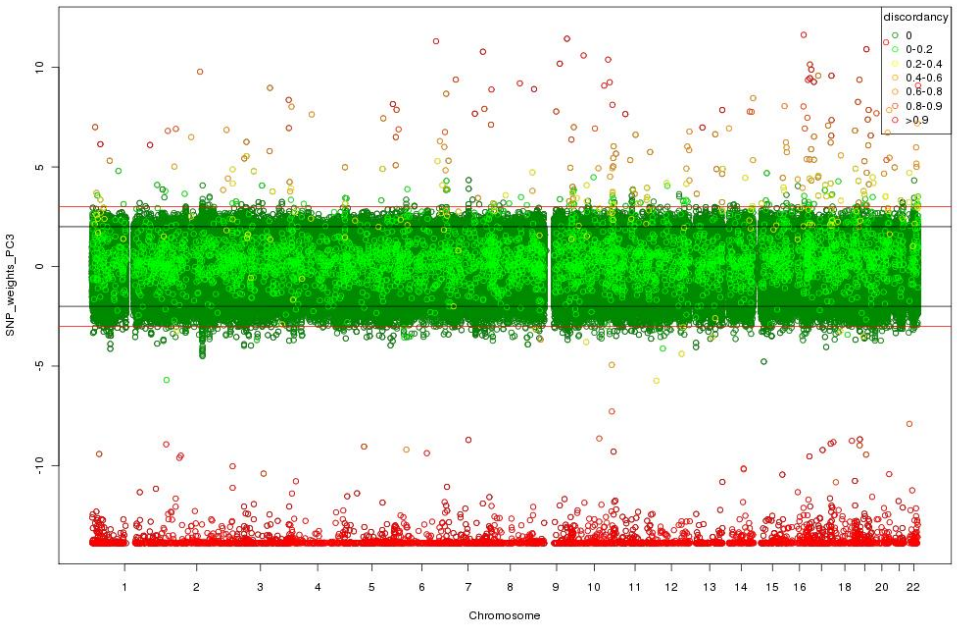
we carried out PCA on quad samples, projecting onto the octo chip, and confirmed that samples were identically projected along PCs. We also carried out clustering analysis in ADMIXTURE to identify chip related clustering of genetic data. Removal of SNPs representing chip effects among principal components 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

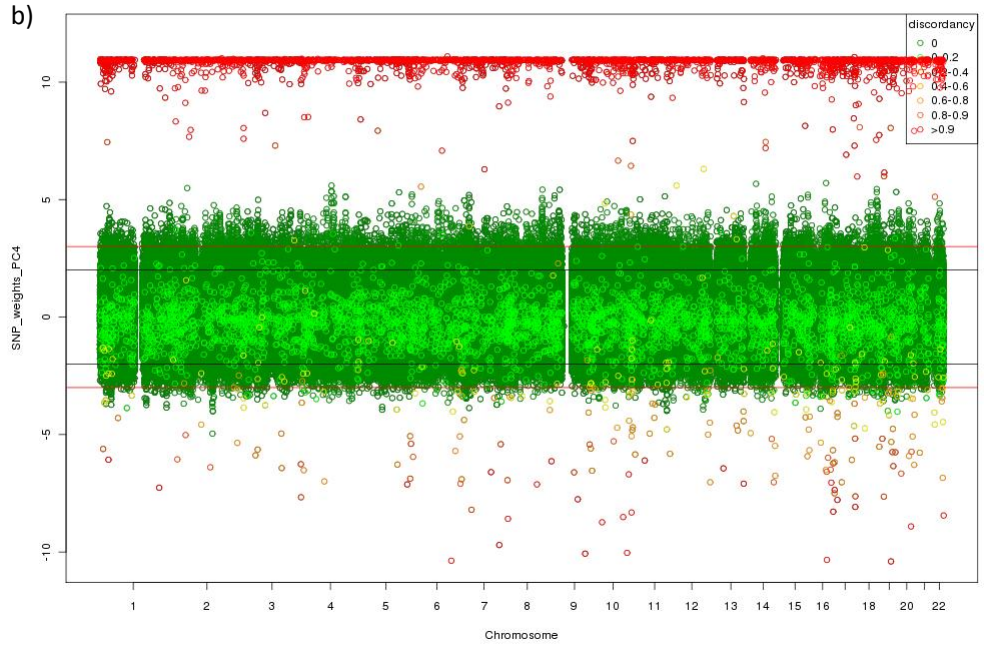
The global dataset was produced by merging genotype data from 33 different populations (**SM. Tables 4 and 5**) after completion of sample and SNP quality based filtering. PCA revealed chip effects between quad and octo genotyped data, evident as separation between chips along PCs 3 and 4 (**SN1 Figure 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)

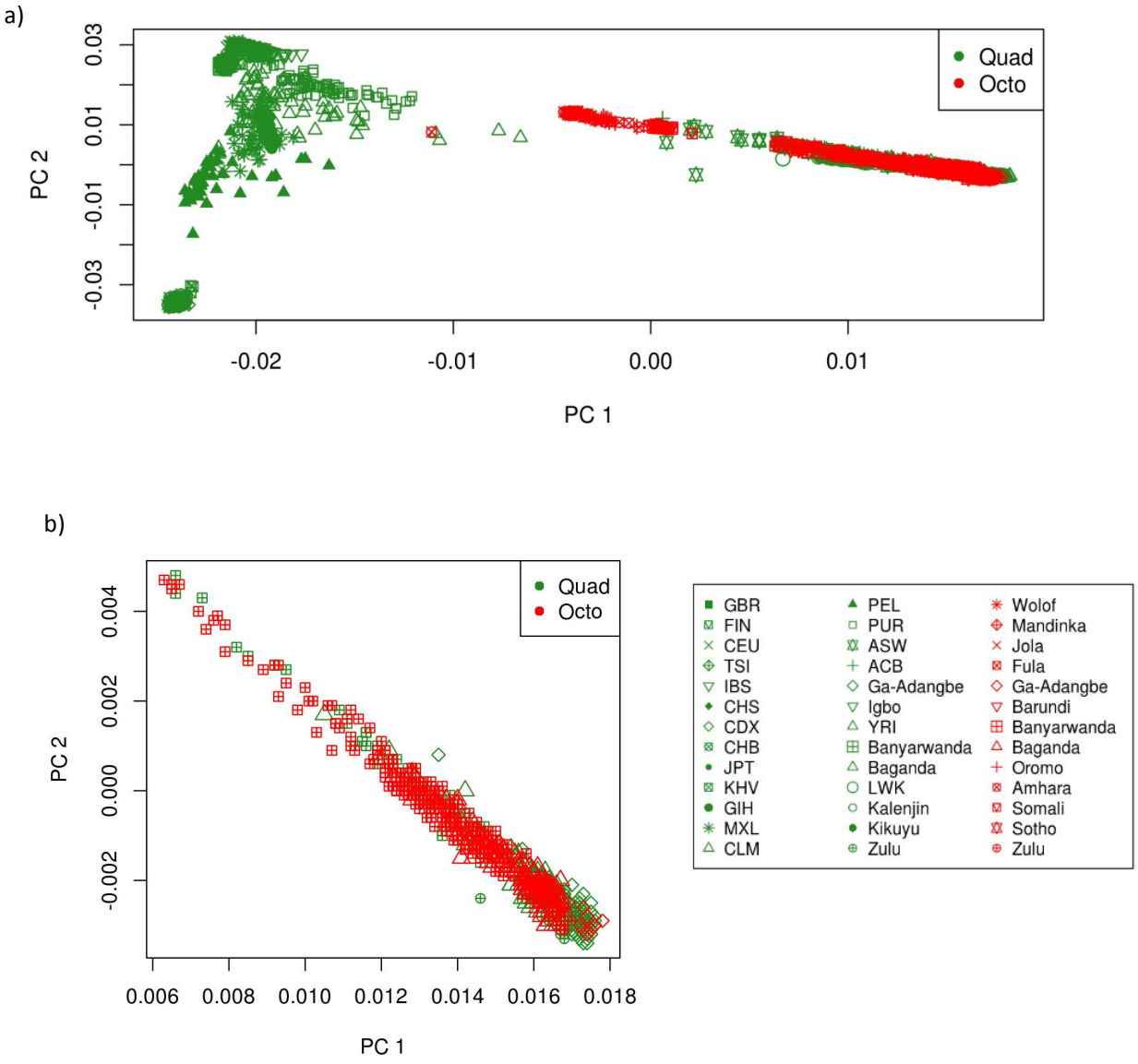


b)



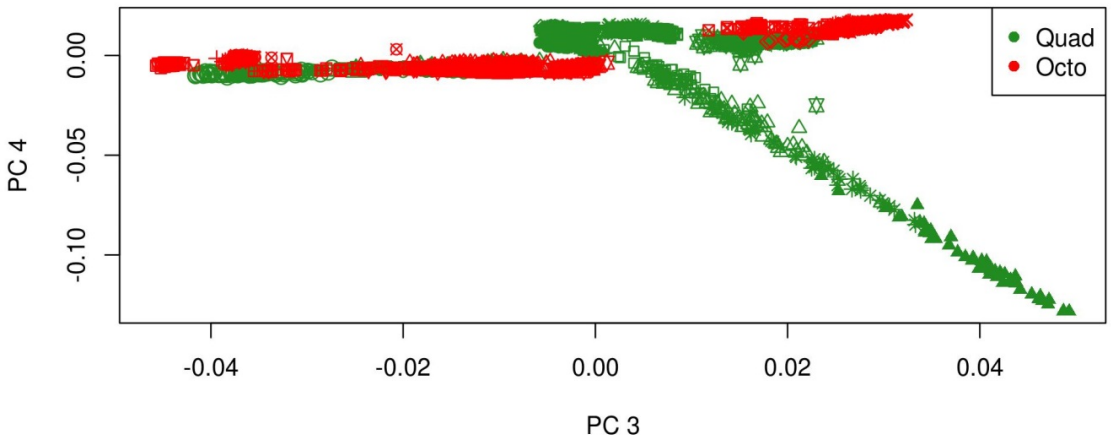
SN1 Figure 3a and 3b represent standardised SNP loadings along PCs 3 and 4 for the global dataset along chromosomes 1-22. The black and red lines represent 2 and 3 SD thresholds from the mean respectively. Sites along chromosomes are coloured by the level of discordancy in genotypes between quad and 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

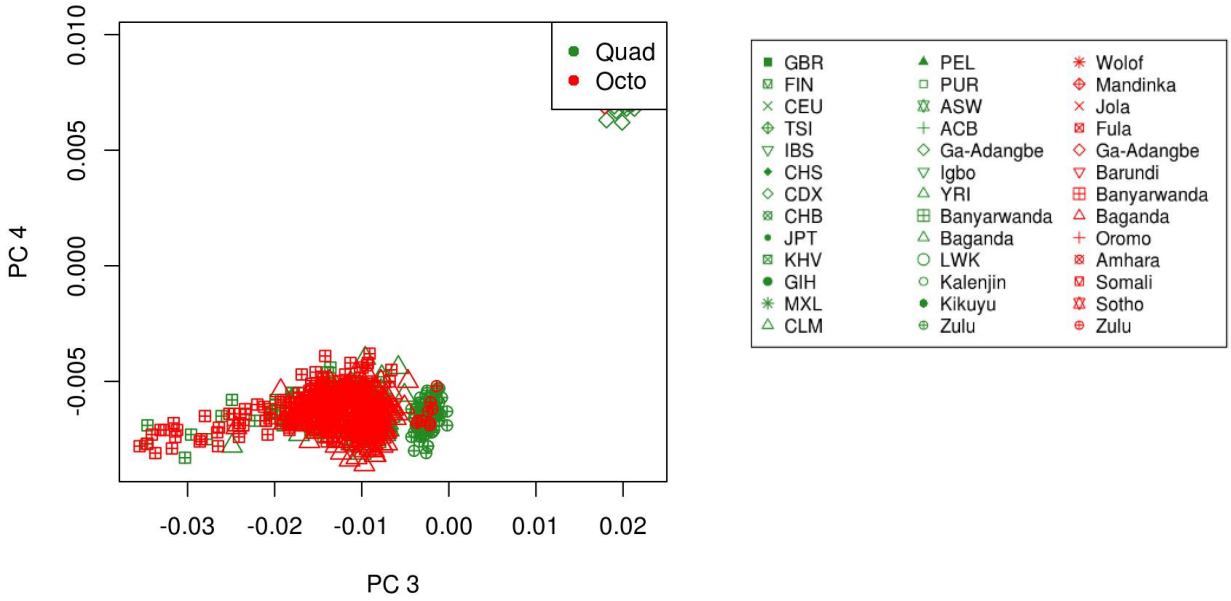


SN1. Figure 4a shows global samples represented along PCs 1 and 2 after removal of SNPs with weights > 3 SD from the mean along PCs 3 and 4. Samples genotyped on the quad chip are shown in green, while those genotyped on the octo platform are shown in red. SN1 Figure 4b depicts PCs 1 and 2 for samples 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.