The American Journal of Human Genetics, Volume *102*

Supplemental Data

Whole-Exome Sequencing Reveals

Uncaptured Variation and Distinct Ancestry

in the Southern African Population of Botswana

Gaone Retshabile, Busisiwe C. Mlotshwa, Lesedi Williams, Savannah Mwesigwa, Gerald Mboowa, Zhuoyi Huang, Navin Rustagi, Shanker Swaminathan, Eric Katagirya, Samuel Kyobe, Misaki Wayengera, Grace P. Kisitu, David P. Kateete, Eddie M. Wampande, Koketso Maplanka, Ishmael Kasvosve, Edward D. Pettitt, Mogomotsi Matshaba, Betty Nsangi, Marape Marape, Masego Tsimako-Johnstone, Chester W. Brown, Fuli Yu, Adeodata Kekitiinwa, Moses Joloba, Sununguko W. Mpoloka, Graeme Mardon, Gabriel Anabwani, Neil A. Hanchard, and for the Collaborative African Genomics Network (CAfGEN) of the H3Africa Consortium

SUPPLEMENTAL DATA

Supplemental data includes 5 figures and 5 tables

Supplemental Figures

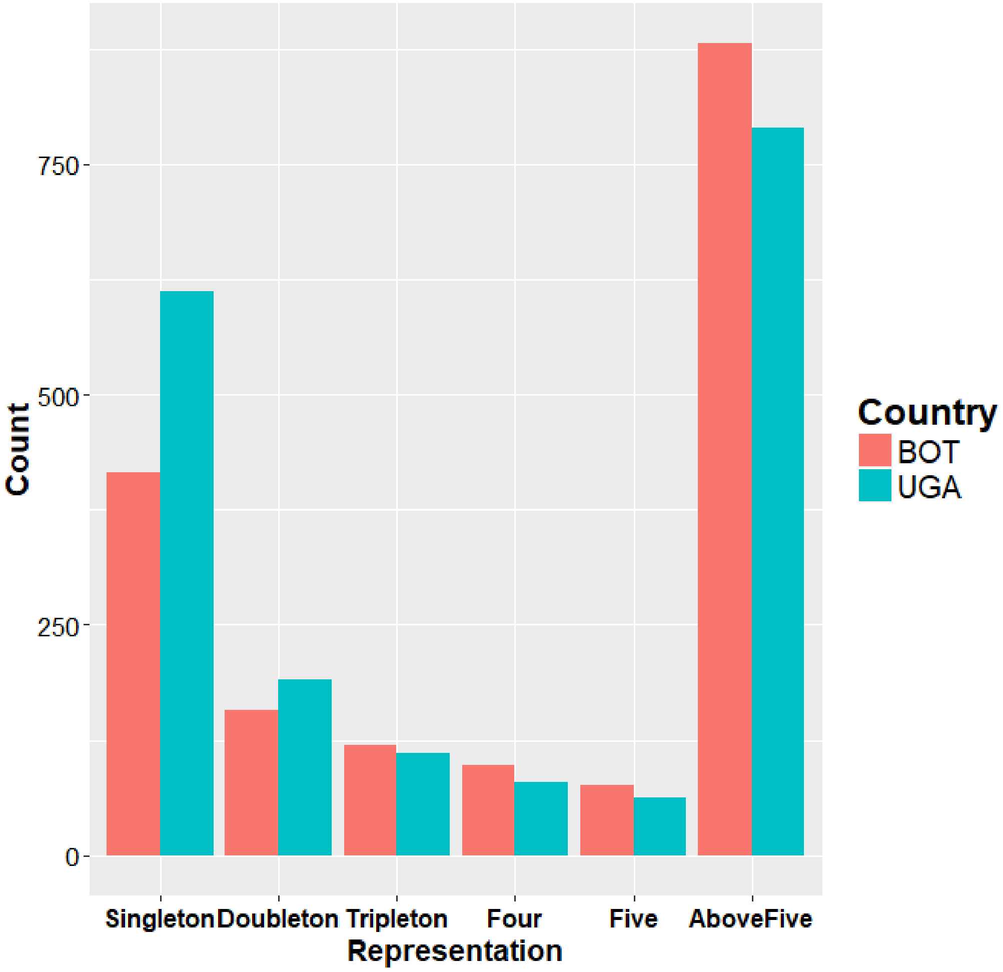

Figure S1. Putative loss-of-function variants in Botswana and Uganda population, annotated using ANNOVAR. Exons were defined using the exon start and end positions as defined within the UCSC KnownGene database using Variant Tools software (v2.6.1).

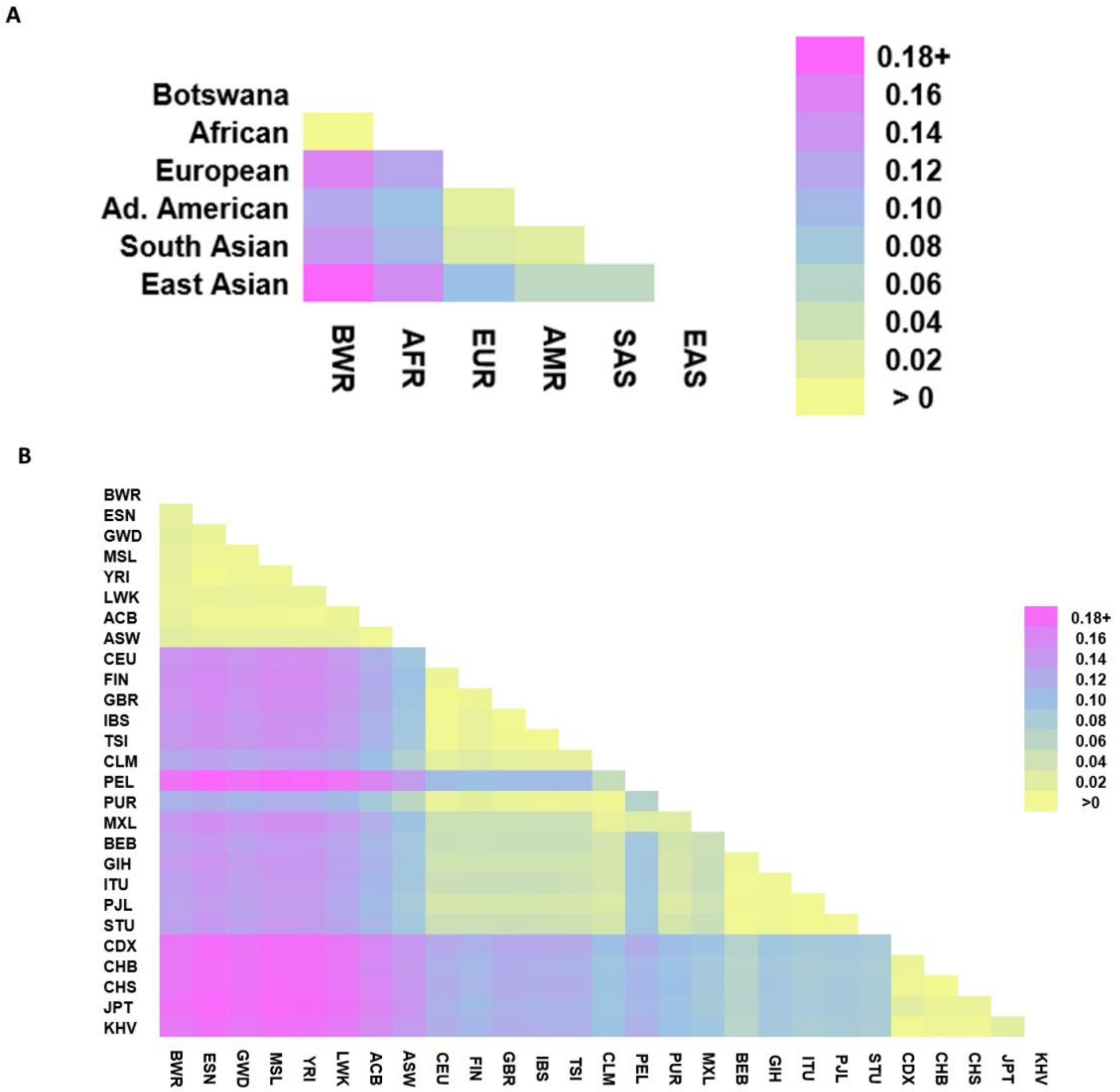

Figure S2 - Continental Weir and Cockerham's F_{ST} comparison. Closely related populations have smaller F_{ST} values, whilst populations with large F_{ST} values are taken to have higher genetic differentiation. **A -** F_{ST} comparison between Botswana and continental groupings of the 1000 Genomes data correlates with PCA results indicating affinity of Batswana with other African populations. **B -** Comparison between Batswana and other populations in the 1000 Genomes.

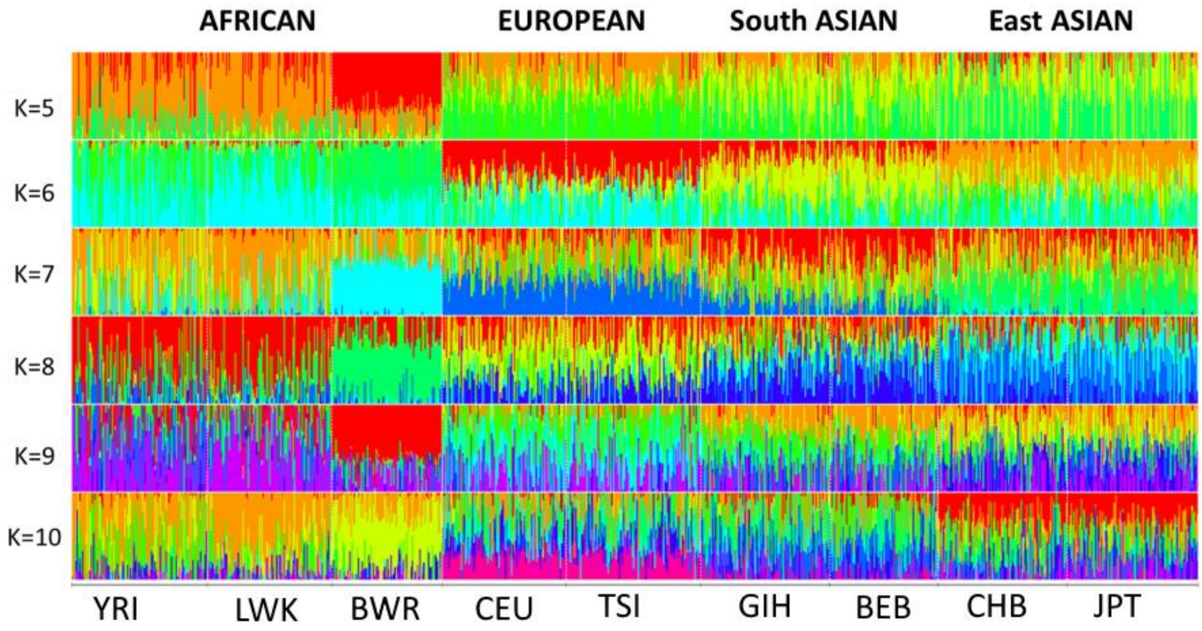

Figure S3 - Estimation of ancestral population admixture within Batswana and 1000 Genomes populations using unsupervised clustering. Runs for K5 - K10 are shown. At K=7 the Botswana population is best characterised by at least 3 ancestral populations for the majority of the individuals. Most ancestral contributions were African with a minimal Eurasian component as well as a component distinct from the two other African populations in the analysis. YRI – Yoruba in Nigeria; LWK - Luhya in Kenya; BWR – Batswana; CEU – Northern Europeans from Utah; TSI – Tuscans from Italy; GIH - Gujarati Indians from Houston, Texas; BEB - Bengali from Bangladesh; CHB - Han Chinese in Beijing, China; JPT - Japanese in Tokyo, Japan.

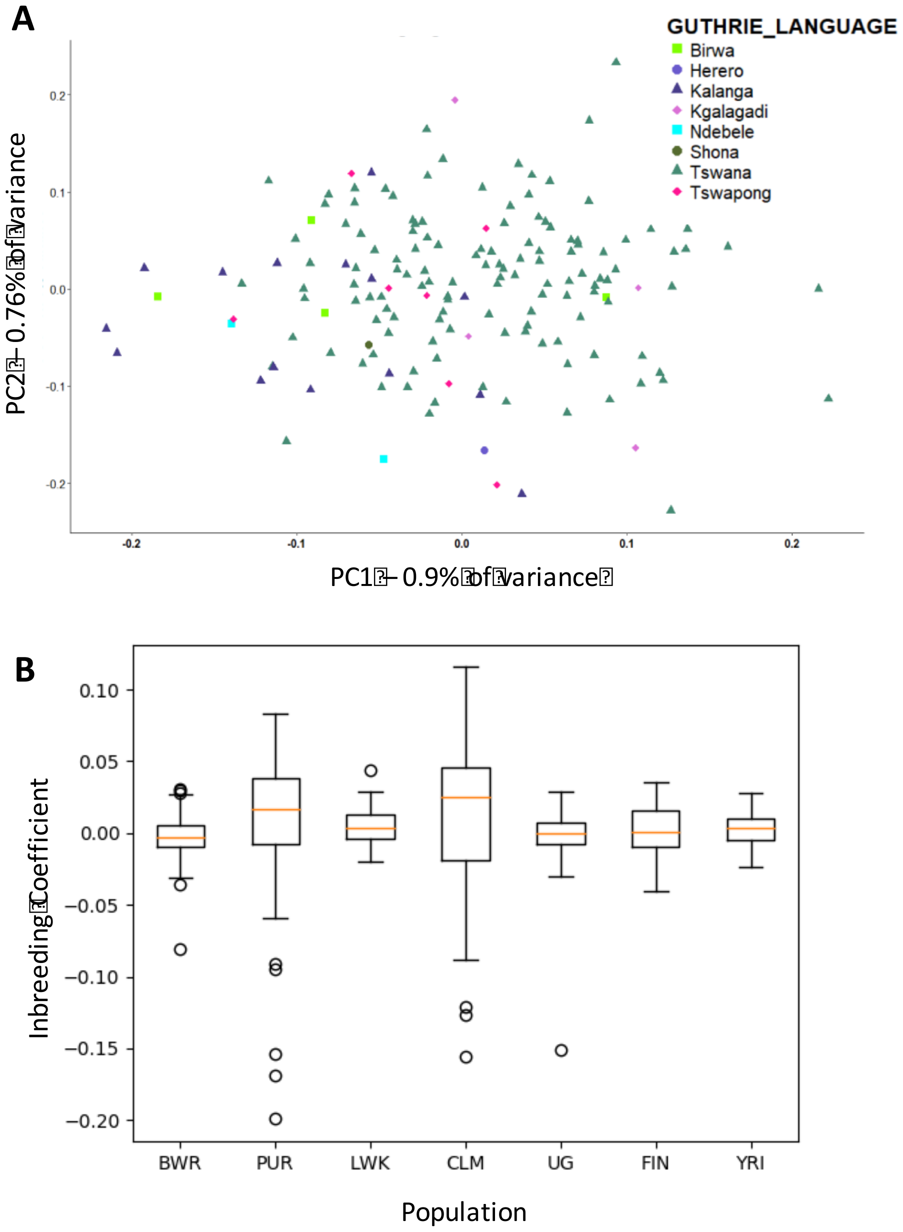

Figure S4 – S4A - population structure in Botswana by Guthrie language group; **S4B** - inbreeding coefficients for Botswana (BWR), Uganda (UG), and 1000 Genomes populations assessed in Figure 5.

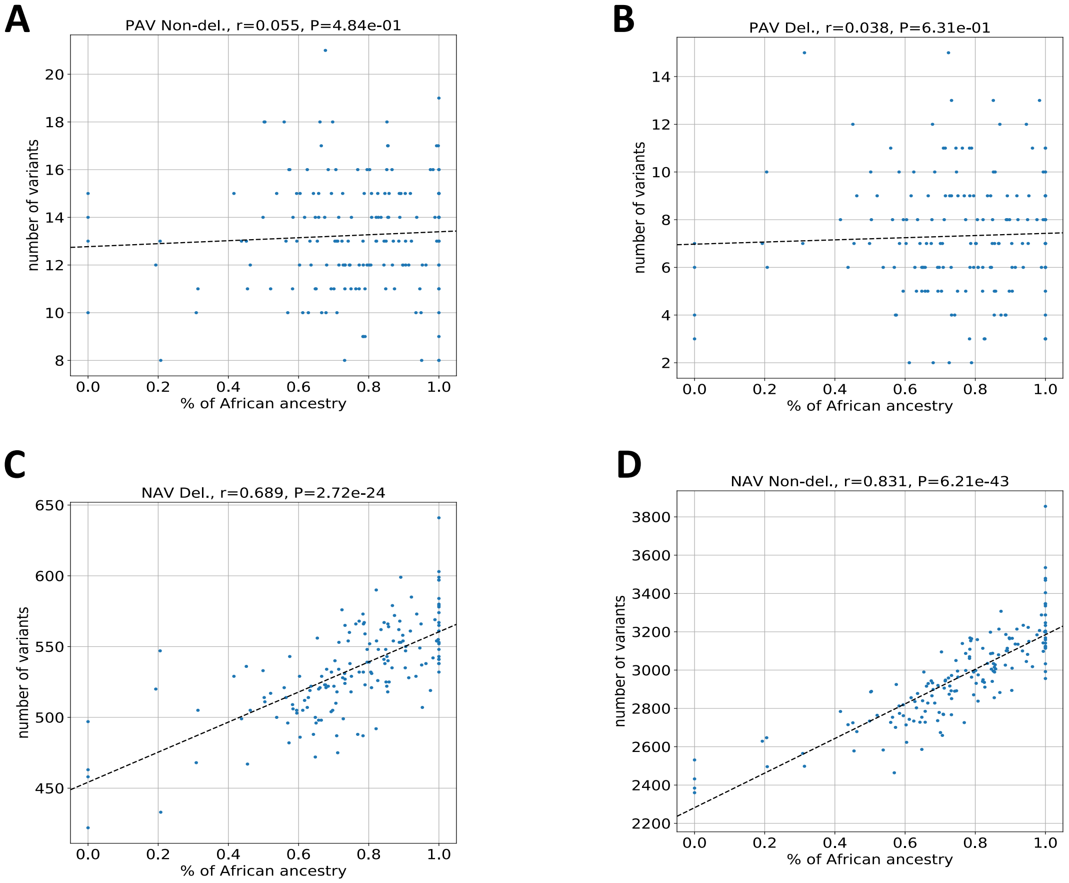

Figure S5 - The correlation between the number of variants per sample in each classification with the proportion of southern African ancestry. The variants were categorized into four groups: **A** - pathogenic and deleterious (PAV Del.); **B** - pathogenic but not deleterious (PAV Non-del.); **C** - non-pathogenic but deleterious (NAV Del.) and **D** - non-pathogenic and non-deleterious (NAV Non-del.) (see Methods for details of variant classification). The proportion of African ancestry was estimated using ADMIXTURE (K=2). Since the ADMIXTURE was performed over African samples from West, East and Southern African, the x-axis refers to the proportion of Southern African ancestry (see Figure 4).

Supplemental Tables

Table S1. Demographics of the participants. Gender and age distribution of the participants by country and HIV disease progression status.

Country	HIV Disease progression status	Male (N)	Female (N)	Median Age
Botswana	Rapid Progressors*	60	42	13
	Long-term Non-Progressors*	23	39	19
Uganda	Rapid Progressors	38	33	8
	Long-term Non-Progressors	33	46	15
	Total	154	160	

*Rapid Progressors – World Health Organization (WHO) clinical and immunological criteria for rapid progression: Anti-Retroviral treatment (ART) within 3 years of birth and/or an AIDS defining illness (WHO stage 3 or 4 OR Centers for Disease Control category 3); Two or more CD4 T cell percentage values below 15% within 3 years of birth. Long-term Non-Progressors – WHO clinical and immunological criteria for long term non-progression: Asymptomatic HIV-infection for 10 years or more after initial infection; not needing ART.

Table S2. Self-reported ancestry and Guthrie language classification of the Botswana participants

Self-Reported Ancestry	Guthrie Language Class	No. of participants
Babirwa	S32*	4
Bahurutshe	S31	10
Bakalanga	S16	15
Bakgatla_Kgafela	S31	15
Bakgatla_Mmanaana	S31	11
Bakwena	S31	24
Balete	S31	18
Bangwaketse	S31	13
Bangwato	S31	27
Barolong	S31	3
Batlokwa	S31	9
Batswapong	S32*	7
Babolaongwe	S311	2
Baphaleng	S311	1
Bashaga	S311	1
Shona	S10	1
Ndebele	S40	2
Herero	R31	1

*The Babirwa and Batswapong are described as speaking two distinct Northern Sotho languages that have come to resemble each other due their proximity rather than their origins[23].

Table S3. Database representation of all WES sequence variants from Botswana and Uganda samples.

	Botswana **N (%)**	**Uganda** **N (%)**
	On-target*	On-target*
Total number of Variants	191,758	190,584
dbSNP141_Uncaptured	36,432(14.4)	32,463(17.0)
ThouGen Uncaptured	50,955(26.6)	40,243(21.1)

*Vcrome v2.1 bed, KnownGene and ENSEMBL exon positions

Table S4. Quality control of autosomal biallelic variant markers used in PCA analysis comparing Batswana to data from 1000 Genomes and African Genome Variation Project.

Population	Variants	Variants removed	Post QC variants
Batswana	600,695	540,815	59,880
1000 Genomes	418,913	396,519	22,394
AGVP Sotho	2,139,912	2,124,068	15,844
AGVP Zulu	2,050,451	2,035,047	15,404