The American Journal of Human Genetics, Volume 91

Supplemental Data

Ethiopian Genetic Diversity Reveals Linguistic Stratification and Complex Influences on the Ethiopian Gene Pool

Luca Pagani, Toomas Kivisild, Ayele Tarekegn, Rosemary Ekong, Chris Plaster, Irene Gallego Romero, Qasim Ayub, S. Qasim Mehdi, Mark G. Thomas, Donata Luiselli, Endashaw Bekele, Neil Bradman, David J. Balding, and Chris Tyler-Smith

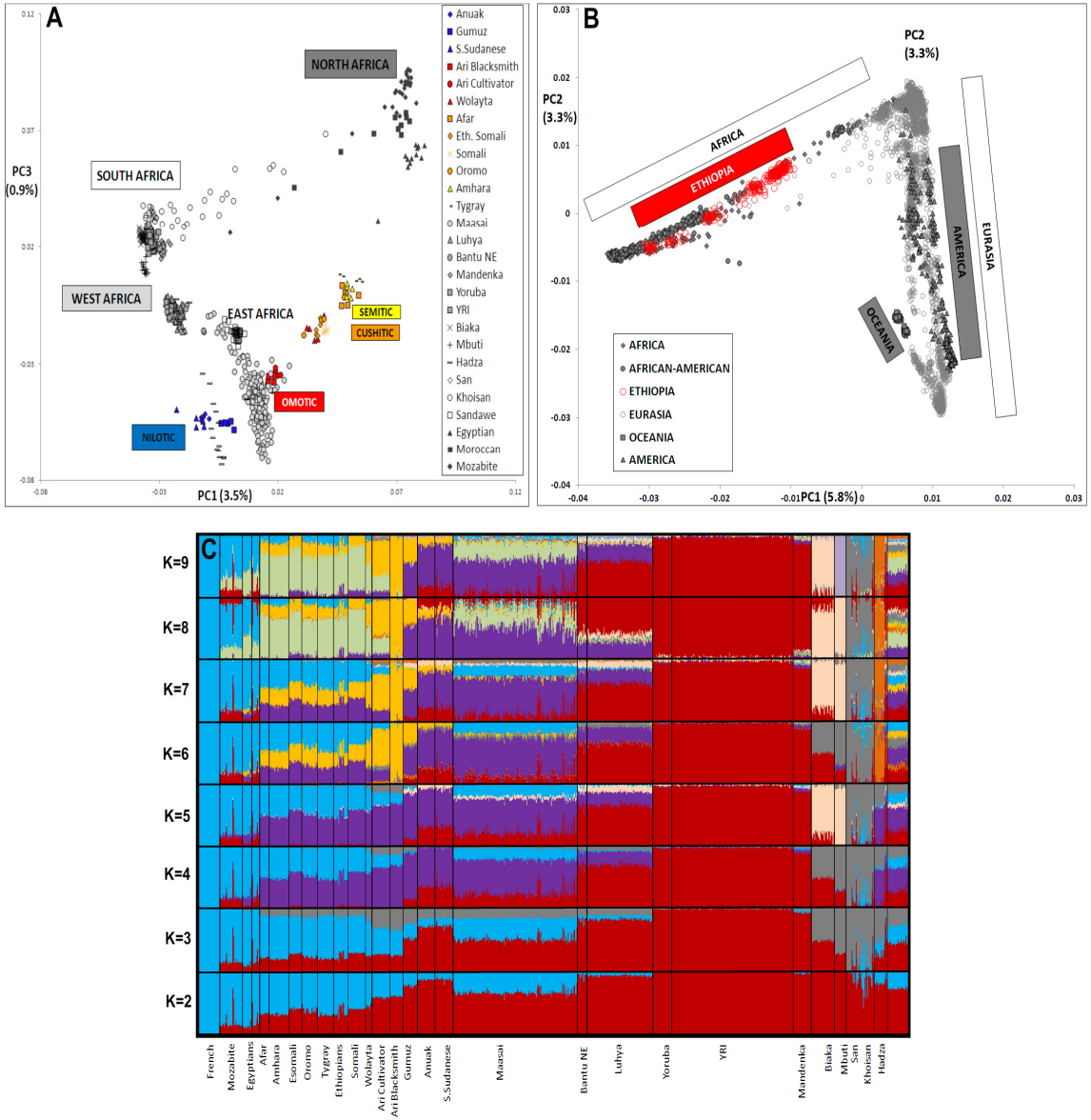

Figure S1. Alternative PCA and STRUCTURE-like Plots

The 1st and 3rd component of the PCA based on the full African panel using only 5 individuals per Ethiopian population is shown in **A.** The African samples are viewed in a worldwide context in **B**, where the first 2 components show the Ethiopian samples (in red) spanning most of the African diversity. **C** shows a series of ADMIXTURE plots using values of K ranging between 2 and 9.

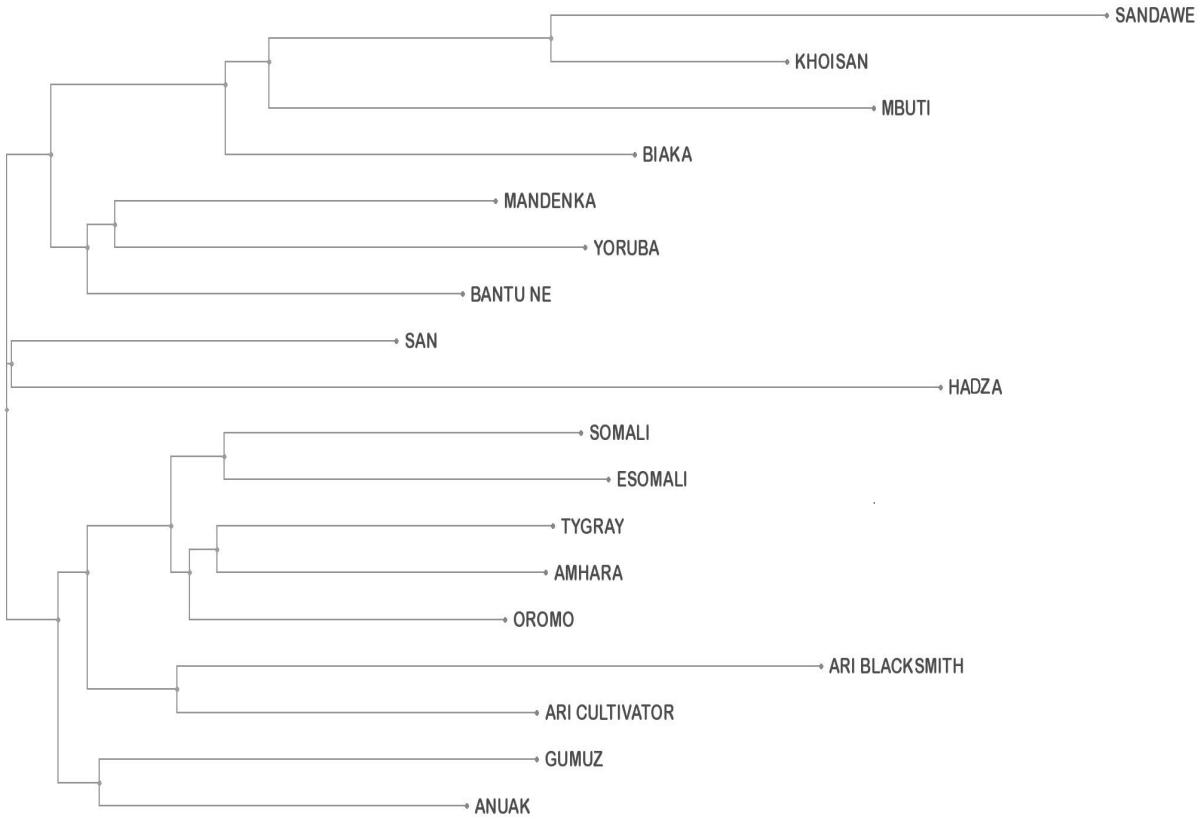

Figure S2. Neighbor-Joining Tree

The genetic distance (Fst) calculated on the African component only was used to draw a neighbor-joining tree of the studied Sub-Saharan and Ethiopian populations.

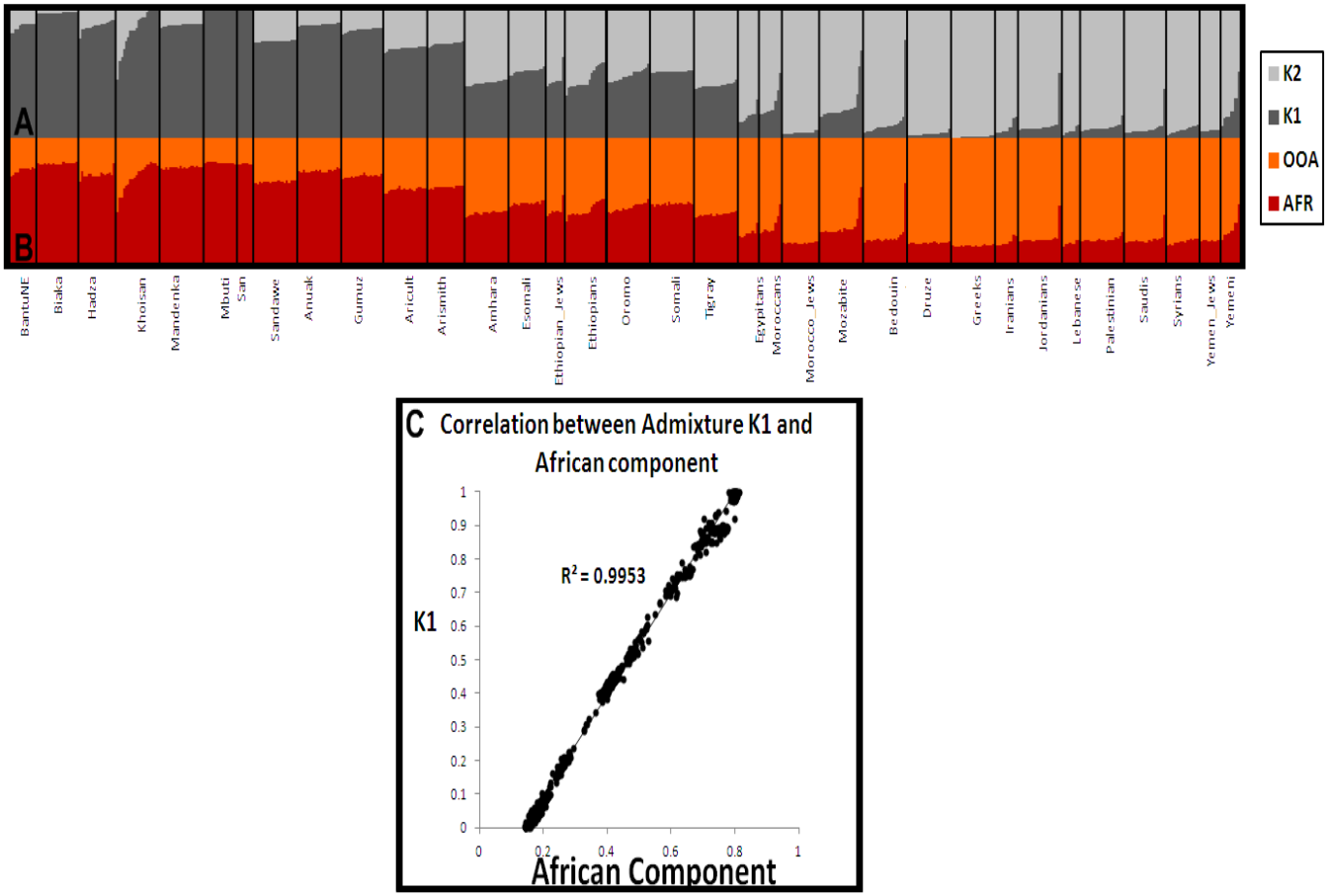

Figure S3. Comparison between the PCA-Based Genome Partitioning and the ADMIXTURE Results

A maximum of 20 pruned samples were processed in ADMIXTURE with K=2 (**A**) and PCA-based genome partitioning (**B**). The correlation (r^2) between the K1 and the African components from these plots is displayed in **C.**

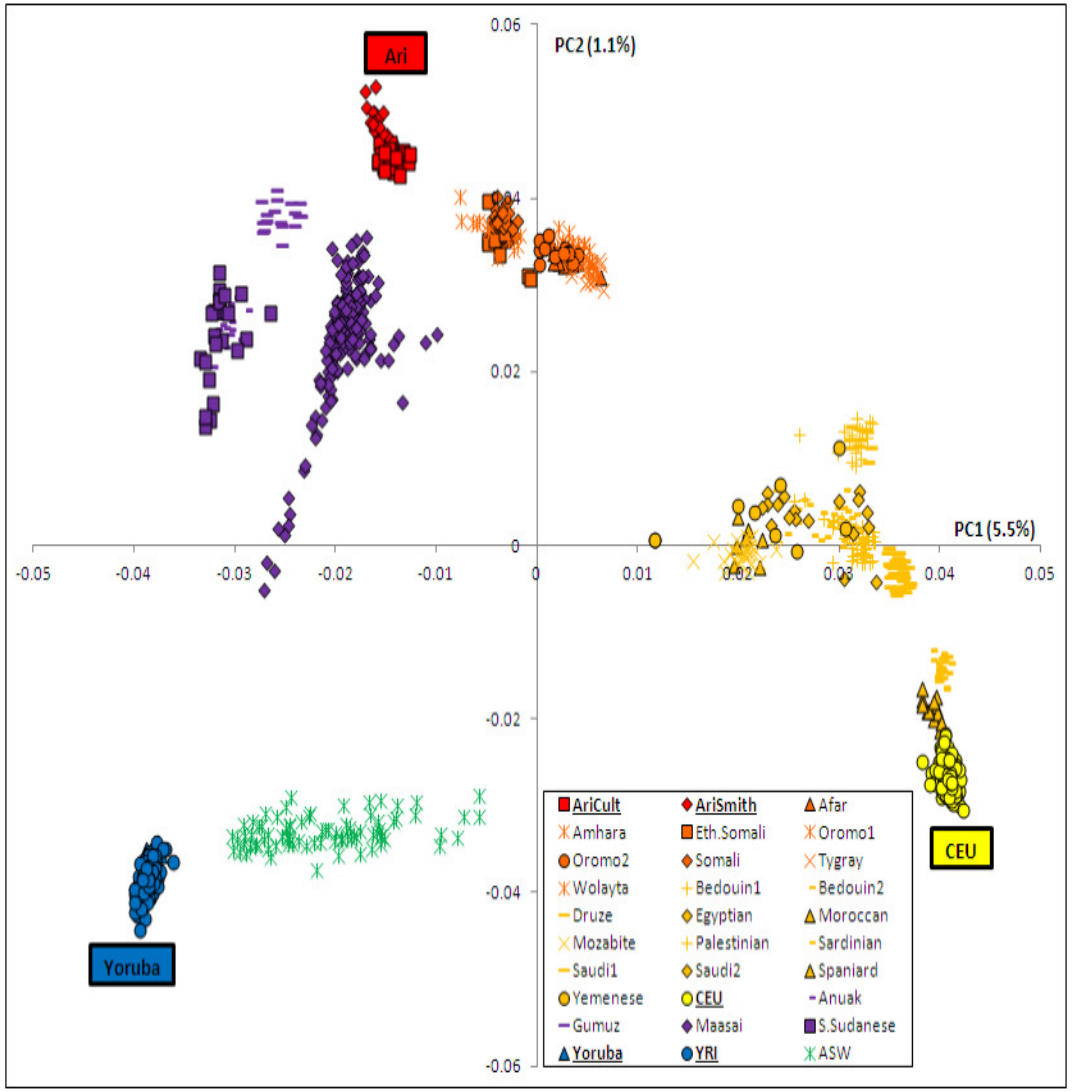

Figure S4. PCA Plot of African, West Asian, and European Samples

The Yoruba (blue), CEU (Yellow) and Ari (Red) define a triangular shape with the other samples analyzed (colored according to their position between the three primary colors) distributed on its three sides.

Table S1. Sample Size, Location, and Sociological Features of the Genotyped Populations

Pop	N	Lat	Long	Elev	Geo. Location	Ling. Group	Language	1998 Census	Endo/ Exogamous	Patri/ Matrilocal	Mono/ Polygamous	Patri/ Matrilineal	High/ Lowland	Food Production
Afar	12	12	41	379	Wag Hemra Zone	Cushitic	Xamtan	143369	Endo	Patrilocal	Poly	Patrilineal	Highland	Agriculturalist
Amhara	26	10	39	2088	Amhara Region	Semitic	Amharic	17372913	Endo	Patrilocal	Mono	Patrilineal	Highland	Agriculturalist
Anuak	23	8	34	500	Gambella	Nilotic	Anuak	45646	Endo	Patrilocal	Poly	Patrilineal	Lowland	Mixed Farming
Ari Blacksmith	17	6	37	1348	South Omo	Omotic	Ari	158857	Endo	Patrilocal	Poly	Patrilineal	Highland	Agriculturalist
Ari Cultivator	24	6	37	1348	South Omo	Omotic	Ari	158857	Endo	Patrilocal	Poly	Patrilineal	Highland	Agriculturalist
Ethiopian Somali	17	9	42	1543	Somali Region	Cushitic	Somali	3334113	Endo	Patrilocal	Poly	Patrilineal	Lowland	Pastoralist
Gumuz	19	NA	NA	NA	Beni-Shangul Gumuz	Nilo-Saharan	Gumuz	120424	Endo	Patrilocal	Poly	Patrilineal	Lowland	Pastoralist
Oromo	21	8	37	1758	Oromia Region	Cushitic	Oromo	Ca. 17000000	Endo	Patrilocal	Mono/Poly	Patrilineal	Highland	Agriculturalist/Mixed Farming/Pastoralist
Somali	23	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
South Sudanese	24	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Tigray	21	9	38	1696	Tigray Region	Semitic	Tigrayan	3224875	Endo	Patrilocal	Mono	Patrilineal	Highland	Agriculturalist
Wolayta	8	6	37	1737	Wolayta Zone	Omotic	Wolayta	1231673	Endo	Patrilocal	Mono	Patrilineal	Highland	Agriculturalist

Table S2. Genomic Average Heterozygosity and F_{ST}

Calculated using the same pruned SNPs as for the genome partitioning. The values here formed the basis of the heat maps in Figure 2.

(This table is available as a supplemental Excel file)

Table S3. Minimum Pairwise Genetic Distance in Whole Genome (A) and Non-African Component Only (B) between Ethiopians and Surrounding Populations to Compare with the F_{ST} Results Reported in Figure 3

A	Whole Genome	Bedouin	Druze	Egyptian	Greek	Iranian	Jordanian	Lebanese	Moroccan	Mozabite	Palestinian	Saudi	Syrian	Yemeni
	AMHARA	0.0382	0.0383	0.0381	0.0387	0.0387	0.0384	0.0383	0.0380	0.0387	0.0385	0.0381	0.0384	0.0383
	ESOMALI	0.0380	0.0383	0.0380	0.0385	0.0385	0.0382	0.0381	0.0377	0.0387	0.0383	0.0381	0.0384	0.0381
	OROMO	0.0386	0.0388	0.0383	0.0391	0.0390	0.0388	0.0387	0.0380	0.0389	0.0388	0.0386	0.0389	0.0385
	SOMALI	0.0390	0.0391	0.0384	0.0394	0.0394	0.0391	0.0389	0.0380	0.0391	0.0391	0.0389	0.0393	0.0388
	TYGRAY	0.0377	0.0379	0.0380	0.0382	0.0382	0.0379	0.0377	0.0380	0.0386	0.0378	0.0377	0.0380	0.0380
	AVERAGE	0.0383	0.0385	**0.0381**	0.0388	0.0388	0.0385	0.0383	0.0380	0.0388	0.0385	0.0383	0.0386	**0.0383**

B	Non-African Component Only	Bedouin	Druze	Egyptian	Greek	Iranian	Jordanian	Lebanese	Moroccan	Mozabite	Palestinian	Saudi	Syrian	Yemeni
	AMHARA	0.0373	0.0374	0.0372	0.0375	0.0376	0.0373	0.0373	0.0377	0.0377	0.0374	0.0372	0.0373	0.0375
	ESOMALI	0.0377	0.0379	0.0376	0.0380	0.0380	0.0377	0.0378	0.0381	0.0381	0.0378	0.0377	0.0379	0.0378
	OROMO	0.0377	0.0378	0.0375	0.0379	0.0380	0.0378	0.0377	0.0380	0.0379	0.0377	0.0376	0.0378	0.0379
	SOMALI	0.0379	0.0381	0.0377	0.0383	0.0383	0.0380	0.0380	0.0381	0.0382	0.0379	0.0379	0.0381	0.0381
	TYGRAY	0.0370	0.0373	0.0368	0.0373	0.0374	0.0372	0.0371	0.0375	0.0375	0.0371	0.0370	0.0372	0.0373
	AVERAGE	0.0375	0.0377	**0.0373**	0.0378	0.0379	0.0376	0.0376	0.0379	0.0379	0.0376	0.0375	0.0376	**0.0377**

Table S4. 40-SNP Windows Showing an Outlier Number of Non-African Chromosomes in the Majority of the Semitic-Cushitic and Ari Blacksmith Populations

	Z-scores bins for the proportion of European haplotypes in each 40 SNP window							
Number of populations showing a given Z score per given region	**-3.5**	**-2.5**	**-1.5**	**-0.5**	**0.5**	**1.5**	**2.5**	**3.5**
1	9	87	362	469	487	384	82	0
2	0	29	217	476	454	231	25	0
3	**1**	**5**	103	348	332	116	**8**	**0**
4	**0**	**4**	51	132	139	42	**6**	**0**
5	**0**	**0**	6	37	24	16	**0**	**0**

Genomic regions showing deficiency (light grey) or excess (dark grey) of non-African component

Chr	**Start**	**End**	**Z-score**
19	61523852	62240101	**-3**
1	202611308	203863493	**-2.5**
1	233121260	234561036	**-2.5**
2	159910201	162832297	**-2.5**
12	114756940	115796582	**-2.5**
13	62291269	65337535	**-2.5**
15	27642703	29438608	**-2.5**
15	31797258	32775283	**-2.5**
17	18231718	21377174	**-2.5**
18	14930293	19200794	**-2.5**
1	238073408	238678237	**2.5**
5	149063400	149839428	**2.5**
5	149875685	150607133	**2.5**
6	154486907	155207161	**2.5**
Chr	**Start**	**End**	**Z-score**
7	95305557	96856821	**2.5**

7	155434633	155960870	**2.5**	
8	132605749	133525623	**2.5**	
11	102147471	103368144	**2.5**	
12	123717947	124441419	**2.5**	
14	21063338	21641738	**2.5**	
15	45473882	47025722	**2.5**	****includes SLC24A****
15	51368181	52099177	**2.5**	
15	68167541	69443763	**2.5**	
19	1290058	2416737	**2.5**	

The number of regions showing a given Z-score in the specified number of populations is reported. Regions showing Z-scores <-2 (light grey) or >2 (dark grey) in at least 3 populations were considered outlier. The genomic coordinates on the highlighted windows are reported below.