Supplementary Materials for

A Genome-Wide Detection of Natural Selection in African Americans Pre-and Post-Admixture

Wenfei Jin, [1] *Shuhua Xu,* [1,*] *Haifeng Wang,* [3] *Yongguo Yu,* [4] *Yiping Shen,* [2,5] *Bailin Wu,* [2,5] *Li Jin,* [1,2,*]

[1] Chinese Academy of Sciences Key Laboratory of Computational Biology, Chinese Academy of Sciences and Max Planck Society (CAS-MPG) Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China;

[2] Ministry of Education (MOE) Key Laboratory of Contemporary Anthropology, School of Life Sciences and Institutes of Biomedical Sciences, Fudan University, Shanghai 200433, China;

[3] Chinese National Human Genome Center at Shanghai 201203, China;

[4] Shanghai Children's Medical Center, Shanghai Jiaotong University School of Medicine, Shanghai 200127, China;

[5] Children's Hospital Boston, Harvard Medical School, Boston, MA 02115, USA.

* To whom correspondence should be addressed. E-mail: xushua@picb.ac.cn (S.X.); lijin.fudan@gmail.com (L.J.)

Supplemental Text

Additional Information on populations and samples

Overall 1,890 unrelated AfA samples with ignorable ancestry outside of Africa and Europe were studied, from which 1,838 individual were download from iControlDB, and 52 African Americans were ASW (individuals of African ancestry from the southwest U.S.) from the International HapMap Project (Altshuler et al. 2010). 113 YRI (Yoruba in Ibadan, Nigeria) from the International HapMap Project and 102 HGDP indigenous Africans from 7 different geographic or linguistic groups (including 21 Yoruba from West Central Africa, 22 Mandenka from West Africa, 11 Bantu speakers from Kenya, 8 Bantu speakers from Southwestern Africa, 22 Biaka Pygmy and 13 Mbuti Pygmy from Central Africa, and 5 San from Southern African) (Li et al. 2008) represent the potential African ancestry of AfA. 113 CEU (Utah residents with northern and western European ancestry from the CEPH collection) from the International HapMap Project and 156 HGDP indigenous Europeans from 8 different geographic and linguistic groups (including 17 Adygei from Russia, 24 Basque from France, 28 French in France, 12 Italian from North Italy, 15 Orcadian from Orkney Islands, 25 Russian in European, 28 Sardinian in Sardinia island and 7 Tuscan from Italy) (Li et al. 2008) represent the potential European ancestry of AfA. Overall 2,648 CAU-GWAS also represent the European ancestry of AfA. The 84 CHB (Han Chinese from Beijing) and 85 CHD (Han Chinese from metropolitan Denver, CO, USA) represent populations in East Asia. In addition, 24 Amerindian samples showing no genetic admixture with European and African from HGDP represent Native American.

Correcting for Linkage disequilibrium and PCA analysis

Principle component analysis (PCA) is a statistics method widely used to identify population genetic variation across geographical location and ethnic background. It is well known that the uncorrected linkage disequilibrium (LD) would seriously distort the results of PCA and thus bias the population structure (Patterson et al. 2006). In order to reduce the LD between markers, we used the parameter "--indep-pairwise 50 5 0.5" in PLINK (Purcell et al. 2007) to remove any SNP that had a pairwise r^2 greater than 0.5 in

window of 50 SNPs, shifted and recalculated every 5 SNPs. The original 491,526 autosomal SNPs were reduced to 341,672 SNPs after this process.

Based on these thinned markers, PCA was performed using *smartpca* in the EIGENSOFT package (Patterson et al. 2006) under default setting with no outlier removal. In the PCA analysis, the first principal component (PC1) is the coordinate drawn in the multidimensional space so that the projections of the points (each point represents an individual) to the coordinate have the largest variance. The second principal component (PC2) is the coordinate drawn in the multidimensional space so that the projections of the points to the coordinate have the second largest variance, and so forth. Intuitively, PC1 and PC2 could reflect the two primary genetic differences between samples. In this study, only PC1 and PC2 were plotted in our PCA plot and have been carefully analyzed.

The projections of individuals onto existing axes not only identified the population structure, but also could identify admixture proportion of each admixed individual (McVean 2009). So we performed PCA on AfA and its two ancestral parental populations (YRI and CEU), and inferred the ancestral proportion of each AfA based on projection of samples on PC1.

YRI and CEU are better choices for ancestral parental populations than the others

Using the 341,672 not high-linked SNPs, the 1,890 AfA samples were studied together with 3320 samples from 19 putative ancestral parental populations. In the PCA plot, populations from the same continents were clustered together, which reflected genetic diversity of populations from the same continent is rather lower compared with that of inter-continent populations. AfA lie on the direct line between European (except Russian and Adygei, which may have some components from Asia) and African, which reflected their varying levels of admixture in AfA between the two ancestral groups (Figure S2A). When samples from Asia and America (as well as Russian and Adygei) were removed, AfA almost distributed only between West African and European, which reflected the West African, especially YRI, contributing to most of the African components of ancestry (Figure S2B). Substructure in European populations showed up when only YRI, AfA and European populations were analyzed. All European populations

except Sardinian contribute to European components of AfA. However, CEU contributed to most of the European components of AfA as only a few samples were beyond the line between YRI-CEU (Figure S2C). Then, we performed PCA on samples from YRI, CEU, and AfA. The PC1 explained 80% of the total variance of top ten PCs while PC2 explained only 2.4%, which reflect that most of the variances were caused by the genetic differentiation between West African and European ancestry, especially like YRI and CEU (Figure S2D).

Population structure analysis using FRAPPE and STRUCUTE

A number of methods and software have been developed to infer the proportional ancestry of each individual using genotype data without prior knowledge of individual ancestral and geographic information. FRAPPE (Tang et al. 2005), which implements an expectation-maximization (EM) algorithm, was used to infer the admixture proportion of AfA in this study. We used FRAPPE not just because of its fastness, but also because we could take full advantage of the extremely high-density SNPs data since it was almost unaffected by Linkage disequilibrium (LD). The program was allowed to run for 10,000 iteration, with prior-specified cluster numbers. First, we run FRAPPE on the 1,890 AfA and all putative ancestral parental populations (7 HGDP African populations, 8 HGDP European populations, CEU and YRI) by taking K=2. Second, we only used CEU and YRI as the ancestral populations of AfA, and ran FRAPPE on AfA, CEU and YRI by setting K=2. The two independent runs yield almost identical results (Figure S3, S4).

STRUCTURE (Pritchard et al. 2000; Falush et al. 2003), implementing a model-based clustering method, was widely used to infer population structure. However, the model in the program is best suited to these markers not highly linked. Therefore, any contiguous SNPs <1M were removed and only a set of 2,654 SNPs was kept for further analysis. We used the admixture-model and assumed that allele frequencies were correlated. The program was run with 100,000 iterations and 100,000 burn-ins.

Comparisons of methods for inferring locus-specific ancestry

Various methods have been developed to infer locus-specific ancestry based on high density SNPs data, such as ANCESTRYMAP (Patterson et al. 2004), SABER (Tang

et al. 2006), LAMP and LAMP-ANC (Sankararaman et al. 2008b), uSWITCH and uSWITCH-ANC (Sankararaman et al. 2008a), HAPAA (Sundquist et al. 2008) and HAPMIX (Price et al. 2009). There is no doubt that these methods outperform older methods that only analyzed a few unlinked markers. However, It is still difficult to distinguish which method performs better than the other since the data used in different studies may vary greatly, and sometimes maybe completely different. In particular, we are interested in which method performs the best on our dataset.

First, the genotypes of the two parental populations were labeled in our simulations so that the ancestral status of the simulated AfA was known. Second, by comparing the inferred ancestral status of AfA with the real ancestral status, the accuracies of five methods (SABER, LAMP-ANC, LAMP, HAPAA, HAPMIX) were calculated. The accuracy was simply calculated by using the number of correct inferences divided the total number or comparisons. The inferred accuracy of each method was: HAPMIX (98.1%), LAMP-ANC (96.5%), LAMP (96.3%), HAPAA (95.7%), SABER (92.3%). It is obvious that the performance of HAPMIX is much better than the other methods in this case. These observations were essentially similar to those of previous studies (Sankararaman et al. 2008a; Sundquist et al. 2008; Price et al. 2009). HAPMIX, which applies an extension of haplotype-based method (Price et al. 2009), was used to infer the locus specific ancestry in this study. However, This does not mean that HAPMIX performs better in other situations (e.g., the performance of HAPMIX was not so good if high-qualified haplotypes of ancestral parental populations were not available; data not shown). Moreover, many other situations, such as multiply-wave admixture and closely related parental populations, were not simulated in this study.

Generation since admixture for each individual

By running HAPMIX in diploid mode, we also detected chromosomal segments of distinct ancestry in each admixed individual. Then, we computed the generation since admixture for each individual (g) following Price et al. (Price et al. 2009):

$$g = \frac{N}{4u(1-u) \times 35.3}$$

where N is the number of ancestry transition observed, and μ is the ancestry proportion of the admixed individual. A fraction of $2u(1-u)$ of all recombination events was expected, and 35.3 Morgan was the overall genetic distance used in our genetic map, which was adapted from HapMap.

Locus-specific ancestry inferred by LAMP-ANC

To investigate the repeatability of selection signals detected by HAPMIX, we also inferred the locus-specific ancestry using LAMP-ANC, which also performs very well. The input genotypic data for LAMP-ANC were the same as those used by HAPMIX except their ancestral allele frequencies were provided directly. The generation since admixture was set to 7 (hybrid isolation model) based on HAPMIX results. Genome-wide distribution of European ancestral contribution in AfA based on CAU-GWAS and YRI was calculated (Figure S6), and the results are essentially the same when other putative ancestral populations were used. The locus-specific European ancestry proportion across the genome of AfA estimated by LAMP-ANC was 21.69% ± 0.82% (mean ± SD). The standard deviation (SD) estimated by LAMP-ANC was higher than that based on HAPMIX, which might be caused by higher statistic noise as the result of higher error rate of LAMP-ANC. We found that all the ancestry deviation regions can be replicated using LAMP-ANC to some extent.

The results are unlikely to be affected by ascertainment bias

The data of this study displayed a deficit of rare variants and an excess of intermediate frequency SNPs. This frequency bias is likely to affect statistics based on allele frequencies including F_{ST} (Clark et al. 2005). However, our goal was not to define precise levels of population differentiation, but to detect differentiations of F_{ST} among different SNP categories. In this context, as the ascertainment of our SNPs was agnostic to any functional annotation, all SNP classes should be equally influenced by

ascertainment bias. These biases should therefore have a minimal effect on our analysis and conclusions. In order to clarify whether ascertained bias has an effect on the enrichment analysis of functional SNPs, we randomly chose several sets of SNPs (1% and 5%), and we found that essentially no functional SNPs classes were significantly enriched in the selected SNPs bin.

References

Altshuler, DM, Gibbs, RA, Peltonen, L, Altshuler, DM, Gibbs, RA, Peltonen, L, Dermitzakis, E, Schaffner, SF, Yu, F, Peltonen, L et al. 2010. Integrating common and rare genetic variation in diverse human populations. *Nature* **467:** 52-58.

Clark, AG, Hubisz, MJ, Bustamante, CD, Williamson, SH, Nielsen, R. 2005. Ascertainment bias in studies of human genome-wide polymorphism. *Genome research* **15:** 1496-1502.

Falush, D, Stephens, M, Pritchard, JK. 2003. Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. *Genetics* **164:** 1567-1587.

Li, JZ, Absher, DM, Tang, H, Southwick, AM, Casto, AM, Ramachandran, S, Cann, HM, Barsh, GS, Feldman, M, Cavalli-Sforza, LL et al. 2008. Worldwide human relationships inferred from genome-wide patterns of variation. *Science (New York, NY* **319:** 1100-1104.

Long, JC. 1991. The genetic structure of admixed populations. *Genetics* **127:** 417-428.

McVean, G. 2009. A genealogical interpretation of principal components analysis. *PLoS genetics* **5**.

Patterson, N, Hattangadi, N, Lane, B, Lohmueller, KE, Hafler, DA, Oksenberg, JR, Hauser, SL, Smith, MW, O'Brien, SJ, Altshuler, D et al. 2004. Methods for high-density admixture mapping of disease genes. *Am J Hum Genet* **74:** 979-1000.

Patterson, N, Price, AL, Reich, D. 2006. Population structure and eigenanalysis. *PLoS genetics* **2:** e190.

Pfaff, CL, Parra, EJ, Bonilla, C, Hiester, K, McKeigue, PM, Kamboh, MI, Hutchinson, RG, Ferrell, RE, Boerwinkle, E, Shriver, MD. 2001. Population structure in admixed populations: Effect of admixture dynamics on the pattern of linkage disequilibrium. *American journal of human genetics* **68:** 198-207.

Price, AL, Tandon, A, Patterson, N, Barnes, KC, Rafaels, N, Ruczinski, I, Beaty, TH, Mathias, R, Reich, D, Myers, S. 2009. Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genet* **5:** e1000519.

Pritchard, JK, Stephens, M, Donnelly, P. 2000. Inference of population structure using multilocus genotype data. *Genetics* **155:** 945-959.

Purcell, S, Neale, B, Todd-Brown, K, Thomas, L, Ferreira, MAR, Bender, D, Maller, J, Sklar, P, de Bakker, PIW, Daly, MJ et al. 2007. Plink: A tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics* **81:** 559-575.

Sankararaman, S, Kimmel, G, Halperin, E, Jordan, MI. 2008a. On the inference of ancestries in admixed populations. *Genome Res* **18:** 668-675.

Sankararaman, S, Sridhar, S, Kimmel, G, Halperin, E. 2008b. Estimating local ancestry in admixed populations. *Am J Hum Genet* **82:** 290-303.

Sundquist, A, Fratkin, E, Do, CB, Batzoglou, S. 2008. Effect of genetic divergence in identifying ancestral origin using hapaa. *Genome research* **18:** 676-682.

Tang, H, Coram, M, Wang, P, Zhu, X, Risch, N. 2006. Reconstructing genetic ancestry blocks in admixed individuals. *American journal of human genetics* **79:** 1-12.

Tang, H, Peng, J, Wang, P, Risch, NJ. 2005. Estimation of individual admixture: Analytical and study design considerations. *Genetic epidemiology* **28:** 289-301.

Figures

Figure S1. Schematic of the two natural selection events (pre-and post-admixture) in African Americans.

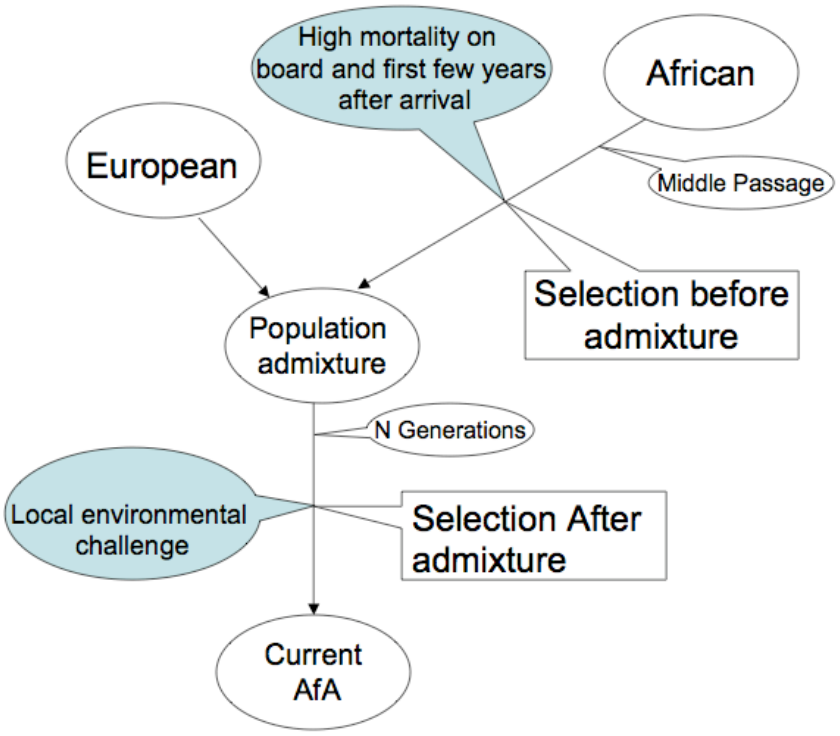


Figure S2. Schematic of the strategy to detect natural selection by comparing the African components of ancestry in AfA with African populations.

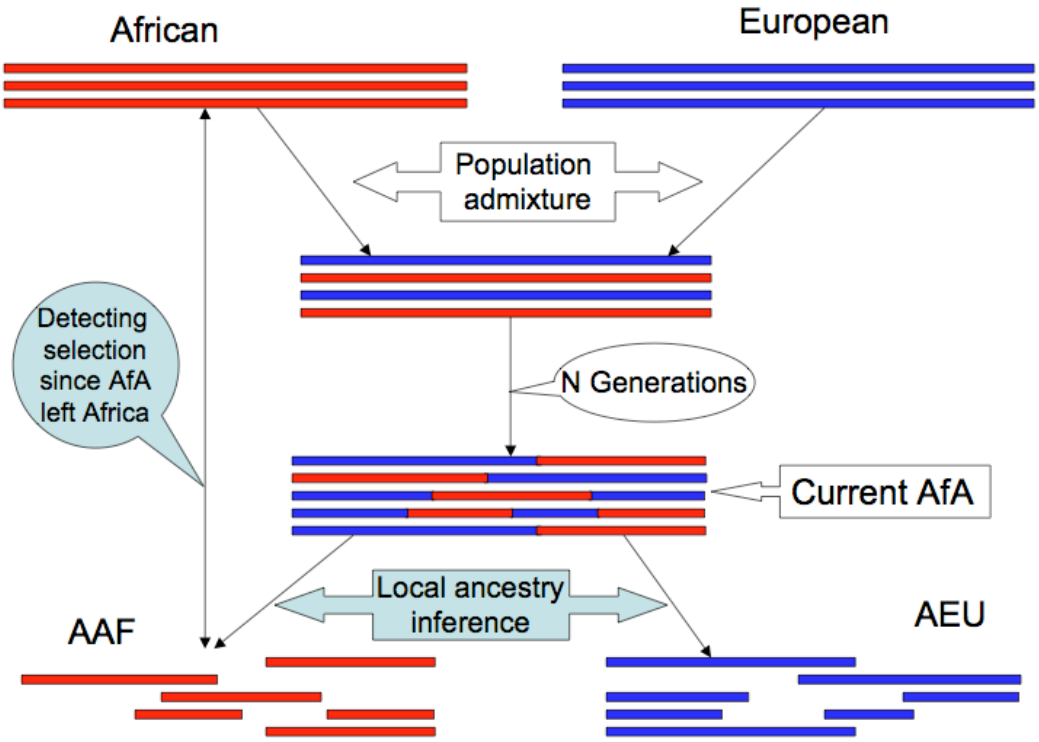


Figure S3. Analyses of the first two principal components. The % Eigenvalue is the percentage of the total variance in the Top ten PCs. (A) 1,890 AfA and 3320 samples from 19 other populations represent global-wide populations. (B) 1,890 AfA and all possible European or African ancestral populations. (C) 1,890 AfA, YRI and all possible European ancestral populations. (D) 1,890 AfA and two putative parental populations (YRI and CEU).

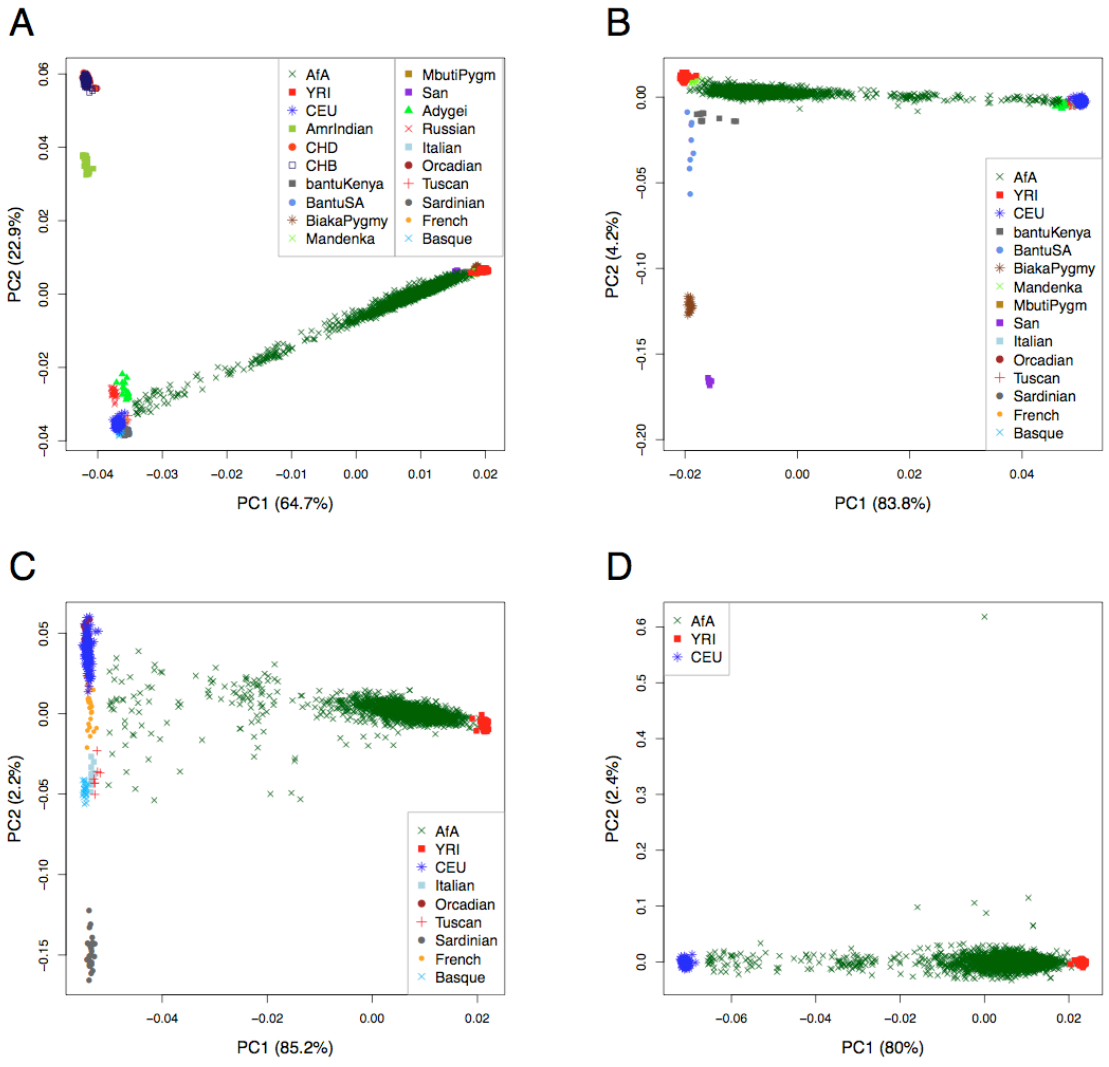


Figure S4. FRAPPE analysis of 8 African populations, 9 European populations and AfA when K=2. Each individual is represented by a vertical line, which is partitioned into two segments corresponding to the inferred membership of the genetic clusters indicated by the colors.

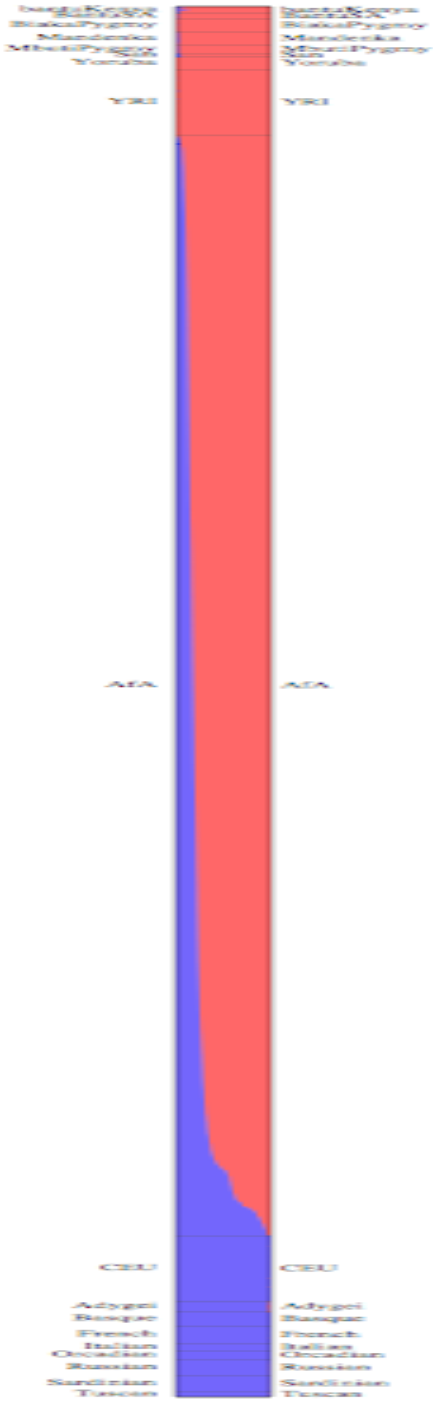


Figure S5. FRAPPE analysis of YRI, AfA and CEU when K=2. Each individual is represented by a vertical line, which is partitioned into two segments corresponding to the inferred membership of the genetic clusters indicated by the colors.

Figure S6. Histogram of generations since admixture for the 1,890 African Americans.

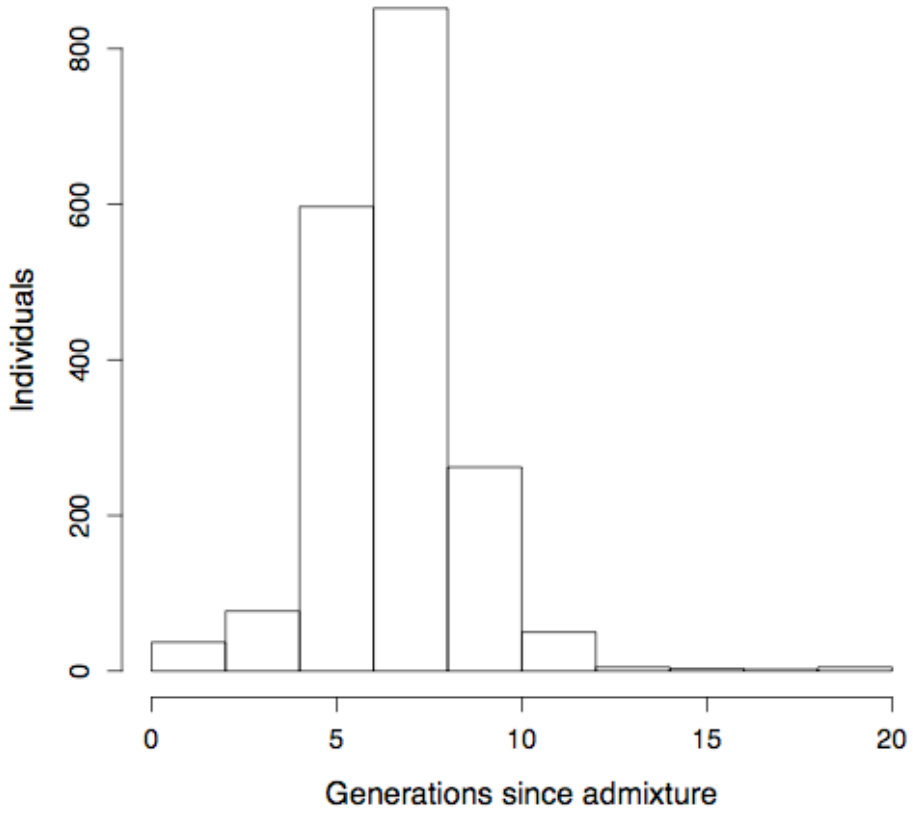


Figure S7. Genome-wide distribution of European ancestral contributions estimated by LAMP-ANC. Mean European ancestral contributions across 1,890 African American individuals at each SNP. Green line is the estimated genome-wide mean European ancestral contributions (21.69%). Blue bands indicate +2 and -2 SDs from the mean ancestral contributions and red Bands indicate +3 and -3 SDs from the mean ancestral contribution.

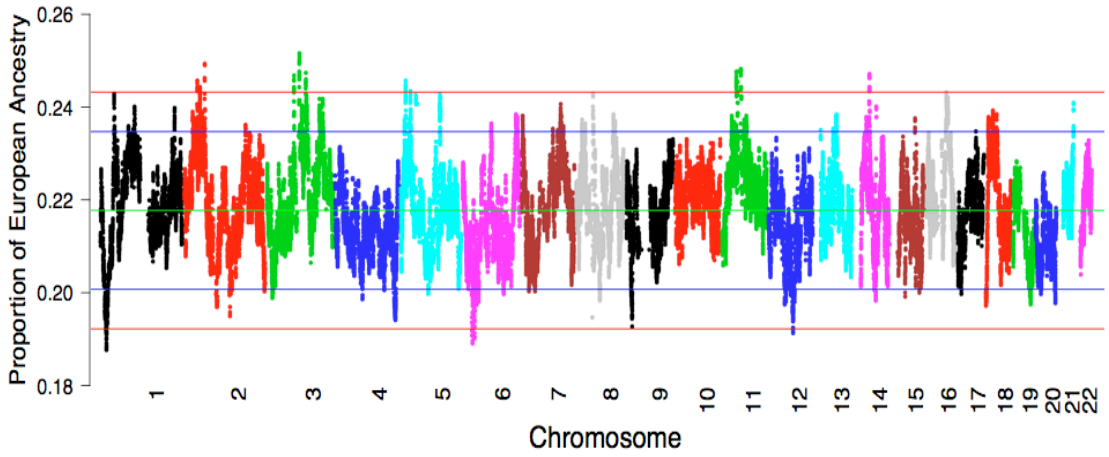


Figure S8. Q-Q plot of locus-specific F_{ST} between AAF and YRI. This Q-Q plot compares empirical values on the vertical axis to simulated values on horizontal axis. Red line shows null distribution that F_{ST} of empirical data are the same as that of simulated.

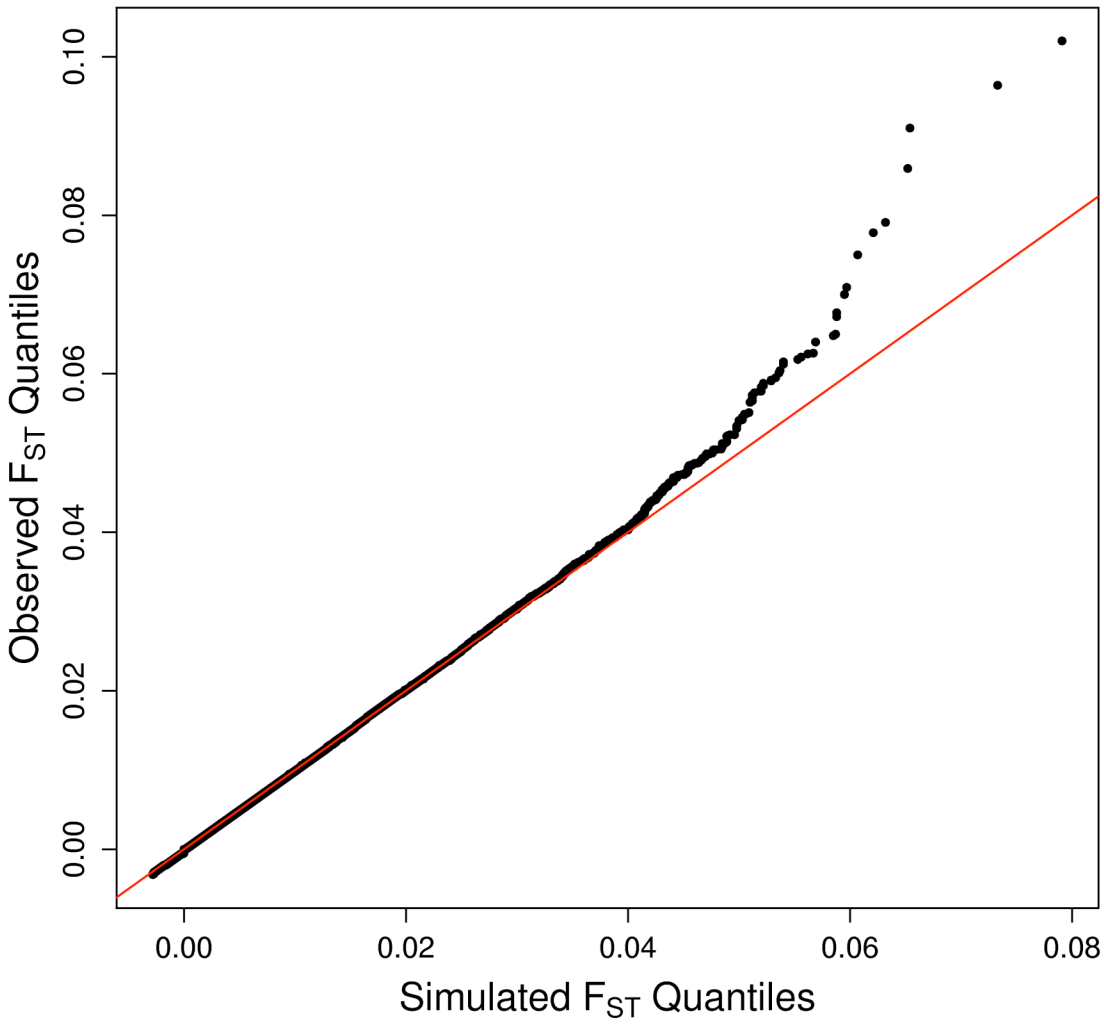


Figure S9. Genome-wide distribution of F_{ST} between AAF and YRI for all autosomal SNPs. The dashed red horizontal line indicates the cutoff threshold (99.99th percentile).

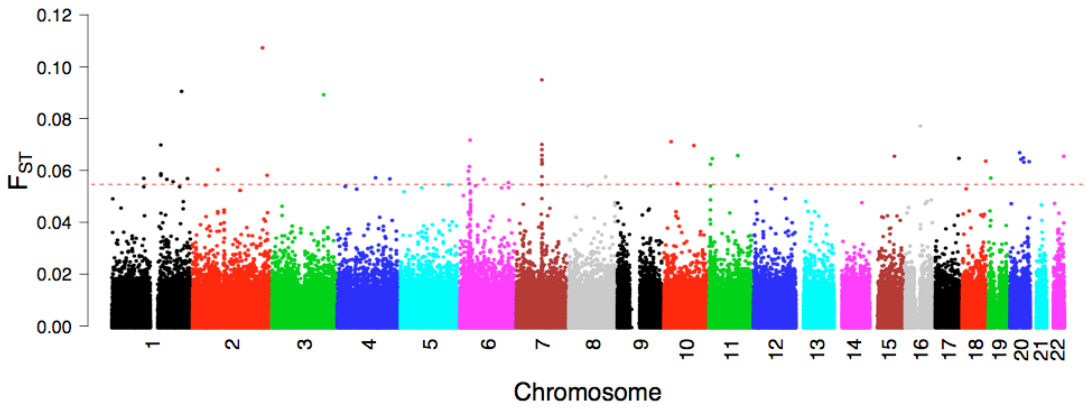


Figure S10. Genome-wide distribution of F_{ST} between AfA and rAfA. rAfA is construct by CAU-GWAS and YRI according to estimated admixture proportion. The dashed red horizontal line indicates the cutoff threshold (99.99th percentile).

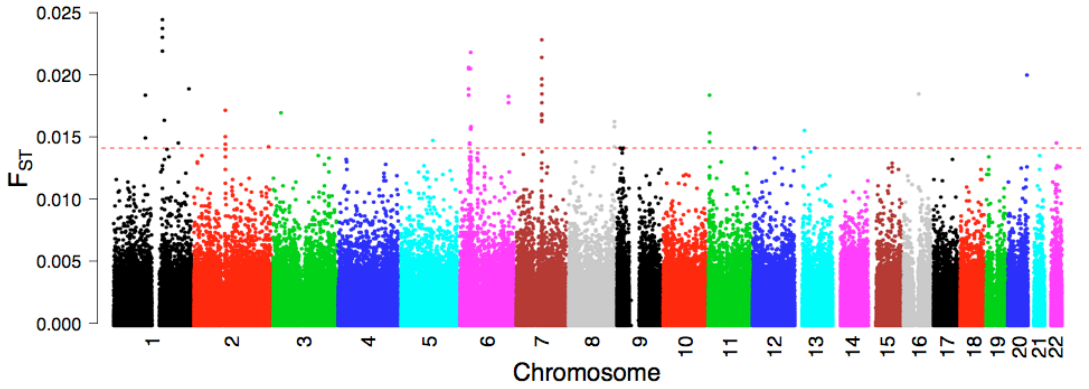


Figure S11. Genome-wide distribution of F_{ST} between AfA and rAfA. rAfA is construct by CAU-GWAS and APP (African parental population of AfA constructed by YRI, Mandenka and Bantu) according to estimated admixture proportion. The dashed red horizontal line indicates the cutoff threshold (99.99th percentile).

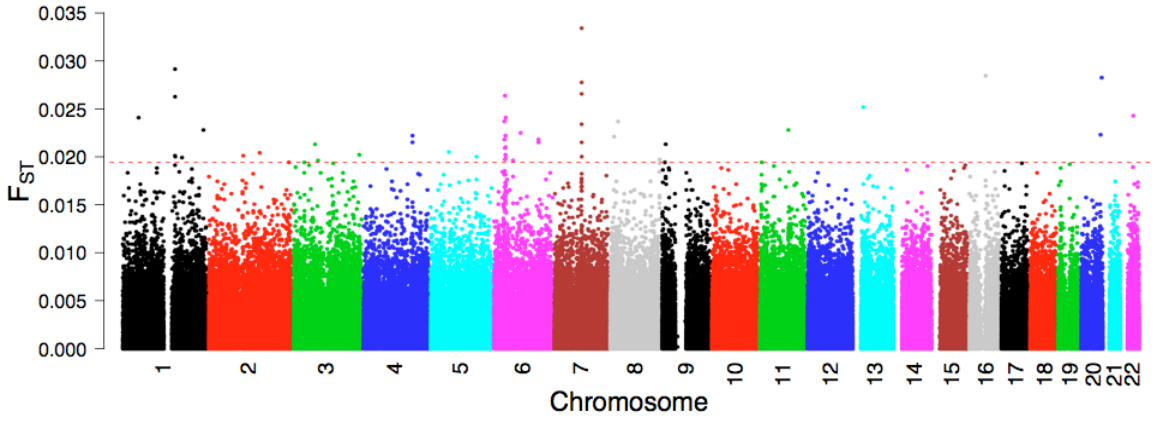


Figure S12. Enrichment of high F_{ST} loci for different SNP categories (95th percentile; F_{ST} >0.0083). Observed excess of high F_{ST} loci for different SNP classes, with respect to nongenic class, among high F_{ST} bin (95th percentile; F_{ST} >0.0083). The values on the bar are p-values of χ^2 tests. "NS" stands for "not significant".

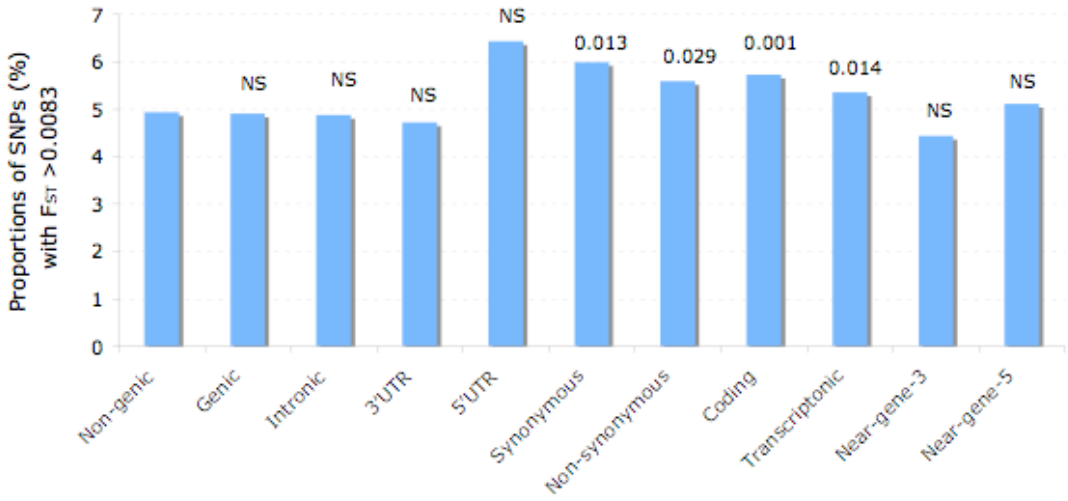


Figure S13. Enrichment of high F_{ST} loci for different SNP categories (99.9th percentile; F_{ST} >0.0304). Observed excess of high F_{ST} loci for different SNP classes, with respect to nongenic class, among high F_{ST} bin (99.9th percentile; F_{ST} >0.0304). The values on the bar are p-values of χ^2 tests. "NS" stands for "not significant".

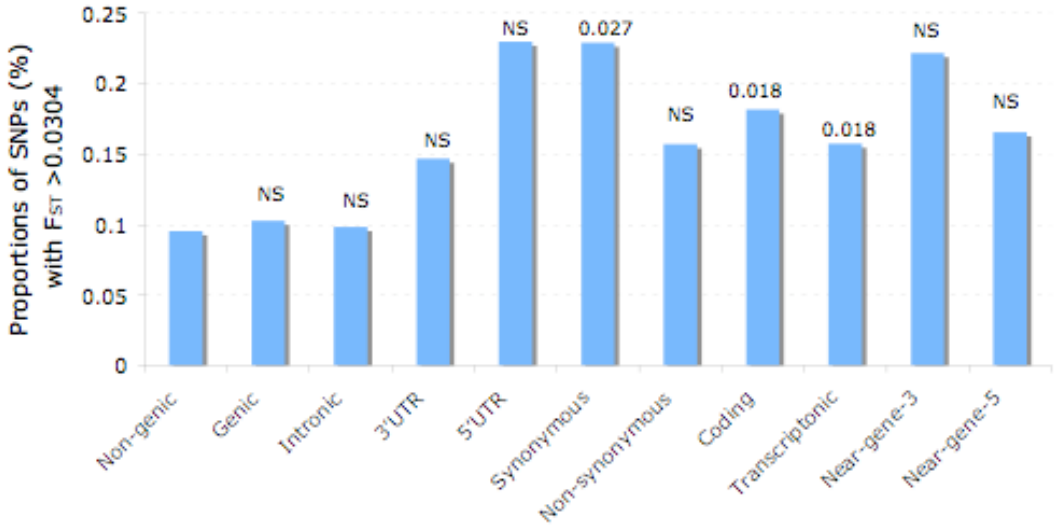


Figure S14. Enrichment of high F_{ST} loci for different SNP categories (MAF>0.05; 99th percentile; F_{ST} >0.01623). Observed excess of high F_{ST} loci for different SNP classes, with respect to nongenic class, among high F_{ST} bin (MAF>0.05; 99th percentile; F_{ST} >0.01623). The values on the bar are p-values of χ^2 tests. "NS" stands for "not significant".

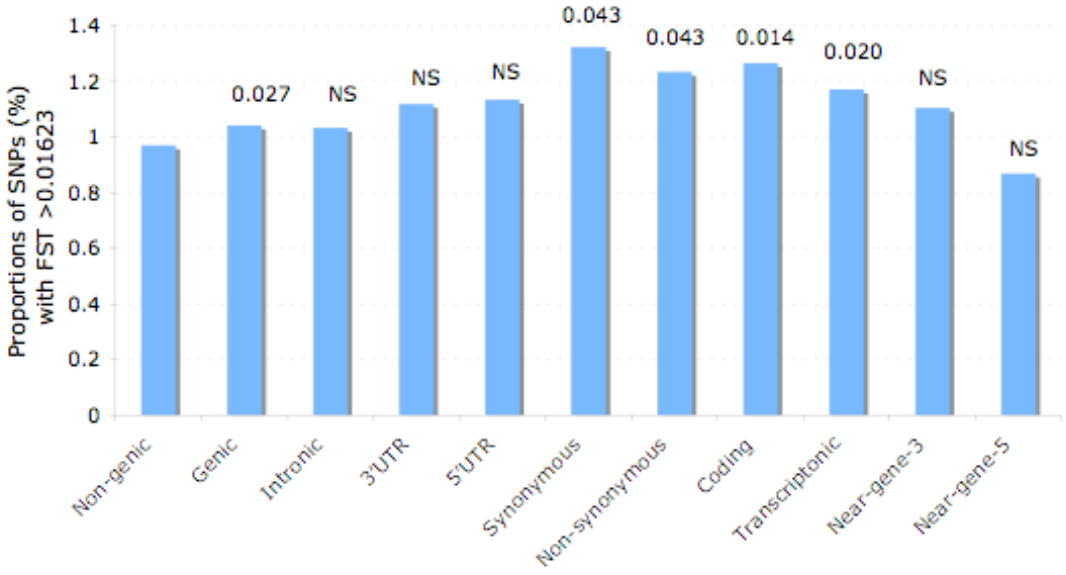
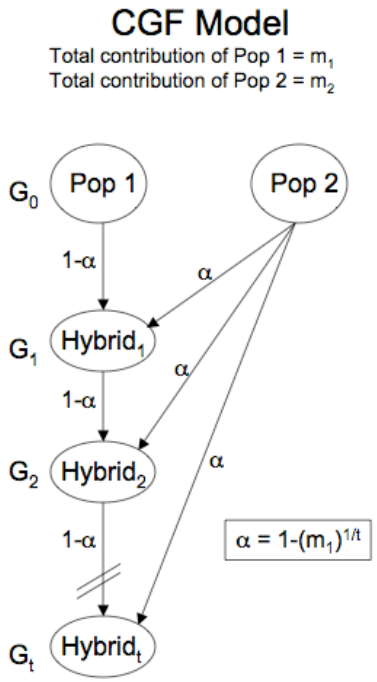


Figure S15. Schematic of continuous-gene-flow (CGF) model adapted from Long (Long 1991) and Pfaff *et al.* (Pfaff et al. 2001).



Tables

Table S1. SNPs showing strong linkage with rs3211938(G) in CD36.

dbSNP_id	Alleles	Allele linked with rs3211938(G)	Frequency (YRI)	Frequency (AAF)	F_{ST} (AAF-YRI)	r^2 (YRI)
rs10216027	C/T	T	0.29	0.16	0.0569	0.514
rs1404315	A/C	C	0.32	0.16	0.0693	0.489
rs1722504	C/T	C	0.38	0.22	0.0652	0.403