Hunter-gatherer genomic diversity suggests a southern African origin for modern humans

Brenna M. Henn[a,1], Christopher R. Gignoux[b], Matthew Jobin[c,d], Julie M. Granka[e], J. M. Macpherson[f], Jeffrey M. Kidd[a], Laura Rodríguez-Botigué[g], Sohini Ramachandran[h], Lawrence Hon[f], Abra Brisbin[i], Alice A. Lin[j], Peter A. Underhill[j], David Comas[g], Kenneth K. Kidd[k], Paul J. Norman[l], Peter Parham[l], Carlos D. Bustamante[a], Joanna L. Mountain[f], and Marcus W. Feldman[e]

[a]Department of Genetics, Stanford University, Stanford, CA 94305; [b]University of California, San Francisco, CA 94158; [c]Department of Anthropology, Santa Clara University, Santa Clara, CA 95050; [d]Department of Anthropology, Stanford University, Stanford, CA 94305; [e]Department of Biological Sciences, Stanford University, Stanford CA 94305; [f]23andMe, Inc., Mountain View, CA 94043; [g]Institute of Evolutionary Biology, Universitat Pompeu Fabra, 08003 Barcelona, Spain; [h]Department of Ecology and Evolutionary Biology, Brown University, Providence, RI 02912; [i]Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, NY 14850; [j]Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford, CA 94305; [k]Department of Genetics, Yale University School of Medicine, New Haven, CT 06520; and [l]Department of Structural Biology, Stanford University, Stanford, CA 94305

This Feature Article is part of a series identified by the Editorial Board as reporting findings of exceptional significance.

Edited by Mary-Claire King, University of Washington, Seattle, WA, and approved February 3, 2011 (received for review November 29, 2010)

Africa is inferred to be the continent of origin for all modern human populations, but the details of human prehistory and evolution in Africa remain largely obscure owing to the complex histories of hundreds of distinct populations. We present data for more than 580,000 SNPs for several hunter-gatherer populations: the Hadza and Sandawe of Tanzania, and the ≠Khomani Bushmen of South Africa, including speakers of the nearly extinct N|u language. We find that African hunter-gatherer populations today remain highly differentiated, encompassing major components of variation that are not found in other African populations. Hunter-gatherer populations also tend to have the lowest levels of genome-wide linkage disequilibrium among 27 African populations. We analyzed geographic patterns of linkage disequilibrium and population differentiation, as measured by F_{ST}, in Africa. The observed patterns are consistent with an origin of modern humans in southern Africa rather than eastern Africa, as is generally assumed. Additionally, genetic variation in African hunter-gatherer populations has been significantly affected by interaction with farmers and herders over the past 5,000 y, through both severe population bottlenecks and sex-biased migration. However, African hunter-gatherer populations continue to maintain the highest levels of genetic diversity in the world.

human evolution | population genetics | Khoisan

African human populations are the most genetically diverse in the world (1–4), but inference about African demographic history, evolution, and disease associations has been limited by relatively few genetic samples and scarce archaeological remains in many regions (5, 6). Modern humans are generally thought to have originated in eastern Africa, where the earliest anatomically modern skulls have been found (5), and because populations outside of Africa carry a subset of the genetic diversity found in eastern Africa (1, 7). Before 5,000 y ago, most of sub-Saharan Africa was sparsely occupied by a collection of linguistically and culturally diverse hunter-gatherer (HG) populations (5). Within the past 5,000 y the majority of HG populations in Africa have disappeared, either through assimilation into expanding agropastoral (farmer and herder) groups or by extinction. The remaining HG groups include the forest Pygmy populations of central Africa, isolated click-speaking populations of Tanzania, and the "Bushmen" of the Kalahari Desert region of southern Africa. Even some of these groups, however, have been experiencing a transition to agricultural subsistence over the past 100 y (8).

The expansion of agropastoralist populations has likely had a major impact on the distribution of genetic variation within Africa (3, 9–12). However, the extent to which these groups interacted with local HG is poorly understood because HG variation has been poorly characterized. Do the current HG populations represent agriculturalists who recently reverted to hunting-gathering or mixed subsistence strategies, as has been seen in other parts of the world (13), or do sub-Saharan HG represent geographically isolated remnants of populations from which other African populations diverged early in human prehistory? A recent study of noncoding autosomal sequences suggests that central African Pygmies and their agriculturalist neighbors diverged approximately 60,000 y ago, but inference of divergence was based on fewer than 500 SNPs (14). Further, a recently published Kalahari "Bushman" genome identified more than 700,000 unique SNPs, suggestive of high and ancient phylogenetic divergence of southern African KhoeSan (15). However, population genetic inference of African demographic history requires both the characterization of intrapopulation variation and the comparison of samples from the numerous geographically dispersed and highly structured populations of this continent. To explore questions about HG demographic history, we present the largest dataset to date of genomic SNPs for the remaining click-speaking (or Khoisan-speaking) HG populations in eastern and southern Africa, analyzed jointly with three other HG and 21 agriculturalist populations, including newly generated data from seven northern African populations.

The Khoisan languages of Africa are unique in incorporating a large variety of click consonants. The Hadza and Sandawe of Tanzania are the only Khoisan-speaking populations to reside outside of southern Africa. However, inclusion of the Hadza and Sandawe in the Khoisan language family is highly contentious among linguists (16, 17). Despite their small population sizes

Author contributions: B.M.H., P.A.U., C.D.B., and M.W.F. designed research; B.M.H., C.R.G., S.R., and A.A.L. performed research; M.J., L.H., A.B., D.C., K.K.K., P.P., and J.L.M. contributed new reagents/analytic tools; B.M.H., C.R.G., M.J., J.M.G., J.M.M., J.M.K., L.R.-B., and P.J.N. analyzed data; and B.M.H., C.R.G., P.J.N., C.D.B., J.L.M., and M.W.F. wrote the paper.

Conflict of interest statement: The authors from 23andMe, Inc. (C.R.G., J.M.M., L.H., and J.L.M.) declare competing financial interests as employees at and stock holders of 23andMe, Inc. SNP arrays designed by 23andMe were used to generate a unique dataset reported in this article. To our knowledge, affiliation with 23andMe, Inc. did not bias the results or discussion of results reported in this article.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

Data deposition: Genotype data will be hosted on a Stanford University website: http://www-evo.stanford.edu/pubs.html.

[1]To whom correspondence should be addressed. E-mail: bmhenn@stanford.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1017511108/-/DCSupplemental.

(the Hadza number only ≈1,000 individuals), the Khoisan-speakers exhibit a large amount of Y-chromosomal and mitochondrial genetic diversity. Many of their Y- and mtDNA lineages are quite rare, but diversity *within* the lineages may date to more than 50,000 y ago. The southern KhoeSan or "Bushmen" populations (see *Materials and Methods* for population nomenclature) today occupy the Kalahari Desert region of Botswana, Namibia, and northern South Africa. Haploid (i.e., Y and mtDNA) genetic studies routinely recognize Khoisan-speaking Ju Bushmen groups as the most divergent population in the world (11, 18, 19). The deepest basal splits in Y-chromosome and mitochondrial phylogenetic trees are between southern Ju Bushmen lineages and eastern African lineages (2). Only a handful of KhoeSan individuals have been characterized in detail for multiple genetic loci (15, 19, 20): these individuals originated from the northern Kalahari/Okavango region. Thus, only a very small amount of the potentially great genetic variation in southern Africa has been examined previously.

Results

We investigated the demographic history and population structure of African HG populations using primarily whole-genome Illumina 650K, 550K, and Affymetrix 500K and 6.0 SNP arrays (*SI Appendix*, Table S1). We also present 550K custom array data for 90 Khoisan-speaking individuals from the Hadza and Sandawe populations of Tanzania and the ≠Khomani Bushmen of South Africa, merged with corresponding data from an additional 24 African populations (*SI Appendix*, *SI Appendix*, Table S1). We use a variety of unique analyses to identify the ancestry of recently admixed genomic segments, severity of population bottlenecks, and the location of ancestral origins within Africa.

Population Structure Within Africa. We identify putative ancestral clusters within sub-Saharan Africa by applying an unsupervised, maximum-likelihood clustering analysis (21) to 12 African and one European population for k possible ancestral populations (k = 2 through 8; Fig. 1). At k = 4, we see a western African/Bantu-speaking cluster, an eastern African cluster, a cluster representing Europeans that likely also signifies ancestral variation maintained in eastern Africa (e.g., Maasai and Sandawe populations), and finally, a cluster that links all our HG populations. Populations in eastern African have the highest proportion of the European ancestral cluster (represented by the Italian Tuscans, HapMap3), supporting prior models of the Out of Africa (OOA) migration originating from a population of eastern African ancestry (1). Higher k values (6 through 8) extract each of the HG populations as a distinct ancestral group [although eastern and western forest Pygmies do not differentiate at these low k (2 through 8) values, they do differentiate at higher k values (3)] (Fig. 1). The Sandawe and Hadza populations contain relatively high amounts of eastern African ancestry (>40%), which at k = 8 splits into Sandawe-specific and Hadza-specific clusters. The high fraction of eastern African ancestry indicates that these groups have likely been part of a greater eastern African population pool for a substantial number of generations. Additionally, at k = 4 through 8 it seems that the South African ≠Khomani Bushmen and the Hadza have both absorbed recent migrants from other clusters; individuals with partial Bantu ancestry in the Hadza, and Bantu and European ancestry in the ≠Khomani Bushmen (Fig. 1). Log-likelihood estimates increased as k was increased, indicating a better fit between the data and model at higher k values (*SI Appendix*, Fig. S1).

We also calculated F_{st} between clusters using the cluster-specific allele frequencies generated from *ADMIXTURE* (Table 1). Because the F_{st} is estimated here from the cluster-based allele frequencies rather than the population-based frequencies, much of the influence of recent migration on our estimates has been removed. Ancient migration and mis-estimation of cluster-based frequencies are potential sources of bias in this approach, but admixture between highly diverged populations, such as the Bushmen and Europeans, is likely to be detected and removed. The southern Bushmen, central forest Pygmies, and the Hadza, compared with Europeans, have F_{st} estimates in excess of 0.23 (Table 1), approximately twice the average F_{st} between other global populations (1). Pairwise comparisons with the Hadza tend to be exceptionally elevated, likely owing to the affect of genetic drift during a recent, severe bottleneck in that population (Fig. 4 *B* and *C*).

Genetic Diversity Within Africa. We merged SNP data for 27 populations across multiple genomic array platforms (*Materials and Methods* and *SI Appendix*, Table S1) and created a set of 55,000 autosomal SNPs common to all platforms. Linkage disequilibrium (LD) was measured using r^2 between pairs of SNPs in sliding 1-Mb windows and estimated from equal sample sizes for

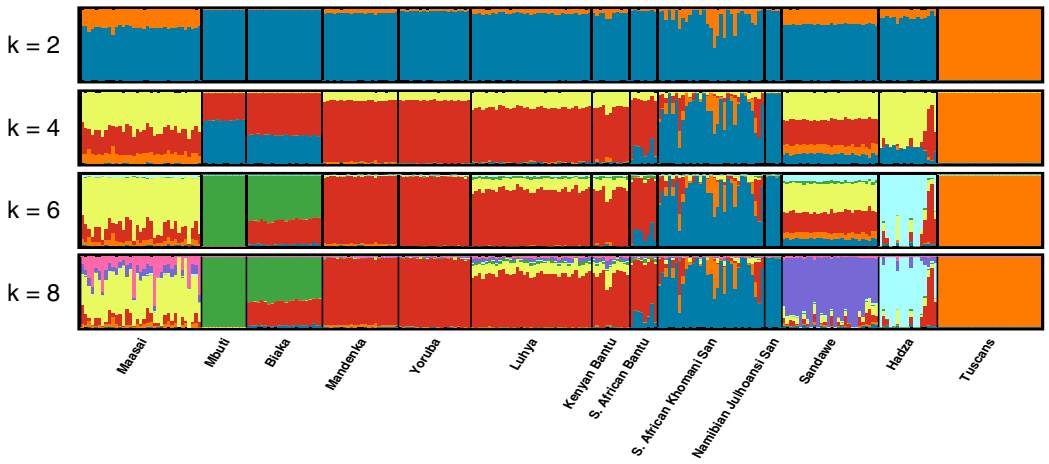


Fig. 1. Ancestral population clusters in sub-Saharan Africa. An unsupervised clustering algorithm, *ADMIXTURE* (21), was used to analyze population structure among 12 sub-Saharan African populations using ≈461K autosomal SNP loci. We plot k: 2, 4, 6, 8 ancestral populations. European Tuscans were included to allow for potential recent European admixture in South Africans. We randomly chose a subset of 30 unrelated Maasai and Luhya for representation in this figure. At k = 4, all HG retain shared ancestry (in blue), and South African Bantu-speakers are likely to have recently absorbed 10–20% KhoeSan ancestry. At k = 8, HG populations emerge with four distinct, ancestral population clusters.

Table 1. F_{st} estimates for k = 8 clusters assigned with *ADMIXTURE*

Cluster*	European	Sandawe	Hadza	Eastern Africa	Maasai†	Western African	Forest Pygmies
European							
Sandawe	0.135						
Hadza	0.256	0.158					
Eastern Africa	0.117	0.054	0.154				
Maasai†	0.172	0.108	0.218	0.104			
Western Africa	0.169	0.053	0.16	0.046	0.103		
Forest Pygmies	0.23	0.102	0.158	0.105	0.167	0.084	
Southern KhoeSan	0.25	0.122	0.222	0.131	0.194	0.115	0.107

*Cluster names correspond to the population or geographic affinities identified in Fig. 1, assuming k = 8 ancestral populations. F_{st} was calculated only on the basis of the allele frequencies estimated within each cluster to exclude recent admixture (e.g., F_{st} between the forest Pygmies and other clusters corresponds to the "green" cluster in Fig. 1).

†Although the majority of Maasai ancestry is assigned to the eastern African (or "yellow") cluster, a second ancestry is also found primarily in the Maasai, indicated in Fig. 1 by the "pink" cluster. Preliminary results suggest that this ancestry is likely of North African origin.

each population. Within Africa, the HG populations tend to have the lowest levels of LD; the five populations with the lowest mean LD are South African ≠Khomani Bushman, Biaka Pygmies, Namibian Bushmen, Fang (not HG), and Tanzanian Sandawe (*SI Appendix*, Fig. S13). LD values are generally indistinguishable between agropastoralist African populations (e.g., Yoruba and Maasai); all northern African populations are elevated relative to sub-Saharan Africa (Fig. 2*A*). LD measured by r^2 is a function of both recombination rate (c) and effective population size (N_e) such that $E(r^2) = 1/(4N_ec)$ (22). Lower values of r^2 in the South African KhoeSan, Biaka, and Sandawe HGs are either the result of larger effective population sizes or higher recombination rates in these populations (23, 24). Population substructure or recent admixture is expected to increase LD (23, 25), and thus our r^2 estimates for African HGs may even be conservatively high given historical gene flow from Bantu-speaking and Europeans into these populations (Fig. 1). Assuming genome-wide recombination rates do not vary greatly between human populations, the LD results are consistent with several African HG groups (e.g., Bushmen, Biaka, and Sandawe) having greater N_e than almost all other African populations. Elevated LD in our samples of the HG Hadza and Mbuti is consistent with recent bottlenecks in these populations [see below and Patin et al. (14)].

To find the best-fit location for an origin of modern humans within Africa, we regressed LD statistics by distance in kilometers from different points on the continent (Fig. 2*B*), in a manner similar to the heterozygosity regressions of Ramanchandran et al. (1). Under a serial founder model, populations further from the origin will have smaller estimated N_e and higher LD on average. Therefore, in this model, positive correlations indicate a good fit with the serial founder model. The highest positive correlations occur when the point of origin is in southwestern Africa (using mean LD at 5 Kb; *SI Appendix*, Fig. S3). Bantu-speaking populations in eastern and southern Africa were removed for this analysis, because they are considered recent migrants and would not reflect the ancient patterns of migration (5, 9, 26), but exclusion of these samples made little difference to the LD maps. Regressions of LD on distance from southwestern Africa were highly statistically significant (at 5-Kb windows, $P \approx 4.9 \times 10^{-6}$) (Fig. 2*C*). Best-fit (*Materials and Methods*) locations based on LD are consistent with a common origin in southern Africa. A point of origin in southwestern Africa was approximately 300–1,000 times more likely than in eastern Africa; detailed likelihood statistics are available in *SI Appendix*, Table S6. The best-fit regression originates at 14°S latitude and 12°E longitude (Fig. 2 *B* and *C*) with an r of 0.78. The geographic pattern in Fig. 2 is driven by exceptionally low LD in the southern Africa KhoeSan and central African Biaka pygmies.

Because LD estimates can be affected by recent admixture, we used a second statistic, F_{st}, to infer the location of origin within Africa. F_{st} was calculated from cluster-specific allele frequencies using the *ADMIXTURE* software at k = 14. Similar to Tishkoff et al. (3), we still detect discrete population membership at k = 14. We used the inferred genetic clusters, rather than the population-based allele frequency estimates, to remove the effect of recent migration on F_{st}. This procedure will, for example, minimize the bias of European gene flow on allele frequency estimates from the South African and Namibian Bushmen samples, and the cluster-based F_{st} estimate will better reflect the divergence of the Bushman population before arrival of recent migrants. This procedure does, however, decrease the number of independent points from 27 in the LD analysis to 13 within Africa. Using F_{st} between each African cluster and Europe (Italy) as a sample OOA population, we performed the regression and permutation analysis of geographic origin as described above for LD. The best-fit regression originates at 20°S latitude and 22°E longitude (Fig. 2*D*), with an r of −0.84. With fewer observations the P value is not as low as in the LD analysis; however, we still find a P value of 0.06 after correcting for multiple tests via 1,000 permutations. Given that the F_{st} estimates are not collinear with the LD decay estimates (because the *ADMIXTURE* model does not take LD into account), this provides an additional line of evidence for a southern origin for modern humans.

To provide an array-independent assessment of heterozygosity we typed the ≠Khomani Bushmen and Hadza at high resolution for *HLA-A*, *B*, and *C*, arguably the most polymorphic of human genes. In comparison with other African populations the Hadza have low *HLA-A*, *B*, and *C* diversity but unusually high frequency (0.375; *SI Appendix*, Table S5) of *HLA-B*44:03*, features often characteristic of small populations that have been subject to pressure from infectious disease (27). In contrast, the ≠Khomani Bushmen have the greatest *HLA-A* heterozygosity (30 alleles, H = 0.95; *SI Appendix*, Table S4), not only among African populations but throughout the world. ≠Khomani Bushmen remain among the most heterozygous populations through the *HLA-B* and *C* loci (H = 0.94 and 0.89, respectively; *SI Appendix*, Table S4). Further attesting to the extraordinary *HLA* diversity and low genome-wide LD of the ≠Khomani Bushmen, we estimated 82 different *HLA-*

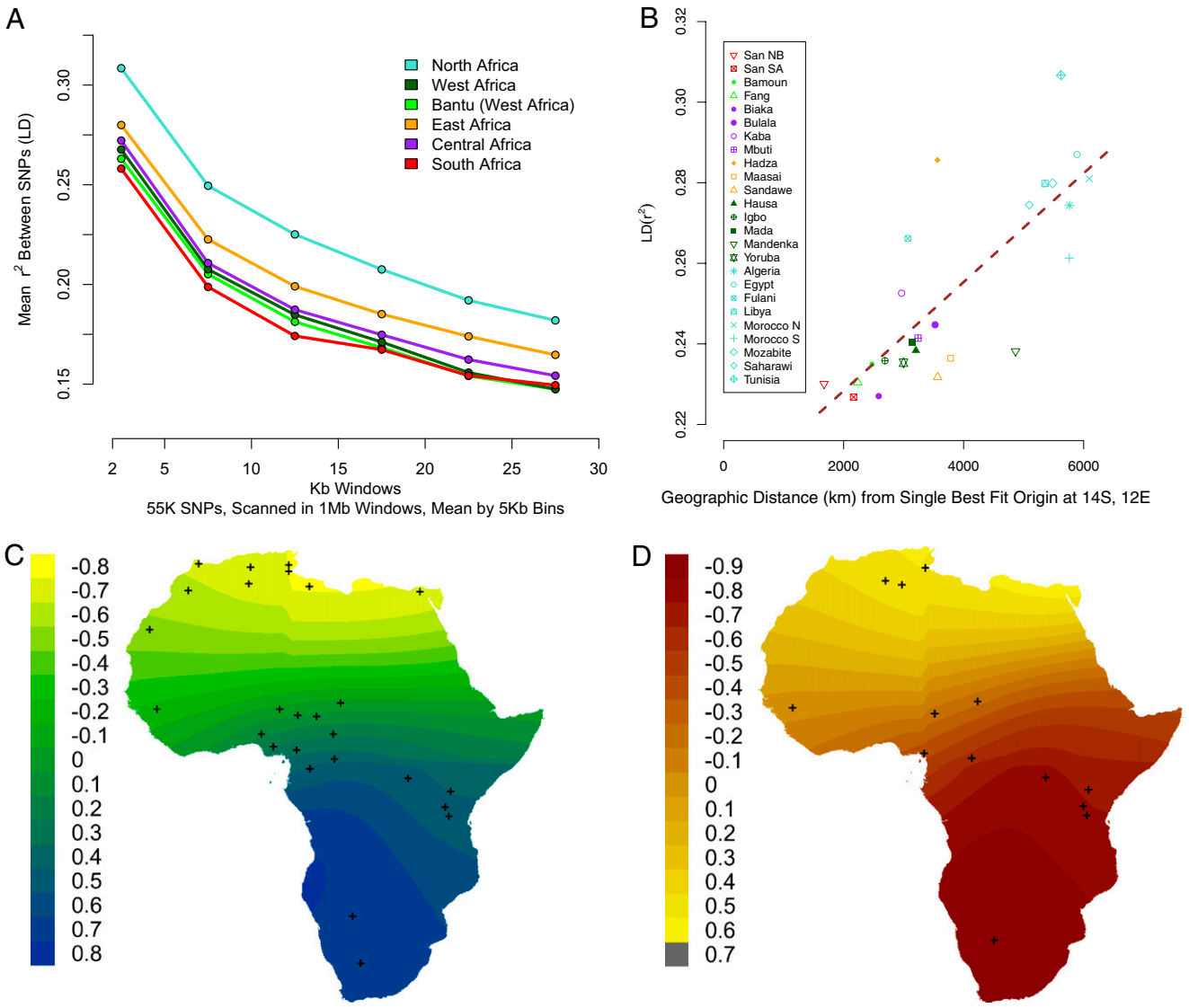


Fig. 2. Genome-wide LD and F_{st} in African populations. (*A*) Each line represents the LD decay averaged across populations within each of six geographic regions; regions are described in *SI Appendix*, Table S1. LD (r^2) between SNPs was calculated in sliding 1-Mb windows. The r^2 estimates were binned by the genetic distance between SNPs in 5-Kb bins. HG populations have the lowest LD curves (*SI Appendix*, Fig. S13 shows population-specific LD decay curves). (*B*) We assessed a possible point of origin by regressing LD onto geographic distance. Regression for the single best fit for geographic origin is shown, with a correlation coefficient of 0.78, to the point 14°S, 12°E. (*C*) Map is shown using mean LD within 0–5 Kb. The highest correlation coefficient in blue indicates the best fit with a potential geographic origin. Crosses indicate the geographic coordinates of the sampled populations. (*D*) We assessed a possible point of origin by regressing F_{st} between African populations and Europe onto geographic distance, using a waypoint in the Near East. Populations were grouped for F_{st} analysis; crosses indicate the geographic mean for each population grouping.

A, *B*, and *C* haplotypes in the panel studied, compared with a mean of 25 haplotypes per population worldwide (*SI Appendix*).

Demographic History of the Khoisan HG. All HG populations in our study segregated into unique population clusters, separate from neighboring agriculturalist populations, indicating that the HG groups derive from distinct ancestral populations (Fig. 1 and *SI Appendix*, Fig. S2). We asked whether there was evidence of recent gene flow between HG and agriculturalists by examining the population affiliation of both the haploid systems and the autosomes. Y-chromosome and mtDNA haplogroups were assigned on the basis of approximately 2,000 SNPs per haploid system. Virtually every individual in the ≠Khomani Bushmen sample carried a highly diverged mtDNA haplogroup that is indigenous to southern Africa, 93% L0d1; two individuals carried basal pre-L0a'f lineages that have an unknown geographic distribution, but given that KhoeSan individuals retain many of the deep divergence mtDNA haplogroups, it is likely that this lineage is also endemic to the KhoeSan (18) (*SI Appendix*, Table S3). However, ≈35% of the paternal lineages were of either European or Bantu origin. Across all three click-speaking populations the contribution of maternal haplogroups from putative agriculturalists varied from 0 to 28%, whereas the corresponding paternal contribution of agriculturalists varied from 30% to 42% (*SI Appendix*, Table S3).

We used phased data from admixed individuals to visualize the distribution of ancestry along each chromosome, assuming the number of ancestral populations was equal to three. A principal components (PCA)-based analysis of genomic local ancestry (28) in 40 SNP windows was implemented for the Khoisan-speaking populations (*Materials and Methods*). Motivated by the *ADMIXTURE* results in Fig. 1, ancestral populations for the potentially

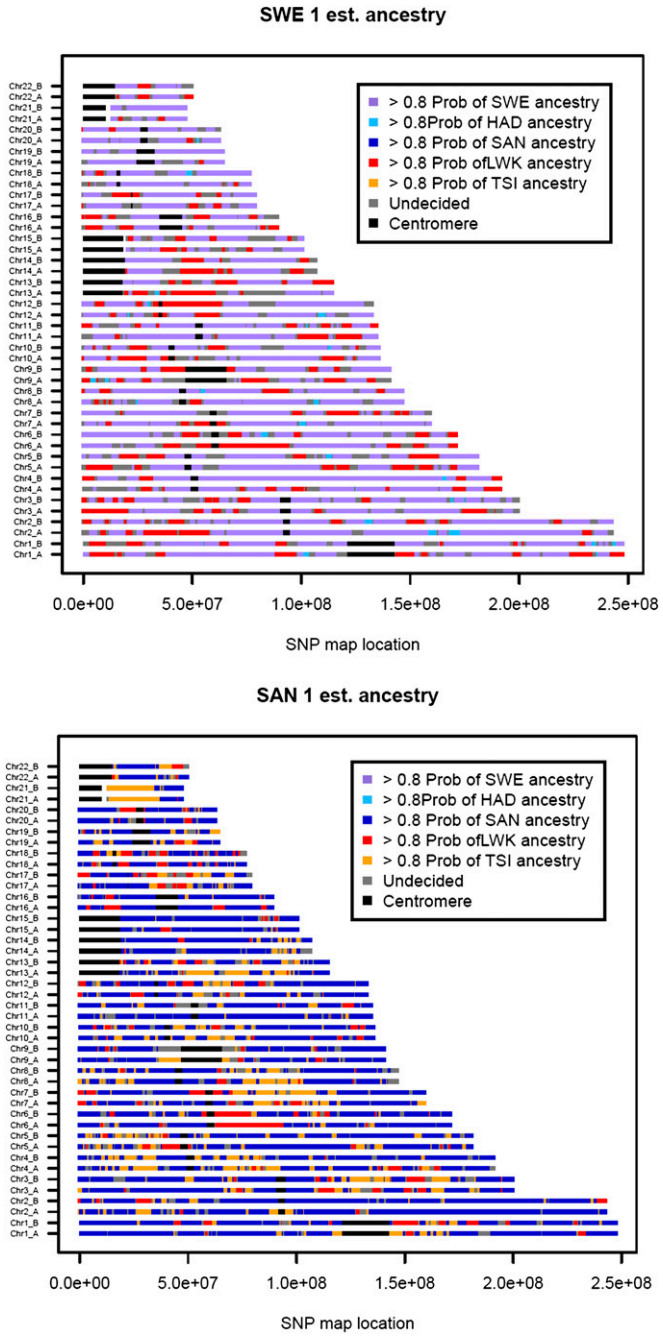


Fig. 3. Local ancestry assignment along phased chromosomes. Two individuals with potential admixture (Fig. 1) were projected onto the principal component space of three putative ancestral populations. The three ancestral populations differed for Sandawe individuals (SWE, Sandawe; HAD, Hadza; LWK, Luhya Bantu) and South African ≠Khomani Bushmen (SAN, KhoeSan; TSI, European Tuscan; LWK, Luhya Bantu). Ancestry was assigned in 40-SNP windows along phased chromosomes (haplotypes A and B) by calculating the minimal distance to an ancestral population (28). Ancestry from Bantu-speaking agriculturalists seems to have occurred relatively recently, as indicated by many ≥10-Mb segments. Switch errors in the phasing could potentially shorten the length of these migrant tracts, but with low levels of admixture, phase switch errors are less likely to lengthen inferred migrant tracts.

admixed Sandawe individuals were represented by the Sandawe, Hadza, and Kenyan Luhya (Bantu) populations. As illustrated in Fig. 3*A*, for one Sandawe genome, segments of probable Bantu ancestry tended to be long, often spanning more than 10 Mb, and very little Hadza ancestry could be detected. In the South African ≠Khomani Bushmen, we assumed three possible source populations: ≠Khomani Bushmen, Kenyan Luhya, or European Tuscans. One example, individual SA1 (Fig. 3*B*), did not report European ancestry among parents or grandparents, yet displays a fairly high amount of apparent European ancestry.

Distributions of the cumulative amount of runs of homozygosity (cROH, where we define a run of homozygosity to be >500 Kb) were calculated for all individuals in the three click-speaking populations (Fig. 4) (*Materials and Methods*). The Tanzanian Hadza population differed strikingly from the other two groups; approximately 65% of Hadza individuals have a cROH >100 Mb, and on average the fraction of the genome in runs of homozygosity (fROH) is 8% in the Hadza. We found no strong correlation between the number of missing genotypes and cROH (*SI Appendix*, Fig. S5). In comparison with the 52 Human Genome Diversity Project (HGDP) populations, higher fROH values are found in four Native American populations, which have been subject to serial founder events.

The elevated mean and variance of cROH in the Hadza population is indicative of a severe and possibly recent population bottleneck. To test this hypothesis we performed a number of demographic simulations using a rejection-based approximate Bayesian algorithm (29) (REJECTOR; *Materials and Methods*), focusing on the summary statistic fROH. On the basis of an observed $fROH_{Hadza}$ of 8%, the most likely estimate of effective population size (N_e) of the Hadza was 2,590 [95% confidence interval (CI) 475–21,222]; current census size of the Hadza is thought to be only 1,000 (30) (Fig. 4*B*). Single population estimates of N_e vary widely (19, 31), but among a global sample of HGDP populations only Native Americans and the Kalash have an N_e <3,000. We estimate the mean severity of a bottleneck in the Hadza to be sixfold (95% CI 0.4–9.67) (Fig. 4*C*). The fROH statistic did not seem to be sensitive to the time of the bottleneck. The corresponding estimates of mean N_e in the neighboring Sandawe are 7,000 (95% CI 3,340–23,200) and 11,670 in the ≠Khomani Bushmen (95% CI 5,760–28,650). Posterior distributions of bottleneck severity were flat for the Sandawe and Bushmen (*SI Appendix*, Fig. S6).

Loci Under Selection. We scanned five African HG populations for ongoing selective sweeps using the iHS statistic (32) (*SI Appendix*, Fig. S8). We identified regions of the genome characterized by extreme iHS values in each population using the procedures described by Pickrell et al. (33). The most extreme 1% regions generally do not overlap between HG groups. We find eight regions that are in the most extreme 10 loci from a single population and are also in the top 5% for at least two additional African HG populations. These regions include genes such as *POT1* and *TUSC3*, associated with carcinogenesis, a section on chromosome 6 (30.2–30.6, Build Hg36) of the HLA region that contains several tripartite motif-containing genes that have been implicated in pathogen response, and a region on chromosome 9 that is in the top 10% in all five African HG populations and contains nine genes, including *IKBKAP* mutations, which result in familial dysautonomia (*SI Appendix*, Fig. S8).

We also surveyed a subset of known loci with functions related to infectious disease and that may have been under recent selection owing to increased population density and contact between groups. One SNP, *rs2395029*, is a missense Val-to-Gly mutation in the HLA complex P5 (HCP5); the function of this protein is not well understood. The G allele, likely ancestral, has been shown to be associated with reduced viral load in HIV-positive individuals, explaining 10% of the variance in HIV viral load among infected Europeans (34). A second genome-wide association study also found that in European individuals the G allele is associated with HIV nonprogression (35) (odds ratio

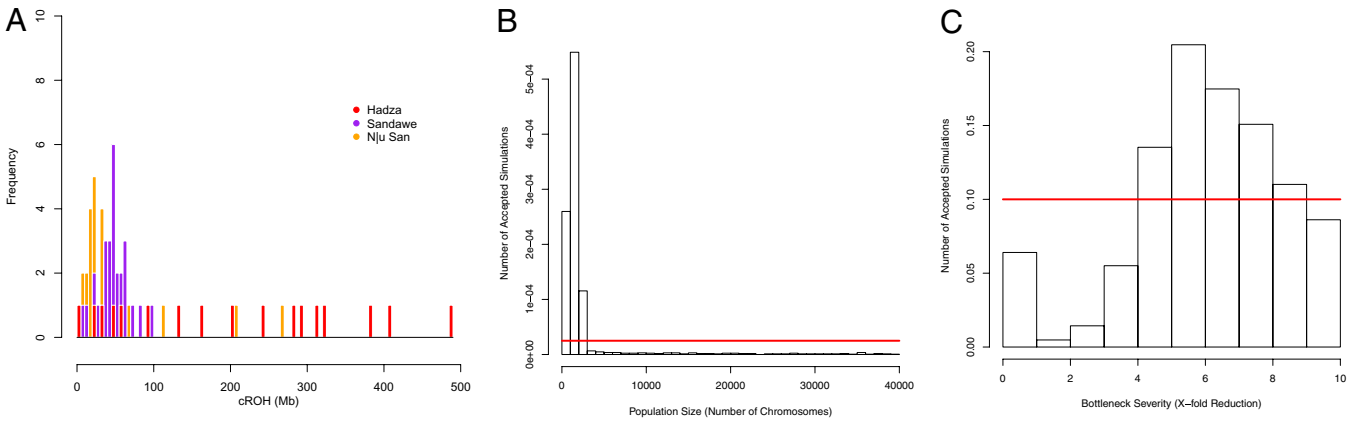


Fig. 4. Runs of homozygosity among Khoisan-speakers. (*A*) Long runs of homozygosity were calculated for individuals in the Hadza, Sandawe, and ǂKhomani Bushmen populations. Runs were constrained to a minimum of 1 Mb, and two missing genotypes were allowed per run. cROH are plotted for all individuals; the *y*-axis represents counts of individuals. The Hadza distribution differs markedly from the other two populations, with 65% (n = 11/17) of individuals having cROH >100 Mb. This distribution is consistent with a severe, recent bottleneck in the Hadza. (*B*) Simulated posterior distribution of effective population size in the Hadza, generated by sampling from a uniform distribution of N_e and keeping simulated parameters within 20% of the observed fROH with REJECTOR (29). (*C*) Simulated posterior distribution of bottleneck severity in the Hadza, as modeled above.

3.47). We found our sample of the Sandawe to carry rs2395029 at an elevated 17% frequency (excluding close relatives), substantially higher than in other populations from HGDP– Centre d'Etude du Polymorphisme Humain (CEPH) (*SI Appendix*, Fig. S7). The SNP is absent in the Hadza and ǂKhomani Bushmen. The iHS scan, meant to measure ongoing or incomplete selective sweeps, gave a score of 2.79 (absolute value) to the rs2395029 allele in the Sandawe population; values >2 are generally indicative of selection (32) (although this window is not within the top 1% of the iHS scores). The cross-population extended haplotype homozygosity XP-EHH and composite likelihood ratio selection statistics also find the HCP5 locus significant at an empirical $P <$ 0.01 level in the Sandawe (but not in other HG populations), with Europeans used for the cross-population comparison.

Discussion

Population Structure. It is unclear to what extent the ancestral African population was structured at the time of the OOA migration by modern humans, ≈60,000 y ago (36, 37). Deep ancestral population structure within Africa would affect inference of divergence time among African and non-African populations, extent of admixture between expanding modern humans and Neanderthals (38), and the causes of divergence during the origin of modern humans (i.e., demographic growth, unique mutations affecting behavior, etc.). By examining genomic data from six HG populations from southern, central, and eastern Africa (i.e., Ju and ǂKhomani Bushmen, Mbuti and Biaka, Hadza and Sandwe) in conjunction with 21 other African populations, we characterized the structure and relationship between these populations. Frequentist population structure analyses (Fig. 1 and *SI Appendix*, Fig. S2) demonstrate that the HG populations are highly differentiated, both among themselves and compared with other African populations (19, 39). Thus, they do not seem to derive from agriculturalist populations and instead represent geographically distinct populations isolated from other groups for thousands of years.

However, all African HG populations do share an ancestral cluster at small k levels (Fig. 1). A similar phenomenon was observed in the Tishkoff et al. (3) microsatellite *structure* analysis and suggests that central African Pygmies, southern KhoeSan, and Tanzanian click-speakers share common ancestry distinct from other African populations, either through a more recent common ancestor or through long-distance migration. Population divergence between the eastern and western Pygmy populations, here represented by the Mbuti and Biaka, respectively, is estimated to have occurred approximately 15,000–27,000 y ago (14, 40, 41). Our Y-chromosome microsatellite data from the highly diverged, basal KhoeSan A3b1-M51 clade, found exclusively in southern Africa, has a most recent common ancestor estimated at 40,000 y ago (*SI Appendix*, Fig. S11) (42). If the time to the most recent common ancestor (TMRCA) *within* the KhoeSan population is 40,000 y ago, then this is among the first genetic evidence that the KhoeSan have continuously occupied southern Africa for at least the entire span of the Upper Paleolithic (43). This conclusion assumes that any past geographic movement of the KhoeSan ancestral population would have spread the distribution of A3b1 lineages, and these lineages would survive in contemporary populations. African HG populations seem to have been highly structured at the time of the OOA migration, but further sophisticated modeling of autosomal data are necessary to confirm this.

Geographic Origins and Population Size. LD decays more rapidly between SNPs in African populations than in non-African populations, and single-gene studies indicate there is variation in levels of LD among African populations (44, 45) (Fig. 2*A*). We present LD decay plots based on genome-wide data for 27 African populations. Of the five populations with the lowest levels of LD, four are HG populations: the ǂKhomani and Namibian Bushmen, Sandawe and Biaka Pygmies; the Fang from Cameroon also have very low LD. This suggests that the largest human effective population sizes have been in African HG populations. We estimate N_e in the ǂKhomani Bushmen to be high, namely 11,600, on the basis of fROH simulations. We note, however, that an estimate of N_e using the fraction of the genome in runs of homozygosity is an indicator of relatively recent effective population size, not of the ancestral population living tens of thousands of years ago.

Recently Tishkoff et al. (3) suggested a potential origin for modern humans in southern Africa, on the basis of heterozygosity estimates from microsatellite data. However, their sample of KhoeSan was small, and the directionality of a southwestern origin of humans based on heterozygosity could have been driven by the inclusion of a highly admixed (and thus highly heterozygous) "Coloured" population. To account for these concerns, we used a large sample from two KhoeSan populations (n = 47) and different statistics: F_{st} and mean genomic LD. At levels of mean LD between 0 and 30 Kb, the best-fit geographic origin is found

in southwestern Africa, likely driven by the exceptionally low LD in the Bushmen (South Africa and Namibia) and Biaka Pygmies (Central African Republic). F_{st} comparisons also support a southern African origin. Geographic dispersal of modern humans from southern Africa is consistent with the earliest archaeological evidence for worked bone awls, inscribed ostrich eggshell, and climatic evidence indicating a more hospitable climate in southern Africa than eastern Africa until 60,000–70,000 y ago (46–48). The large number of unique SNPs not present in dbSNP but identified in the Namibian KB1 (15) "Bushman" genome is also suggestive of the greater diversity present in these KhoeSan populations, but resequencing of larger sample sizes from many geographically separated African populations is required to confirm this conclusion.

Extremely elevated LD, increased runs of homozygosity, and very low HLA and haplotype heterozygosity estimates in comparison with other sub-Saharan African populations (Figs. 2 and 4 and *SI Appendix*, Tables S2 and S4) all indicate a severe population bottleneck in the Hadza. We estimate their N_e to be only 2,500 individuals, which corresponds to an ancestral population size of 15,000 (given the estimated sixfold bottleneck). Strikingly, our N_e estimate is more than twice the current census size of the Hadza, suggesting that the bottleneck is ongoing and that historically the Hadza had a much larger population size. Either our current population sample reflects a substantial founder effect from a larger eastern African HG population, or the entire population has experienced a bottleneck that narrowed its geographic distribution to present day Lake Eyasi. The geographic extent of the historic Hadza population is unknown and is complicated by the lack of evidence for substantial gene flow between the Hadza and neighboring Sandawe, who live only 150 km away (Fig. 3*A*).

Selection in HG. We find little overlap between HG populations in ongoing or incomplete selective sweeps (*SI Appendix*, Fig. S8), even between populations in similar environments such as the central African Pygmies or the Tanzanian Hadza and Sandawe. This is consistent with the ancient and continuing separation of these groups (Figs. 1 and 2) and the global observation in Pickrell et al. (33) that ongoing selective sweeps tend to be regionally specific. Direct comparison of haplotype-based statistics, such as iHS, between samples can be confounded by differences in SNP selection and phasing (*SI Appendix*). The five HG populations examined here (*SI Appendix*, Fig. S8) were jointly phased for the same 461K SNPs with a diverse set of trio-based haplotype seeds from Yoruba, Maasai, and Europeans and duo-seeds from the Sandawe and ≠Khomani Bushmen.

We identified evidence for positive selection on the HCP5 locus in the Tanzanian Sandawe together with an increased frequency of the rs2395029 [G] allele. There was no evidence of a recent demographic bottleneck in the Sandawe that could explain such a high allele frequency by drift (*SI Appendix*). The rs2395029 [G] allele has been associated with protection from HIV by virtue of LD with *HLA-B*57:01*, where putative causal amino acid variants have been identified (34). We found that the G allele is *not* in LD with *HLA-B*57:01* in the Sandawe; instead it tends to be on the background of *HLA-B*49:01*, which encodes none of the HIV-protective amino acids of B*57. Additionally, in Europeans rs2395029 [G] is in nearly complete LD with *HLA-B*57:01* and is used as a diagnostic marker for hypersensitivity to abacavir, a drug prescribed during HIV therapy (49). Our results suggest both that the HPC5 locus is under strong selection in the Sandawe independently of B*57:01 and that rs2395029 [G] is not a diagnostic marker for B*57:01 in individuals with African ancestry, among whom this allele has recombined onto a variety of haplotypes. It remains to be determined whether the selection has been due to an increase in HIV prevalence in Tanzania or another infectious disease.

Conclusions

We present a targeted analysis of African HG genomes, including genomic SNP array data for the Tanzanian Hadza, Sandawe, and the South African ≠Khomani Bushmen. We find that the six sub-Saharan HG populations share common ancestry *distinct* from agriculturalists (Fig. 1 and *SI Appendix*, Fig. S2). Analyses of LD and heterozygosity suggest that HGs, especially the click-speaking ≠Khomani and Namibian Bushmen, are among the most diverse of all human populations. LD and F_{st} patterns support the hypothesis that all human populations originated from southern Africa (Fig. 2 and Table 1), although we caution that sampling in Africa remains sparse; our sample size of Bushmen numbers just 47 individuals, and other highly variable populations may be discovered. Additionally, the interaction between agropastoralist populations and HG over the past 5,000 y has markedly affected HG populations in a variety of ways, including sex-biased gene flow from agropastoralists into HG groups, demographic bottlenecks [in the Hadza and central Pygmies (41)], and possibly unique selection pressures, for example those associated with infectious disease. We suggest that HG groups in Africa are currently experiencing rapid evolutionary change, as reflected in their patterns of genomic diversity.

Materials and Methods

Samples. Sampling of the ≠Khomani Bushmen in Upington, South Africa and neighboring villages occurred in 2006. Institution review board approval was obtained from Stanford University. ≠Khomani Nlu-speaking individuals, local community leaders, traditional leaders, nonprofit organizations, and a legal counselor were all consulted regarding the aims of this research before collection of DNA. All individuals consented orally to participation. DNA via saliva (Orogene kits) and ethnographic information regarding ancestry, language, and parental place of birth were collected for all participants. Hadza individuals were collected in the Arusha district of Tanzania; the Commission for Science and Technology and the National Institute for Medical Research of Tanzania approved the sample collection (50). Sandawe individuals were collected in 2004 [informed consent described in Li et al. and Donnelly et al. (51, 52)]. Samples from the ≠Khomani Bushmen, Hadza, and Sandawe were typed on the Illumina Beadchip 550K custom v2 chip designed by 23andMe, Inc. and processed through 23andMe's partner laboratory. Genomic data for comparable loci were also obtained from HGDP-CEPH (19) and HapMap3 (http://hapmap.ncbi.nlm.nih.gov/), publicly available resources. These genotype data were merged with samples typed on the Affymetrix 6.0, Affymetrix 500K, Illumina 1M for a total of 27 African populations (using PLINK v1.07; *SI Appendix*, Table S1). Data for western African populations genotyped on the Affymetrix 500K were obtained from Bryc et al. (28) (available in the database of Genotypes and Phenotypes), and North African population data generated for this manuscript, shown in *SI Appendix*. A total of 55,000 SNPs were identified as being common to all platforms (available at www-evo.stanford.edu/pubs.html).

Nomenclature. We use the term "Khoisan" to refer to the linguistic family consisting of click-speaking populations as defined by Greenberg (53). The term "KhoeSan" refers to both the pastoralist and HG Khoisan-speakers of southern Africa, including the Nama, Jul'hoansi, and Nlu. "San" is indicative of a historic HG subsistence lifestyle, regardless of the specific Khoisan southern African subfamily language affiliation and has been often used in the literature interchangeably with "Bushmen." Many individuals in our sample preferred to be referred to as "Bushmen."

Population Structure. An unsupervised clustering algorithm, *ADMIXTURE* (21), was run on our three unique Khoisan-speaking populations, HGDP-CEPH sub-Saharan Africans, HapMap3 Kenyan Luhya, Maasai, and Italian Tuscans. Eight ancestral clusters (k = 1 through 14) in total were tested successively. Log-likelihoods for each k clusters are shown in *SI Appendix*. Replicate runs of k ancestral populations did not result in substantially different individual ancestral frequencies. F_{st} based on allele frequencies was calculated in *ADMIXTURE* for each identified cluster at k = 8 and k = 14.

Phasing. Parent-offspring (PO) pairs were first identified using segments sharing identity by descent; PO pairs in the Sandawe and ≠Khomani San (absent in the Hadza) were phased separately using BEAGLE software (54). The phased Khoisan PO pairs were then combined with HapMap3 trio-

phased Yoruba, Maasai, and European population samples (n = 30) for a larger seed ancestral haplotype pool. HGDP African populations, Tanzanian Hadza and Sandawe, and South African ≠Khomani San were then phased together in BEAGLE.

Runs of Homozygosity and Rejection Algorithm. Long runs of homozygosity were calculated with PLINK v1.07 software (55) (http://pngu.mgh.harvard.edu/~purcell/plink/download.shtml) under the following parameters: SNP density of 50 SNPs per window, minimum ROH length of 1 Mb, two missing genotypes per window allowed, one heterozygote per window allowed, sliding window of 1 Mb, minor allele frequency 5%. These parameters were similar to those chosen by Nalls et al. (56) and Auton et al. (57) for their calculations of the accumulated ROH. Estimates of the fROH for rejection algorithm simulations used the same parameters as above, except that all missing genotypes had been imputed in BEAGLE. We investigated sensitivity to the number of missing genotypes and correlation between cROH and the number of ROH segments (*SI Appendix*, Fig. S5). Using a rejection algorithm program we inferred demographic history by comparing simulated cROH estimates with observed cROH estimates. Extensive documentation of the rejection algorithm method, accuracy, and the software REJECTOR (29) package are available online (*SI Appendix* and www.rejector.org).

Linkage Disequilbrium. LD, r^2, was calculated between all SNPs in sliding 1-Mb windows. Pairs of SNPs were then binned by their genomic distance in 5-Kb bins. Each population was randomly subsampled 10 times for 12 individuals, LD decay was calculated independently, and mean LD for each bin was averaged across the 10 subsamples for the final population estimate (Fig. 2). To infer goodness-of-fit of a serial founder effect model across the continent of Africa using LD, we first created a 1° × 1° grid of latitude/longitude points across the continent. For each of these, we calculated great circle distances from our potential point-of-origin to every population. For distances between western and southern Africa we used a waypoint in Cameroon to reflect realistic land-based migration. We then calculated the regression of LD on distance, estimated via smoothing the empirical LD decay measurements with exponential decay functions. Best-fit was identified by the explained variance "r." This method was similarly applied to the pairwise F_{st} estimates between African populations and Europe, using nonadmixed allele frequencies estimated from *ADMIXTURE*, k = 14, and with an additional waypoint in the Near East for the OOA populations. For clusters consisting of multiple populations, such as western African Bantu-speakers, the geographic location of the cluster F_{st} was calculated as the average distance to the centroid for the longitude and the sine average distance to the centroid for the latitude. We used MapViewer (http://www.goldensoftware.com) to create the Kriging interpolation plot of the correlation coefficients.

ACKNOWLEDGMENTS. We thank Andy Reynolds and Mike Palmer for technical assistance; Dara Torgerson, Lindsey Roth, and Saunak Sen for discussion; the University of California, San Francisco (UCSF) Biostatistics High Performance Computing System for computing resources; and the Tanzanian and South African participants who generously contributed DNA. This work was funded by the Center for Human Origins and Evolution, the Morrison Institute for Population and Resource Studies at Stanford University, a UCSF Chancellor's Graduate Research Fellowship (to C.R.G.), and National Institute of Health Grants 3R01HG003229 and 3R01GM028016.

1. Ramachandran S, et al. (2005) Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proc Natl Acad Sci USA* 102:15942–15947.
2. Semino O, Santachiara-Benerecetti AS, Falaschi F, Cavalli-Sforza LL, Underhill PA (2002) Ethiopians and Khoisan share the deepest clades of the human Y-chromosome phylogeny. *Am J Hum Genet* 70:265–268.
3. Tishkoff SA, et al. (2009) The genetic structure and history of Africans and African Americans. *Science* 324:1035–1044.
4. Vigilant L, Stoneking M, Harpending H, Hawkes K, Wilson AC (1991) African populations and the evolution of human mitochondrial DNA. *Science* 253:1503–1507.
5. Phillipson DW (2005) *African Archaeology* (Cambridge Univ Press, Cambridge, UK).
6. Mitchell P (2010) Genetics and southern African prehistory: An archaeological view. *J Anthropol Sci* 88:73–92.
7. Gonder MK, Mortensen HM, Reed FA, de Sousa A, Tishkoff SA (2007) Whole-mtDNA genome sequence analysis of ancient African lineages. *Mol Biol Evol* 24:757–768.
8. Newman JL (1970) *The Ecological Basis for Subsistence Change Among the Sandawe of Tanzania* (National Academies, Washington, DC).
9. Beleza S, Gusmão L, Amorim A, Carracedo A, Salas A (2005) The genetic legacy of western Bantu migrations. *Hum Genet* 117:366–375.
10. Destro-Bisol G, et al. (2004) Variation of female and male lineages in sub-Saharan populations: The importance of sociocultural factors. *Mol Biol Evol* 21:1673–1682.
11. Tishkoff SA, et al. (2007) History of click-speaking populations of Africa inferred from mtDNA and Y chromosome genetic variation. *Mol Biol Evol* 24:2180–2195.
12. Sikora M, Laayouni H, Calafell F, Comas D, Bertranpetit J (2011) A genomic analysis identifies a novel component in the genetic structure of sub-Saharan African populations. *Eur J Hum Genet* 19:84–88.
13. Waters T (2005) Comment on "Recent origin and cultural reversion of a hunter-gatherer group". *PLoS Biol*, 3: e269, author reply e270.
14. Patin E, et al. (2009) Inferring the demographic history of African farmers and pygmy hunter-gatherers using a multilocus resequencing data set. *PLoS Genet* 5:e1000448.
15. Schuster SC, et al. (2010) Complete Khoisan and Bantu genomes from southern Africa. *Nature* 463:943–947.
16. Güldemann T, Stoneking M (2008) A historical appraisal of clicks: A linguistic and genetic population perspective. *Annu Rev Anthropol* 37:93–109.
17. Sands B, Güldemann T (2009) What click languages can and can't tell us about language origins. *The Cradle of Language*, eds Botha R, Knight C (Oxford University Press, Oxford), p 204.
18. Behar DM, et al.; Genographic Consortium (2008) The dawn of human matrilineal diversity. *Am J Hum Genet* 82:1130–1140.
19. Li JZ, et al. (2008) Worldwide human relationships inferred from genome-wide patterns of variation. *Science* 319:1100–1104.
20. Jakobsson M, et al. (2008) Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* 451:998–1003.
21. Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* 19:1655–1664.
22. Hill W, Robertson A (1968) Linkage disequilibrium in finite populations. *Theor Appl Genet* 38:226–231.
23. Pritchard JK, Przeworski M (2001) Linkage disequilibrium in humans: Models and data. *Am J Hum Genet* 69:1–14.
24. Serre D, Nadon R, Hudson TJ (2005) Large-scale recombination rate patterns are conserved among human populations. *Genome Res* 15:1547–1552.
25. Wilson JF, Goldstein DB (2000) Consistent long-range linkage disequilibrium generated by admixture in a Bantu-Semitic hybrid population. *Am J Hum Genet* 67: 926–935.
26. Berniell-Lee G, et al. (2009) Genetic and demographic implications of the Bantu expansion: Insights from human paternal lineages. *Mol Biol Evol* 26:1581–1589.
27. Gendzekhadze K, et al. (2009) Co-evolution of KIR2DL3 with HLA-C in a human population retaining minimal essential diversity of KIR and HLA class I ligands. *Proc Natl Acad Sci USA* 106:18692–18697.
28. Bryc K, et al. (2010) Genome-wide patterns of population structure and admixture in West Africans and African Americans. *Proc Natl Acad Sci USA* 107:786–791.
29. Jobin MJ, Mountain JL (2008) REJECTOR: Software for population history inference from genetic data via a rejection algorithm. *Bioinformatics* 24:2936–2937.
30. Blurton Jones NG, Smith LC, O'Connell JF, Hawkes K, Kamuzora CL (1992) Demography of the Hadza, an increasing and high density population of Savanna foragers. *Am J Phys Anthropol* 89:159–181.
31. Conrad DF, et al. (2006) A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nat Genet* 38:1251–1260.
32. Voight BF, Kudaravalli S, Wen X, Pritchard JK (2006) A map of recent positive selection in the human genome. *PLoS Biol* 4:e72.
33. Pickrell JK, et al. (2009) Signals of recent positive selection in a worldwide sample of human populations. *Genome Res* 19:826–837.
34. Fellay J, et al. (2007) A whole-genome association study of major determinants for host control of HIV-1. *Science* 317:944–947.
35. Limou S, et al.; ANRS Genomic Group (2009) Genomewide association study of an AIDS-nonprogression cohort emphasizes the role played by HLA genes (ANRS Genomewide Association Study 02). *J Infect Dis* 199:419–426.
36. Harding RM, McVean G (2004) A structured ancestral population for the evolution of modern humans. *Curr Opin Genet Dev* 14:667–674.
37. Plagnol V, Wall JD (2006) Possible ancestral structure in human populations. *PLoS Genet* 2:e105.
38. Green RE, et al. (2010) A draft sequence of the Neandertal genome. *Science* 328: 710–722.
39. Zhivotovsky LA, Rosenberg NA, Feldman MW (2003) Features of evolution and expansion of modern humans, inferred from genomewide microsatellite markers. *Am J Hum Genet* 72:1171–1186.
40. Quintana-Murci L, et al. (2008) Maternal traces of deep common ancestry and asymmetric gene flow between Pygmy hunter-gatherers and Bantu-speaking farmers. *Proc Natl Acad Sci USA* 105:1596–1601.
41. Batini C, et al. (2011) Insights into the demographic history of african pygmies from complete mitochondrial genomes. *Mol Biol Evol* 28:1099–1110.
42. Cruciani F, et al. (2002) A back migration from Asia to sub-Saharan Africa is supported by high-resolution analysis of human Y-chromosome haplotypes. *Am J Hum Genet* 70: 1197–1214.
43. Mitchell P (2002) *The Archaeology of Southern Africa* (Cambridge Univ Press, Cambridge, UK).
44. Tarazona-Santos E, Tishkoff SA (2005) Divergent patterns of linkage disequilibrium and haplotype structure across global populations at the interleukin-13 (IL13) locus. *Genes Immun* 6:53–65.
45. Tishkoff SA, et al. (1996) Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. *Science* 271:1380–1387.
46. Henshilwood CS, d'Errico F, Marean CW, Milo RG, Yates R (2001) An early bone tool industry from the Middle Stone Age at Blombos Cave, South Africa: implications for

the origins of modern human behaviour, symbolism and language. *J Hum Evol* 41: 631–678.
47. Scholz CA, et al. (2007) East African megadroughts between 135 and 75 thousand years ago and bearing on early-modern human origins. *Proc Natl Acad Sci USA* 104: 16416–16421.
48. Texier PJ, et al. (2010) From the Cover: A Howiesons Poort tradition of engraving ostrich eggshell containers dated to 60,000 years ago at Diepkloof Rock Shelter, South Africa. *Proc Natl Acad Sci USA* 107:6180–6185.
49. Colombo S, et al.; Swiss HIV Cohort Study (2008) The HCP5 single-nucleotide polymorphism: A simple screening tool for prediction of hypersensitivity reaction to abacavir. *J Infect Dis* 198:864–867.
50. Knight A, et al. (2003) African Y chromosome and mtDNA divergence provides insight into the history of click languages. *Curr Biol* 13:464–473.
51. Li H, et al. (2007) Geographically separate increases in the frequency of the derived ADH1B*47His allele in eastern and western Asia. *Am J Hum Genet* 81:842–846.
52. Donnelly MP, et al. (2010) The distribution and most recent common ancestor of the 17q21 inversion in humans. *Am J Hum Genet* 86:161–171.
53. Greenberg JH (1963) *The Languages of Africa* (Indiana University, Bloomington, IN).
54. Browning SR, Browning BL (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet* 81:1084–1097.
55. Purcell S, et al. (2007) PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81:559–575.
56. Nalls MA, et al. (2009) Measures of autozygosity in decline: Globalization, urbanization, and its implications for medical genetics. *PLoS Genet* 5:e1000415.
57. Auton A, et al. (2009) Global distribution of genomic diversity underscores rich complex history of continental human populations. *Genome Res* 19:795–803.

HUNTER-GATHERER GENOMIC DIVERSITY SUGGESTS A

SOUTHERN AFRICAN ORIGIN FOR MODERN HUMANS

Authors: Brenna M. Henn[1*], Christopher R. Gignoux[2], Matthew Jobin[3,4], Julie M. Granka[5], J. Michael Macpherson[6], Jeffrey M. Kidd[1], Laura Rodríguez-Botigué[7], Sohini Ramachandran[8], Lawrence Hon[6], Abra Brisbin[9], Alice A. Lin[10], Peter Underhill[10], David Comas[7], Kenneth K. Kidd[11], Paul J. Norman[12], Peter Parham[12], Carlos D. Bustamante[1], Joanna L. Mountain[6], Marcus W. Feldman[5]

Supporting Appendix

Supplementary Methods:

Local Ancestry Assignment for Admixed Individuals

We applied a principal component analysis-based (PCA) method of assigning local ancestry across chromosomes to a subset of our click-speaking individuals (1). Each "admixed" individual was projected onto a PCA plot for three ancestral populations, representing the three populations from which the individual showed recent ancestry. The ancestral populations included: Bantu Luhya, Sandawe, Hadza, South African ≠Khomani Bushmen and Italian Tuscans. Each ancestral population of interest contained the fifteen individuals (n=10 for Hadza) with the least admixture from neighboring populations; putative recent admixture was estimated from the cluster-based *ADMIXTURE* analysis (Fig. 1). Sets of three ancestral populations of interest were used in the creation of the PCA plots. SNPs were thinned to exclude pairs of SNPs with r^2>0.8. Phased chromosomes were then strung together to create a single, extended genome for each individual for the PCA. By scanning along the chromosome of an "admixed" individual in 40-SNP non-overlapping windows, we scored the closest ancestral population for each window in PC space.

REJECTOR

We estimated population size, bottleneck severity and timing of bottleneck by using the rejection-based approximate Bayesian inference software REJECTOR (www.stanford.edu/~mjobin/rejector). The method employed by this software invokes a coalescent simulation (in this case MaCS, http://www-hsc.usc.edu/~garykche/) over numerous iterations, and with each iteration uses parameter values drawn from prior distributions supplied by the user. Summary statistics are calculated for each iteration and compared to the statistics calculated on the observed data, with the parameter values retained in a posterior distribution if that iteration's summary statistics fall within a tolerance window of the observed statistics. The summary statistics used for the parameters are first tested for both sensitivity to changes in parameter values and accuracy through recovery tests. We used fROH, the proportion of the genome within runs of homozygosity to the total sampled genome, as a summary statistic in this study. We used a sliding window of at least 500Kb and 50 SNPs across the genome to identify runs of homozygosity (ROHs), allowing a maximum of 5 missing and 1 heterozygous SNPs per window, while disallowing the extension of any window over a gap longer than 500Kb. We included a SNP in an ROH if it forms part of a run of homozygous SNPs at

least 25 SNPs and 100Kb in length. These parameters were similar to those chosen by Nalls et al. (2) and Auton et al. (3) for their calculations of the cumulative amount of runs of homozygosity. A sensitivity test of the fROH statistic shows accurate recovery of population size and bottleneck severity (Fig. S12).

HLA

HLA-A,-B and *-C* are highly polymorphic loci, whose genotyping enabled an independent assessment of diversity that is less biased by non-African population data than the genome-wide array design. The extensive body of published work on HLA-A, B and C (4-9) also allowed comparison with sub-Saharan African populations not included in the whole genome study, but having greatest HLA diversity worldwide (9).

HLA-A,-B,-C loci were genotyped using bead-based sequence specific oligonucleotide probe hybridization that was detected using a Luminex-100 instrument (Luminex Corp., Austin, TX). The assays were performed using LABType® SSO reagents (One Lambda, Canoga Park, CA with lots #11 (HLA-A), #14 (HLA-B) and #9 (HLA-C)). Ten percent of individuals were selected at random and DNA sequenced for exons 2 and 3 of *HLA-A, -B*, *-C* as previously described (5). HLA alleles are annotated by a series of digits, with the first two digits identifying alleles that encode broad serologically-distinct allotypes, the first four digits identifying unique protein sequences and subsequent digits identifying synonymous or intronic differences. A full description of the nomenclature is available (10). The three *HLA class I* loci were genotyped to four-digit resolution. In cases where these four-digit alleles are not distinguishable by the SSO genotyping method, the lowest number in the series was taken (e.g A*68:01 represents the string of possible alleles A*68:01/68:22/68:25/68:35/68:43 and A*68:02 represents A*68:02/68:18N/68:28/68:34/68:40/68:44/68:48). Genetic diversity was calculated using Nei's unbiased estimator for HLA haplotypes at each locus (11).

Selection

For 5 hunter-gatherer populations, we calculated iHS scores across the genome (461K SNPs common to all populations after filtering for quality) following a similar method in Pickrell et al. (33) for choice of window and binning scheme. Empirical p-values were calculated by binning windows by the number of SNPs within a window in increments of 20, up to 100 SNPs. Windows with greater than 100 SNPs were combined into a single bin due to the small number of windows with >100 SNPs. iHS scores for all loci in these 5 populations are freely available on request.

Supplementary Results and Discussion:

Haplotype Heterozygosity

We calculated haplotype heterozygosity across the genome using samples from HapMap3, HGDP-CEPH African populations and our click-speaking populations of Tanzania and South Africa (Table S1) with non-overlapping windows of different sizes. Window size substantially changes the estimate of haplotype heterozygosity (12) and window sizes of less than 20Kb are likely affected by severe ascertainment bias in array SNPs, particularly for these African populations. Bantu-speaking populations, which have recently absorbed migrants during the Bantu expansion, have the highest haplotype

heterozygosity at 100Kb windows; the hunter-gatherer Sandawe and ≠Khomani San groups have elevated haplotype heterozygosity, possibly in part due to recent gene flow from agriculturalist into these populations (Fig. 1). The Hadza have the lowest heterozygosity within Africa. Because there was very little difference between most African populations in heterozygosity, regressions of haplotype heterozygosity by distance as in Ramachandran et al. (13) were not significant within Africa. However, given the low LD measured in these African populations, we would argue that 100Kb windows may be too large and all populations would likely have a breakdown in LD. Therefore this type of analysis may not be as informative in African populations as it is for populations outside of Africa.

PCA and Population Structure

Principal component analysis provides an alternative method to discriminate genetically-based population clusters and to examine the distance between different clusters. We first performed PCA on a random sample of 15 individuals from each African population (Fig. S2a,b); PCA of all individuals is available in Supplemental Material. The first principal component (PC1) differentiates southern African hunter-gatherers (South African and Namibian KhoeSan) from eastern Africans (Fig. S2a). The Pgymy populations fall closer to the southern KhoeSan, but the HG Sandawe population lies near the Maasai. PC2 isolates only the Tanzanian Hadza individuals, possibly due to a recent population bottleneck, and PC3 differentiates the central African Pygmy populations from other Africans (Fig. S2b). Our new sample of South African KhoeSan individuals is generally dispersed around the Ju|hoansi Namibian San (14, 15) (Fig. S2). Even after excluding individuals with greater than 5% admixture from European and/or Bantu-speakers (estimated from *admixture* k=8, Fig. 1), the South African KhoeSan have greater spread along the PC1 axis than most other African populations analyzed here (Fig. 2).

We included the European Tuscans (Fig. S2) in order to compare *ADMIXTURE* with PCA; when European individuals are included, the largest distance along PC1 occurs between southern KhoeSan and European Tuscans. Eastern African populations, such as the Sandawe and Maasai, are the closest African populations to the Europeans, which is consistent with shared variation between these populations, apparent at k= 2 through 6 (Fig. 1).

The distribution of admixture varies between the hunter-gatherer populations according to the cluster analysis in Figure 1. At k=8 about a third of the Hadza and ≠Khomani Bushmen are inferred to share ancestry with Bantu, Maasai or Europeans, whereas the remainder of the individuals show little detectable recent admixture (Fig. 1). The variation in admixture between individuals within these populations indicates that the gene flow may have occurred only during the past few generations, and random mating within the population has not had time to equilibrate admixed allele frequencies. In contrast, the Sandawe and Biaka populations show less intra-population variation in the extent of Bantu admixture, suggesting gene flow into these populations occurred over longer periods of time.

Gene Flow

The difference between mtDNA and Y-chromosome haplogroup affiliations suggests that gene flow into hunter-gatherers is sex-biased towards higher male migration. Our results are more pronounced, but consistent with sex-biased migration estimates in South African "Coloured" populations (16). We also implemented a local ancestry assignment method for autosomal data in order to assess the genomic length of different putative ancestries in the Sandawe and ≠Khomani Bushmen (Fig. 4). Long segments of Bantu ancestry are consistent with some recent migration into the Sandawe population (also Fig. 1). Results for the ≠Khomani Bushmen are more difficult to interpret as there are both long and very short segments with European ancestry (Fig. 4). It is possible that the full extent of KhoeSan variation is not captured in our South African sample, leading to spurious European segments; alternatively, contributions from a fourth highly diverged population such as eastern Nilotic pastoralists could explain the short, interspersed "European" ancestral segments (17). Patterns for other putatively admixed individuals are shown in Fig. S4.

iHS Selection Scan

We compared the most extreme regions of the genome identified using the iHS statistic calculated with our data set for the HGDP Biaka samples with the regions previously reported for the same samples by Pickrell et al. (18). Surprisingly, we found that only 15 out of the top 100 regions reported by Pickrell et al. (18) were also among the most extreme 1% of regions we identified (Figure S8), with 42/100 Pickrell regions present in the most extreme 5% identified by this study (Figure S9). We considered several possibilities that could contribute to this difference. First, we used a slightly modified binning strategy in which we combined windows with >100 SNPs for the purpose of determining empirical p-values since few windows contained more than 100 SNPs in our dataset. However, reanalysis of the phased haplotypes used by Pickrell et al. (obtained from http://hgdp.uchicago.edu/Phased_data/) indicates that the modified binning procedure acts to only rerank the most extreme genomic regions: 72/100 regions are still identified as the most extreme 1% and 99/100 regions are found among the most extreme 5%. This suggests that the difference in results is not a result of the calculation of the iHS statistic but likely reflects differences in processing genotype data.

The samples considered in this study were typed on a different SNP array platform (Illumina 550K) from the HGDP samples analyzed by Pickrell et al.(18). For comparisons, we limited our analysis to the typed SNPs common to both arrays, after filtering for quality. Notably, the Pickrell et al. study, which used genotyping data from Li et al. (14), used Illumina 650K arrays that contain additional SNPs designed to better tag haplotypes in the HapMap YRI. Reanalyzing the previously phased Biaka haplotypes from Pickrell et al. (18) with this reduced set of common SNP positions reduced the overlap to 44/100 top 1% regions and 67/100 regions for the top 5% (Figure S10). The difference in set of analyzed SNPs is not enough to account for the poor overlap among extreme regions identified using iHS, indicating that differences in phasing may have a substantial effect on the regions identified as having extreme iHS values. It is expected that phase switch errors will act to reduce the apparent length of shared haplotypes, resulting in reduced iHS scores.

Pickrell et al. (18) phased all HGDP populations together using fastPHASE with known haplotypes from HapMap2 YRI and CEU trios. Our analysis of African groups

uses haplotypes phased with BEAGLE (19) and with a more diverse set of phased haplotype seeds obtained from both HapMap3 trios and 6 parent-offspring pairs from our Khoisan populations (Methods). To our knowledge, uncertainty in phasing when calculating haplotype-based genomic selection scans has been largely ignored (18, 20, 21). Our results suggest that statistics such as iHS, and likely, XP-EHH are sensitive to phase switch errors, which are expected to decrease the length of shared haplotypes. This would be especially relevant when haplotype scans are computed in diverse populations where large sets of well-phased trios are not readily available.

Sensitivity to Ascertainment Bias

Our data consist primarily of SNPs from Illumina and Affymetrix array platforms; these arrays were designed to genotype known SNPs, many which had been initially discovered in Eurasian populations. In order to minimize the effect of genotyping SNPs that do not reflect the full spectrum of genetic diversity in African populations, we focused on analyses that are less sensitive to ascertainment bias, such as: clustering algorithms, PCA, mean LD, ROH and haplotype statistics. Conrad et al. (12) found no systematic difference in population-based estimates of LD using different ascertainment schemes in an HGDP SNP dataset. In practice, we expect higher LD in a sample with poor ascertainment because of an increase in homozygous loci; furthermore, the ascertained loci are in general common, older variants if they are polymorphic in both Europeans and Africans. In principle, runs of homozygosity might be longer than expected with an inappropriately ascertained sample, as novel SNPs that are heterozygous in the population would not be identified and hence would not break up a run of homozygosity. However, our estimates are based on dozens, if not hundreds of markers per window, many of which have common variants across Africa. Empirically, populations from HGDP were surveyed for fROH and many populations, even those for which the Illumina platform represents a poorly ascertained SNP sample, had levels of ROH similar to European populations. For example, Sardinians, Basque and Tuscans have a fROH of 2-3% and Yorubans, Mbuti and Biaka Pymies also have a fROH of 1-3%.

Supplementary Tables

Table S1

Description of Dataset

Population	Country	Region	Sample Size	Platform	Reference
≠Khomani Bushman	South Africa	South	35	Illumina 550K	*Present study*
!Xhosa	South Africa[1]	Migrant[5]	13	Affymetrix 500K	Bryc 2010, Li 2008
Julhoansi Bushman	Namibia[2]	South	5	Illumina 650K	Li 2008
Juu San	Namibia[2]	South	12	Illumina 1M	Schuster 2010
Hadza	Tanzania	East	20	Illumina 550K	*Present study*
Sandawe	Tanzania	East	35	Illumina 550K	*Present study*
Maasai	Kenya[3]	East	133	HapMap3 rel2	HapMap3
Luhya	Kenya[3]	Migrant[5]	90	HapMap3 rel2	HapMap3
Mbuti	DRC	Central	13	Illumina 650K	Li 2008
Biaka	CAR	Central	22	Illumina 650K	Li 2008
Bulala	Chad	Central	15	Affymetrix 500K	Bryc 2010
Kaba	Chad	Central	17	Affymetrix 500K	Bryc 2010
Bamoun	Cameroon	West	18	Affymetrix 500K	Bryc 2010
Fang	Cameroon	West	15	Affymetrix 500K	Bryc 2010
Mada	Cameroon	West	12	Affymetrix 500K	Bryc 2010
Yoruba	Nigeria	West	21	Illumina 650K	Li 2008
Hausa	Nigeria	West	12	Affymetrix 500K	Bryc 2010
Igbo	Nigeria	West	15	Illumina 650K	Li 2008
Fulani	Nigeria	North	13	Affymetrix 500K	Bryc 2010
Mandenka	Senegal	West	22	Illumina 650K	Li 2008
Saharawi	West Sahara	North	18	Affymetrix 6.0	*Present study*[4]
South Moroccan	Morocco	North	16	Affymetrix 6.0	*Present study*[4]
North Moroccan	Morocco	North	18	Affymetrix 6.0	*Present study*[4]
Mozabite Berber	Algeria	North	29	Illumina 650K	Li 2008
Algerian	Algeria	North	19	Affymetrix 6.0	*Present study*[4]
Tunisian Berber	Tunisia	North	18	Affymetrix 6.0	*Present study*[4]
Libyan	Libya	North	17	Affymetrix 6.0	*Present study*[4]
Egyptian	Egypt	North	19	Affymetrix 6.0	*Present study*[4]

[1] !Xhosa samples (n=5) were combined with South African Bantu from HGDP-CEPH (n=8) for linkage disequilbrium analysis.

[2] The two Northern Juu-speaking Bushmen samples were combined to form a larger sample for PCA and LD analysis.

[3] We randomly chose 30 unrelated Maasai and Luhya for representation in population structure analysis in Figure 1.

[4] North African samples were utilized only in the linkage disequilbrium decay analysis. Data from the intersection of 55,000 SNPs common to all platforms and used for LD analysis will be made publicly available. The other North African genotype data are in preparation (Henn, Rodriguez-Botigue et al., *in prep*.).

[5] The Bantu-speaking Luhya from Kenya and !Xhosa from South Africa were considered recent geographic migrants and were not included in the primary LD analysis shown in Figure 2. Removal of these populations, however, does not change the LD regression significantly.

Table S2

Haplotype Heterozygosity Estimated from Genome-wide SNPs

Population	Sample Size	100Kb[1]	20Kb[1]	Sample Set
South African Bantu	8	0.967	0.862	HGDP
Kenyan Bantu[2]	12	0.964	0.858	HGDP
South African Khomani San	30	0.963	0.851	present
Sandawe	27	0.962	0.863	present
Yoruba[2]	21	0.961	0.852	HGDP
Maasai	30	0.961	0.861	HapMap3
Biaka Pygmy	22	0.960	0.847	HGDP
Mandenka	22	0.960	0.848	HGDP
Namibian Julhoansi San	5	0.947	0.808	HGDP
Mbuti Pygmy	13	0.947	0.815	HGDP
Mozabite	29	0.927	0.817	HGDP
Hadza	17	0.920	0.807	present
Tuscan (Italy)	30	0.912	0.785	HapMap3

[1] Haplotype heterozygosity was estimated in non-overlapping windows of 100 Kb and 20 Kb (citation for het equation). Each window was constrained to contain a minimum of 20 SNPs and 5 SNPs for 100 Kb and 20 Kb windows, respectively.

[2] HapMap3 samples from similar populations, Kenyan Luhya and Nigerian Yoruba, were not included in the haplotype heterozygosity calculation.

Table S3

mtDNA and Y-chromosome haplogroup frequencies and distributions

Population	mtDNA[a]	Y-Chromosome	Primary (Endemic) Distribution[b]
Hadza	L0a2*: 6% (1)	B2b: 10% (1)[b]	Eastern African
	L3h: 11% (2)	B2b4*: 50% (5)	
	L4g: 56% (10)	E1b1b1: 10% (1)	
	L2a: 22% (4)	E1b1a7a3a: 30% (3)	Western African
	L3b: 6% (1)		
Sandawe	L0a2*: 20% (6)	A3b2*: 12% (2)	Eastern African
	L3x1: 13% (4)	B2b4*: 29% (5)[b]	
	L4g: 37% (11)	E1b1b1: 18% (3)	
	L2a: 10% (3)	E2b1: 6% (1)	Western African
		E1b1a7a3a: 24% (4)	
		E1b1a8a: 12% (2)	
	L3e3: 17% (5)		Eastern/Western
≠Khomani San	L0d1a: 43% (14)	A3b: 26% (5)	Southern African
	L0d1b: 50% (16)	A3b1: 32% (6)	
		B2b4*: 5% (1)[b]	
		E2b1: 5% (1)	Western African
		E1b1a7a3a: 10% (2)	
		E1b1a8a: 10% (2)	
		R1b1b2a1a: 10% (2)	Eurasian
	L0a'b'f*: 7% (2)		Unknown

[a] Frequency of mitochondrial (mtDNA) and Y-chromosome haplogroups are shown in each column; individual counts are indicated in parentheses. Haplogroup assignments were based on approximately 2,000 mtDNA and 2,000 Y-chromsome SNPs using a customized algorithm designed by 23andMe, Inc.

[b] The primary geographic distribution of each haplogroup is denoted by region within Africa. These primary geographic regions tend to be associated with the origin and an ancient endemic presence of the haplogroup. Y-chromosome haplogroup B2b4-M112 is distributed throughout central Africa, eastern Africa and southern Africa almost exclusively in hunting-gatherering populations (22). Its precise origin is not known. For this reason, B2b4 is considered endemic to the hunter-gatherer populations listed here.

Table S4:

Heterozygosity estimates from HLA[4]

Country	Region or Ethnic Group	N	HLA-A		HLA-B		HLA-C	
			K[1]	H[2] (s.e.)	K[1]	H[2] (s.e.)	K[1]	H[2] (s.e.)
Burkina Faso	Fulani	49	18	0.92 (0.009)	17	0.93 (0.007)	-	-
Burkina Faso	Mossi	51	15	0.90 (0.012)	18	0.90 (0.013)	13	0.87 (0.014)
Burkina Faso	Rimaibe	49	16	0.89 (0.013)	16	0.88 (0.015)	14	0.87 (0.017)
Cameroon	Baka*[3]	10	11	0.92 (0.041)	10	0.88 (0.005)	7	0.86 (0.042)
Cameroon	Bamileke	77	22	0.93 (0.007)	31	0.94 (0.007)	16	0.91 (0.009)
Cameroon	Beti	174	27	0.93 (0.004)	37	0.94 (0.000)	21	0.91 (0.006)
Cameroon	Cameroon	91	29	0.92 (0.009)	32	0.95 (0.005)	-	-
Cameroon	Sawa	13	14	0.93 (0.033)	13	0.90 (0.003)	9	0.89 (0.028)
Cape Verde	Northwestern	56	24	0.93 (0.009)	34	0.94 (0.010)	-	-
Cape Verde	Southeastern	55	27	0.94 (0.008)	35	0.95 (0.007)	-	-
Ghana	Ga-Adangbe	55	21	0.91 (0.014)	27	0.93 (0.012)	18	0.89 (0.015)
Guinea Bissau	Guinea Bissau	65	20	0.92 (0.010)	31	0.94 (0.009)	-	-
Kenya	Kenya	143	39	0.94 (0.004)	46	0.95 (0.003)	25	0.90 (0.008)
Kenya	Luo	265	30	0.94 (0.003)	47	0.94 (0.003)	22	0.90 (0.005)
Kenya	Nandi	241	28	0.93 (0.004)	39	0.95 (0.003)	21	0.89 (0.006)
Mali	Doggon	138	21	0.90 (0.009)	31	0.92 (0.007)	19	0.84 (0.013)
Senegal	Mandeka	93	25	0.92 (0.008)	36	0.95 (0.005)	19	0.91 (0.005)
South Africa	**≠Khomani San***	**55**	**30**	**0.95 (0.009)**	**31**	**0.94 (0.008)**	**20**	**0.89 (0.018)**
South Africa	Natal (Tamil)	55	16	0.89 (0.012)	23	0.94 (0.009)	21	0.92 (0.012)
South Africa	Zulu	186	28	0.94 (0.004)	33	0.93 (0.004)	18	0.89 (0.009)
Sudan	Central and South	209	30	0.93 (0.005)	55	0.97 (0.002)	24	0.91 (0.006)
Tanzania	**Hadza***	**44**	**18**	**0.88 (0.001)**	**15**	**0.80 (0.001)**	**14**	**0.86 (0.001)**
Uganda	Kampala 1	161	34	0.94 (0.007)	50	0.97 (0.002)	24	0.92 (0.006)
Uganda	Kampala 2	178	33	0.94 (0.004)	40	0.94 (0.004)	22	0.90 (0.007)
Zambia	Lusaka	43	20	0.90 (0.018)	30	0.95 (0.010)	12	0.90 (0.009)
Zimbabwe	Shona	225	30	0.92 (0.004)	39	0.94 (0.003)	21	0.91 (0.004)

[1] Number of distinct haplotypes.

[2] Heterozygosity, corrected for sample size.

[3] "*" Indicates hunter-gatherer populations.

[4] Data obtained from: (4-9, 23-25)

Table S5

*Hadza HLA-B*44:03-containing haplotypes*

Haplotype	HLA-A	HLA-B	HLA-C	Frequency
1	A*01:02	B*44:03	C*17:01	0.159
2	A*24:02	B*44:03	C*04:01	0.091
3	A*74:01	B*44:03	C*17:01	0.045
4	A*24:88	B*44:03	C*04:01	0.034
5	A*02:14	B*44:03	C*17:01	0.023
6	A*03:03	B*44:03	C*04:01	0.011
7	A*74:01	B*44:03	C*12:02	0.011

Table S6

Δ Log Likelihood for each Region versus South

Region	lat	long	LD 0-5kb	LD 5-10kb	LD 10-15kb
South	-14	12	0	0	0
Cameroon	7	12	-12.33	-11.49	-9.90
Ethiopia	9	39	-13.17	-12.39	-10.91
Tanzania	-7	39	-8.96	-8.44	-7.47

Southern origin is X-fold more likely than an origin in…

Region	lat	long	LD 0-5kb	LD 5-10kb	LD 10-15kb
South	-14	12	1	1	1
Cameroon	7	12	227,000	97,500	19,900
Ethiopia	9	39	525,000	240,000	54,500
Tanzania	-7	39	7,750	4,620	1,750

Likelihood comparisons of regression models demonstrate support for our best-fit model of an origin in southern Africa. We extracted the likelihood values from each linkage disequilbrium (LD) regression from a grid of possible origin (latitude/longitude) points in Africa. Here, we compare our best-fit regression in southern Africa to three other locations, in eastern and central Africa. As the regression-fitting surface across Africa is relatively smooth (see Fig. 2b) these point estimates are highly predictive of nearby estimates from the same region. For Table S6, regressions are based on mean population LD estimated from SNPs within 0-5kb, 5-10kb and 10-15kb.

We also evaluated the fit of points in western and northwestern Africa (Senegal and Morocco, respectively). However, in these regions, there was a negative correlation of LD with geographic distance, violating the assumption that LD correlates positively with distance from the origin.

Supplementary Figures

Figure S1:

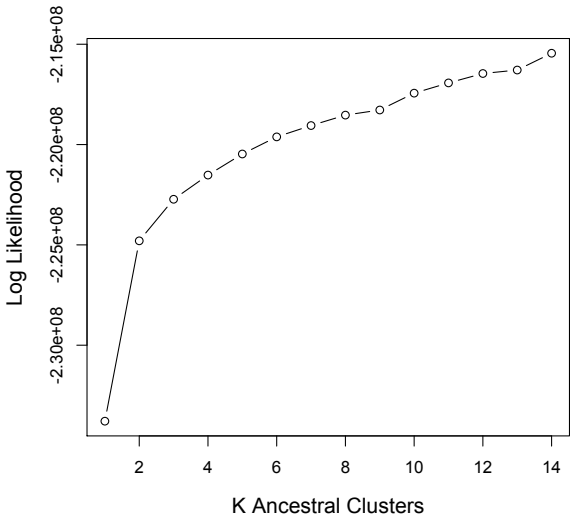
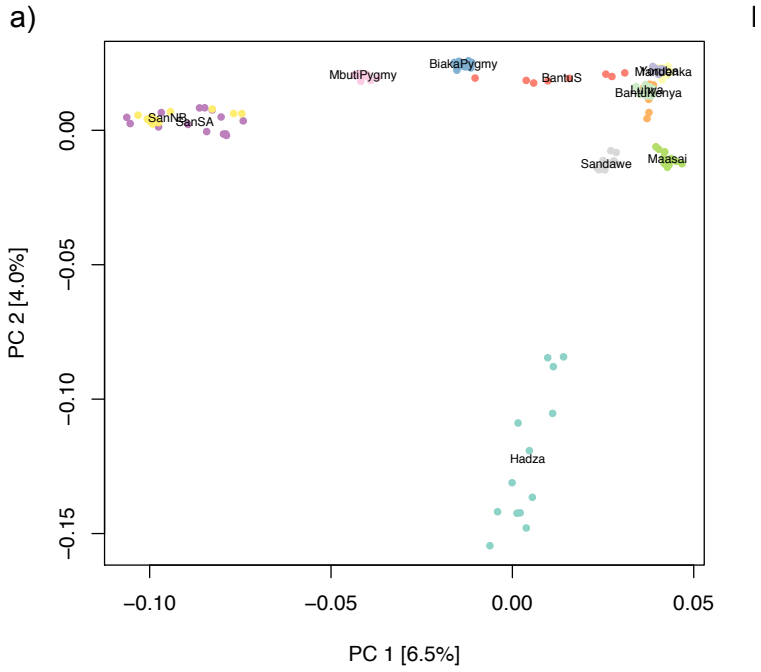
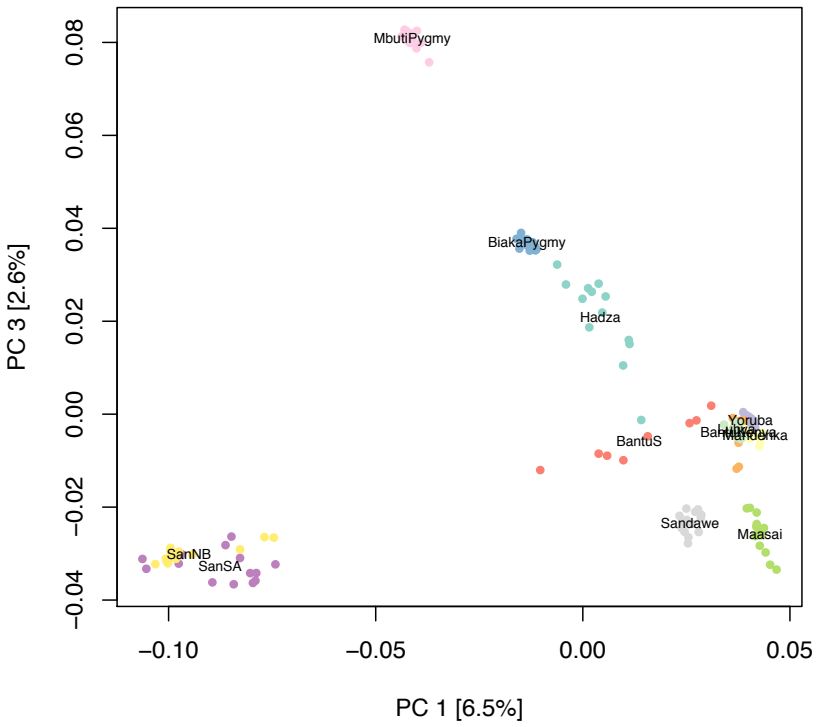


Figure S1: *Log Likelihoods for admixture k=1:14*. Log likelihoods (Y-axis) are plotted for the number of ancestral clusters k=1 through 14 (X-axis) generated with the *ADMIXTURE* program (26). Clusters k=2,4,6,8 are presented in Figure 1, main text.

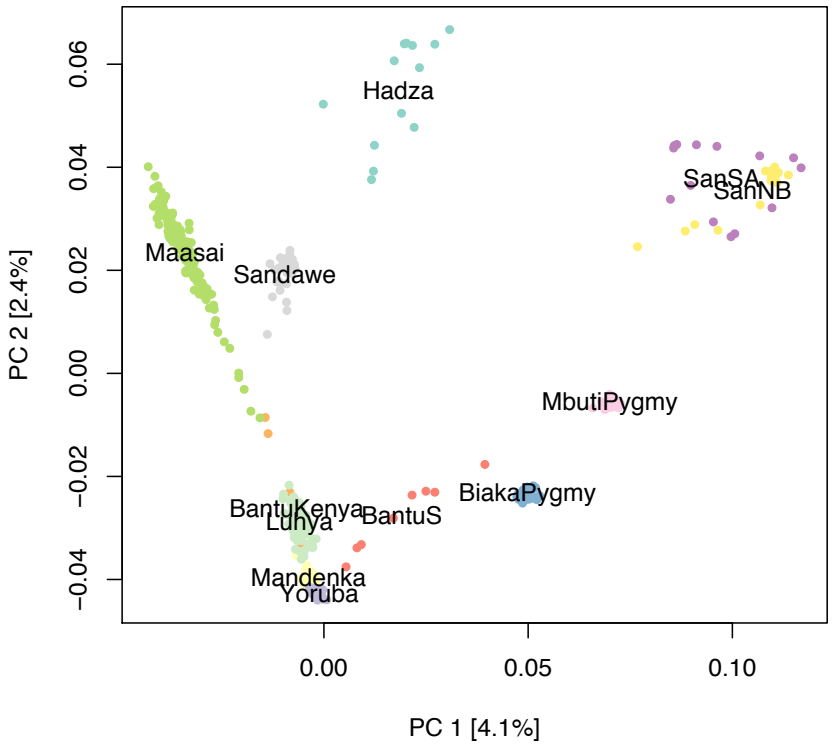
Figure S2:



b)


MbutiPygmy
BiakaPygmy
Hadza
BantuS
Yoruba
Mandenka
Sandawe
Maasai
SanNB
SanSA
0.08
0.06
0.04
0.02
0.00
−0.02
−0.04
PC 3 [2.6%]
−0.10
−0.05
0.00
0.05
PC 1 [6.5%]

c)


Hadza
SanSA
SanNB
Maasai
Sandawe
MbutiPygmy
BiakaPygmy
BantuKenya
Luhya
BantuS
Mandenka
Yoruba
0.06
0.04
0.02
0.00
−0.02
−0.04
PC 2 [2.4%]
0.00
0.05
0.10
PC 1 [4.1%]

d)

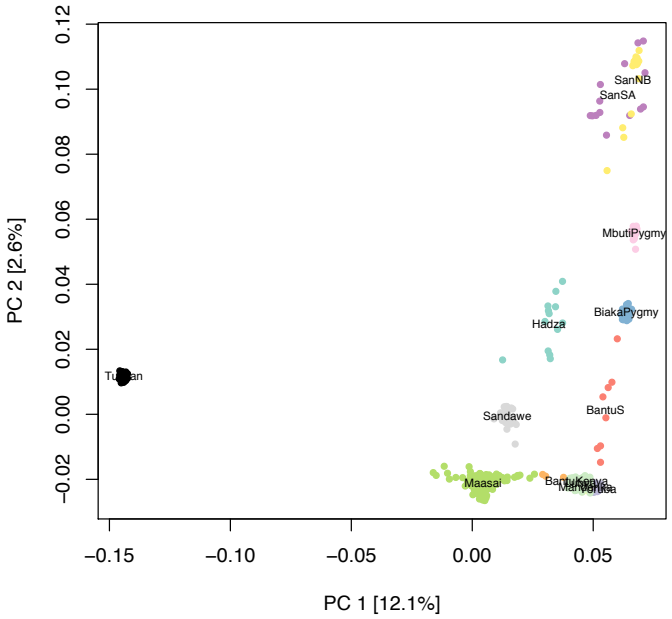


Figure S2: *Population structure of current African populations inferred from PCA*. We used principal component analysis to discriminate clusters of genetic variation within Africa. a) Population samples are displayed along the PC1 and PC2 axes of variation. Fifteen individuals from each population were randomly sampled for this analysis. "San_SA" in purple represents the ≠Khomani Bushmen from South Africa; "San_NB" in yellow represents Ju-speaking Bushmen from Namibia, a combined sample of individuals from HGDP and Schuster et al. (15) b) PC1 versus PC3 axes of variation. PC3 pulls out central African forest Pygmies and the Hadza hunter-gatherers. c) For comparison, we include an additional 15 European Tuscan individuals (HapMap3). c) PC1 by PC2 for All individuals from 12 African populations, except for Hadza and Khomani San [San_SA] with >5% Bantu or European admixture. d) For comparison, we include an additional 30 European Tuscan individuals (HapMap3).

Figure S3

a) b)

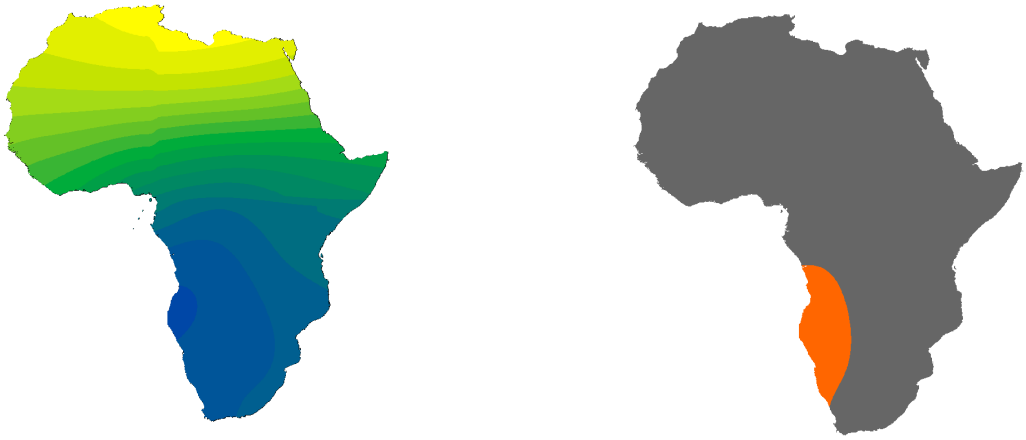
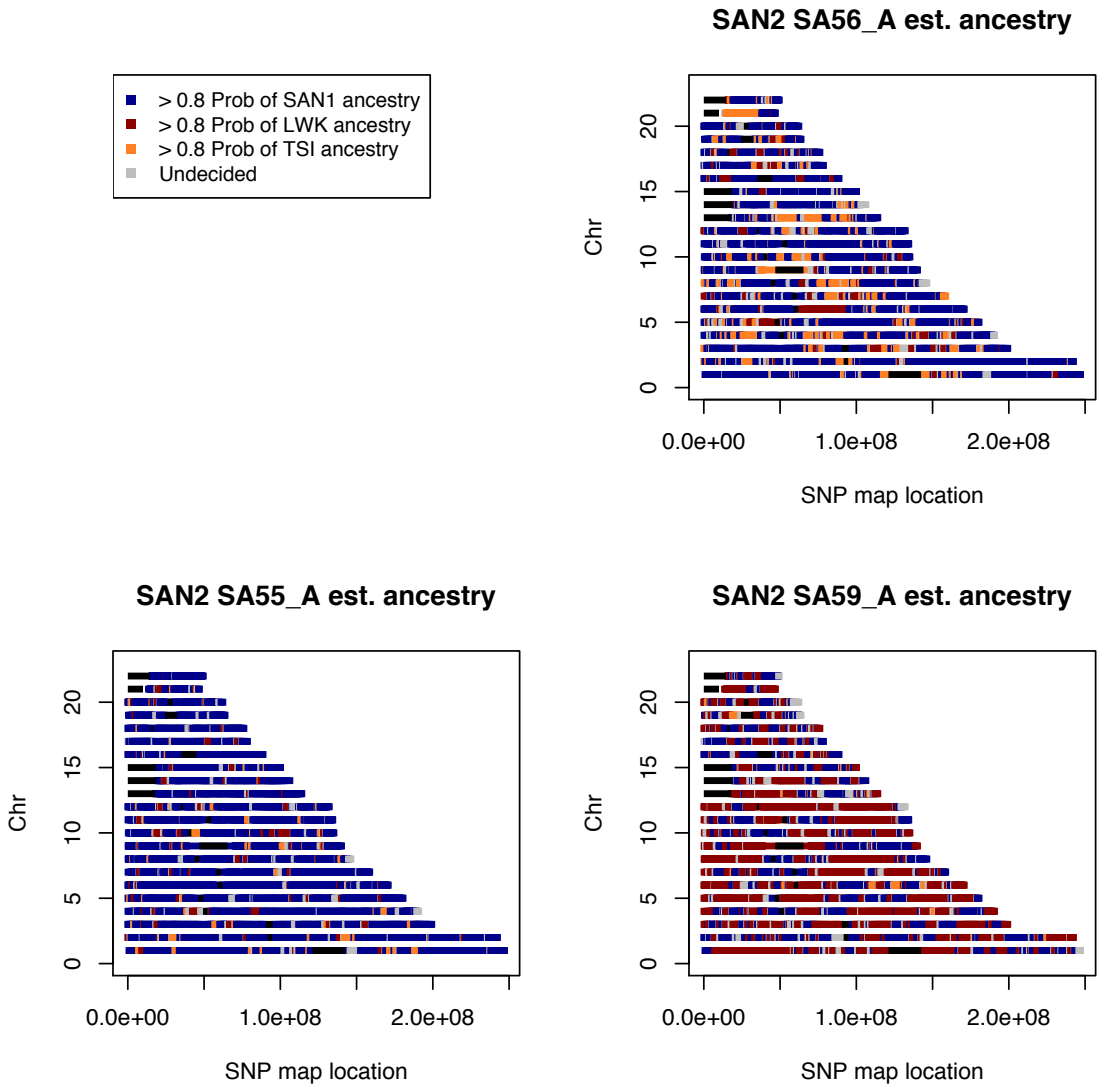


Figure S3: *Genome-wide linkage disequilbrium in 26 African populations (including admixed Bushmen)*. Regressions of LD on geographic distance were calculated assuming a grid of origin locations. All southern African Bushmen were included in these LD samples, regardless of their degree of admixture. a) Map is shown using mean LD at the 0-5Kb bins. Similar results were obtained with mean LD within 10Kb, 20Kb and the area under the curve between 5-50Kb. The highest correlation coefficient indicates the best fit with a potential geographic origin. We used MapViewer (http://www.goldensoftware.com) to create the Kriging interpolation plot of the correlation coefficients. b) We assessed a confidence interval around our best point estimate, here shown in orange; plotted are all the points in the grid with an adjusted $p <$ 0.05 after 1000 permutations.

Figure S4:


> 0.8 Prob of SAN1 ancestry
> 0.8 Prob of LWK ancestry
> 0.8 Prob of TSI ancestry
Undecided
SAN2 SA56_A est. ancestry
SAN2 SA55_A est. ancestry
SAN2 SA59_A est. ancestry
Chr
SNP map location
0.0e+00
1.0e+08
2.0e+08
0
5
10
15
20

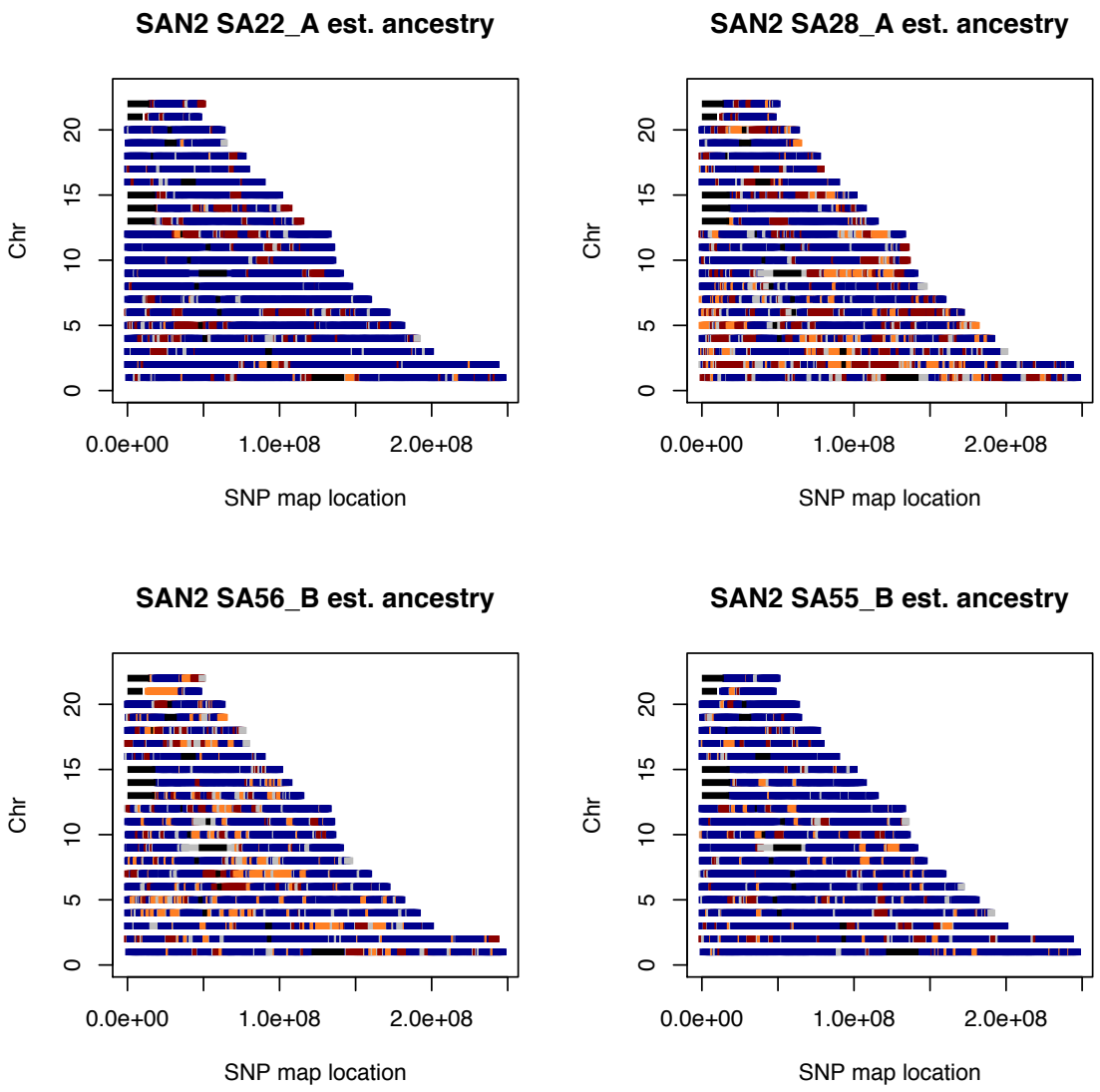


Figure S4 Legend: *Local ancestry assignment along phased chromosomes*. Individuals with potential admixture (Fig. 1) were projected into the principal component space of three putative ancestral populations. South African ≠Khomani San: SAN, European Tuscan: TSI, Luhya Bantu: LWK. Ancestry was assigned in 40-SNP windows along phased chromosomes (haplotypes A and B) by calculating the minimal distance to an ancestral (1, 27). Chromosomes A and B are separated for ease of visualization and do not correspond exactly to the separate maternal and paternal haplotypes because we do not have trios for most individuals, (i.e. chromosome 1A could be paternal and chromosome 2A could be maternal). Along the chromosomes, blue represents KhoeSan ancestry, orange represents European (Tuscan) ancestry, and red represents Bantu (Luhya) ancestry.

Figure S5:

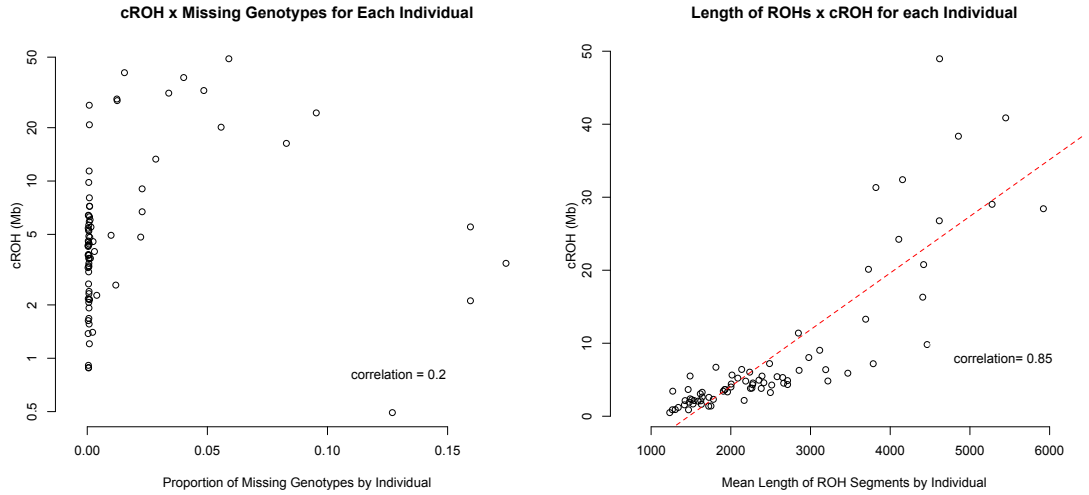
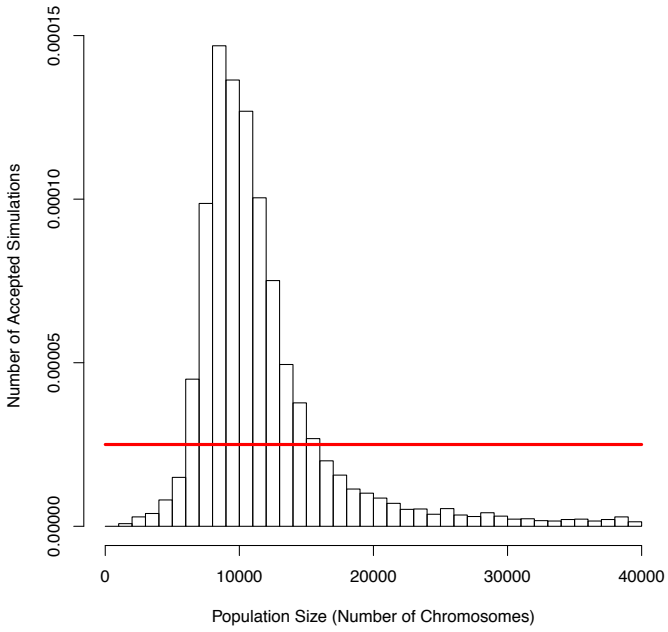


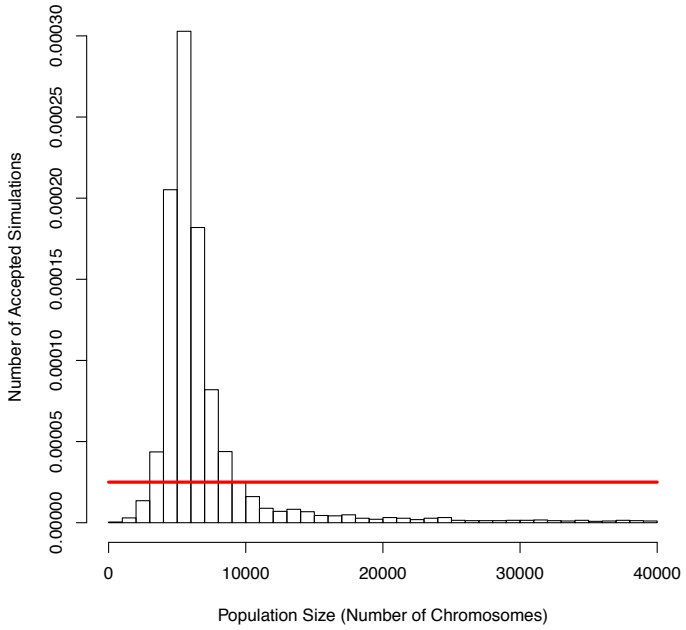
Figure S5 Legend: Cumulative runs of homozygosity (cROH) by missingness and segment length.

Figure S6:

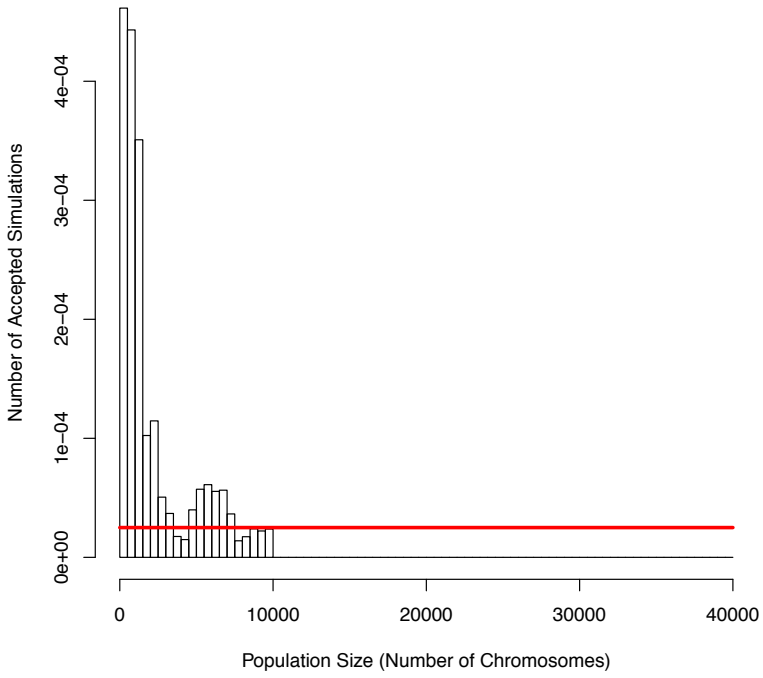
a)



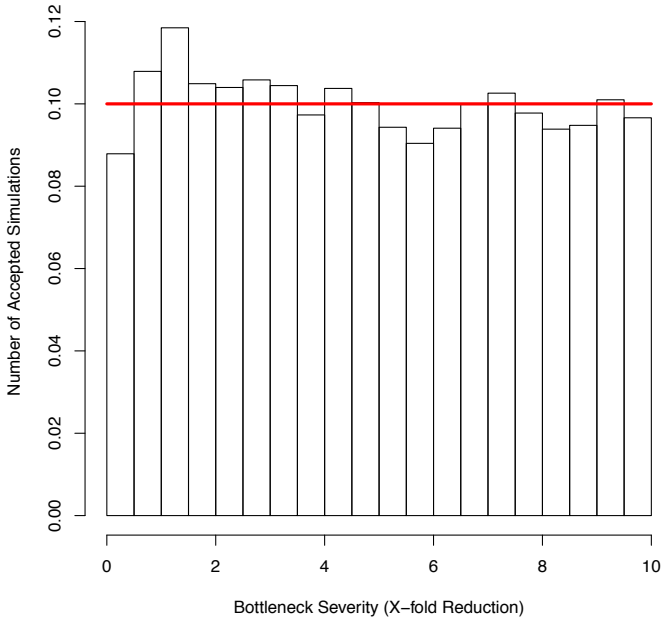
b)


Number of Accepted Simulations
0.00000
0.00005
0.00010
0.00015
0.00020
0.00025
0.00030
0
10000
20000
30000
40000
Population Size (Number of Chromosomes)

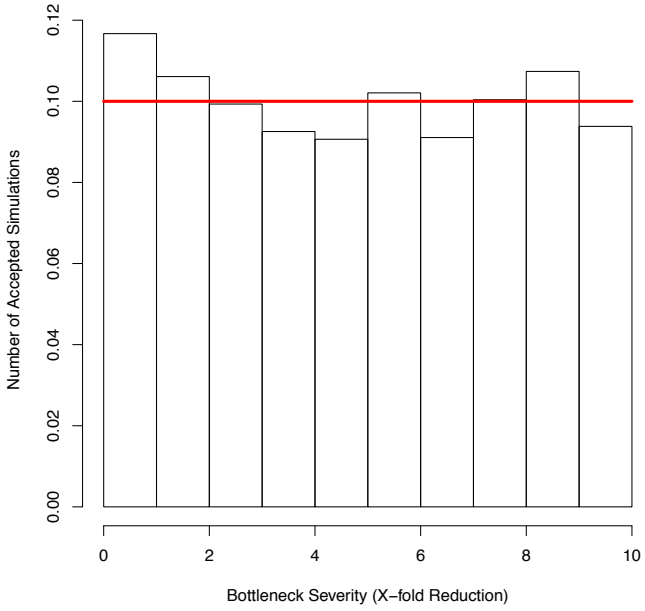
c)


Number of Accepted Simulations
0e+00
1e−04
2e−04
3e−04
4e−04
0
10000
20000
30000
40000
Population Size (Number of Chromosomes)

d)


Number of Accepted Simulations
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0
2
4
6
8
10
Bottleneck Severity (X−fold Reduction)

e)


Number of Accepted Simulations
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0
2
4
6
8
10
Bottleneck Severity (X−fold Reduction)

f)

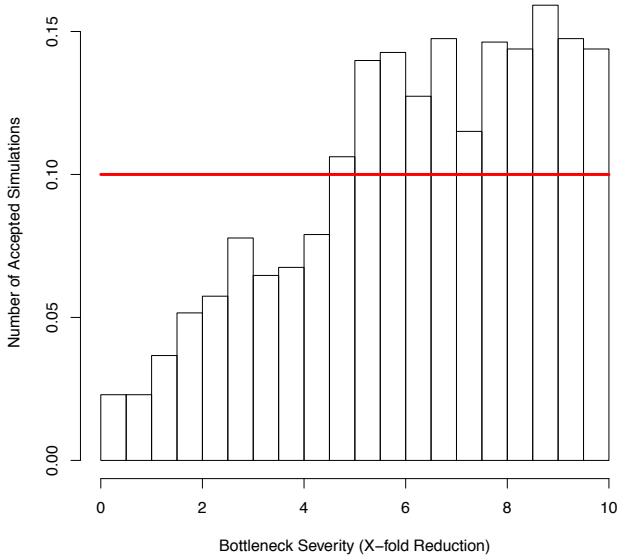


Figure S6: Simulated posterior distribution of effective population size in the a) Khomani San, b) Sandawe, c) Mbuti generated by sampling from a uniform distribution of N_e and keeping simulated parameters within 20% of the observed fROH (fraction of the genome in ROH) with REJECTOR (28). Simulated posterior distributions of bottleneck severity in the d) Khomani San, e) Sandawe, f) Mbuti as modeled above. Red lines reflect the expected number of accepted runs under a uniform distribution.

Figure S7:

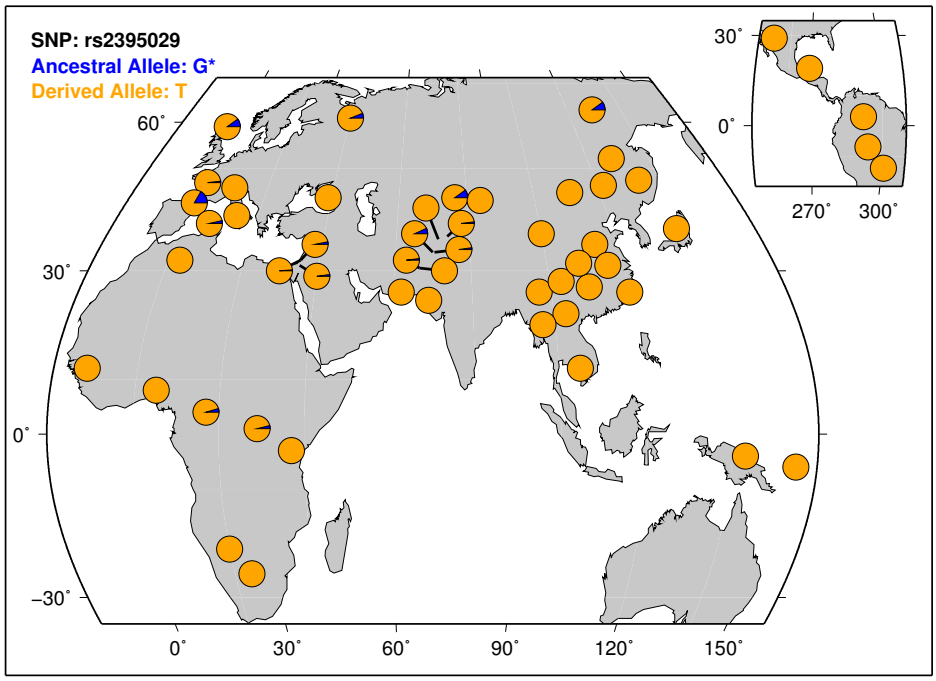


Figure S7 Legend: *Allele frequencies for rs2395029*. Allele frequencies of rs2395029, a missense Val to Gly mutation in the HPC5 gene, for 52 of the Human Genome Diversity Project populations. The Sandawe population carries the G allele at 19% frequency, higher than almost all other world populations. Other populations with elevated G allele frequencies tend to have experienced a recent bottleneck (e.g. Basque, Yakut, Orcadinian, Uygur, Hazara.) Figure was generated using the HGDP Selection Browser (http://hgdp.chicago.edu).

Figure S8

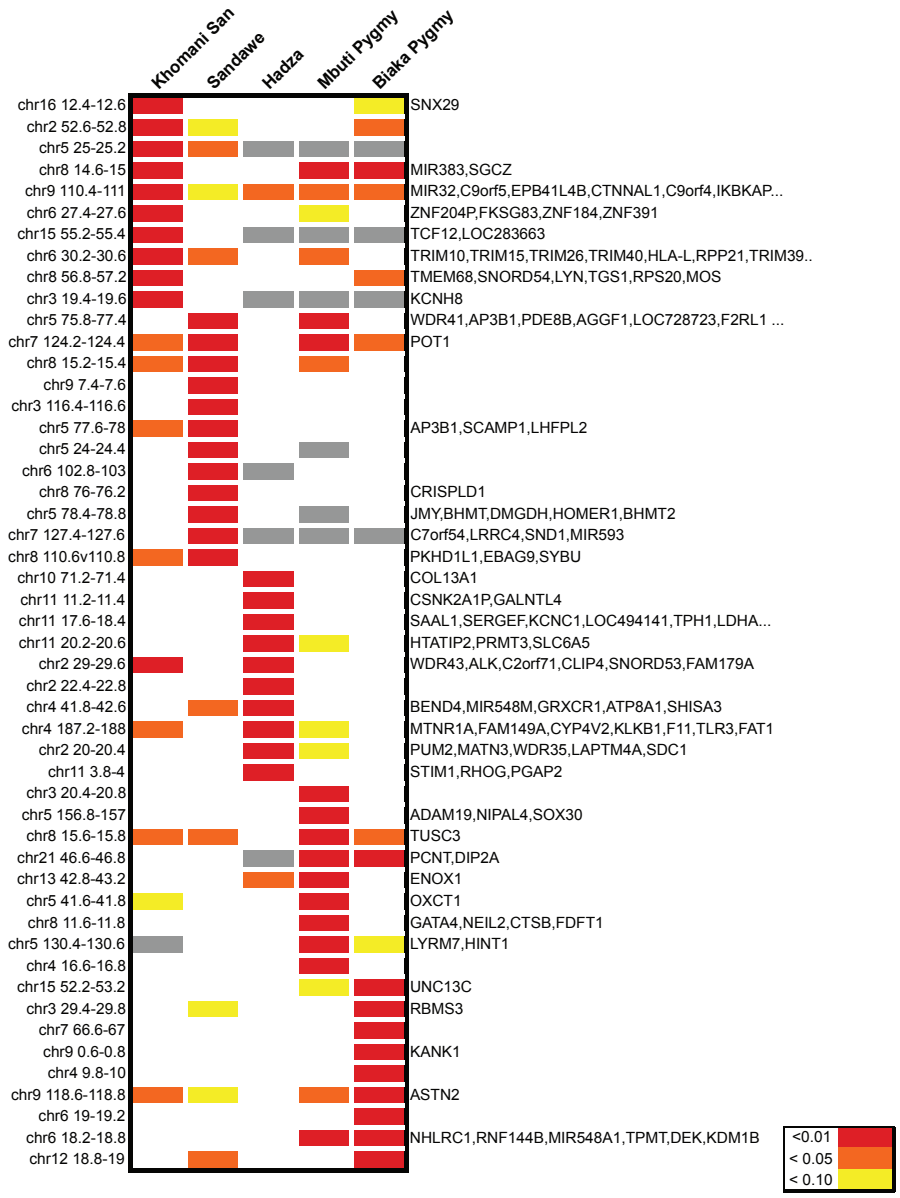


Figure S8 Legend: *Haplotype-based scan (iHS) for selection in 5 hunter-gatherer populations*.
A haplotype-based statistic (*iHS*) was used to identify ongoing selective sweeps in 5 hunter-gatherer populations using a common set of 461K SNPs. Empirical p-values for each population were calculated following Pickrell et al. (18) (*Methods*). For each population, the top ten most significant windows are highlighted in red; if the window was significant at the 0.05 or 0.10 level in another population, those loci are also marked with orange or yellow boxes respectively. Genes in the window are listed on the right side of the figure; adjacent highly significant windows were collapsed and genes for the entire set are also listed.

Figure S9

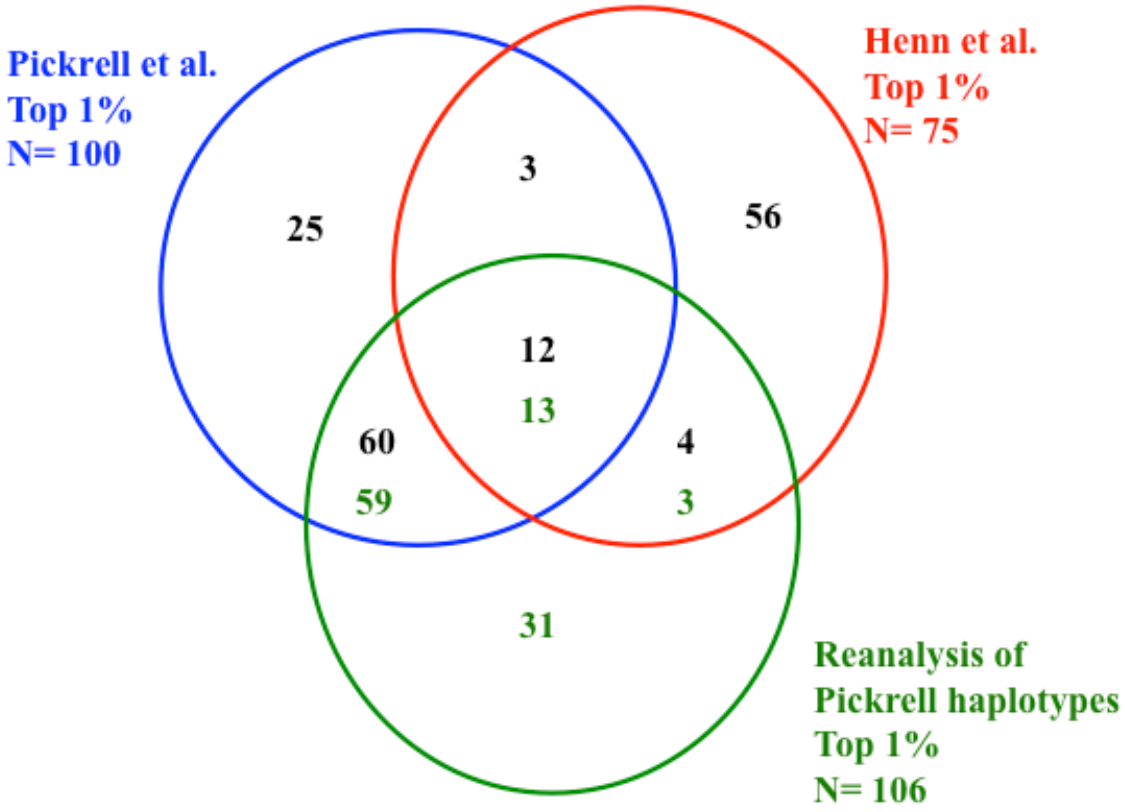


Figure S9 Legend. Comparison of extreme iHS regions identified for the HGDP Biaka. The overlap among the most extreme regions identified by Pickrell et al. (18) (in blue), in this study (Henn et al., in red), and found by applying the binning strategy employed in this study to the previously phased Biaka haplotypes from Pickrell et al. (18) (in green) is shown.

Figure S10

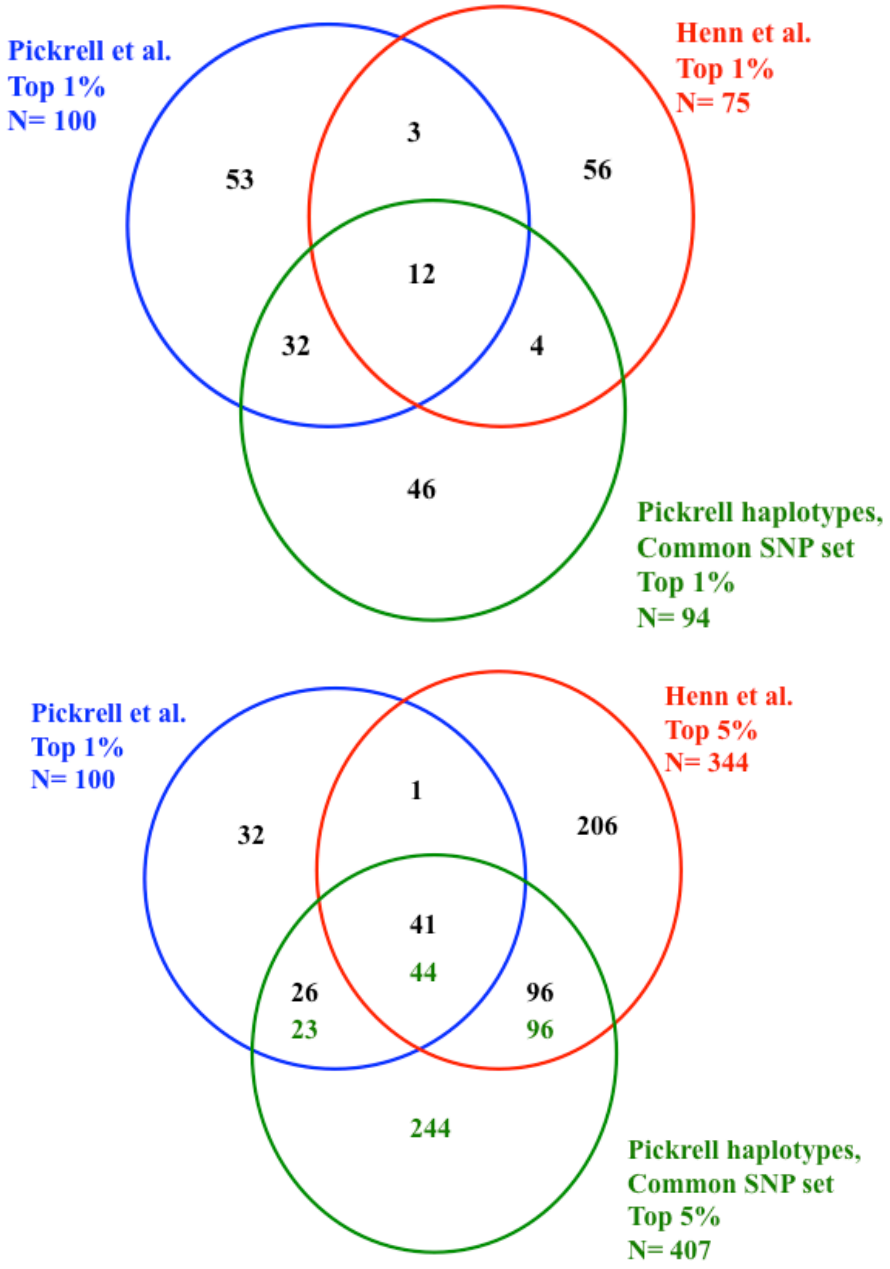


Figure S10. Impact of reduced SNP set on identified regions of extreme iHS. Analysis was performed on the set of common SNPs used for the populations analyzed in this study but using the phased haplotypes obtained from Pickrell et al. (18)

Figure S11:

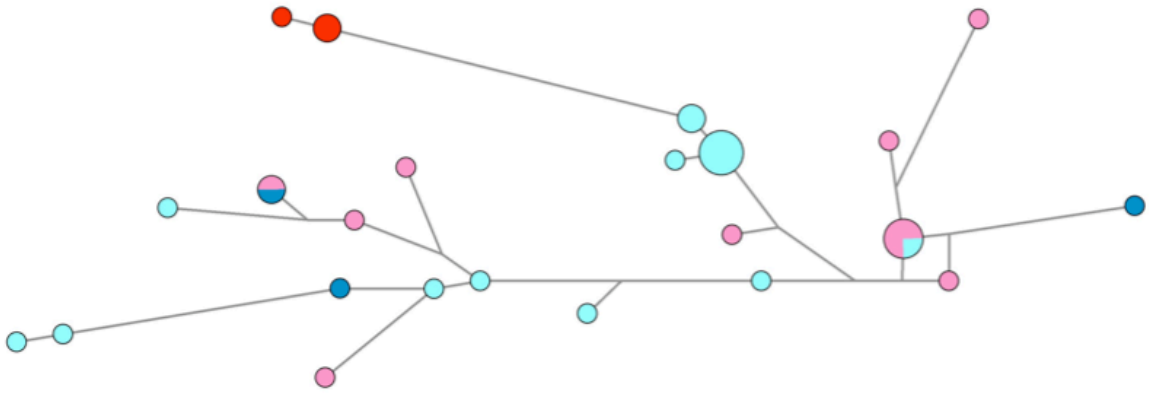
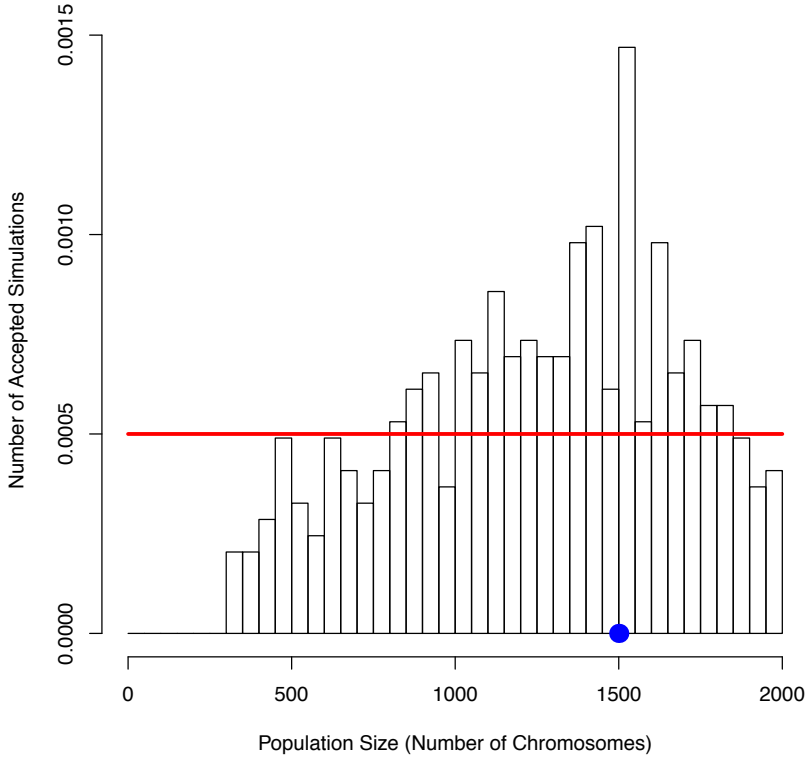


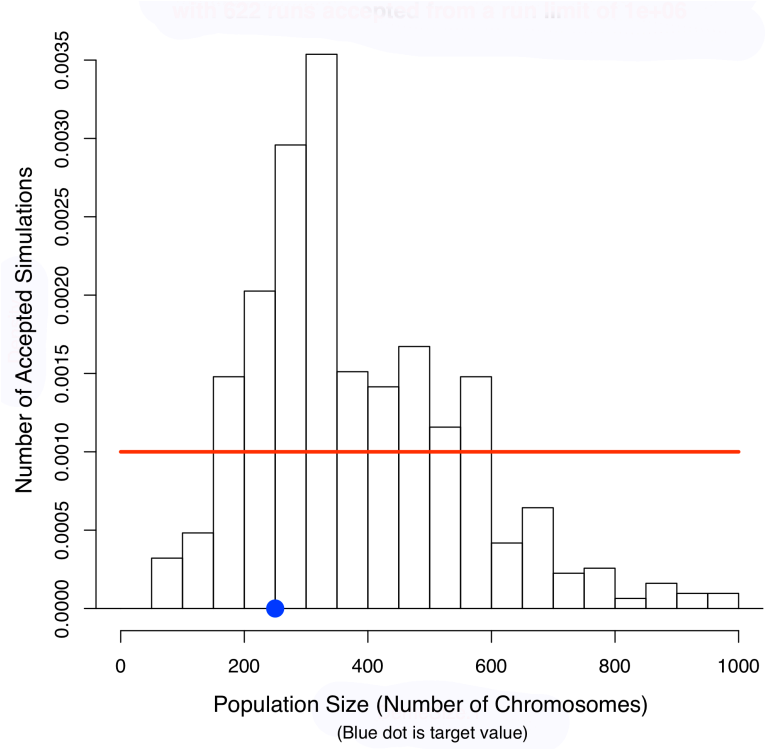
Figure S11 Legend: Median-joining network of 10 Y-chromosome STRs for individuals belonging to haplogroup A-M51. Each circle represents a distinct haplotype, and node sizes are proportional to the number of individuals. Branch length is proportional to the number of mutations. Colors represent different paternal ethnicities (turquoise: !Kung San of southern Angola, blue: N|u-speaking San of South Africa, pink: Khoe-speaking San of South Africa, red: Kxoe from the Caprivi Strip, Namibia). Combined with data from Cruciani et al. (2002).

Figure S12

a)



b)



c)

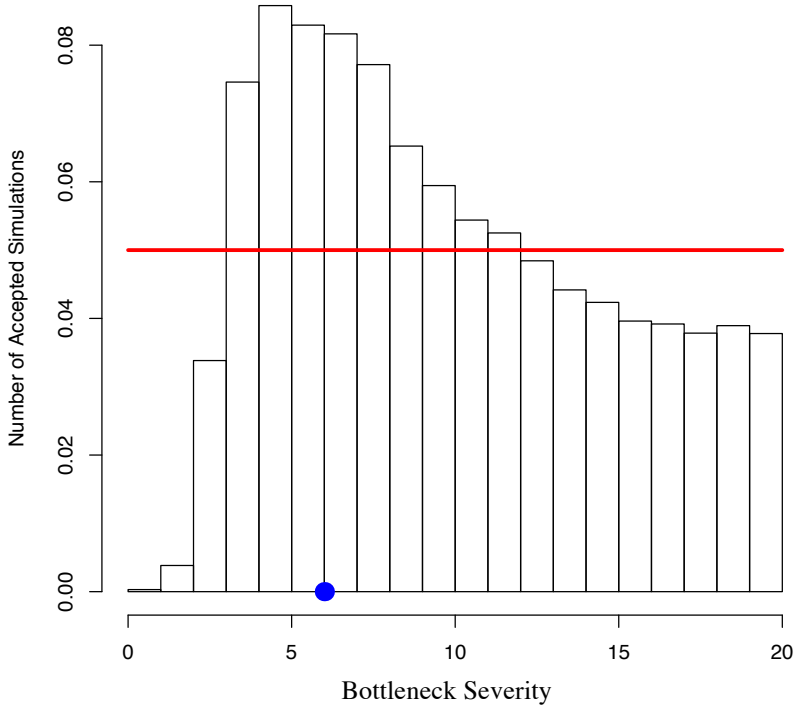


Figure S12 Legend: We performed simulations to demonstrate the accuracy of recovering effective population size and bottleneck severity estimates using the fROH statistic in REJECTOR. Blue dots indicate the known, simulated parameter. Populations with N_e of a) 1500 and b) 250, and c) a 6-fold bottleneck were simulated. The fROH for all individuals in the sample was calculated in REJECTOR. For the rejection algorithm, we sampled from a uniform distribution and kept simulated parameters within 5% of the observed fROH with REJECTOR (28). Red lines reflect the expected number of accepted runs under a uniform distribution.

Figure S13:

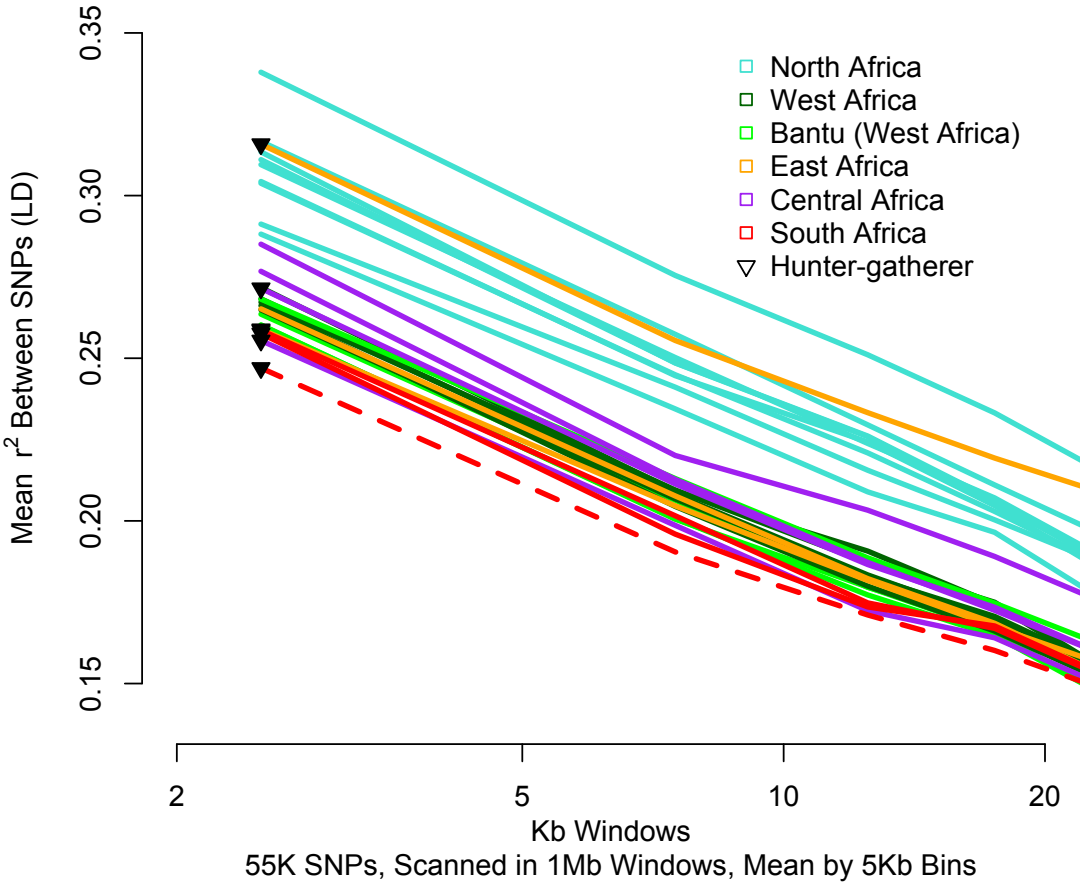


Figure S13 Legend: *Genome-wide linkage disequilbrium in 26 African populations*. Each line represents the population-specific LD decay. LD (r^2) between SNPs calculated in sliding 1Mb windows. r^2 estimates were binned by the genetic distance between SNPs, in 5Kb bins. Hunter-gatherers have the lowest LD curves (marked with 6 black triangles). LD calculated with a sample of all ≠Khomani Bushmen are indicated by a dashed red line; LD calculated after removing "admixed" individuals (<90% inferred Bushman ancestry) is indicated by a solid red line. Namibian Bushmen are indicated with a separate red line.

References:

1. Bryc K, *et al*. (2010) Genome-wide patterns of population structure and admixture in West Africans and African Americans. *Proc Natl Acad Sci USA* 107(2):786-791.
2. Nalls MA, *et al*. (2009) Measures of autozygosity in decline: globalization, urbanization, and its implications for medical genetics. *PLoS Genet* 5(3):e1000415.
3. Auton A, *et al*. (2009) Global distribution of genomic diversity underscores rich complex history of continental human populations. *Genome Res* 19(5):795-803.
4. Cao K, *et al*. (2004) Differentiation between African populations is evidenced by the diversity of alleles and haplotypes of HLA class I loci. *Tissue Antigens* 63(4):293-325.
5. Ellis JM, *et al*. (2000) Diversity is demonstrated in class I HLA-A and HLA-B alleles in Cameroon, Africa: description of HLA-A*03012, *2612, *3006 and HLA-B*1403, *4016, *4703. *Tissue Antigens* 56(4):291-302.
6. Kijak GH, *et al*. (2009) HLA class I allele and haplotype diversity in Ugandans supports the presence of a major east African genetic cluster. *Tissue Antigens* 73(3):262-269.
7. Middleton D, Menchaca L, Rood H, & Komerofsky R (2003) New allele frequency database: http://www.allelefrequencies.net. *Tissue Antigens* 61(5):403-407.
8. Modiano D, *et al*. (2001) HLA class I in three West African ethnic groups: genetic distances from sub-Saharan and Caucasoid populations. *Tissue Antigens* 57(2):128-137.
9. Solberg OD, *et al*. (2008) Balancing selection and heterogeneity across the classical human leukocyte antigen loci: a meta-analytic review of 497 population studies. *Hum Immunol* 69(7):443-464.
10. Marsh S, *et al*. (2010) An update to HLA Nomenclature, 2010. *Bone Marrow Transplant* 45(5):846-848.
11. Nei M (1987) *Molecular evolutionary genetics* (Columbia Univ Pr).
12. Conrad D, *et al*. (2006) A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nat Genet* 38(11):1251-1260.
13. Ramachandran S, *et al*. (2005) Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proc Natl Acad Sci USA* 102(44):15942-15947.
14. Li JZ, *et al*. (2008) Worldwide human relationships inferred from genome-wide patterns of variation. *Science* 319(5866):1100-1104.
15. Schuster SC, *et al*. (2010) Complete Khoisan and Bantu genomes from southern Africa. *Nature* 463(7283):943-947.
16. Quintana-Murci L, *et al*. (2010) Strong Maternal Khoisan Contribution to the South African Coloured Population: A Case of Gender-Biased Admixture. *Am J Hum Genet* 86(4):611-620.
17. Henn B, *et al*. (2008) Y-chromosomal evidence of a pastoralist migration through Tanzania to southern Africa. *Proc Natl Acad Sci USA*.

18. Pickrell J, *et al*. (2009) Signals of recent positive selection in a worldwide sample of human populations. *Genome Res* 19(5):826-837.
19. Browning SR & Browning BL (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet* 81(5):1084-1097.
20. Coop G, *et al*. (2009) The role of geography in human adaptation. *PLoS Genet* 5(6).
21. Sabeti P, *et al*. (2007) Genome-wide detection and characterization of positive selection in human populations. *Nature* 449(7164):913-918.
22. Tishkoff S, *et al*. (2007) History of Click-Speaking Populations of Africa Inferred from mtDNA and Y Chromosome Genetic Variation. *Mol Biol Evol* 24(10):2180-2195.
23. Mack SJ & Erlich HA (2007) Anthropology/Human Genetic Diversity Joint Report. *Immunobiology of the MHC: Proceedings of the 13th International Histocompatibility Workshop and Conference*, ed Hansen JA (IHWG Press, Seattle, WA), Vol 1, pp 557-766.
24. Spinola H, Bruges-Armas J, Middleton D, & Brehm A (2005) HLA polymorphisms in Cabo Verde and Guine-Bissau inferred from sequence-based typing. *Hum Immunol* 66(10):1082-1092.
25. Torimiro JN, *et al*. (2006) HLA class I diversity among rural rainforest inhabitants in Cameroon: identification of A*2612-B*4407 haplotype. *Tissue Antigens* 67(1):30-37.
26. Alexander DH, Novembre J, & Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* 19(9):1655.
27. Brisbin A (2010) Linkage Analaysis for Categorical Traits and Ancestry Assignment in Admixed Individuals. Ph.D. (Cornell University, Ithaca).
28. Jobin M & Mountain J (2008) REJECTOR: Software for Population History Inference from Genetic Data via a Rejection Algorithm. *Bioinformatics* 24(24):2936-2937.