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:

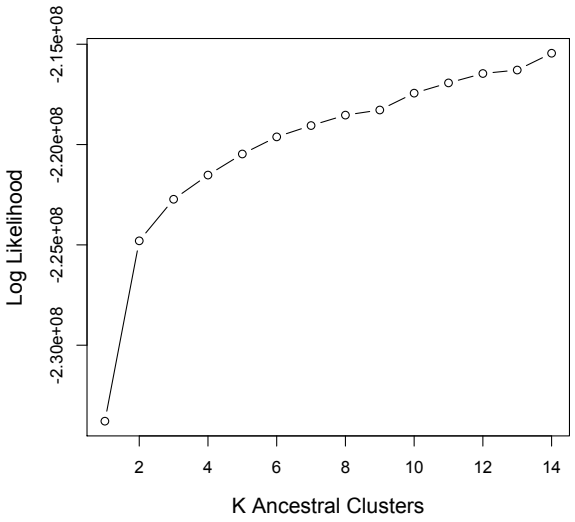
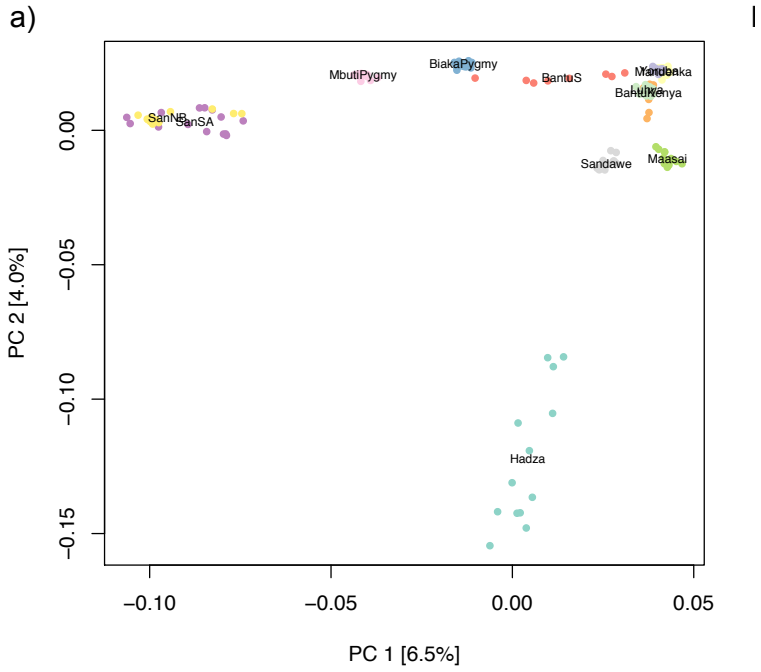
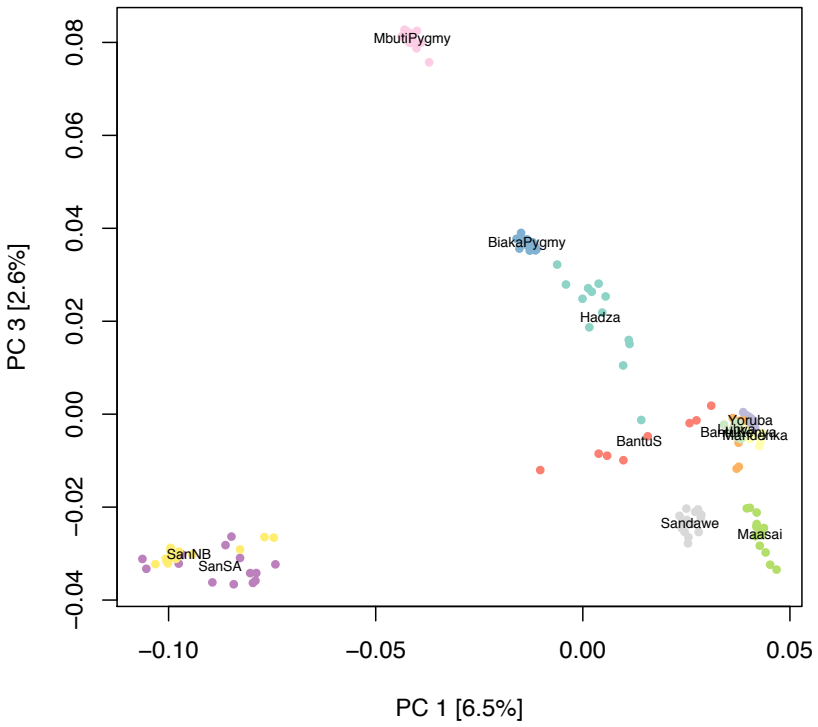


Figure S1: *Log Likelihoods for admixture k=1:14*. Log likelihoods (Y-axis) are plotted for the number of ancestral clusters k=1 through 14 (X-axis) generated with the *ADMIXTURE* program (26). Clusters k=2,4,6,8 are presented in Figure 1, main text.

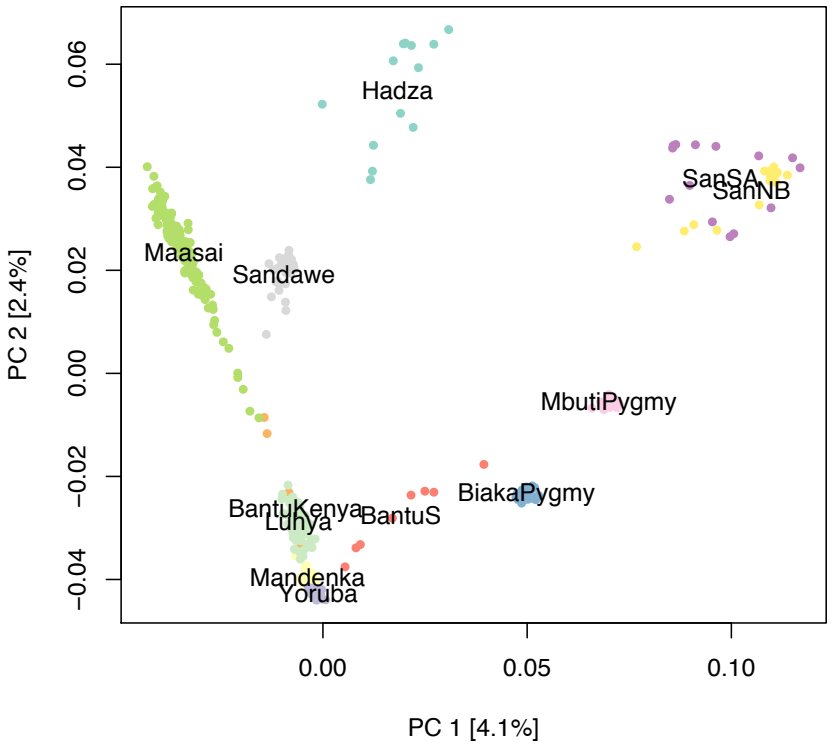
Figure S2:



b)


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

c)


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

d)

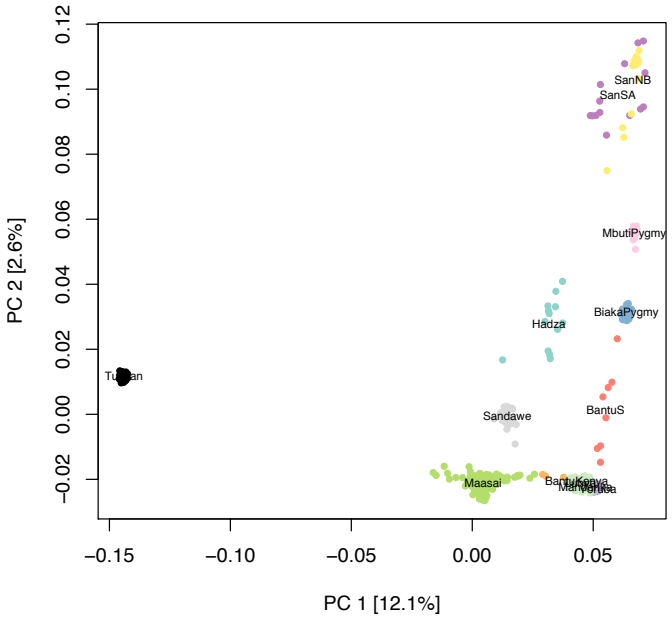


Figure S2: *Population structure of current African populations inferred from PCA*. We used principal component analysis to discriminate clusters of genetic variation within Africa. a) Population samples are displayed along the PC1 and PC2 axes of variation. Fifteen individuals from each population were 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)

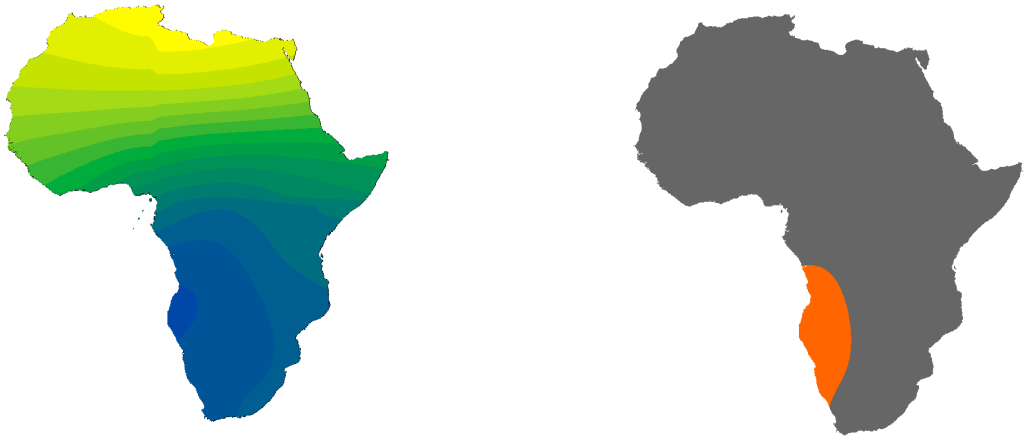
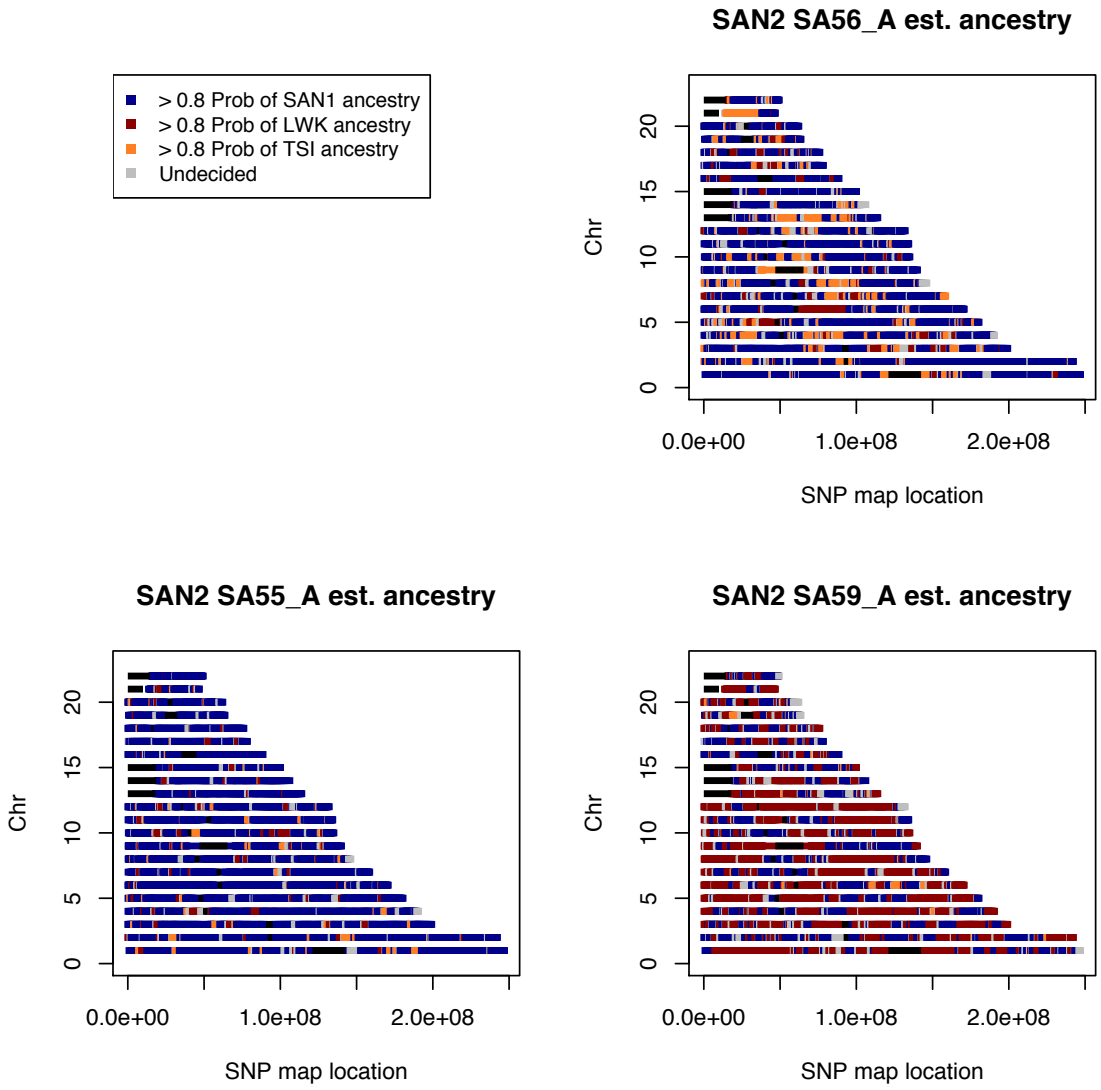


Figure S3: *Genome-wide linkage disequilbrium in 26 African populations (including admixed Bushmen)*. Regressions of LD on geographic distance were calculated assuming a grid of origin locations. All southern African Bushmen were included in these LD samples, regardless of their degree of admixture. 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

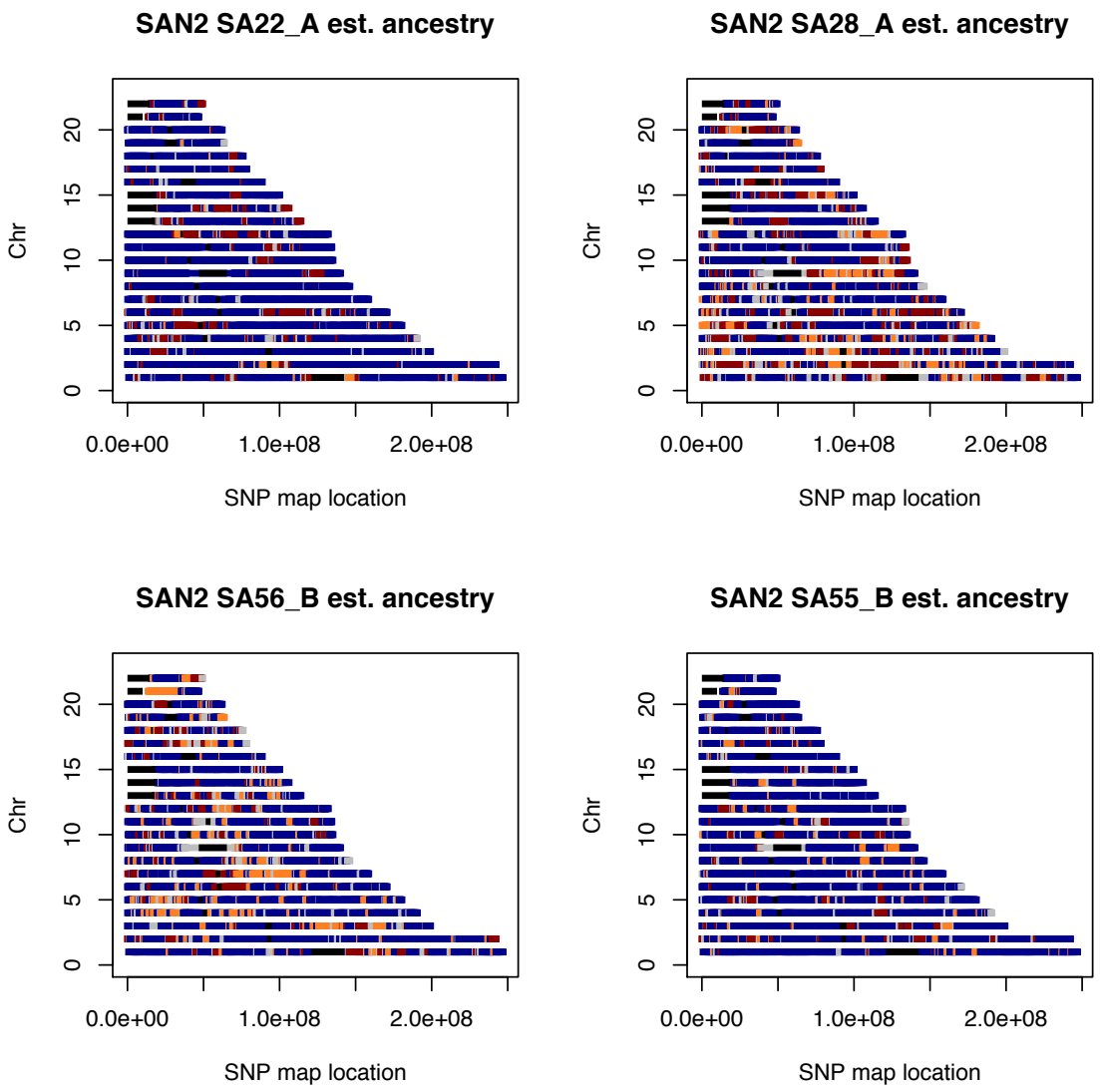


Figure S4 Legend: *Local ancestry assignment along phased chromosomes*. Individuals with potential admixture (Fig. 1) were projected into the principal component space of three putative ancestral populations. South African ≠Khomani San: SAN, European Tuscan: TSI, Luhya Bantu: LWK. Ancestry was assigned 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:

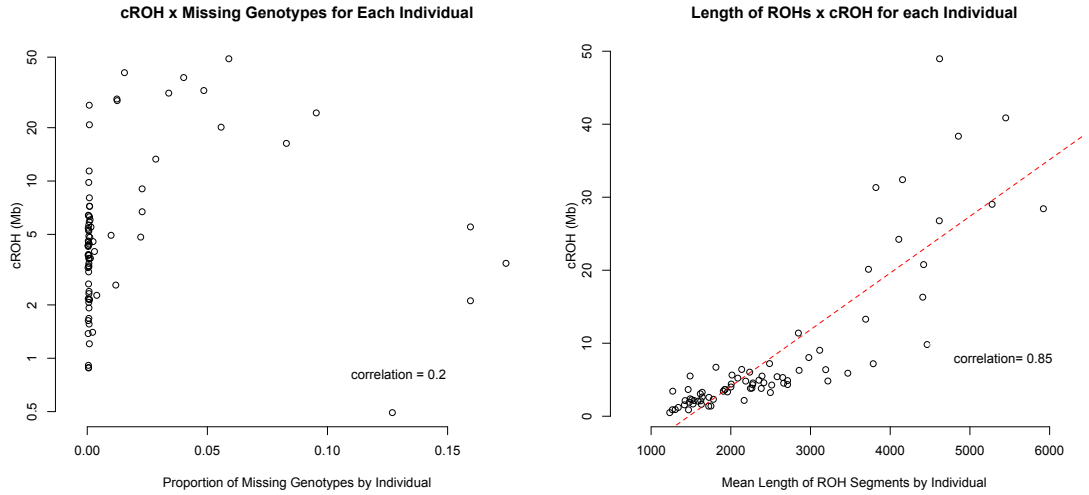
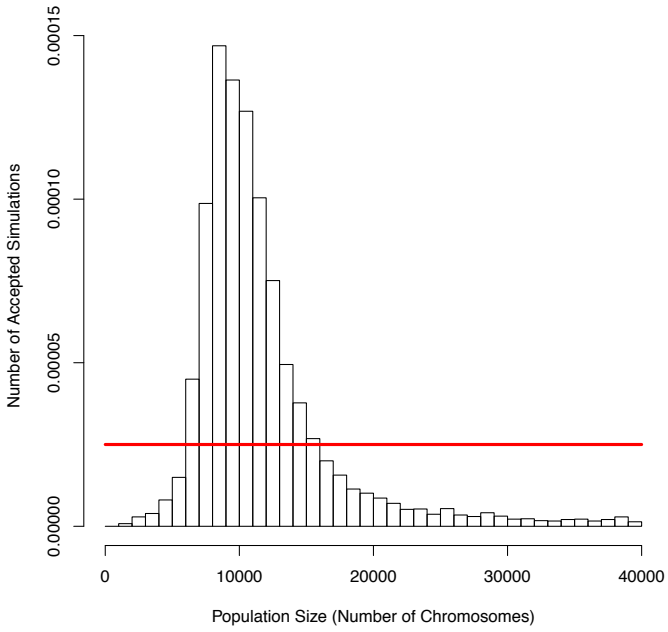


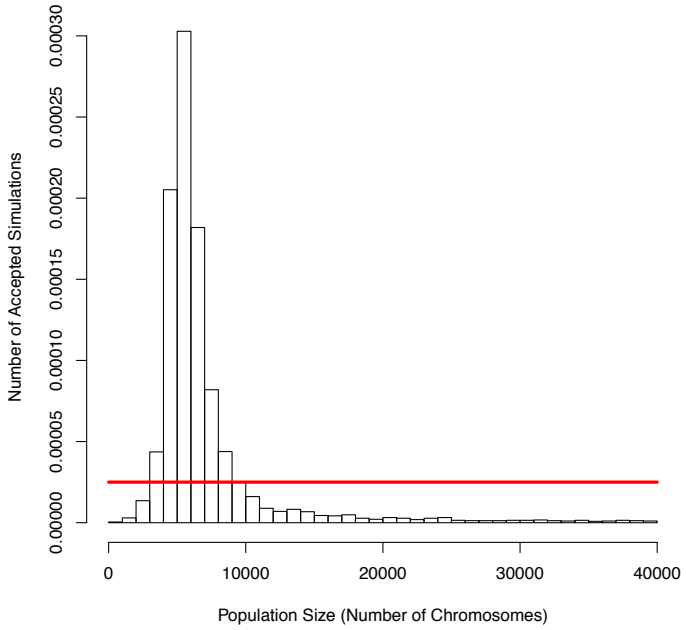
Figure S5 Legend: Cumulative runs of homozygosity (cROH) by missingness and segment length.

Figure S6:

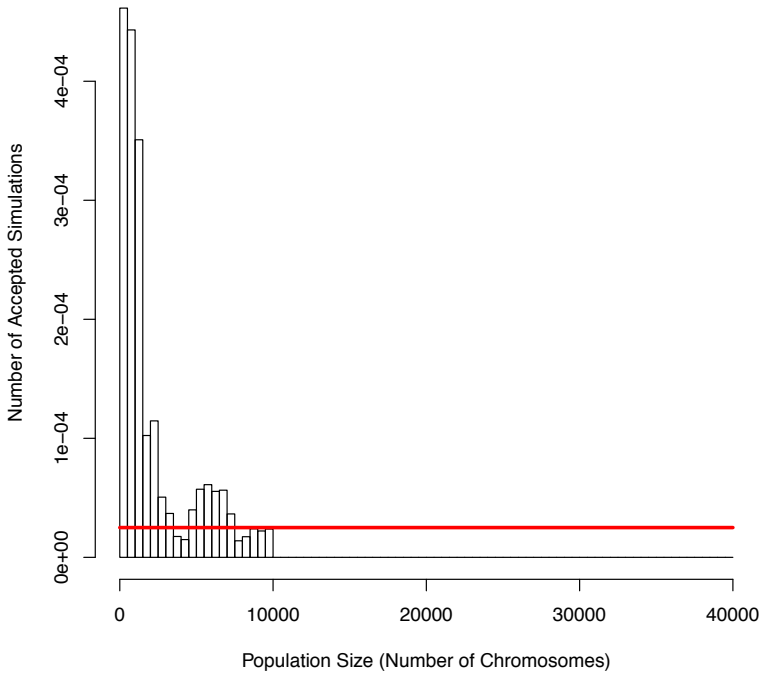
a)



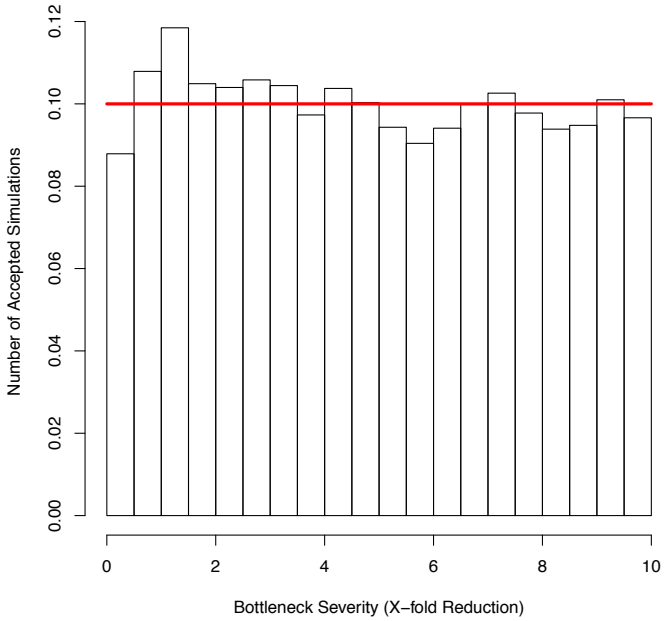
b)


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)

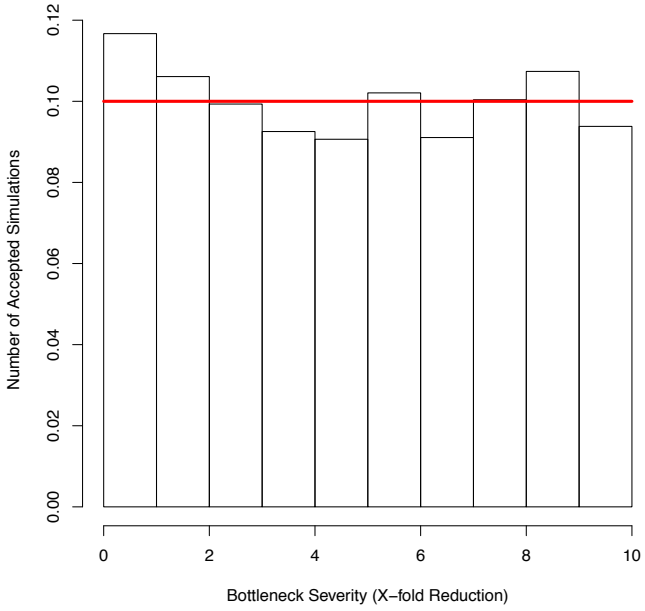
c)


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)

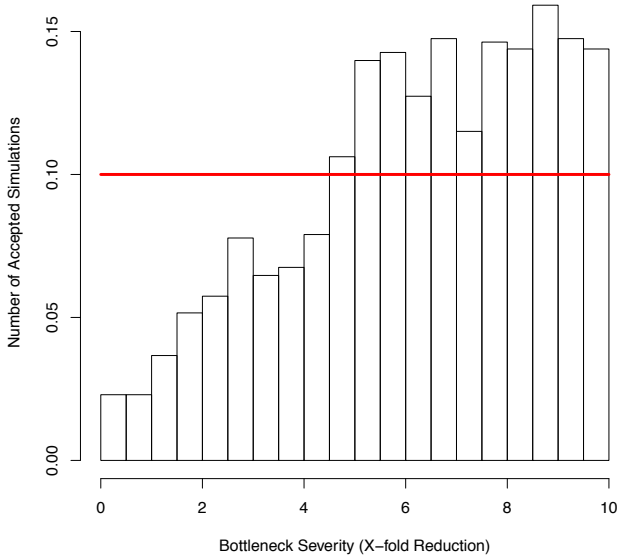


Figure S6: Simulated posterior distribution of effective population size in the a) Khomani San, b) Sandawe, c) Mbuti generated by sampling from a uniform distribution of N_e and keeping simulated parameters within 20% of the observed fROH (fraction of the genome in ROH) with REJECTOR (28). Simulated 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:

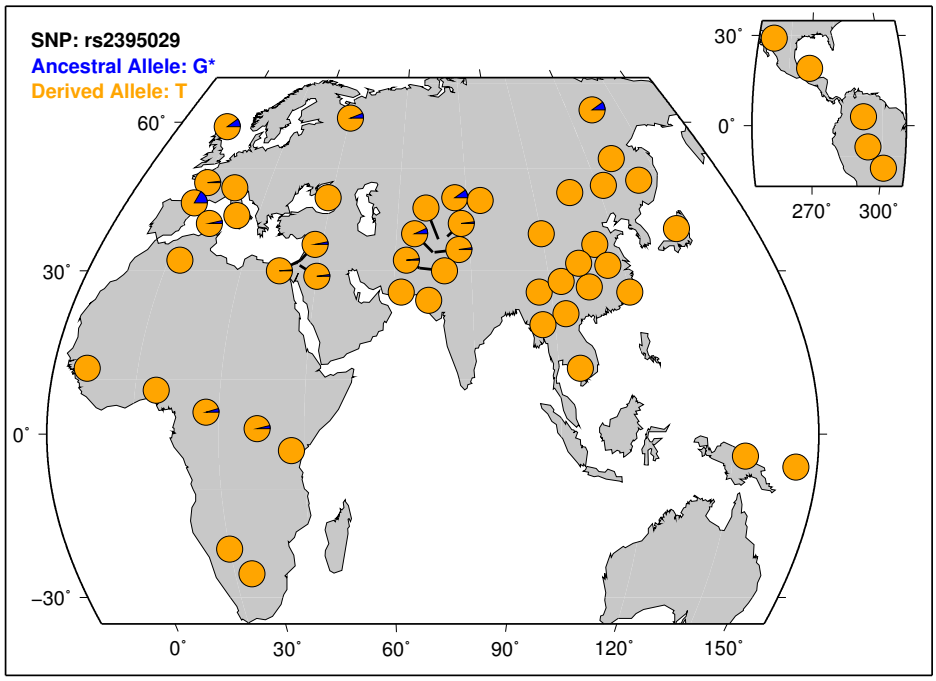


Figure S7 Legend: *Allele frequencies for rs2395029*. Allele frequencies of rs2395029, a missense Val to Gly mutation in the HPC5 gene, for 52 of the Human Genome Diversity Project populations. The Sandawe population carries the G allele at 19% frequency, higher than almost all other world populations. 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

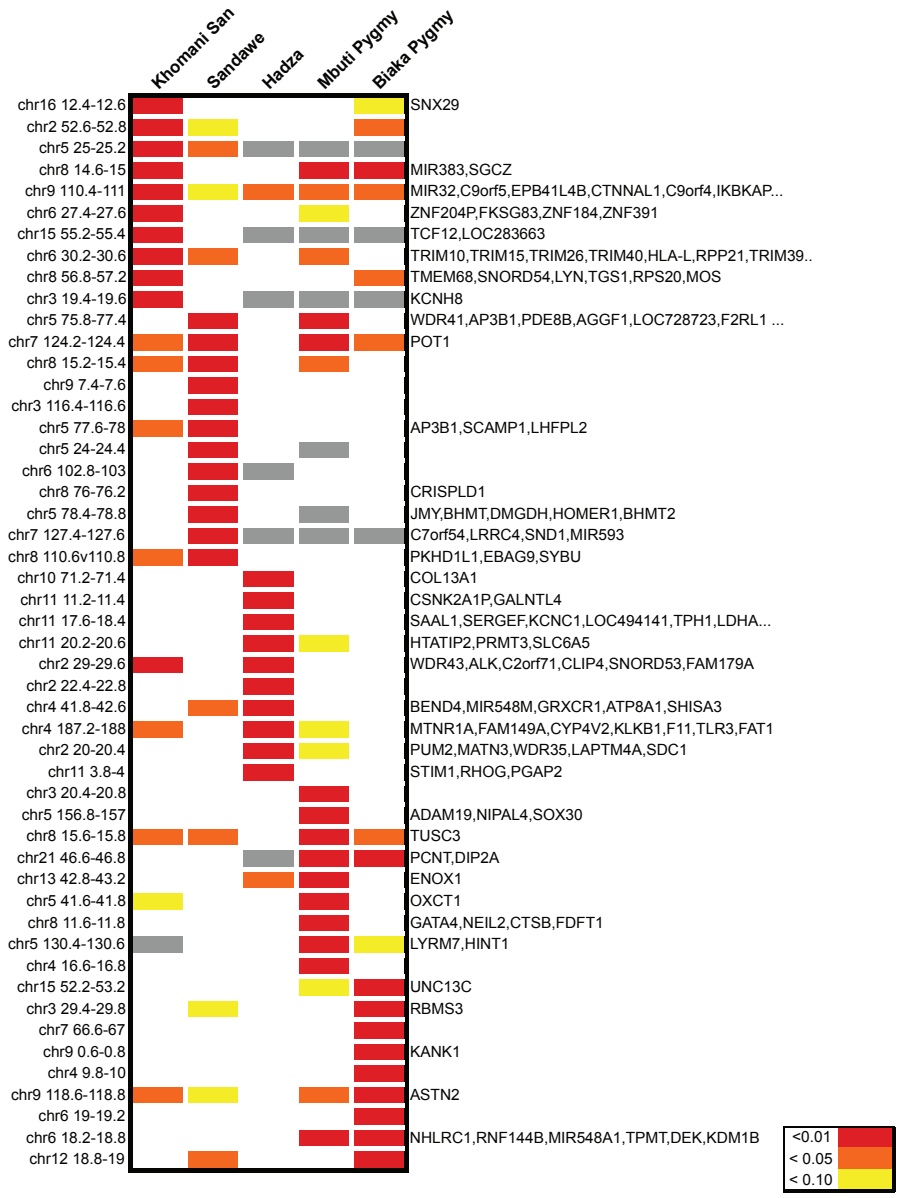


Figure S8 Legend: *Haplotype-based scan (iHS) for selection in 5 hunter-gatherer populations*.
A haplotype-based statistic (*iHS*) was used to identify ongoing selective sweeps in 5 hunter-gatherer populations using a common set of 461K SNPs. Empirical p-values for each population were calculated following 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

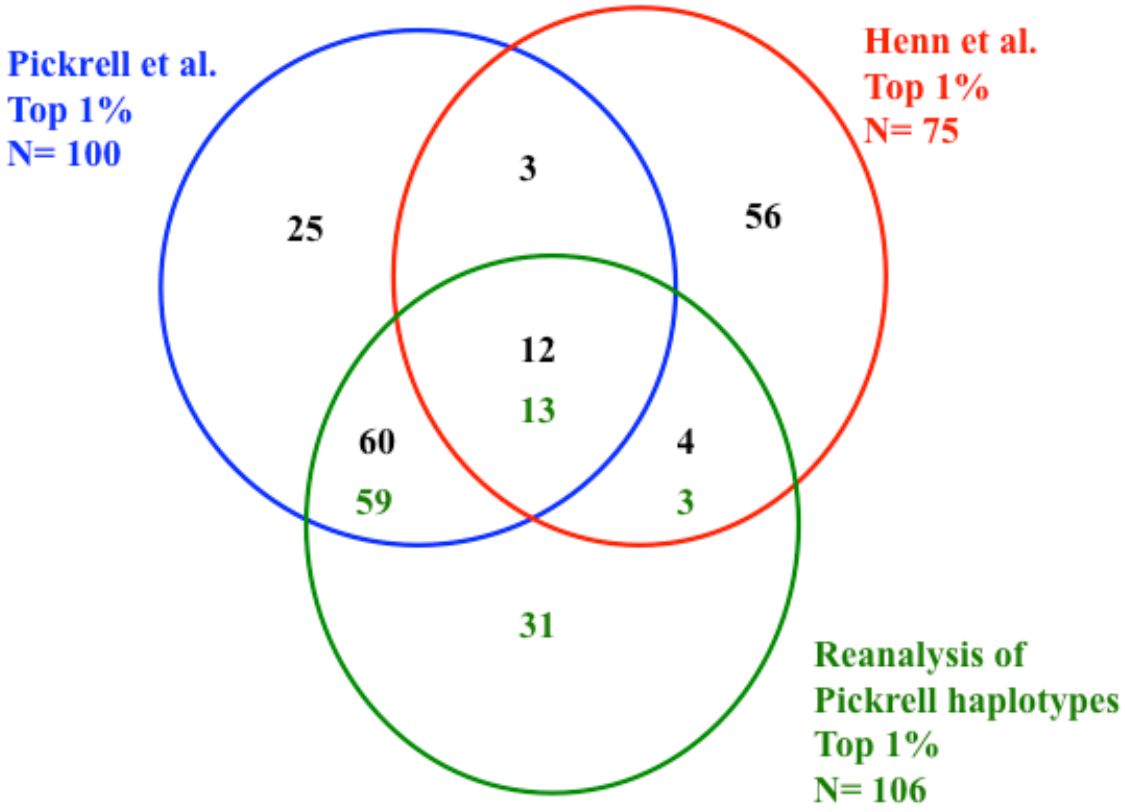


Figure S9 Legend. Comparison of extreme iHS regions identified for the HGDP Biaka. The overlap among the most extreme regions identified by Pickrell et al. (18) (in blue), in this study (Henn et al., in red), and found by applying the binning strategy employed in this study to the previously phased Biaka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

Figure S10

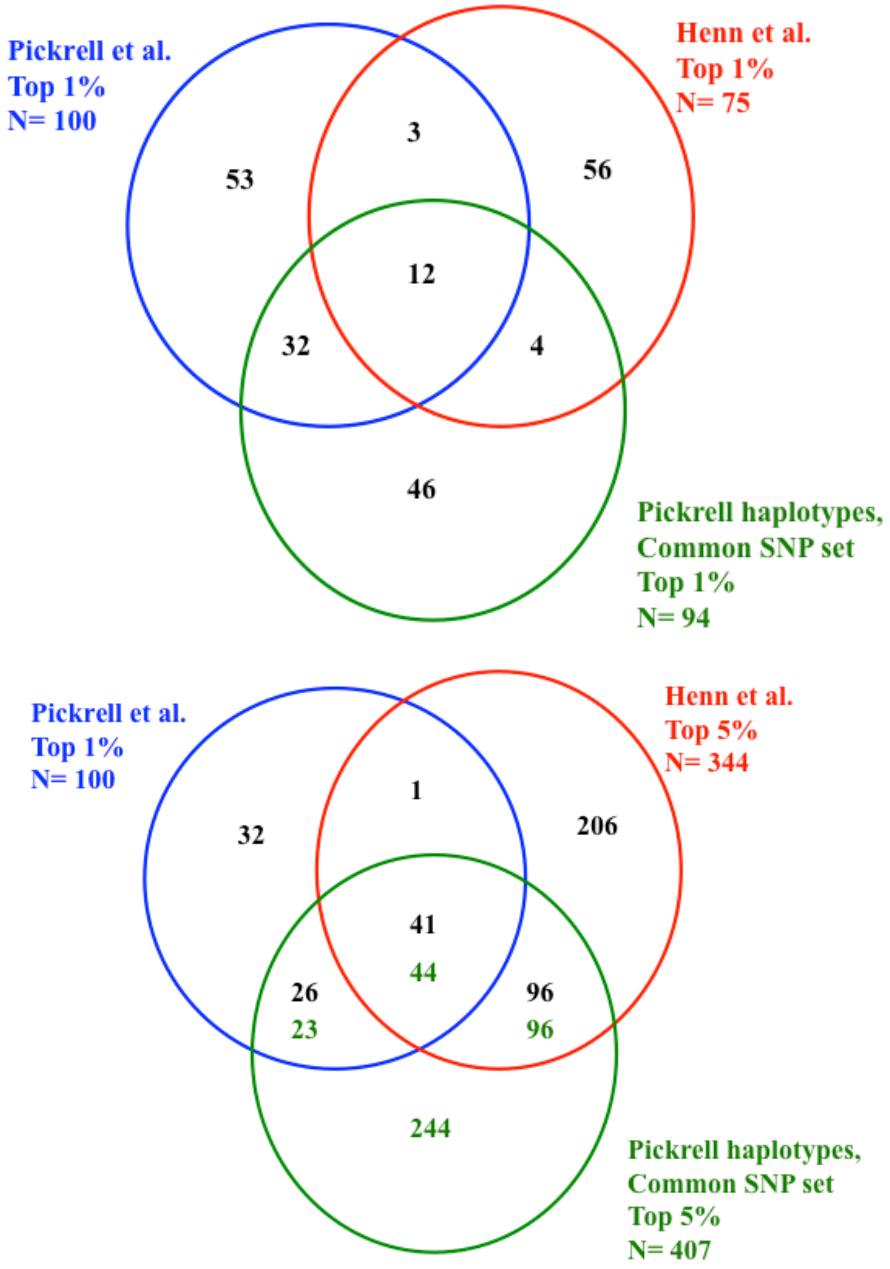


Figure S10. Impact of reduced SNP set on identified regions of extreme iHS. Analysis was performed on the set of common SNPs used for the populations analyzed in this study but using the phased haplotypes obtained from Pickrell et al. (18)

Figure S11:

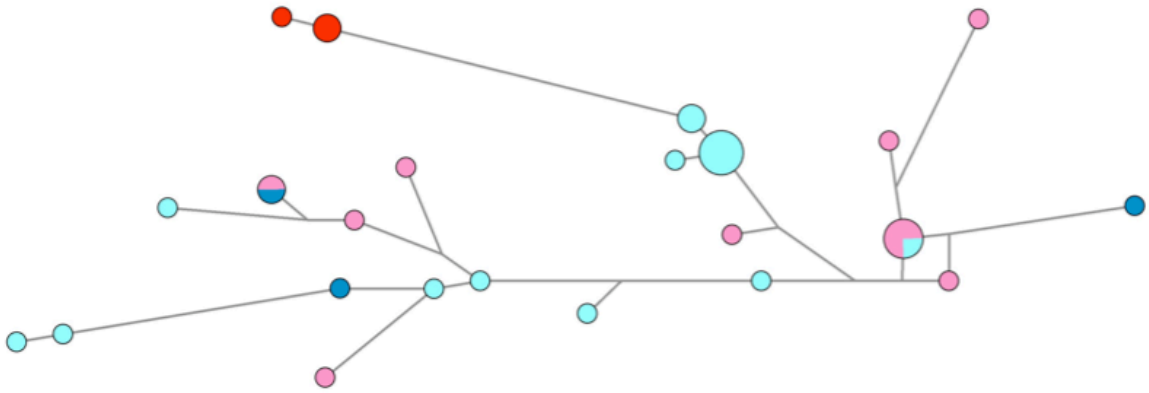
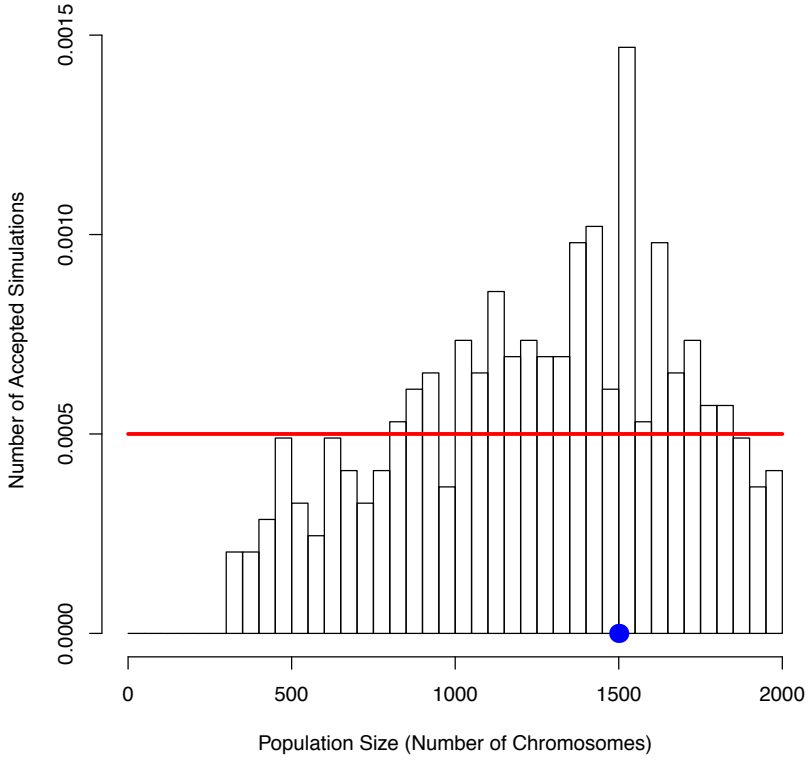


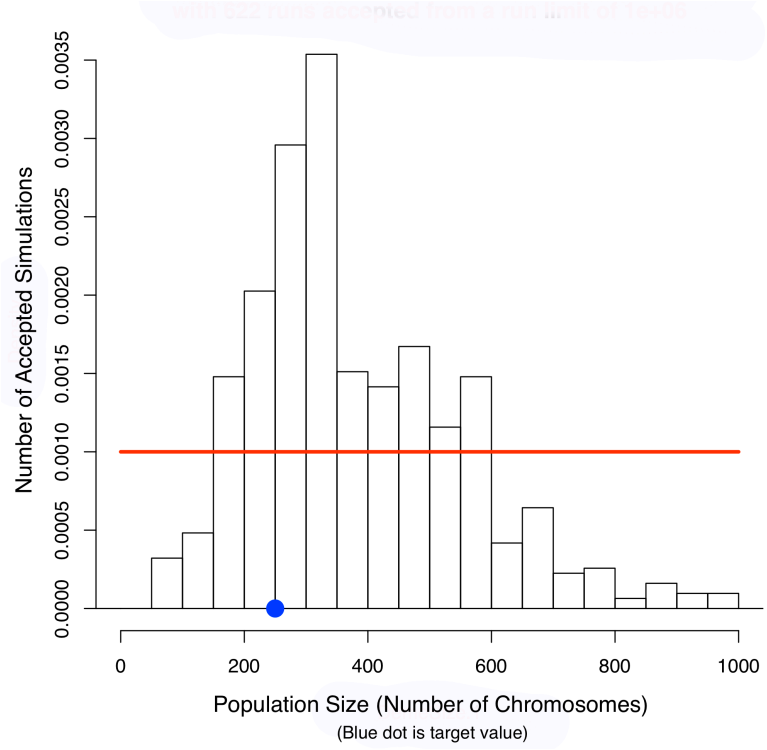
Figure S11 Legend: Median-joining network of 10 Y-chromosome STRs for individuals belonging to haplogroup A-M51. Each circle represents a distinct haplotype, and node sizes are proportional to the number of individuals. Branch length is proportional to the number of mutations. Colors represent different 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)

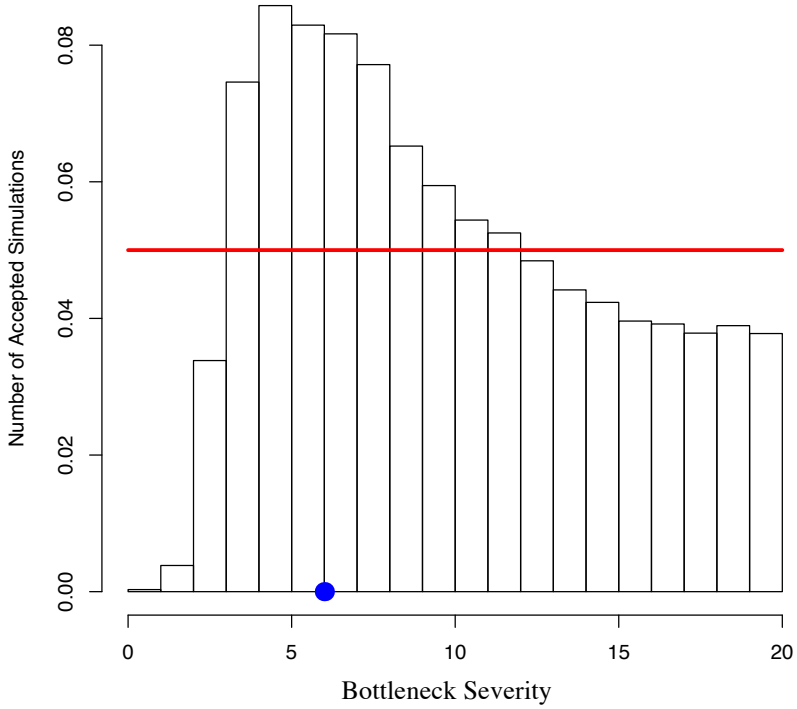


Figure S12 Legend: We performed simulations to demonstrate the accuracy of recovering effective population size and bottleneck severity estimates using the fROH statistic in REJECTOR. Blue dots indicate the known, simulated parameter. Populations with N_e of a) 1500 and b) 250, and c) a 6-fold bottleneck 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:

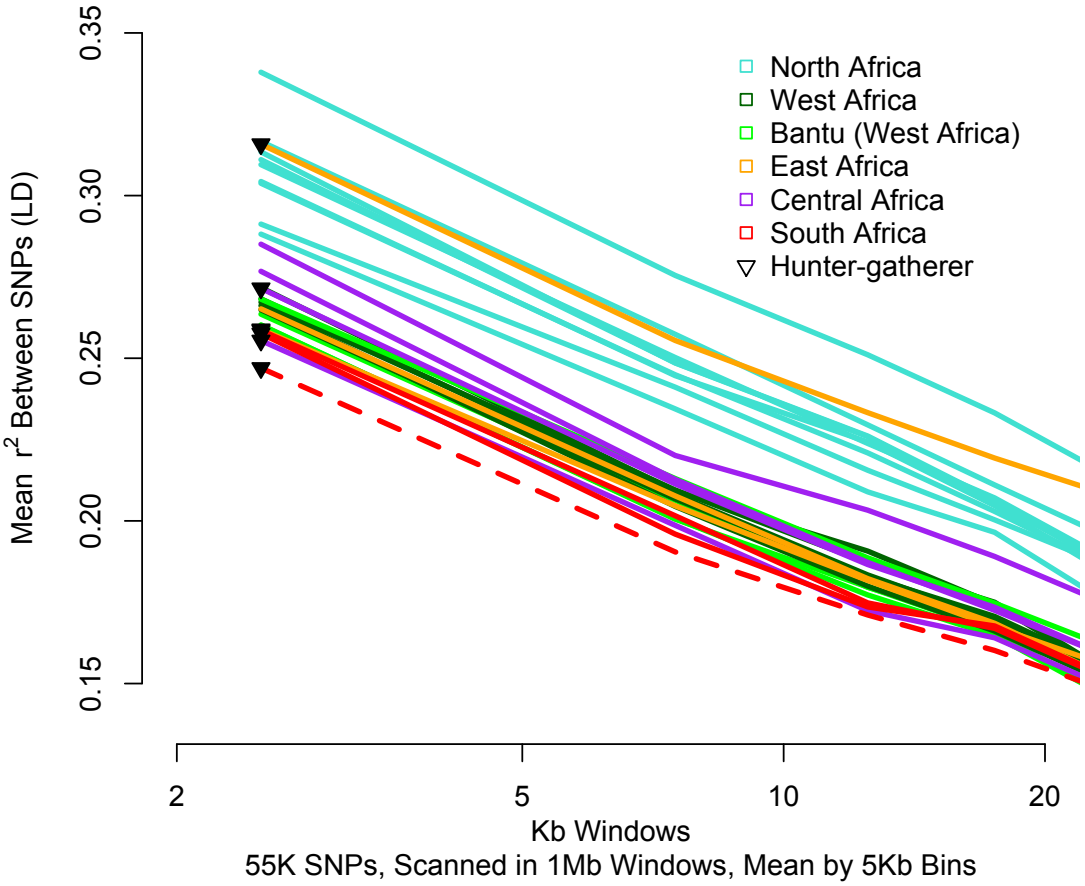


Figure S13 Legend: *Genome-wide linkage disequilbrium in 26 African populations*. Each line represents the population-specific LD decay. LD (r^2) between SNPs calculated in sliding 1Mb windows. r^2 estimates were binned by the genetic distance between SNPs, in 5Kb bins. Hunter-gatherers have the lowest 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.