Supporting Information

Skoglund and Jakobsson 10.1073/pnas.1108181108

SI Materials and Methods

Genotype Data Collection and Quality Filtering. We obtained phased Hapmap 3 genotypes from seven populations (Utah residents with Northern and Western European ancestry from the Centre de'Etude du Polymorphism Humain collection, CEU; Toscani in Italy, TSI; Gujarati Indians in Houston, TX, GIH; Japanese in Tokyo, Japan, and Han Chinese in Beijing, China, JPT + CHB; Maasai in Kinyawa, Kenya, MKK; Luhya in Webuye, Kenya LWK, and Yoruba in Ibadan, Nigeria, YRI; ftp://ftp.ncbi.nlm.nih.gov/hapmap/phasing/2009-02_phaseIII/HapMap3_r2/ 30 nov 2010) (1), additional Finnish HapMap data (FIN_HAPMAP; ftp://ftp.fimm.fi/pub/FIN_HAPMAP3/) (2), and phased genotypes for 938 individuals from the Human Genome Diversity Project (HGDP; http://hgdp.uchicago.edu/Phased_data/) (3, 4). In geographic analyses, we used waypoint distances from Addis Ababa from the work in ref. 5.

To resolve strand issues in the datasets, we flipped SNPs to hg18 strand orientation for the HGDP and the HapMap/FIN_HAPMAP data separately and intersected the 1,163,280 SNPs common to HapMap and FIN_HAPMAP with the HDGP Illumina data to a total of 502,931 remaining SNPs. For our analysis of population structure, we used a dataset created by randomly sampling one allele from each individual at each position to allow comparison with the single-pass ancient data. We intersected the dataset above with chimpanzee nucleotides from a multiple alignment of human genomes (6, 7), resulting in a dataset of 491,388 overlapping SNPs.

Sequence reads aligned to the hg18/NCBI36 from Denisova (8) (ftp://hgdownload.cse.ucsc.edu/gbdb/hg18/denisova) and Neandertal (9) (ftp://ftp.ebi.ac.uk/pub/databases/ensembl/neandertal/) were downloaded from the authors. We extracted autosomal loci from Neandertal and Denisova, and we removed bases with quality <40 and from within 5 and 1 bp from the 5′ end of the sequence reads from Neandertal and Denisova, respectively (8). We filtered reads with mapping quality <90 (Neandertal) and <37 (Denisova) (8) and randomly chose a single read from positions covered by multiple reads.

This dataset consisted of 228,984 SNPs, of which 4,006 were triallelic and excluded (most likely because of postmortem mutations or sequencing error in the archaic genomes), leaving 224,978 SNPs. To avoid the effect of postmortem nucleotide misincorporations in analyses that included ancient genomes, we followed the works in refs. 8 and 9 by removing all transitions (C/T and G/A), resulting in 40,656 SNPs with no triallelic SNPs identified. However, we obtained similar results when transitions were included (Fig. S1).

We also excluded one SNP from each pair of SNPs with $r^2 > 0.2$. In the dataset that included only transversion SNP, the number of SNPs used was 38,848 (1,808 excluded because of $r^2 > 0.2$), but the full dataset comprised 183,166 SNPs (49,741 excluded).

Genome Sequence Data Collection and Quality Filtering. We analyzed the 7.5× genome from the work by Levy et al. (10) obtained with capillary sequencing technology together with four ∼30× genome assemblies obtained with Illumina sequencing technology from two individuals from Utah of European descent (11), one individual of Chinese descent (12) and one individual of Korean descent (13). We added two Yoruba individuals from Ibadan, Nigeria, that had also been sequenced to ∼30× coverage (11) to this dataset. We identified SNPs between the eight genomes using genotypes called by the 1,000 Genomes Project (11) (ftp://ftp.sanger.ac.uk/pub/rd/humanSequences/) that were filtered for mapping uniqueness and consensus quality. We excluded hypermutable CpG sites and transition substitutions and retained SNPs for which there was both Neandertal and Denisova genome data, using the same genotyping criteria for the ancient genomes as described above.

Projecting Modern Human Data on Principal Components Defined by Archaic and Chimpanzee Genomes. Principal component analysis (PCA) relates conceptually to covariance in coalescent times between gene copies (14) and has been shown to capture fundamental properties of genetic ancestry (15). The method of projecting samples on the principal components (PCs) defined by single samples of chimpanzee, Neandertal, and Denisova individuals (9) circumvents some of the problems associated with PCA (14) in that each extant human individual is projected independently onto the defined axes of variation. Using haploid data from each individual, we expected that demographic changes within human populations (after their divergence from archaic populations) should not affect the projection. This prediction follows from the prediction that each single genetic lineage from the modern human population is analyzed separately, allowing no influence from other modern human samples on the PCA (Fig. 2 *D–F*). Thus, the effects of sample size, genetic drift, and isolation by distance on PCA (14) are expected to be alleviated by the projection approach. Projecting extant humans onto predefined axes of variation will not capture the major components of extant human genetic variation but can, instead, detect heterogeneities in similarity to the two archaic human genomes (9). The simulation results reported above confirm that this approach is insensitive to relatedness between groups in the absence of admixture (Fig. 2*D*). However, our simulations also show that, although this expectation seems to hold true for unbiased genetic data, the joint effect of ascertainment bias and genetic drift can cause strong but qualitatively predictable deviations from the expected pattern.

SI Results

To investigate putative signs of intraregional differences in archaic ancestry, we compared the projection of modern humans on the two top PCs in the archaic analyses (PC1 separates chimpanzee from the two archaic hominins and PC2 separates Denisova and Neandertal) with PC loadings computed directly on genotypes of modern humans in different continents (using the full dataset of 183,166 SNPs). In the PCA of Eurasian population structure, we computed the top 10 PCs using only European and East Asian populations and thereafter, projected the remaining populations (Central/South Asians) on the obtained components. For all other PC analyses of African, European, and Native American populations, we computed the PCs using all available populations from the region, and for Africa and America, we also included HapMap Tuscans (TSI) to investigate the effect of European admixture in the data. We then tested for each obtained PC in each analyses if it was correlated to the loading of an individual in the archaic PCA.

We found significant correlation for both datasets in the case of Eurasian population structure and additionally, in regions of the world where recent admixture has introduced European genetic variation (see below). Excluding Native Americans and Oceanians, we find that the PC separating chimpanzee and archaic hominins is significantly correlated with population differentiation within Eurasia (Table S2). For African populations, we also

find a trend to correlation with PC1 (not significant after Bonferroni correction for multiple tests), separating hunter–gatherer populations (San, Mbuti, and Biaka) from Bantu-speaking populations. However, we find that, when Maasai from Kenya (MKK) are excluded, the correlation does not remain (Table S2), in line with the suggestion that the MKK population is the result of admixture with a European-related population (1). Similarly, a significant correlation between archaic PC1 and population structure within America was weakened when we removed individuals that were skewed to having European ancestry (Table S2), and the remaining pattern of increased signs of archaic ancestry in American populations more distant from Africa is in line with the joint effect of ascertainment bias and genetic drift described in the text. We found no correlation between the archaic components of variation and the top 10 PCs of intra-Europe population differentiation.

1. Altshuler DM, et al. (2010) Integrating common and rare genetic variation in diverse human populations. *Nature* 467:52–58.
2. Surakka I, et al. (2010) Founder population-specific HapMap panel increases power in GWA studies through improved imputation accuracy and CNV tagging. *Genome Res* 20:1344–1351.
3. Li JZ, et al. (2008) Worldwide human relationships inferred from genome-wide patterns of variation. *Science* 319:1100–1104.
4. Pickrell JK, et al. (2009) Signals of recent positive selection in a worldwide sample of human populations. *Genome Res* 19:826–837.
5. 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.
6. Chimpanzee Sequencing and Analysis Consortium (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* 437:69–87.
7. Schuster SC, et al. (2010) Complete Khoisan and Bantu genomes from southern Africa. *Nature* 463:943–947.
8. Green RE, et al. (2010) A draft sequence of the Neandertal genome. *Science* 328:710–722.
9. Reich D, et al. (2010) Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* 468:1053–1060.
10. Levy S, et al. (2007) The diploid genome sequence of an individual human. *PLoS Biol* 5: e254.
11. Durbin RM, et al. (2010) A map of human genome variation from population-scale sequencing. *Nature* 467:1061–1073.
12. Wang J, et al. (2008) The diploid genome sequence of an Asian individual. *Nature* 456: 60–65.
13. Ahn S-M, et al. (2009) The first Korean genome sequence and analysis: Full genome sequencing for a socio-ethnic group. *Genome Res* 19:1622–1629.
14. McVean G (2009) A genealogical interpretation of principal components analysis. *PLoS Genet* 5:e1000686.
15. Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. *PLoS Genet* 2:e190.

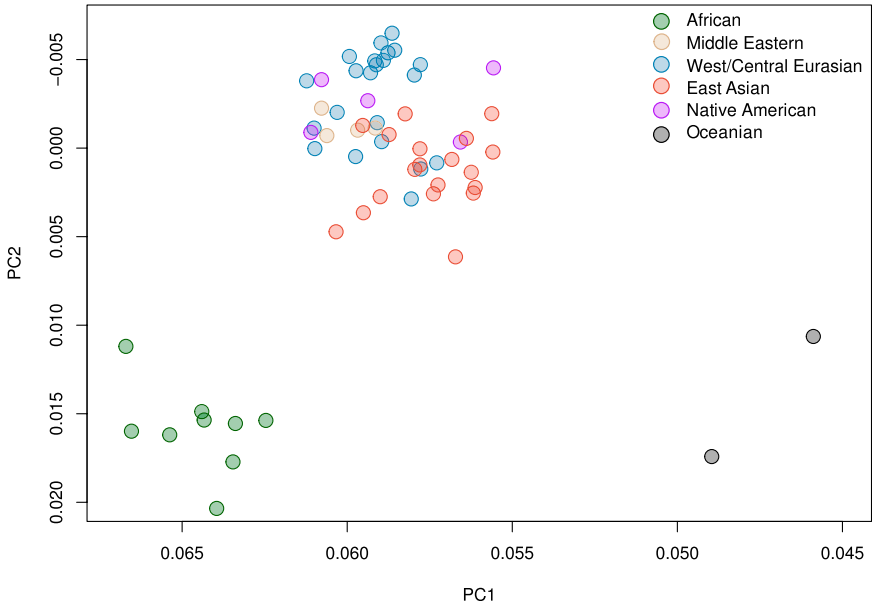

Fig. S1. Modern human variation projected on axes of variation defined by Denisova, Neandertal, and chimpanzee using all available SNPs including transition polymorphisms (C/T and G/A). Axes are inverted for comparison with Fig. 1*B*. PC1 and PC2 loadings for Denisova, Neandertal, and chimpanzee were (−0.32, 0.75), (−0.49, −0.65), and (0.81, −0.10), respectively.

Table S1. Two-sided *t* tests for difference between humans from different regions in the PCA of chimpanzee, Denisova, and Neandertal

PC	pop1	pop2	Mean pop1	Mean pop2	*P*
PC1 (chimp–archaic)	East Asia	West Eurasia	−0.06465564	−0.06768811	**0.0000007331**
PC1 (chimp–archaic)	East Asia	Europe	−0.06465564	−0.0672409	**0.0003080**
PC1 (chimp–archaic)	East Asia	America	−0.06465564	−0.06103651	**0.0008398**
PC1 (chimp–archaic)	America	West Eurasia	−0.06103651	−0.06768811	**0.000000003703**
PC2 (Denisova–Neandertal)	East Asia	West Eurasia	−0.0017002564	0.0002794992	**0.005914**
PC2 (Denisova–Neandertal)	East Asia	Europe	−0.0017002564	0.0008832836	**0.002934**
PC2 (Denisova–Neandertal)	East Asia	America	−0.001700256	0.005192063	**0.00002059**
PC2 (Denisova–Neandertal)	America	West Eurasia	0.0051920635	0.0002794992	**0.001568**

Significant tests are indicated in bold.

Table S2. Results of Spearman's rank correlation tests between the projection on PC1 and PC2 in the chimpanzee–Neandertal–Denisova analyses (archaic PC) and PCs representing population structure in different regions of the world (intraregional PC)

Archaic PC	Intraregional PC	r_s	*P*
PC1	Eurasian PC1	0.138	**2.293×10^{-6}***
PC1	Eurasian PC2	0.090	**0.002***
PC2	Eurasian PC1	−0.073	0.016
PC2	Eurasian PC2	0.050	0.089
PC1	African PC1	−0.142	0.016
PC2	African PC1	−0.001	0.988
PC1	African PC2	−0.085	0.148
PC2	African PC2	−0.020	0.737
PC1	African PC1 (excluding MKK)	0.000	0.998
PC2.	African PC1 (excluding MKK)	0.074	0.293
PC1	African PC2 (excluding MKK)	0.000	0.995
PC2	African PC2 (excluding MKK)	0.015	0.835
PC1	European PC1	−0.026	0.614
PC2	European PC1	0.042	0.412
PC1	European PC2	−0.018	0.733
PC2	European PC2	−0.010	0.847
PC1	American + TSI PC1	0.250	**0.002***
PC2	American + TSI PC1	0.112	0.170
PC1	American + TSI PC2	0.069	0.399
PC2	American + TSI PC2	0.007	0.927
PC1	American + TSI PC1 [excluding admixed (PC1 < 0.09)]	0.249	0.121
PC2	American + TSI PC1 [excluding admixed (PC1 < 0.09)]	−0.020	0.903
PC1	American + TSI PC2 [excluding admixed (PC1 < 0.09)]	0.350	0.027
PC2	American + TSI PC2 [excluding admixed (PC1 < 0.09)]	−0.028	0.864

Eurasian analyses include samples from Europe, Central/South Asia, and East Asia. Putatively admixed individuals in American populations were excluded based on a cutoff value (PC1 < 0.09) determined by visual inspection of PC1 separating European TSI and American populations.
**P* values below 0.00208 (Bonferroni-corrected cutoff for 24 tests) are bold.

Table S3. Average frequency of the Denisova allele at SNPs where Denisova differs from both Neandertal and chimpanzee

Population	Average frequency of the Denisova allele
Papuan	0.534685512191
Yizu	0.530125345231
Melanesian	0.528999362651
Colombian	0.528149564478
Karitiana	0.527283424033
She	0.526811132356
Miaozu	0.526726152539
Naxi	0.52642341194
Yakut	0.526263012535
Tujia	0.526258763544
French	0.525880906856
Tu	0.525855109412
Dai	0.525727639686
Daur	0.525293298397
Tuscan	0.525112785713
CHB	0.524894533977
Han	0.524721401394
Xibo	0.524537922243
Mongola	0.524367962609
North Italian	0.524148431414
Hezhen	0.524065812147
Russian	0.52402804334
Cambodian	0.523985553431
French Basque	0.523659525094
Adygei	0.523513165623
Orcadian	0.523135755258
CEU	0.523063272473
Sardinian	0.522974900604
Lahu	0.522944550669
Maya	0.522899025767
Pathan	0.522712787531
Palestinian	0.522565836266
Bedouin	0.52240870571
Japanese	0.522292027072
TSI	0.52200784132
Sindhi	0.522006231853
Kalash	0.521988527725
Uygur	0.52181856809
Oroqen	0.521681656162
Burusho	0.521640110474
Makrani	0.521614616529
GIH	0.521599841629
JPT	0.521254835697
LSFIN	0.521170203558
Pima	0.520941454976
Hazara	0.520897309616
FIN	0.520538559592
Druze	0.520359747893
Balochi	0.520280079314
Brahui	0.519226683663

Han Chinese in Beijing, China, CHB; Utah residents with Northern and Western European ancestry from the CEPH collection, CEU; Toscani in Italy, TSI; Gujarati Indians in Houston, TX, GIH; Japanese in Tokyo, Japan, JPT; late-settlement founder population in Finland, LSFIN; and Finnish from Finland, FIN.

Table S4. Results of 4-population-tests for admixture between seven regions

		(Pop1, Pop2, Denisova, chimpanzee)			(Pop1, Pop2, Denisova, Neandertal)		
pop1	pop2	*D* (%)	SE (%)	*Z*	*D* (%)	SE (%)	*Z*
Oceania	Africa	5.85	1.08	**5.41**	2.19	1.31	1.67
Oceania	C/S Asia	3.59	0.90	**3.98**	2.70	1.06	**2.53**
Oceania	Middle East	3.71	0.97	**3.84**	2.55	1.12	**2.27**
America	Africa	3.93	1.06	**3.71**	−1.67	1.30	−1.29
SE Asia	Africa	3.40	0.92	**3.67**	−0.53	1.03	−0.52
C/S Asia	Africa	2.64	0.76	**3.47**	−0.25	0.91	−0.28
Oceania	Europe	3.38	0.98	**3.46**	2.30	1.13	**2.04**
Europe	Africa	2.77	0.82	**3.38**	0.07	1.01	0.07
Oceania	NE Asia	3.40	1.03	**3.32**	3.78	1.28	**2.96**
NE Asia	Africa	2.96	0.90	**3.31**	−1.06	1.00	−1.06
Middle East	Africa	2.46	0.77	**3.20**	−0.18	0.97	−0.19
Oceania	SE Asia	2.89	1.06	**2.74**	3.17	1.25	**2.54**
SE Asia	NE Asia	0.55	0.23	**2.40**	0.66	0.30	**2.22**
Oceania	America	2.20	1.27	1.73	4.44	1.45	**3.07**
America	Middle East	1.64	1.03	1.60	−1.62	1.19	−1.36
America	C/S Asia	1.47	0.99	1.49	−1.59	1.11	−1.44
America	NE Asia	1.21	0.87	1.39	−0.76	1.07	−0.71
Europe	Middle East	0.40	0.29	1.36	0.29	0.36	0.80
America	Europe	1.29	1.04	1.24	−1.93	1.18	−1.64
SE Asia	Middle East	1.04	0.88	1.18	−0.38	0.97	−0.39
SE Asia	C/S Asia	0.84	0.78	1.08	−0.31	0.82	−0.38
America	SE Asia	0.68	0.88	0.77	−1.38	1.11	−1.24
SE Asia	Europe	0.67	0.88	0.76	−0.66	0.97	−0.68
C/S Asia	Middle East	0.23	0.31	0.73	−0.08	0.43	−0.19
NE Asia	Middle East	0.57	0.87	0.66	−0.94	0.94	−1.00
NE Asia	C/S Asia	0.36	0.77	0.47	−0.90	0.78	−1.14
Europe	C/S Asia	0.16	0.36	0.46	0.37	0.45	0.82
NE Asia	Europe	0.19	0.86	0.22	−1.23	0.93	−1.32

Significant tests ($|Z| > 2$) are indicated in bold. A significant positive value in the test (Pop1, Pop2, Denisova, chimpanzee) is interpreted as a correlation in allele frequencies between Pop1 and Denisova and/or Pop2 and chimpanzee, whereas a significant negative value indicates the opposite case. A significant positive value in the test (Pop1, Pop2, Denisova, Neandertal) is interpreted as a correlation in allele frequencies between Pop1 and Denisova and/or Pop2 and Neandertal, whereas a significant negative value indicates the opposite case. Population comparisons are presented in rank order of *Z*(Pop1, Pop2, Denisova, chimpanzee) for comparison with Fig. 3 in the main text. SEs where computed with a block jackknife over 114 contiguous blocks with the same number of SNPs.

Table S5. Tests for Denisovan ancestry in East Asia using seven complete genomes

H3	H4	n_{AABB}	n_{ABAB}	n_{ABBA}	*D* (%)	SE (%)	*Z*	*N* blocks
SJK (Korean)	YH (Chinese)	152,493	6,649	6,524	0.95	0.99	0.96	46
JCV (European)	YH (Chinese)	125,557	6,913	6,766	1.07	1.32	0.81	49
NA12891 (CEU)	YH (Chinese)	135,898	7,685	7,545	0.92	1.25	0.73	57
NA12892 (CEU)	YH (Chinese)	135,120	7,644	7,623	0.14	1.20	0.11	57
NA19238 (YRI)	YH (Chinese)	101,782	9,337	8,035	7.49	0.99	**7.58**	67
NA19239 (YRI)	YH (Chinese)	103,415	9,754	8,474	7.02	0.93	**7.52**	71
YH (Chinese)	SJK (Korean)	152,428	6,550	6,669	−0.90	1.18	−0.76	80
JCV (European)	SJK (Korean)	125,829	6,840	6,916	−0.55	1.28	−0.43	82
NA12891 (CEU)	SJK (Korean)	136,386	7,684	7,635	0.32	1.15	0.28	80
NA12892 (CEU)	SJK (Korean)	135,715	7,628	7,855	−1.47	1.16	−1.26	81
NA19238 (YRI)	SJK (Korean)	102,303	9,179	8,228	5.46	1.10	**4.95**	82
NA19239 (YRI)	SJK (Korean)	104,235	9,801	8,649	6.24	1.05	**5.97**	81

The *D* statistic is computed for (Denisova, Neandertal, H3, H4). n_{AABB} is the number of sites where Denisova and Neandertal share allele A and H3 and H4 share allele B. n_{ABBA} is the number of sites where Denisova and H4 share allele A and Neandertal and H3 share allele B. n_{ABAB} is the number of sites where Denisova and H3 share allele A and Neandertal and H4 share allele B. SEs where computed with a block jackknife of windows of 200 informative SNPs.