Evolutionary history and adaptation from high-coverage whole-genome sequences of diverse African hunter-gatherers

(Supplementary Information)

Joseph Lachance[1], Benjamin Vernot[2], Clara C. Elbers[1], Bart Ferwerda[1], Alain Froment[3], Jean-Marie Bodo[4], Godfrey Lema[5], Wenqing Fu[2], Thomas B. Nyambo[5], Timothy R. Rebbeck[6], Kun Zhang[7], Joshua M. Akey[2], Sarah A. Tishkoff[1*]

[1]Departments of Biology and Genetics, University of Pennsylvania, Philadelphia, PA 19104 USA.

[2]Department of Genome Sciences, University of Washington, Seattle, WA 98185 USA.

[3]IRD-MNHN, Musée de l'Homme, 75116 Paris, France

[4]Ministère de la Recherche Scientifique et de l'Innovation, BP1457, Yaoundé, Cameroon.

[5]Department of Biochemistry, Muhimbili University College of Health Sciences, Dar es Salaam, Tanzania.

[6]Perelman School of Medicine Abramson Cancer Center, University of Pennsylvania, Philadelphia, PA 19104 USA.

[7]Department of Bioengineering, Institute for Genomic Medicine and Institute of Engineering in Medicine, University of California at San Diego, San Diego, CA 92093 USA.

*Corresponding author (tishkoff@mail.med.upenn.edu)

Samples for Whole-genome Sequencing, Quality Control, and Identification of Variants

Prior to sample collection, informed consent was obtained from all research participants, and permits were received from the Ministry of Health and National Committee of Ethics in Cameroon and from COSTECH and NIMR in Dar es Salaam, Tanzania. In addition, appropriate IRB approval was obtained from both the University of Maryland and the University of Pennsylvania. Although the term ‘Pygmy’ has historically been pejorative, it has recently been used by indigenous groups themselves as well as activist groups working on their behalf (Ballard, 2006; Leonhardt, 2006; Pelican, 2009). Acknowledging this recent trend and the absence of a better term that encompasses the hunting and gathering peoples from Cameroon, we use the word ‘Pygmy’ to collectively refer to Baka, Bakola, and Bedzan individuals in our study. Hadza samples were collected at sites near Lake Eyasi in the Arusha region and Sandawe samples were collected in the Kondoa district in the Dodoma region of Tanzania. Individuals were chosen to be unrelated based on microsatellite (Tishkoff et al., 2009) and genome-wide SNP (Jarvis et al., 2012) data analyses, including a pi-hat filter of 0.25 using PLINK (Purcell et al., 2007). However, we note that the small population size of the Hadza means that token amounts of relatedness are shared between samples. White blood cells were isolated in the field from whole blood with a salting out procedure modified from (Miller et al., 1988) and DNA was extracted in the lab with a Purgene™ DNA extraction kit (Gentra Systems Inc., Minneapolis, MN). Because DNA was obtained from whole blood, we avoid possible artifacts that can arise from use of cell lines

(Maitra et al., 2005).

Hunter-gatherer genomes were sequenced at >60x coverage (Table 1) using the combinatorial probe-anchor ligation and DNA nanoarray technology of Complete Genomics. The standard Complete Genomics bioinformatics pipeline (Assembly Pipeline version 1.10 and CGA™ Tools 1.4) was used for sequence alignment, read mapping, assembly, and data analysis. This pipeline uses stringent criteria to make variant calls (likelihood ratios >100:1 are required to make homozygous variant calls and likelihood ratios >10,000:1 are required to make heterozygous variant calls), and published error rates are less than 1/100,000bp (Drmanac et al., 2010). Importantly, calls for different individuals were independent (otherwise, our allele frequency distributions would underestimate rare alleles). To assess genotyping accuracy we sequenced two additional genomes as technical replicates (each technical replicate was a duplicate of one of the five sequenced Hadza genomes). Data from our technical replicates revealed low error rates: 26,415 and 28,292 discordant variant calls are found for each pair of technical replicates.

As an additional test of genotyping accuracy we compared calls from whole-genome sequencing and the Illumina1M-duo BeadChip array (of which we had data for 14 of 15 hunter-gatherers). Ignoring A/G and C/T sites (to avoid strand flipping issues in our Illumina1M-duo dataset), calls at total of 743,516 SNPs were compared. For each individual, concordance between platforms was very high (mean=0.999564, range=0.999463 to 0.999635). In practice, only one out of every 2294 variant calls differed between platforms. Median coverage was similar for concordant (48x) and discordant (49x) SNPs, suggesting that errors were not due to poor coverage during whole-genome sequencing. Furthermore, a total of 152 SNPs were classified as highly-discordant (>50%) between genotyping technologies, and 32/152 highly-discordant SNPs were found to be tri-allelic after cross-referencing with dbSNP.

Prior to quality control filters, we observed 15,748,468 variants in hunter-gatherer genomes. We then filtered our data based on “missingness” and departure from Hardy- Weinberg filters. Some genomic regions (such as centromeres and telomeres) are less likely to be successfully sequenced, and we find that sites called successfully in only a subset of individuals are more likely to be discordant (Figure S1). Because of this, we used a “missingness” filter, whereby sites called in <80% of all individuals were excluded from analysis, eliminating 2,311,725 variants (Figure S1). Because genotyping error can yield abnormal proportions of heterozygotes and homozygotes, we also used quality control filter to detect departures from Hardy-Weinberg proportions. We note that even sites under strong selection are expected to pass a departure from Hardy-Weinberg proportion filter (Lachance, 2009a). To avoid artifacts due to population stratification (such as Wahlund effects) we summed Chi-square statistics for each hunter-gatherer population and excluded all sites with Chi-square values ≥13. In practice this involved excluding sites where every individual was heterozygous, and this quality control filter eliminated a further 16,425 variants (Figure S1). We then merged data from pairs of genomes that contained technical replicates. This involved resolving discordant calls, and eliminating an additional 12,801 variants. After all quality control filters, a total of 13,407,517 variants remained.

We note the following additional details: Genomic coordinates used in this paper refer to build37/hg19 of the human genome. For population genetics analyses (including calculation of θ, allele frequencies, and F_{ST}) we treated partially called sites as missing data. We note that the following analyses were restricted to SNPs (as opposed to SNPs and indels): PCA, NJ tree, Neutrality Index calculations, T_{MRCA} scans, archaic introgression, and LSBL scans. In a previous study (Lam et al., 2012), the sequencing technology of Complete Genomics was found to be highly accurate in detecting indels (22 of 23 successfully amplified indels were validated). However, Lam et al. also note that indel detection by Complete Genomics lacks sensitivity, indicating that the number of indels discovered in hunter-gatherer genomes may be an underestimate.

Variants in Different Global Populations and Genomic Locations

As additional genomes are sequenced there are diminishing returns in the number of observed variants. To quantify how these diminishing returns vary by population we analyzed up to five genomes per population. In addition to hunter-gatherer genomes sequenced in this study, we included YRI (NA18501, NA18502, NA18504, NA18505, NA18508), CEU (NA06985, NA06994, NA07357, NA10851, NA12004), and ASN (NA18526, NA18537, NA18555, NA18940, NA18942) genomes from the Complete Genomics public data release. There are $\frac{5!}{(5-k)!k!}$ ways to select *k* different genomes from a set of five genomes, and for each value of *k* we calculated the number of variants for every possible combination of genomes. Mean values of the number of variants observed per *k* genomes are plotted in Figure S2A (with variants absent from dbSNP131 labeled as novel). Because some data points involve single genomes, “missingness” and departure from Hardy-Weinberg filters were not used for Figure 2A-B. Data from multiple genomes can be fit to a power law distribution, and this distribution can in turn be used to predict the number of variants that will be observed as additional genomes are sequenced.

$$v_k = \alpha k^{\beta} \tag{S1}$$

where α and β are scaling parameters, and v_k is the total number of variants observed in *k* diploid genomes. Rearranging terms in Equation S1 allows us to find the number of diploid genomes that need to be sequenced in order to observe a particular number of variants.

$$k = \sqrt[\beta]{\frac{v_k}{\alpha}} \tag{S2}$$

Least squares fitting of observed data yields parameter estimates (Figure S2). Note that the number of novel variants observed is not a linear function of the number of sequenced genomes ($\beta < 1$). We note that different patterns can arise when the number of variants per sequenced genome does not follow a power law distribution (Gravel et al., 2011).

Among hunter-gatherer populations analyzed in this paper, Pygmies contain the most genetic variation, followed by the Sandawe and the Hadza. After controlling for sample size, we find that Pygmy and Yoruba genomes have comparable amounts of variation, and that Hadza and Sandawe genomes contains more variants than non-African (Northern European, Chinese, and Japanese) genomes but fewer variants than West African (Pygmy and Yoruba) genomes (Figure S2).

We also calculated the number of autosomal variants per non-overlapping 1Mb window for each hunter-gatherer population. Variant density varies across the genome for many reasons, including local differences in mutation rate, selective sweeps, and repetitive sequences (as multiple copies can map to the same region of the reference genome). Distributions of the number of variants per Mb are similar for each population (Figure S2), and numbers of variants per window are highly correlated between populations (Figure S2). This observation indicates that differences in the number of variants found in each population (Pygmy > Sandawe > Hadza) reflect a broad genome-wide pattern. Spikes in Figure S2 correspond to the MHC region on chromosome 6, a region near the *CSMD1* gene on chromosome 8 that is known to contain CNVs (Shaikh et al., 2009), and multiple regions on chromosome 16. Although the number of variants per Mb varies across the genome, we find that genomic regions that contain a large number of variants in one population also contain a large number of variants in other populations (Figure S2). This trend is most pronounced over large spatial scales (sliding-window comparisons of the number of variants found in multiple populations yield $R^2 > 0.83$ for 1Mb windows, as opposed to $R^2 < 0.59$ for 10kb windows).

Y Chromosome and mtDNA Lineages

Y chromosome and mtDNA give additional insight into the demographic history of paternal and maternal lineages (Table S2). All three Pygmy, Hadza, and Sandawe hunter-gatherer populations contain Y chromosome haplogroups that are common among other hunter-gatherer populations in East, Central and Southern Africa (B2b and B2b*) and a lineage that is a signature of the Bantu expansion (E1b1a1) (Berniell-Lee et al., 2009). In addition, the Sandawe contain a Y chromosome haplogroup lineage associated with the expansion of Afro-Asiatic speakers (E1b1b1). All five Pygmies sequenced in this paper contain a mtDNA lineage that is common among Central African Pygmies (L1c1a).

Runs of Homozygosity

One notable feature of genomes is that they can contain large runs of homozygosity (ROH). Possible causes of runs of homozygosity include inbreeding, selective sweeps, reduced mutation rates, and population bottlenecks. We calculated the number of runs of homozygosity and the cumulative size (in base pairs) of runs of homozygosity for each hunter-gatherer genome.

One complication is that some genomic regions are poorly sequenced, particularly centromeres. To ensure that poorly sequenced regions were not incorrectly labeled as runs of homozygosity, we required that at least one successful call be made per 10kb. Edges of poorly sequenced regions were treated as endpoints of runs of homozygosity. Another possible complication is genotyping error. This is because false heterozygotes can interrupt runs of homozygosity. To accommodate genotyping error we allowed for one false heterozygote per 50kb. We then identified all runs of heterozygosity above 100kb and 1Mb in size. We also calculated the cumulative number of base pairs contained within runs of homozygosity (cROH) for each size threshold.

Each hunter-gatherer genome contained many runs of homozygosity, with >1% of each genome belonging to large runs of homozygosity (Figure S5). These findings are consistent with previous studies of global populations using SNP data (Hunter-Zinck et al., 2010; Kirin et al., 2010; Nothnagel et al., 2010). Estimates of cROH in African hunter-gatherers from whole-genome sequencing (this paper) are similar to estimates from SNP arrays (Henn et al., 2011). We observe more runs of homozygosity and larger cROH in Hadza genomes compared to Pygmy and Sandawe genomes (Figure S5). This pattern arises regardless of whether ROH are required to be >100kb or >1Mb. A previous study used a theoretical model to suggest that the cause of large ROH in the Hadza was due to a severe population bottleneck (Henn et al., 2011). An additional explanation for the observed patterns is inbreeding among the Hadza. Recent common ancestry among an individual’s parents (inbreeding) can occur often in small populations (Lachance, 2009b), and inbreeding does not have to involve sibling or first cousin mating to result in Mb-sized genomic regions that are identical by descent (Chapman and Thompson, 2003). We note that there is high variance in cROH observed in Hadza genomes (Figure S5), and that three Hadza individuals contain many large ROH, a pattern that is consistent with inbreeding (Pemberton et al., 2012). In addition, a previous study using genealogical data found evidence of inbreeding in 163 of 931 Hadza individuals (Stevens et al., 1977). Note that population bottlenecks and inbreeding can both contribute to observed patterns; they are not mutually exclusive (all five Hadza individuals have greater cROH than Pygmy or Sandawe individuals).

Principal Components Analysis (PCA) and Construction of a Neighbor Joining Tree

Principal component analysis was run on a set of 68 high-coverage genomes sequenced using the same technology. These genomes included the 15 Pygmy, Hadza, and Sandawe hunter-gatherers and 43 unrelated individuals from the Complete Genomics public data release (http://www.completegenomics.com/sequence-data/download-data/). We only used three of the four publicly available Maasai genomes because there is evidence that two are related, picking NA21732 instead of NA21737. Analyzed SNPs were autosomal, fully called in all 68 individuals, and not in LD. The prcomp() command in R was used to generate PCA plots from 50,000 randomly chosen genomic SNPs (Figure S4).

To complement PCA analyses and infer evolutionary history of African hunter-gatherers and other global populations we generated a neighbor joining tree using PHYLIP (Felsenstein, 2005). Analyses were run using 1,260,982 autosomal SNPs obtained from whole-genome sequencing (i.e. SNPs free of ascertainment bias). The chimpanzee genome was used as an outgroup and 61 individuals were chosen for the neighbor joining tree (Table S1, PUR and MXL individuals were excluded due to complex patterns of admixture). In the cladogram shown in Figure 1F, Pygmies are basal (but not monophyletic) to other African and non-African populations sequenced by Complete Genomics. Furthermore, we find that Hadza and Sandawe populations cluster together. Non-African populations are embedded within the part of the tree that contains populations from East Africa (MKK, Maasai from Kenya). We caution that although the neighbor joining tree in Figure 1F is a useful way to show the hierarchal clustering of genomic data, a tree structure cannot fully describe complex evolutionary histories (because of gene flow). Bootstrap values were generated from 1000 replicates, and we found 90.8% support for the split between Pygmies and other sequenced populations, 100% support for the split between Hadza/Sandawe populations and Maasai/non-African populations, and 61.9% support for the split between Hadza and Sandawe populations.

Structural Variation

Structural variants (SVs) were called with Complete Genomics' standard pipeline. To identify overlapping SVs, junctions in the highConfidenceJunctionsBeta files were compared using the *junctiondiff* function in Complete Genomics' CGA™ Tools. A binary matrix was generated for the hunter-gatherer genomes sequenced in this study (15 unique genomes plus 2 technical replicates) and 48 genomes present in the Complete Genomics diversity panel. In this matrix, 1 indicates the presence of a SV junction, and 0 indicates the absence of a SV junction. A neighbor-joining tree was generated from the matrix.

Clustering patterns based on structural variants (translocations, inversions, etc.) are more complex than patterns based on SNPs: individuals in the same hunter-gatherer population do not cluster together with respect to structural variation. The complexity of these patterns may be due to the difficulty in identifying structural variants from short read sequencing technologies (Alkan et al., 2011). Indeed, technical replicates do not cluster together with respect to SV, suggesting that technological improvements will be needed to successfully identify population-level patterns of SV (including improved bioinformatics and sequencing of longer reads).

Comparisons with a South African San genome (KB1)

In a previous study the whole-genome of a San hunter-gatherer from South Africa (KB1) was sequenced using Roche/454 and Illumina technologies (Schuster et al., 2010). After converting coordinates from hg18 (GRCh36) to hg19 (GRCh37), a list of positions where KB1 differs from the human reference genome was downloaded from the GALAXY server (http://main.g2.bx.psu.edu/). We then looked for KB1 variants that were also present in the 15 hunter-gatherer genomes sequenced in our study. Of these shared variants, approximately half (51.76%) were found in all three hunter-gatherer populations analyzed in this paper. Note, however, that many variants shared between hunter-gatherer populations are also present in other global populations (i.e. populations with different subsistence patterns).

Shared ancestry and/or gene-flow between the San and other hunter-gatherer populations can also be estimated via the D-test. This test has been used to identify ancient admixture between human populations and archaic populations, including Neanderthals and Denisovans (Durand et al., 2011; Green et al., 2010; Reich et al., 2010). The D-test uses counts of shared derived alleles to infer relative levels of gene flow and/or common ancestry.

$$D = \frac{d_{ABBA} - d_{BAAB}}{d_{ABBA} + d_{BAAB}} \tag{S3}$$

In Equation S3, d_{ABBA} is the number of shared derived alleles between P1 and P3, and d_{BABA} is the number of shared derived alleles between P2 and P3 (P1, P2, and P3 are genomes from three different populations). A positive D-statistic occurs if there is greater gene flow between P1 and P3, and a negative D-statistic occurs if there is greater gene flow between P2 and P3. Because of the possibility of incomplete lineage sorting, we advise some caution in interpreting the results of D-tests involving the San and other hunter-gatherers. Standard errors were calculated by a block jackknife procedure (Efron, 1981). As per (Green et al., 2010) we used 100 genomic blocks to calculate standard errors and generate Z-scores. We have whole-genome sequences from five individuals per population, which allowed us to compute 25 D-tests per pair of hunter-gatherer populations. Mean values of tests for each population are listed in Table S3. Chimpanzee, orangutan, and rhesus macaque genomes were used as outgroups to infer derived allele states via maximum likelihood. There appears to be slightly more gene flow and/or closer ancestry between the San and Pygmy individuals than between San and Hadza or Sandawe individuals.

Derived Allele Frequency Distributions

We inferred derived and ancestral alleles at each polymorphic SNP using maximum likelihood. Chimpanzee, orangutan, and rhesus macaque genomes were used as outgroups. We used PHYLIP (Felsenstein, 2005) to infer ancestral states via maximum likelihood (settings: ti/tv ratio of 2 with base frequencies calculated from human sequence data). We considered two models, one that does not allow for incomplete lineage sorting (non-ILS) and one that allows for incomplete lineage sorting (ILS). The latter model involved considering multiple trees (all three possible topologies involving variant, reference, and chimpanzee alleles). However, ancestral states were inferred with higher confidence using the non-ILS model, as opposed to the ILS model. Because of this, we opted to use the non-ILS model in this paper. Only ancestral states that were identified with > 95% confidence were retained, and we required that each SNP be fully called in all 15 hunter-gatherers. Derived allele frequency distributions were qualitatively similar when ancestral and derived states were inferred via parsimony, as opposed to maximum likelihood (data not shown).

After inferring derived states we were left with 10.4 million SNPs (many sites in the human genome lack calls in other primates). Variant allele frequencies were then polarized by identifying whether reference alleles are ancestral (in which case derived allele frequencies equal variant allele frequencies) or derived (in which case derived allele frequencies equal reference allele frequencies). Derived allele frequency distributions in Figure 2D only consider polymorphic sites (monomorphic derived or ancestral alleles in each population are ignored). Neutral expectations in

Figure 2D follow from population genetics theory. Assuming an infinite sites model and constant population size, the probability of observing a particular SNP is inversely proportional to the derived allele frequency at that SNP (Lachance, 2010; Sethupathy and Hannenhalli, 2008). *x* is the derived allele frequency and *C* is a normalizing constant in Equation S4.

$$P(x = X) = \frac{C}{x} \tag{S4}$$

Joint allele frequencies were obtained for pairs of hunter-gatherer populations and depicted in 2D histograms (Figure S6). In each panel of Figure S6 data are normalized (i.e. probabilities sum to one), and SNPs were required to be polymorphic in at least one population. Allele frequency distributions for ascertained SNPs were determined by using only SNPs that are found on the Illumina1M-Duo BeadChip. SNPs on the Illumina1M array are enriched for intermediate frequency alleles in multiple populations. Regardless of genotyping platform, the majority of SNPs contain low frequency derived alleles in both populations (Figure S6). However, many SNPs have high frequency derived alleles in both populations. This secondary peak is unexpected, and is best explained by mis-inference of ancestral states due to hyper-mutability of CpG sites (Hernandez et al., 2007) and/or by GC-biased gene conversion (Duret and Galtier, 2009). Because of this, mean derived allele frequencies in Table 2 and Figure S3 use only non-CpG SNPs that are free from biased gene conversion (i.e. only A/T and C/G SNPs are considered).

Comparisons between allele frequency distributions from whole-genome sequencing and the Illumina1M-Duo platform (Figure S6) reveal that the ascertainment bias of SNPs in genotyping arrays cause intermediate frequency alleles to be over-represented (Figure S6), emphasizing the importance of using whole-genome sequence data to obtain accurate estimates of allele frequency spectra.

Population Genetic Statistics for Different Functional Regions of Genomes

Annotations from the UCSC table browser (http://genome.ucsc.edu/cgi-bin/hgTables) were used to label variants (track: RefSeq Genes). Variants were classified as intergenic, 10kb upstream, 5′ UTR, exon, intron, 3′ UTR, or 10kb downstream. Note that it is possible for variants to belong to multiple classifications (including alternatively spliced sites that can be labeled as exonic or intronic). For each type of variant we calculated a number of population genetics statistics (Table 2).

Within-population genetic statistics for each type of variant include θ, mean derived allele frequency (DAF), and Tajima's D. Only autosomal variants that were fully called in all 15 individuals were considered. Watterson's θ measures the proportion of polymorphic sites (Watterson, 1975), where:

$$\theta_{per\ base\_pair} = \frac{S}{T\sum_{i=1}^{n-1}\frac{1}{i}} \tag{S5}$$

In Equation S5, *S* is the number of segregating sites, *T* is the total number of base pairs belonging to a particular sequence class (e.g. the total number of base pairs found in exons), and *n* is the sample size (in terms of haploid genomes). *T* was calculated by multiplying the total number of base pairs belonging to a particular sequence class by 0.953 (the average proportion of each hunter-gatherer genome that was successfully called). Using intergenic estimates of θ (Table 2), mutation rates of 1.1×10^{-8} (Roach et al., 2010) and 2.5×10^{-8} (Nachman and Crowell, 2000), and the relationship $\theta = 4N_e\mu$, we calculated the effective population size for each hunter-gatherer population.

F_{ST} measures the genetic distance between pairs of populations, and we calculated this statistic from allele frequency data. Only fully called autosomal variants were considered for F_{ST} calculations. Note that F_{ST} calculations only consider polymorphic sites.

$$F_{ST} = \frac{Var(p)}{\bar{p}(1-\bar{p})} \tag{S6}$$

In Equation S6 $Var(p)$ denotes to the variance in allele frequencies across populations and $\bar{p}$denotes mean allele frequency. However, small sample size can lead to biases (overestimates) in F_{ST}. Because of this we used the following equation (Akey et al., 2002; Weir and Cockerham, 1984):

$$F_{ST(corrected)} = \frac{MSP - MSG}{MSP + (n_c - 1)MSG)} \tag{S7}$$

In Equation S7 *MSP* denotes the mean squared error between populations, *MSG* denotes the mean squared errors for loci within populations, and n_c is the corrected sample size. This sample size correction can result in negative values of F_{ST}, and we replaced these negative values with zeroes. The majority of segregating sites have low values of Tajima's D, and mean F_{ST} was highest for comparisons involving the Hadza.

Neutrality Index

We conducted McDonald- Kreitman-type analyses for African hunter-gatherers using the chimpanzee (PanTro3) genome as an outgroup (Table S4). These tests compare relative levels of polymorphism and fixed differences between site classes (e.g. neutral intergenic sites, synonymous sites, non-synonymous sites, sites in DNase I footprint regions, and sites in transcription factor (TF) motifs). Neutral intergenic sites are defined here as intergenic sites that are at least 50kb apart from genes, have GERP++ (Davydov et al., 2010) scores < 1, and are not located in DNase I footprint regions or TF motif regions. The neutrality index (NI) was calculated as previously described (Rand and Kann, 1996), and NI significantly >1 signifies purifying selection. Only sites that were successfully called in hunter-gatherer and chimpanzee genomes were considered.

DNase I Footprints

We observed 674,808 variants in DNase I footprints (Rosenbloom et al., 2012), and 149,072 variants in *cis*-regulatory motifs located within footprints. We also observed 37,797 nonsynonymous and 35,747 are synonymous variants. Thus, the number of putatively functional regulatory variants in hunter-gatherer genomes is an order of magnitude more than putatively functional coding variants, findings that are consistent with (Vernot et al., 2012).

The 3p14.3 (*HESX1* containing) Pygmy AIM cluster contains five AIMs that are in DNase I hyper-sensitive sites (DHS) (chr3:57230332, chr3:57230341, chr3:57263461, chr3:57295123, and chr3:57370649). Two of these AIMs are less than 2 kb upstream of *HESX1* and one AIM is located between *HESX1* and *APPL1.* DHS indicate nucleosome-free regions of DNA that are accessible to the endonuclease DNase I. Such regions are thought to be available for binding by regulatory proteins. In addition, the chr3:57263461 AIM SNP overlaps with a DNase I footprint and is found within the first intron of the APPL1 gene. A DNase I footprint is an experimentally predicted protein binding site within a region of otherwise exposed DNA. This footprint overlaps a computationally predicted binding site for the TAL1/GATA1 complex.

Tajima’s D

Allele frequency distributions were used to calculate Tajima’s D. Positive values of Tajima’s D indicate that there is an excess of intermediate frequency alleles and negative values of Tajima’s D indicate that there is an excess of very high or very low frequency alleles. Population expansions result in genome-wide decreases in Tajima’s D and population bottlenecks result in genome-wide increases of Tajima’s D. The effects of natural selection are expected to affect local regions of the genome, with balancing selection yielding higher values of Tajima’s D and purifying or positive selecting yielding lower values of Tajima's D. The equation for Tajima's D is given by:

$$\text{Tajima' s } D = \frac{\hat{k} - S/a_1}{\sqrt{e_1 S + e_2 S(S-1)}} \qquad \text{(S8)}$$

where $\hat{k}$ is the average pairwise difference between sequences and *S* is the number of segregating sites. a_1, e_1, and e_2 are described in (Tajima, 1989).

To calculate $\hat{k}$ from allele frequency distributions we need to obtain the probability that each site is heterozygous. Given a sample of *n* alleles, n_A of which are allele *A* and n_B of which are allele *B*, the probability that randomly selected diploid individuals are heterozygotes is:

$$P(het) = \frac{2n_A n_B}{n(n-1)} \qquad \text{(S9)}$$

Equation S9 reduces to 2*pq* for large values of *n*. At each polymorphic site n_i is the number of variant alleles observed and *n* is the sample size (10 in our case, as only fully called autosomal sites were considered). Let P_i be the proportion of SNPs where n_i variant alleles are observed. Assuming that a large number of sites are sequenced the average number of pairwise differences is:

$$\hat{k} = S\sum_{i=1}^{n-1} P_i \frac{2n_i(1-n_i)}{n(n-1)} \qquad \text{(S10)}$$

Given a set of *S* polymorphic sites and the allele frequency distribution for these sites, the above equations allow Tajima's D to be calculated.

Using non-overlapping 100kb sliding windows, we find that Pygmy and Sandawe populations have lower values of Tajima's D than the Hadza ($p < 0.0001$ using Z-tests). We also find that genomic regions with low values of Tajima's D tend to have reduced heterozygosity, a pattern that is consistent with both background selection and recent selective sweeps. Although Tajima's D and heterozygosity per 100kb window were positively correlated, values of R^2 were low for all three populations (Pygmies: R^2= 0.103, Hadza: R^2= 0.064, Sandawe: R^2= 0.108).

For each population, we list the 20 windows with the lowest values of Tajima's D and the 10 windows with the highest values of Tajima's D for each population (Table S5). Inspection of calls for each individual reveals that many of these outlier 100kb windows contain what have been called "yin-yang" haplotypes, i.e. two haplotypes that differ at a large number of SNP positions (Curtis and Vine, 2010; Zhang et al., 2003). For example, nine copies of one haplotype and one copy of another divergent haplotype can result in many sites with a minor allele frequency of 10% (which yields low values of Tajima's D).

Signals of Purifying Selection in Hunter-gatherer and Other Global Populations

We tested whether genic signatures of purifying selection vary by subsistence pattern or geographic location by comparing allele frequency distributions, predicted phenotypic effects, and the proportion of nonsynonymous variants. Three groups of populations were considered: African hunter-gatherers (Pygmy, Hadza, and

Sandawe), African agriculturalists and pastoralists (YRI, LWK, MKK), and non-African populations (CEU, CHB, and JPT). To control for sample size differences we analyzed four diploid genomes per population (the minimum number of genomes per population in the Complete Genomics public data release). Selection is expected to shift the derived allele frequency (DAF) distribution in genes relative to noncoding regions of the genome (with the exception of regulatory sequences). Using whole-genome data, we calculated the mean frequency of derived alleles for every intergenic and exon SNP in each population. Only fully called autosomal sites were retained, and we accounted for CpG hyper-mutability and GC-based gene conversion by ignoring SNPs in CpG dinucleotides and only considering A/T or C/G SNPs. To control for demographic effects, such as population bottlenecks, we compared the ratio of mean DAF in exons to intergenic regions for each population. The effects of amino acid changes in each population were assessed by calculating the proportion of sites classified as “benign”, “possibly damaging”, and “probably damaging” by PolyPhen-2 (Adzhubei et al., 2010). For consistency with a previous study (Lohmueller et al., 2008), Polyphen-2 analysis was restricted to SNPs where variant alleles are derived alleles. In addition, Polyphen-2 analysis was restricted to missense SNPs present in dbSNP build131 (i.e. SNPs with PolyPhen-2 data in (Adzhubei et al., 2010)). For each population, we also obtained the number of nonsynonymous and synonymous variants per genome. We then calculated the ratio of nonsynonymous to synonymous variants. Under a null hypothesis of equal selective constraint in each population this ratio should not vary across populations. Tests of selection in our paper were similar to those of (Lohmueller et al., 2008) which looked for differences between African-Americans and European-Americans. Although signals of selection are broadly similar for all nine populations (Figure S3), statistically significant differences still occur. To test whether signals of selection differ for different types of populations we used one-way ANOVA and Tukey's HSD tests. However, because we only have three replicates for each type of population, it is unknown whether assumptions of normality hold. Mean derived allele frequencies were not significantly different between groups of populations ($p > 0.4$, Tukey's HSD). However, we caution that small sample sizes limit our ability to detect rare derived alleles. Although Polyphen-2 data were broadly similar for each group of populations, we observed significantly higher proportions of "probably damaging alleles" in non-African populations ($p < 0.01$, Tukey's HSD). Similarly, the proportions of nonsynonymous SNPs were higher in non-African populations ($p < 10^{-4}$, Tukey's HSD).

T_{MRCA} Calculations

We estimated the Time to Most Recent Common Ancestor (T_{MRCA}) for a set of samples for 50kb sliding windows (20kb step) across all autosomes. We use the method of (Thomson et al., 2000), which computes the average T_{MRCA} in nucleotide substitutions for a set of sequences. Using Equation 1 of (Hudson, 2007) we calculated the T_{MRCA} for each 50kb window. This value was then converted to an estimate of T_{MRCA} in years by computing the divergence between chimpanzee and human for this region (*D*), and setting the molecular clock to 12My/*D* (i.e. we normalize by window-specific mutation rates assuming that humans and chimpanzees split 6Mya).

Human/chimpanzee alignments were downloaded from the UCSC Genome Browser, and the more conservative syntenicNet alignments were used (reference versions GRCh37 and panTro2, http://hgdownload.cse.ucsc.edu/goldenPath/hg19/vsPanTro2/syntenicNet/). For each autosomal variant identified using whole-genome sequencing, ancestral and derived states were calculated using a maximum likelihood method. We used DNAmL (Felsenstein, 2005) to evaluate each site, using a single base alignment consisting of the human reference base, the alternative allele found in our dateset, chimpanzee, orangutan, and macaque bases. DNAmL can handle missing data, and missing bases were marked as such. To allow for the possibility of incomplete lineage sorting between chimpanzee and human, three trees were considered for each site, and the highest likelihood tree was selected. Only states with probability > 0.95 were selected. If the ancestral state was unable to be inferred due to alignment difficulties or low probability, then the most parsimonious state was selected based on the 68 unrelated hunter-gatherer and Complete Genomics genomes (i.e. the minor allele was assumed to be the derived allele).

Archaic Introgression

Only fully called autosomal sites were considered for introgression analyses. In addition, CpG sites and regions identified by RepeatMasker were removed from all S*, T_{MRCA} and STRUCTURE analyses (1.2Gb of sequenced retained after filtering). We identified CpG sites by analyzing the human reference genome (hg19), genomes sequenced by Complete Genomics (15 hunter-gatherer genomes and the public data release), and three additional primate genomes (chimpanzee, orangutan, or macaque). To additionally account for ancestral CpGs, we counted any position that is C or G in one genome, and adjacent to a G or C (respectively) in another genome. This CpG metric is less conservative than the metric used for placental mammals, but more conservative than the metric used for primates in (McVicker et al., 2009).

To test the effects of various demographic parameters on our method, we simulated archaic introgression into either European or African populations. As a starting point, we used demographic parameters from (Plagnol and Wall, 2006), which models Europeans and Yorubans. In addition, we also used the following base parameters: a fixed mutation rate of 1.1×10^{-8}, a fixed recombination rate of 1×10^{-8}, 700kya split time between archaic and modern human populations, and 25kya time of introgression. Simulations were then run for a range of parameter values (Figure S7A, varying a single parameter and keeping other parameters at the base value).

S* was computed using a dynamic programming algorithm, with a running time of $O(n \cdot s^2)$, where *n* is the number of individuals and *s* is the number of variants in a region. For each parameter set we simulated 50kb windows over a range of archaic-to-modern-human migration parameters (0-2.4%, corresponding to 0-4% introgressed sequence per individual). We then identified the top 0.5% of each simulation, as ranked by S*. These are analogous to the 350 top candidate regions used for several of our introgression analyses. To show that these regions are likely to contain introgressed sequence, we determined the false discovery rate (FDR) for each parameter setting (Figure S7A). For introgression levels above 1%, the FDR for most parameters is close to zero. Higher FDR is seen for very low recombination rates, some hotspot models with heterogeneous recombination rates, and for a recent time of split with the archaic population. We also calculated the distribution of normalized S* for simulated datasets and populations with whole-genome sequence data. In simulated data, the extent of positive skew in the distribution of normalized S* was correlated with the extent of archaic introgression (Figure S7D, simulated data with heterogeneous recombination rate: 80% 0.2×10^{-8}, 20% 4.2×10^{-8}). We see a similar positive skew in each of the studied populations, and patterns were similar for each population (Figure S7E), consistent with broadly similar levels of archaic introgression. However, several demographic parameters can affect the distribution of S*. For this reason, we further investigated the characteristics of putatively introgressed regions.

In addition to determining the false discovery rate (FDR) for top candidate regions, we compared top candidate regions for each population to the draft Neanderthal sequence. Due to the low coverage of the draft Neanderthal sequence (Green et al., 2010), we used a comparison method that is less sensitive to errors in the Neanderthal sequence than direct sequence comparison. Specifically, for each region we performed a 2x2 Chi-square test for enrichment of variants matching the Neanderthal sequence that are unique to the target population, in comparison with variants that are not unique to the target population (i.e., present in the reference population). Variants are only considered if they are covered by at least two Neanderthal reads. Top candidate regions in non-African populations show a clear excess of regions with significant enrichment of Neanderthal variants (Figure 3B); top candidate regions from African populations do not show this enrichment (Figure 3A), demonstrating that our top candidate regions are enriched for archaic introgression.

To increase the chance that analyzed genomic regions include entire introgressed haplotypes, we identified a subset of the 350 top candidate regions in which all introgressed variants for a given 50kb region are found in a single individual. This restricted top-candidate dataset (usually about 50% of the top 350 regions) was used for T_{MRCA} and STRUCTURE analyses of introgressed regions. Note that the T_{MRCA} values given in Figure 2A, Figure 2D, and Figure S7C are calculated on single individuals from this subset (to better estimate the time of divergence of the two haplotypes contained in a single individual), whereas T_{MRCA} estimates in the **T_{MRCA} estimates** section of the main text are calculated on entire populations.

To identify population substructure in putatively introgressed regions, we performed a model based clustering analysis using STRUCTURE (Pritchard et al., 2000). Putatively introgressed regions should primarily consist of a single individual containing one introgressed haplotype and one modern human haplotype, with the remaining individuals in the population containing entirely modern human haplotypes. In this situation, the differences between archaic and modern sequence should be more pronounced than the differences between Pygmy, Sandawe and Hadza. To test this hypothesis, we analyzed the top putative introgressed regions from each hunter-gatherer population, as well as a similar number of random sequences from each hunter-gatherer population using STRUCTURE (Pritchard et al., 2000). Because introgressed sequences are likely to be at low frequency, no single individual will contain more than a small amount of introgressed sequence, and STRUCTURE is not well suited to identifying subpopulations that have no extant, or representative, individuals. To combat this problem, we created a "virtual genome" for each hunter-gatherer population, where each region is composed of genotypes from a single individual identified as containing the putatively introgressed haplotype (Figure 2). If a region is a candidate introgressed region in multiple populations, each of those populations’ virtual genome contains the introgressed region. If a region is not a candidate in a given population, the virtual genome for that region and population is set to missing data. These virtual genomes are expected to consist of roughly 50% archaic and 50% modern human haplotypes.

We then performed a STRUCTURE analysis using standard settings, with a burn-in of 100k and run length of 100k. To reduce run time and the affect of LD, we performed each analysis on a random 10% of the selected sites. For each parameter setting, we computed three runs each of K=1-4, and selected the highest log likelihood of the three runs for each K. For random regions (the negative control), K=3 and K=4 have the highest log-likelihood, but all settings of K produce results where each virtual genome is composed almost entirely of a single population, i.e. there is no evidence of a substantial archaic component in these regions. For the

350 top candidate regions, K = 2 has the highest log-likelihood. However, all settings of K produce results where each virtual genome is composed of a ~50/50 mixture of two populations, supporting the hypothesis that these regions are significantly enriched for introgressed archaic sequence. As noted above, rare introgressed haplotypes are expected to be heterozygous, fitting with the observed ~50/50 mixture.

Simulations suggest that the relative time of split between archaic and modern human populations can be recovered via a T_{MRCA} analysis. Specifically, we varied the time of split from 300kya to 1000kya, simulated migration levels from 0.01 to 0.024, and selected the top 50 regions (0.5%) from each simulation. Simulated archaic-modern human split time vs. recovered T_{MRCA} is given in Figure 3C. Empty boxes show simulated introgression events involving Europeans and grey boxes show simulated introgression events involving Yorubans.

To test whether putatively introgressed sequences are enriched in coding regions, we compared the distribution of top candidate regions for each population with the distribution of coding sequences, using the hypergeometric distribution. To minimize the effects of overlapping windows, we used every third 50kb window. Gene definitions were obtained from the UCSC Table Browser, RefSeq Genes track, refFlat table. Exons were extended by 2 bp, and overlapping exons were merged using BEDOPS (Neph et al., 2012). Top candidate regions were not enriched for coding sequence compared to the rest of the genome ($p > 0.05$). Rather, for multiple populations we found that top candidate regions were significantly depleted for coding sequence ($p < 0.01$ for Hadza, Yoruban, CEPH, and Tuscani; $p < 0.05$ for

Massai, Luhya, Chinese and Japanese). Top candidate regions in Pygmy, Sandawe and Gujarati populations were neither significantly enriched nor depleted for coding sequence, although in all three populations there were fewer overlaps between top candidate regions and coding sequence than expected by chance.

To determine if putatively introgressed sequences are clustered in the genome, we performed permutation tests to obtain the distribution of the expected number of 50kb introgressed windows within a 2Mb region. To minimize the effects of overlapping windows, we used every third 50kb window. Although the actual and expected distributions for each hunter-gatherer population are not significantly different (Wilcoxon rank-sum test, $p > 0.05$), we did see a few regions in each population with more top candidates than expected. In all three hunter gatherer populations, the region with the most top candidates is at chr8:3Mb-5Mb. This could be due to shared ancestral introgression, an increased proclivity for introgression involving this region, or enrichment of false positives due to an excess of CNVs in that region (Shaikh et al, 2009), which could lead to sequencing errors. However, several regions in this window from each non-African population are significantly enriched for Neanderthal variants. These enrichments are unlikely to be due to errors in the Neanderthal sequence caused by the complex structure of this window, as none of the hunter-gatherer regions in this window are enriched for Neanderthal variants.

Overlap between putatively introgressed regions was found for each pair of hunter-gatherer populations, and T_{MRCA} was estimated for introgressed regions found in a single population or shared between two populations (FigureS7). We also examined whether putatively introgressed regions are found in regions of high or low recombination (Figure S7B). While there is a slight shift towards lower recombination rates for top candidate regions, the distributions overlap to a large extent, suggesting that low recombination rates are not a major feature of top candidate regions (recombination rates from (Kong et al., 2002)).

Genomic Regions Enriched for LSBL Outliers

We used locus-specific branch lengths (LSBL) to identify 100kb windows that are highly divergent between African hunter-gatherers and populations with agricultural or pastoral subsistence patterns. Locus-specific branch lengths at each polymorphic site were calculated using genetic distances between pairs of populations (F_{ST}):

$$LSBL_{Pygmy} = \frac{F_{ST(Pygmy,YRI)} + F_{ST(Pygmy,MKK)} - F_{ST(YRI,MKK)}}{2} \quad \text{(S11)}$$

$$LSBL_{Hadza} = \frac{F_{ST(Hadza,YRI)} + F_{ST(Hadza,MKK)} - F_{ST(YRI,MKK)}}{2} \quad \text{(S12)}$$

$$LSBL_{Sandawe} = \frac{F_{ST(Sandawe,YRI)} + F_{ST(Sandawe,MKK)} - F_{ST(YRI,MKK)}}{2} \quad \text{(S13)}$$

Locus-specific branch lengths are largest when hunter-gatherer populations have allele frequencies that are very different from agricultural and pastoral populations. The reason that LSBL were calculated instead of simply using F_{ST}, is that high values of F_{ST} do not indicate which of the two populations has diverged the most. To identify which population is divergent it is necessary to triangulate using a third population (Shriver et al., 2004). African population trios involved one hunter-gather population (Pygmy, Hadza, or Sandawe), one agricultural population (YRI), and one pastoral population (MKK). Yoruba and Maasai sequence data come from the public data release from Complete Genomics. Only polymorphic autosomal sites were considered. We also required that sites be fully called in all populations. It is possible for an allele to be present in one population and absent in two populations. This causes complications with respect to Equations S11-S13, as F_{ST} is undefined if a site is monomorphic in a pair of populations. Because of this, pairwise genetic distances were set equal to zero if an allele was absent from a pair of population.

Data from single sites is noisy when sample sizes are small. Because of this, we divided the genome into 100kb windows and looked for regions that contain many sites with large LSBL. Over 26,798 of these windows contain at least one polymorphism. Sites were classified as LSBL outliers if their locus-specific branch lengths were among the top 1% for each population. LSBL outlier cutoffs in each population were: 0.261 (Pygmy), 0.373 (Hadza), and 0.256 (Sandawe). For each genomic 100kb window we calculated how many variants were observed and how many of these variants were LSBL outliers. Many 100kb windows do not contain any outliers and some windows contain >100 outliers. Statistical tests (Chi-square and Z-tests) were used to identify 100kb windows that were most enriched for LSBL outliers. However, the distribution of LSBL variants per 100kb window doesn’t follow any known distribution. After ranking windows by Chi-square scores, we list the top 25 divergent windows for each population in Table S5. Subsequent pathway analyses focused on the top 1% (268) autosomal 100kb windows. In addition, Table S5 contains curated lists of functionally interesting genes in the top windows for each population.

Clusters of Ancestry Informative Markers (AIMs)

Variant alleles were flagged as ancestry informative markers (AIMs) if they were absent from dbSNP 131, found in a single hunter-gatherer sample of 10 genomes, and had a frequency >50%. Only autosomal alleles were considered and we required that each site be fully called in all 15 hunter-gatherers.

Ancestry informative markers appear to cluster in genomes. We further examined this pattern by defining a cluster as a set of at least 10 AIMs, each of which is within 25kb of another AIM. Multiple clusters were observed in each population: 25 Pygmy AIM clusters (including 383 of 1,283 Pygmy AIMs), 281 Hadza AIM clusters (including 6,977 of 12,546 Hadza AIMs), and 5 Sandawe AIM clusters (including 53 of 173 Sandawe AIMs). We used Monte Carlo computer simulations to test whether this clustering was greater than expected by chance. Each simulation run involved taking the observed number of high frequency private alleles, randomly scattering them across the genome, and looking to see how many clusters were observed. Simulations were run 10,000 times for each population. In each case, p were <0.0001 (simulations always yielded fewer clusters than what was empirically observed). Genes found in or near AIM clusters are listed for each population (Tables S5).

Pathway Analyses

Genes identified from outlier approaches (such as LSBL scans) may share biological and/or functional characteristics and we performed pathway analyses to identify these shared characteristics. However, pathway analysis results should be treated with some caution (Elbers et al., 2009; Jia et al., 2011; Wang et al., 2010). This is because we lack a full understanding of the underlying biology of many processes, and many genes do not have pathway annotations. In addition, because genomic scans of selection can identify genes that evolved due to different causes (e.g. selection for immune function or dietary adaptation), statistical power to detect individual pathways can be reduced. Pathway analysis requires a minimum number of genes to be effective (Elbers et al., 2009). This means that genome scans that only identify a small number of genes as outliers are unlikely to be enriched for biological pathways. Despite these caveats, pathway analysis offers a way to make sense out of long lists of outlier loci.

In addition, there is evidence that regulatory variants are evolutionary important (Jeong et al., 2008; Wray, 2007) and long range regulation has been observed for eQTLs that are over 100kb distant from genes (Degner et al., 2012). Because of this, we opted to include 200kb windows flanking each side of highly-divergent (LSBL) genomic regions for our pathway analyses.

Examining LSBL data, we analyzed the 268 (top 1%) most-divergent windows for each population using DAVID 6.7 (Huang da et al., 2009) at default settings (minimum number of genes per term=2, maximum EASE score=0.1). DAVID calculates statistical significance with a modified Fisher's Exact Test, generating an enrichment p (EASE score). Through DAVID it is possible to analyze KEGG and PANTHER databases (Huang da et al., 2009; Kanehisa et al., 2012; Thomas et al., 2003).

KEGG analyses of LSBL results show that both Pygmy-divergent and Hadza-divergent regions are significantly enriched for genes involved in olfactory transduction (Table S6). Additional analyses point to various overrepresented metabolism pathways in Pygmy-divergent regions, and an overrepresented 'taste transduction' pathway in Sandawe-divergent regions (Table S6). Furthermore, many immune related pathways were found to be overrepresented in Hadza-divergent and Sandawe-divergent regions. However, these signals were mainly driven by the HLA-region at 6p21. This region encodes protein of classical HLA class I and II genes in the major histocompatibility complex (MHC) and is essential in immune recognition. This region is highly polymorphic and its LD extends across multiple HLA and non-HLA genes. Genomic regions that contain functionally related genes can bias pathways analyses. Because of this, pathway analyses for Hadza and Sandawe populations were done including and excluding the HLA region (Table S6). In this study we defined the HLA region as chromosome 6: 20,000,00-40,000,000 (GRCh37/hg19). PANTHER analyses of highly-divergent LSBL windows did not reveal statistically significant enrichment after Benjamini-Hochberg corrections ($p > 0.05$, Table S6).

For pathway analysis of AIM data we identified all genes within 50kb of a population-specific AIM. Once again, DAVID 6.7 was used to test whether there was significant enrichment for genes in KEGG or PANTHER pathways (Table S6).

We also tested whether genes identified by LSBL scans of Pygmy genomes are enriched for genes associated with height (as identified largely European GWAS). 318 Height-associated genes were found using the catalog of published genome-wide association studies (http://www.genome.gov/gwastudies/, access date: June 1, 2012), and a 2x2 Chi-square test of independence was used to determine statistical significance (assuming a total of 23,000 genes in the human genome). Considering only genes present in the top 268 Pygmy-divergent windows (the top 1%) we found 6/318 height-associated genes ($p = 0.888$). When the top 268 Pygmy-divergent windows plus 200kb flanking regions were considered we found 11/318 height-associated genes ($p = 0.077$).

We also use a Chi-square test of independence to examine whether genes identified in LSBL scans of Pygmy genomes are enriched for genes associated with pituitary function (either genes involved in early pituitary development or genes expressed in the pituitary). For this analysis we included 22 genes listed in (Lee and Lavin, 2009) plus *FSHR*, *LHCGR*, *TRH*, *GH2*, and *GHR.* Considering only genes present in the top 268 Pygmy-divergent windows we found 3/27 genes involved in pituitary function ($p < 0.0001$). When the top 268 Pygmy-divergent windows plus

200kb flanking regions were considered we found 4/27 genes involved in pituitary function (p = 0.0082).

Association Tests Between Pygmy AIMs and Height

As the chr3:45Mb-60Mb region has previously been associated with Pygmy height (Jarvis et al., 2012) and the *HESX1* and *POU1F1* genes which play a role in pituitary development are plausible candidates to play a role in short height, we genotyped a larger sample of individuals at 15 Pygmy AIM SNPs encompassing AIM clusters at 3p21.13, 3p14.3 and 3p11.2 (Figure 4). Genotyped samples included 95 Pygmy and Bantu samples analyzed in (Jarvis et al., 2012) and 10 Pygmy samples from the Coriell Institutute for Medical Research (five Biaka and five Mbuti samples, see Table S7).

Samples were genotyped using TaqMan® assays (Applied Biosystems), and two replicates of each assay were run. When data was missing in both replicates or calls were discordant between replicates we treated data as missing. Out of 1425 (15*95) total genotyped SNPs, 61 had missing data (37 of which involved the centromeric SNP located at chr3:87681226). Allele frequencies of all 15 AIM SNPs are higher in Pygmy populations than neighboring Bantu populations, including 10/15 SNPs which were absent from Bantu populations (Table S7).

Height and ancestry data were also available for 94 of 95 genotyped individuals. Association between AIM SNPs and height was determined using EMMAX (Efficient Mixed-Model Associated eXpedited), a mixed-model linear regression approach that corrects for both relatedness within populations and structure between them via a pair-wise matrix of genetic relationships among individuals (Kang et al., 2010). We treated ancestry as a covariate (using an identity by state matrix generated from Illumina1M-duo genotyping). Using the same set of individuals as our study, Jarvis et. al. showed that EMMAX is adequately able to account for population structure in Pygmy and Bantu populations (see Figure S7 of (Jarvis et al., 2012)). Dominance of alleles was assumed to be additive. Association between Pygmy AIM SNPs and height were calculated for males, females, and both sexes pooled together (sex as a covariate). We note that sample sizes are smaller for females (n=39) than males (n=55) and that statistical power to detect associations is a function of allele frequency (i.e. power is low for low frequency alleles).

Additional Candidate Loci Related to Height and Pituitary Function

In addition to the 3p14.3 (*HESX1*) and 3p11.2 (*POU1F1*) AIM clusters, we identified other interesting candidate loci that may play a role in short height and pituitary function in Pygmies. For example, one of the top Pygmy-LSBL hits (overlapping *WBSCR27* and *WBSCR28*) and a Pygmy AIM cluster (overlapping *MLXIPL*) are found at 7q11.23, a region associated with Williams Syndrome. Although the Pygmy phenotype differs from that of Williams Syndrome, it is notable that individuals with Williams Syndrome have an abbreviated growth spurt at puberty (Partsch et al., 1999). Also, the largest Pygmy AIM cluster (3p14.3) contains *APPL1* which encodes a protein that directly interacts with the intracellular region of adiponectin receptors (Deepa and Dong, 2009), and adiponectin is known to regulate pituitary function (Rodriguez-Pacheco et al., 2007). In addition, the fourth strongest Pygmy-LSBL hit overlaps the *TRHR* locus, consistent with a prior study of SNP variation (Jarvis et al., 2012). This locus codes for thyrotropin releasing hormone receptor, is expressed in the anterior pituitary and promotes the release of thyroid stimulating hormone from the anterior pituitary. Thyroid function plays a critical role in linear bone growth, sexual maturation, thermo-regulation, and immune function (Kamath et al., 2009). Prior studies indicate that Eastern (Dormitzer et al., 1989) and Western Pygmies (personal communication, B. Hewlett and L. Cordes), who live in a low iodine environment but have low levels of Goiter compared to neighboring Bantu populations, may have a biological adaptation influencing thyroid function. Additionally, the *FSHR* locus is among the top 1% most divergent Pygmy 100kb windows, as identified by LSBL outliers. *FSHR* encodes the follicle-stimulating hormone receptor, and is critical for gonad development.

References

Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. Nat Methods *7*, 248-249.

Akey, J.M., Zhang, G., Zhang, K., Jin, L., and Shriver, M.D. (2002). Interrogating a high-density SNP map for signatures of natural selection. Genome research *12*, 1805-1814.

Alkan, C., Coe, B.P., and Eichler, E.E. (2011). Genome structural variation discovery and genotyping. Nature reviews Genetics *12*, 363-376.

Ballard, C. (2006). Strange Alliance: Pygmies in the Colonial Imaginary. World Archaeology *38*, 133-151.

Berniell-Lee, G., Calafell, F., Bosch, E., Heyer, E., Sica, L., Mouguiama-Daouda, P., van der Veen, L., Hombert, J.M., Quintana-Murci, L., and Comas, D. (2009). Genetic and demographic implications of the Bantu expansion: insights from human paternal lineages. Molecular biology and evolution *26*, 1581-1589.

Chapman, N.H., and Thompson, E.A. (2003). A model for the length of tracts of identity by descent in finite random mating populations. Theoretical population biology *64*, 141-150.

Curtis, D., and Vine, A.E. (2010). Yin yang haplotypes revisited - long, disparate haplotypes observed in European populations in regions of increased homozygosity. Hum Hered *69*, 184-192.

Davydov, E.V., Goode, D.L., Sirota, M., Cooper, G.M., Sidow, A., and Batzoglou, S. (2010). Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput Biol *6*, e1001025.

Deepa, S.S., and Dong, L.Q. (2009). APPL1: role in adiponectin signaling and beyond. Am J Physiol Endocrinol Metab *296*, E22-36.

Degner, J.F., Pai, A.A., Pique-Regi, R., Veyrieras, J.B., Gaffney, D.J., Pickrell, J.K., De Leon, S., Michelini, K., Lewellen, N., Crawford, G.E.*, et al.* (2012). DNase I sensitivity QTLs are a major determinant of human expression variation. Nature *482*, 390-394.

Dormitzer, P.R., Ellison, P.T., and Bode, H.H. (1989). Anomalously low endemic goiter prevalence among Efe pygmies. Am J Phys Anthropol *78*, 527-531.

Drmanac, R., Sparks, A.B., Callow, M.J., Halpern, A.L., Burns, N.L., Kermani, B.G., Carnevali, P., Nazarenko, I., Nilsen, G.B., Yeung, G.*, et al.* (2010). Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. Science *327*, 78-81.

Durand, E.Y., Patterson, N., Reich, D., and Slatkin, M. (2011). Testing for ancient admixture between closely related populations. Molecular biology and evolution *28*, 2239-2252.

Duret, L., and Galtier, N. (2009). Biased gene conversion and the evolution of mammalian genomic landscapes. Annu Rev Genomics Hum Genet *10*, 285-311.

Efron, B. (1981). Nonparametric estimates of standard error: the jackknife, the bootstrap and other methods. Biometrika *68*, 589-599.

Elbers, C.C., van Eijk, K.R., Franke, L., Mulder, F., van der Schouw, Y.T., Wijmenga, C., and Onland-Moret, N.C. (2009). Using genome-wide pathway analysis to unravel the etiology of complex diseases. Genet Epidemiol *33*, 419-431.

Felsenstein, J. (2005). PHYLIP (Phylogeny Inference Package). Cladistics *5*, 164-166.

Gravel, S., Henn, B.M., Gutenkunst, R.N., Indap, A.R., Marth, G.T., Clark, A.G., Yu, F., Gibbs, R.A., and Bustamante, C.D. (2011). Demographic history and rare allele sharing among human populations. Proc Natl Acad Sci U S A *108*, 11983-11988.

Green, R.E., Krause, J., Briggs, A.W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M.H.*, et al.* (2010). A draft sequence of the Neandertal genome. Science *328*, 710-722.

Henn, B.M., Gignoux, C.R., Jobin, M., Granka, J.M., Macpherson, J.M., Kidd, J.M., Rodriguez-Botigue, L., Ramachandran, S., Hon, L., Brisbin, A.*, et al.* (2011). Hunter-gatherer genomic diversity suggests a southern African origin for modern humans. Proc Natl Acad Sci U S A *108*, 5154-5162.

Hernandez, R.D., Williamson, S.H., and Bustamante, C.D. (2007). Context dependence, ancestral misidentification, and spurious signatures of natural selection. Molecular biology and evolution *24*, 1792-1800.

Huang da, W., Sherman, B.T., and Lempicki, R.A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc *4*, 44-57.

Hudson, R.R. (2007). The variance of coalescent time estimates from DNA sequences. J Mol Evol *64*, 702-705.

Hunter-Zinck, H., Musharoff, S., Salit, J., Al-Ali, K.A., Chouchane, L., Gohar, A., Matthews, R., Butler, M.W., Fuller, J., Hackett, N.R.*, et al.* (2010). Population genetic structure of the people of Qatar. Am J Hum Genet *87*, 17-25.

Jarvis, J.P., Scheinfeldt, L.B., Soi, S., Lambert, C., Omberg, L., Ferwerda, B., Froment, A., Bodo, J.M., Beggs, W., Hoffman, G.*, et al.* (2012). Patterns of Ancestry, Signatures of Natural Selection, and Genetic Association with Stature in Western African Pygmies. PLoS genetics *8*, e1002641.

Jeong, S., Rebeiz, M., Andolfatto, P., Werner, T., True, J., and Carroll, S.B. (2008). The evolution of gene regulation underlies a morphological difference between two Drosophila sister species. Cell *132*, 783-793.

Jia, P., Wang, L., Meltzer, H.Y., and Zhao, Z. (2011). Pathway-based analysis of GWAS datasets: effective but caution required. Int J Neuropsychopharmacol *14*, 567-572.

Kamath, J., Yarbrough, G.G., Prange, A.J., Jr., and Winokur, A. (2009). The thyrotropin-releasing hormone (TRH)-immune system homeostatic hypothesis. Pharmacol Ther *121*, 20-28.

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., and Tanabe, M. (2012). KEGG for integration and interpretation of large-scale molecular data sets. Nucleic acids research *40*, D109-114.

Kang, H.M., Sul, J.H., Service, S.K., Zaitlen, N.A., Kong, S.Y., Freimer, N.B., Sabatti, C., and Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. Nat Genet *42*, 348-354.

Kirin, M., McQuillan, R., Franklin, C.S., Campbell, H., McKeigue, P.M., and Wilson, J.F. (2010). Genomic runs of homozygosity record population history and consanguinity. PLoS One *5*, e13996.

Kong, A., Gudbjartsson, D.F., Sainz, J., Jonsdottir, G.M., Gudjonsson, S.A., Richardsson, B., Sigurdardottir, S., Barnard, J., Hallbeck, B., Masson, G.*, et al.* (2002). A high-resolution recombination map of the human genome. Nature genetics *31*, 241-247.

Lachance, J. (2009a). Detecting selection-induced departures from Hardy-Weinberg proportions. Genet Sel Evol *41*, 15.

Lachance, J. (2009b). Inbreeding, pedigree size, and the most recent common ancestor of humanity. J Theor Biol *261*, 238-247.

Lachance, J. (2010). Disease-associated alleles in genome-wide association studies are enriched for derived low frequency alleles relative to HapMap and neutral expectations. BMC Med Genomics *3*, 57.

Lam, H.Y., Clark, M.J., Chen, R., Natsoulis, G., O'Huallachain, M., Dewey, F.E., Habegger, L., Ashley, E.A., Gerstein, M.B., Butte, A.J.*, et al.* (2012). Performance comparison of whole-genome sequencing platforms. Nat Biotechnol *30*, 78-82.

Lee, P.D.K., and Lavin, N. (2009). Pituitary Disorders and Tall Stature in Children. In Manual of Endocrinology and Metabolism, N. Lavin, ed. (Baltimore, MD, Lippincott Williams & Wilkins), pp. 76-78.

Leonhardt, A. (2006). Baka and the Magic of the State: Between Autochthony and Citizenship. African Studies Review *49*, 69-94.

Lohmueller, K.E., Indap, A.R., Schmidt, S., Boyko, A.R., Hernandez, R.D., Hubisz, M.J., Sninsky, J.J., White, T.J., Sunyaev, S.R., Nielsen, R.*, et al.* (2008). Proportionally more deleterious genetic variation in European than in African populations. Nature *451*, 994-997.

Maitra, A., Arking, D.E., Shivapurkar, N., Ikeda, M., Stastny, V., Kassauei, K., Sui, G., Cutler, D.J., Liu, Y., Brimble, S.N.*, et al.* (2005). Genomic alterations in cultured human embryonic stem cells. Nature genetics *37*, 1099-1103.

McVicker, G., Gordon, D., Davis, C., and Green, P. (2009). Widespread genomic signatures of natural selection in hominid evolution. PLoS genetics *5*, e1000471.

Miller, S.A., Dykes, D.D., and Polesky, H.F. (1988). A simple salting out procedure for extracting DNA from human nucleated cells. Nucleic acids research *16*, 1215.

Nachman, M.W., and Crowell, S.L. (2000). Estimate of the mutation rate per nucleotide in humans. Genetics *156*, 297-304.

Neph, S., Kuehn, M.S., Reynolds, A.P., Haugen, E., Thurman, R.E., Johnson, A.K., Rynes, E., Maurano, M.T., Vierstra, J., Thomas, S.*, et al.* (2012). BEDOPS: High performance genomic feature operations. Bioinformatics.

Nothnagel, M., Lu, T.T., Kayser, M., and Krawczak, M. (2010). Genomic and geographic distribution of SNP-defined runs of homozygosity in Europeans. Human molecular genetics *19*, 2927-2935.

Partsch, C.J., Dreyer, G., Gosch, A., Winter, M., Schneppenheim, R., Wessel, A., and Pankau, R. (1999). Longitudinal evaluation of growth, puberty, and bone maturation in children with Williams syndrome. J Pediatr *134*, 82-89.

Pelican, M. (2009). Complexities of Indigeneity and Autochthony: An African Example. American Ethnologist *36*, 52-65.

Pemberton, T.J., Absher, D., Feldman, M.W., Myers, R.M., Rosenberg, N.A., and Li, H. (2012). Genomic patterns of homozygosity in worldwide human populations. Am J Hum Genet *in press*.

Plagnol, V., and Wall, J.D. (2006). Possible ancestral structure in human populations. PLoS genetics *2*, e105.

Pritchard, J.K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. Genetics *155*, 945-959.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J.*, et al.* (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet *81*, 559-575.

Rand, D.M., and Kann, L.M. (1996). Excess amino acid polymorphism in mitochondrial DNA: contrasts among genes from Drosophila, mice, and humans. Molecular biology and evolution *13*, 735-748.

Reich, D., Green, R.E., Kircher, M., Krause, J., Patterson, N., Durand, E.Y., Viola, B., Briggs, A.W., Stenzel, U., Johnson, P.L.*, et al.* (2010). Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature *468*, 1053-1060.

Roach, J.C., Glusman, G., Smit, A.F., Huff, C.D., Hubley, R., Shannon, P.T., Rowen, L., Pant, K.P., Goodman, N., Bamshad, M*., et al.* (2010). Analysis of genetic inheritance in a family quartet by whole-genome sequencing. Science *328*, 636-639.

Rodriguez-Pacheco, F., Martinez-Fuentes, A.J., Tovar, S., Pinilla, L., Tena-Sempere, M., Dieguez, C., Castano, J.P., and Malagon, M.M. (2007). Regulation of pituitary cell function by adiponectin. Endocrinology *148*, 401-410.

Rosenbloom, K.R., Dreszer, T.R., Long, J.C., Malladi, V.S., Sloan, C.A., Raney, B.J., Cline, M.S., Karolchik, D., Barber, G.P., Clawson, H*., et al.* (2012). ENCODE whole-genome data in the UCSC Genome Browser: update 2012. Nucleic Acids Res *40*, D912-917.

Schuster, S.C., Miller, W., Ratan, A., Tomsho, L.P., Giardine, B., Kasson, L.R., Harris, R.S., Petersen, D.C., Zhao, F., Qi, J*., et al.* (2010). Complete Khoisan and Bantu genomes from southern Africa. Nature *463*, 943-947.

Sethupathy, P., and Hannenhalli, S. (2008). A tutorial of the poisson random field model in population genetics. Adv Bioinformatics, 257864.

Shaikh, T.H., Gai, X., Perin, J.C., Glessner, J.T., Xie, H., Murphy, K., O'Hara, R., Casalunovo, T., Conlin, L.K., D'Arcy, M*., et al.* (2009). High-resolution mapping and analysis of copy number variations in the human genome: a data resource for clinical and research applications. Genome research *19*, 1682-1690.

Shriver, M.D., Kennedy, G.C., Parra, E.J., Lawson, H.A., Sonpar, V., Huang, J., Akey, J.M., and Jones, K.W. (2004). The genomic distribution of population substructure in four populations using 8,525 autosomal SNPs. Hum Genomics *1*, 274-286.

Stevens, A., Morissette, J., Woodburn, J.C., and Bennett, F.J. (1977). The inbreeding coefficients of the Hadza. Ann Hum Biol *4*, 219-223.

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics *123*, 585-595.

Thomas, P.D., Kejariwal, A., Campbell, M.J., Mi, H., Diemer, K., Guo, N., Ladunga, I., Ulitsky-Lazareva, B., Muruganujan, A., Rabkin, S*., et al.* (2003). PANTHER: a browsable database of gene products organized by biological function, using curated protein family and subfamily classification. Nucleic acids research *31*, 334-341.

Thomson, R., Pritchard, J.K., Shen, P., Oefner, P.J., and Feldman, M.W. (2000). Recent common ancestry of human Y chromosomes: evidence from DNA sequence data. Proc Natl Acad Sci U S A *97*, 7360-7365.

Tishkoff, S.A., Reed, F.A., Friedlaender, F.R., Ehret, C., Ranciaro, A., Froment, A., Hirbo, J.B., Awomoyi, A.A., Bodo, J.M., Doumbo, O*., et al.* (2009). The genetic structure and history of Africans and African Americans. Science *324*, 1035-1044.

Vernot, B., Stergachis, A.B., Maurano, M.T., Vierstra, J., Neph, S., Thurman, R.E., Stamatoyannopoulos, J.A., and Akey, J.M. (2012). Personal and population genomics of human regulatory variation. Genome Res *(in press)*.

Wang, K., Li, M., and Hakonarson, H. (2010). Analysing biological pathways in genome-wide association studies. Nature reviews Genetics *11*, 843-854.

Watterson, G.A. (1975). On the number of segregating sites in genetical models without recombination. Theor Popul Biol *7*, 256-276.

Weir, B.S., and Cockerham, C.C. (1984). Estimating F-Statistics for the Analysis of Population Structure. Evolution *38*, 1358-1370.

Wray, G.A. (2007). The evolutionary significance of cis-regulatory mutations. Nature reviews Genetics *8*, 206-216.

Zhang, J., Rowe, W.L., Clark, A.G., and Buetow, K.H. (2003). Genomewide distribution of high-frequency, completely mismatching SNP haplotype pairs observed to be common across human populations. Am J Hum Genet *73*, 1073-1081.

Supplemental Tables

Table S1. Individuals analyzed in PCA and/or neighbor joining tree, related to Figure 1

Population	Individuals analyzed in PCA	Individuals in the NJ tree
Pygmy (Cameroon)	5	5
Hadza (Tanzania)	5	5
Sandawe (Tanzania)	5	5
YRI (Yoruba from Ibadan, Nigeria)	9	9
LWK (Luhya from Webuya, Kenya)	4	4
MKK (Maasai fom Kinyawa, Kenya)	3	3
ASW (African-Americans from the Southwest USA)	5	5
CEU (Northern and Western European ancestry from the CEPH collection)	9	9
TSI (Toscans in Italy)	4	4
CHB (Chinese from Beijing, China)	4	4
JPT (Japanese from Tokyo, Japan)	4	4
GIH (Gijurati Indians in Houston, USA)	4	4
PUR (Puerto Rican from Puerto Rico)	2	0
MXL (Mexican ancestry in Los Angeles, USA)	5	0

Table S2. Y chromosome and mtDNA haplogroups, related to Figure 1

Individual	Y chromosome haplogroup	mtDNA haplogroup
Pygmy 1 (Baka)	B2b	L1c1a1a1b1
Pygmy 2 (Baka)	E1b1a1a1f	L1c1a1a1a
Pygmy 3 (Baka)	B2b4	L1c1a1a1a
Pygmy 4 (Bakola)	E1b1a1a1f	L1c1a1a1b1
Pygmy 5 (Bedzan)	E1b1a1a1g	L1c1a2a1
Hadza 1	E1b1a1a1f	L2a1
Hadza 2	B2b*	L2a1
Hadza 3	B2b*	L4b2a2b1a
Hadza 4	B2b*	L3h1a2a2
Hadza 5	B2b	L4b2a2b1a
Sandawe 1	E1b1a1a1g	L3e1d
Sandawe 2	E1b1b1	L2a1i
Sandawe 3	E1b1b1f	L0a2d1
Sandawe 4	B2b*	L0a3a
Sandawe 5	B2b*	L3x1a

Table S3. D-tests (shared derived alleles), related to Figure 1

P1	P2	P3	Shared derived (P1 and P3)	Shared derived (P2 and P3)	D	Std. error	Z-score
Pygmy	Hadza	San (KB1)	608925.4	593355.4	1.30%	0.44%	2.86
Pygmy	Sandawe	San (KB1)	608925.4	592439	1.37%	0.32%	4.24
Hadza	Sandawe	San (KB1)	593355.4	592439	0.08%	0.29%	0.27

Table S3. Number of shared derived alleles, values of D, and standard errors are means from 25 tests.

Table S4. **Neutrality Index tests, related to Table 2.** This table is available as a Microsoft Excel file (TableS4_NeutralityIndex.xlsx). It includes pooled and population-specific Neutrality Index calculations for hunter-gatherers.

Table S5. Genome Scans, related to Figure 4. This table is available as a Microsoft Excel file (TableS5_GenomeScans.xlsx). It contains lists of highly-divergent genomic regions (LSBL scans), AIM clusters, regions with high or low values of Tajima's D, and T_{MRCA} outliers.

Table S6. Pathway Analyses, related to Figure 4. This table is available as a Microsoft Excel file (TableS6_PathwayAnalyses.xlsx). It contains KEGG and PANTHER analyses of genes identified by LSBL scans (200kb flanking the top 1% most-divergent regions) and genes within 50kb of AIM SNPs.

Table S7. AIM frequencies and height associations, related to Figure 5. This table is available as a Microsoft Excel file (TableS7_AIM_Frequency_Height.xlsx). It contains a list of SNP positions, allele frequency data for five Pygmy and two Bantu populations, and p-values for associations with height.

Supplemental figure legends

Figure S1. Quality control, related to Table 1. A) Venn diagram showing the variants in dbSNP 131 (23.4M variants) and hunter-gatherer genomes (13.4M variants). All shapes are drawn to scale, and overlap between each of these two sets amounts to 7.9 million variants. Panel B) Fully called sites have lower error rates. Sites were binned according to whether they were fully called or called in a subset of 15 hunter-gatherer genomes. Discordance rates are for variant positions in technical replicates, and overestimate actual error rates. Panel C) Most variant sites are fully called in all 15 hunter-gatherers. Panel D) Summed Chi-square statistics from Pygmy, Hadza, and Sandawe populations (departure from Hardy-Weinberg proportions filter).

Figure S2. Variants from whole-genome sequencing, related to Figure 1. A) Variants per sequenced genome for different populations. Non-hunter-gatherer populations analyzed are Yoruba (YRI), Asian (ASN, i.e. CHB and JPT), and European (CEU). “Novel” refers to variants absent from dbSNP131. B) Power-law parameters for the number of variants observed in each sequenced genome. Panels C, E, G, and I) Histograms showing the number of variants per Mb. Each histogram contains 150 bins. Panels D, F, H, and J) genomic distribution of variants per Mb. Panels C and D refer to the pooled set of all 15 hunter-gatherer genomes.

Figure S3. Signatures of purifying selection in geographically diverse populations with different subsistence patterns, related to Table 2. To control for sample size, four diploid genomes were analyzed for each population. Hunter-gatherer (HG) populations are Pygmies, Hadza, and Sandawe. African agriculturalists and pastoralists (AP) are YRI, LWK, and MKK. Non-African agriculturalists (NA) are CEU, CHB, and JPT. A) Mean derived allele frequencies (DAF) for intergenic and exon SNPs. B) DAF in exon SNPs relative to intergenic SNPs. C) PolyPhen-2 data indicating the proportion of nonsynonymous variants classified as benign, possibly damaging, or probably damaging. D) Number of nonsynonymous variants per genome divided by the number of synonymous variants per genome. E) Statistical tests between groups of populations. ** indicates $p < 0.01$ and *** indicates $p < 10^{-4}$.

Figure S4. PCA plots, related to Figure 1. In each panel the x-axis corresponds to PC1. The proportion of the variance explained by each PC is indicated along each axis, and individuals are represented by population name. Pygmies are labeled green, Hadza are labeled blue, and Sandawe are labeled red.

Figure S5. Runs of homozygosity (ROH), related to Table 2. In each panel cROH corresponds to the cumulative number of base pairs observed in runs of homozygosity. A run of homozygosity is defined as a 100kb region lacking heterozygote calls (Panel A) or a 1Mb region lacking heterozygote calls (Panel B). One genotyping error is tolerated per 50kb. Colors of points differ for Pygmies (red),

Hadza (blue), and Sandawe (red). Panel C) Statistics for runs of homozygosity. Population means ± one standard deviation are listed. cROH refers to the cumulative number of base pairs found in runs of homozygosity (ROH). CV refers to the coefficient of variation (standard deviation / mean).

Figure S6. Derived allele frequency distributions for hunter-gatherer populations, related to Table 2. A) Derived allele frequency (DAF) distributions for hunter-gatherer populations and null expectations from the neutral theory (infinite sites model with constant population size). B-D): Allele frequency distributions for pairs of populations using whole-genome sequencing data. E-G) Allele frequency distributions for pairs of populations using Illumina1M genotyping array data. In each panel, x- and y-axes correspond to the derived allele frequency in a particular population and the z-axis corresponds to the proportion of SNPs with a particular set of allele frequencies.

Figure S7. S* statistics are robust at detecting introgression, related to Figures 2 and 3. A) False Discovery Rate (FDR) for top 0.5% of simulated 50kb regions, ranked by S*, under several demographic parameters. Panel B) Recombination rate of top candidate regions (red dashed line), and of all regions, in three hunter gatherer populations. C) T_{MRCA} estimates for shared and unique top candidate regions in Pygmy, Hadza and Sandawe. D) Tail of simulated S* distribution over 0-3.8% introgressed sequence per individual. E) Tail of observed S* distribution for 11 populations sequenced by Complete Genomics.