SUPPLEMENTARY METHODS

STRUCTURE Analysis

Analysis of population structure was performed with the program STRUCTURE (Pritchard et al. 2000) using the 1,019 sites that have a called genotype in at least one of our samples and that are included in the Illumina 650Y platform. On average about 40% of these sites have a called genotype in any one of our samples. We used a burn-in period of 2000 iterations followed by 2000 iterations. Our reference genotype data were from the 938 unrelated individuals of the 51 populations in the Human Genome Diversity Panel (Cann et al. 2002) generated on the Illumina 650Y arrays (Li et al. 2008), 288 unrelated individuals from 4 HapMap Phase 3 populations (LWK, MKK, GIH, TSI), and 204 unrelated individuals from 4 populations (Naukan Yupik, Chukchi, Native Australian, Khoisan, Amhara) genotyped in-house (Hancock et al. 2010). The genotypes for HGDP samples and our in-house samples were generated using the Illumina 650Y arrays, while HapMap Phase 3 samples were genotyped using the Illumina Human1M platform. Graphs of STRUCTURE results were produced using DISTRUCT ver. 1.1.

Calculating a correction factor to take into account underestimation of nucleotide diversity due to restriction site polymorphism

In this section we provide the derivations for the formulas used to estimate the correction factor. To calculate:

Prob(site called heterozygous)/Prob(site called),

we first proceed to calculate the numerator. We consider the probability of a restriction fragment of a specified size occurring at a given position of a single sequence (a single chromosome). We then consider two homologous sequences and the possibility of heterozygous sites. For both alleles at a heterozygous nucleotide site (the focal site) to be detected by our method, each of the alleles must be in a restriction fragment approximately 70 base pairs long. In the following, for mathematical simplicity, we assume that the fragment must be exactly 70 base pairs long. The extension to a range of fragment sizes will not change our main result. In the following we calculate the probability that a focal nucleotide site is at a particular position within a 70 base pair restriction fragment. This is also only for convenience and does not affect the final result.

To generate a 70 base pair restriction fragment, we consider a stretch of DNA 74 base pairs long, with the positions numbered, 5’ to 3’, from 1 to 74. From a single diploid individual we have two, possibly distinct, sequences of this segment present in our sample, which we refer to as paternal and maternal copies. We assume a four-cutter restriction enzyme, with recognition sequence AGCT. And say that a restriction site occurs at position j, if this recognition sequence appears at positions j, j+1, j+2, j+3, resulting in a cut between position j+1 and j+2. We assume that the probability of a restriction site at any position is P_{rs}. If the recognition sequence does not occur at position i, we assume that this does not affect the probability that the recognition sequence occurs at position i+1. In this case the probability that the paternal copy has a restriction site at position one, and that there is no recognition sites at positions 5-70, and there is a restriction site at position 71, is

$$P_{rs}(1-P_{rs})^{66}P_{rs} \ .$$

(Note that with our recognition sequence, the presence of the recognition sequence at position one precludes it being present at positions 2, 3 or 4.) This expression gives the probability that the paternal (or equivalently, the maternal) sequence will generate a restriction fragment of length 70, consisting of positions 3-72. If the maternal and paternal sequences are identical, then the quantity above is the probability that both sequences generate the appropriate size fragment. If there is a single heterozygous site in the middle of the sequence, then the probability that both sequences generate a 70 base pair fragment is

$$P_{rs}(1-P_{rs})^{66}P_{rs}(1-P_{rs})^{4}$$

where the extra factor of $(1- P_{rs})^4$ arises because the heterozygous site introduces four new four base pair sequences that might generate a restriction site. For both copies to generate the 70 base pair restriction fragment, there must be no restrictions sites in the interior of the sequence of either the maternal or paternal sequence.

Now we focus attention on a particular nucleotide site, say at position 30, which may or may not be heterozygous in our sample individual. (We assumed that the equation we derived for position 30 can be applied to all positions and we ignored minor changes that would be required for positions adjacent to the restriction sites at the edges of the fragment.) To get the required probabilities we need to consider several possible patterns of polymorphism, namely,

A) Both our focal site and our restriction sites at the ends of the fragment are homozygous.
B) Both the focal site and one of the restriction sites are heterozygous.
C) The focal site is heterozygous and the restriction sites are homozygous.
D) The focal site is homozygous and one of the restriction sites is heterozygous.

Case B results in allele drop out. To calculate these probabilities we assume an equilibrium panmictic infinite-sites neutral model, but expect that these results will apply approximately for more realistic models with mild geographic structure and the possibility of multiple hits at sites. (Similarly, we assume no recombination within the 74 base pair segment.) With these assumptions the probabilities of the polymorphism patterns A – D are:

$$P_A = \frac{1}{1+9\theta} \approx 1-9\theta+81\theta^2$$
$$P_B = \frac{\theta}{1+9\theta}\frac{8\theta}{1+8\theta} + \frac{8\theta}{1+9\theta}\frac{\theta}{1+\theta} \approx 16\theta^2$$
$$P_C = \frac{\theta}{1+9\theta}\frac{1}{1+8\theta} \approx \theta-17\theta^2$$
$$P_D = \frac{8\theta}{1+9\theta}\frac{1}{1+\theta} \approx 8\theta-80\theta^2.$$

The symbol θ is the scaled per base pair neutral mutation rate ($4N_e\mu$). In calculating these probabilities we are just focusing on the 8 positions of the bounding restriction sites, and the single interior site, thus a total of nine sites.

The probability that the focal site is heterozygous and both alleles are in 70 bp restriction fragments that have coverage larger than t is:

$$\text{Prob(site called heterozygous)} = P_c P_{rs}^2 (1-P_{rs})^{66} (1-P_{rs})^4 f_2$$
$$\approx \theta(1-17\theta-4P_{rs}) P_{rs}^2 (1-P_{rs})^{66} f_2 ,$$

where f_2 is the probability that a site has coverage larger than t, given that both the paternal and maternal copies generate a 70 base pair restriction fragment. We will denote by f_1 the probability that a site has coverage larger than t, given that only one copy generates a 70 base pair restriction fragment (i.e. when drop out occurs.)

To calculate Prob (site called), we need to sum the probabilities of each of the events A-D, multiplying each term by the probability of appropriately placed restriction sites in each case:

$$\begin{aligned}\text{Prob(site called)} = {} & P_A P_{rs}^2 (1-P_{rs})^{66} f_2 + \\ & P_B 2P_{rs}^2 (1-P_{rs})^{62} f_1 + \\ & P_C (P_{rs}^2 (1-P_{rs})^{66} (1-P_{rs})^4 f_2 + \\ & \quad P_{rs}^2 (1-P_{rs})^{62} (1-(1-P_{rs})^8) f_1) + \\ & P_D 2P_{rs}^2 (1-P_{rs})^{66} f_1.\end{aligned}$$

This is approximately,

$$\text{Prob(site called)} \approx (1-8\theta+16\theta\frac{f_1}{f_2}) P_{rs}^2 (1-P_{rs})^{66} f_2.$$

So finally,

$$\frac{\text{Prob(site called heterozygous)}}{\text{Prob(site called)}} \approx \theta(1-4P_{rs}-9\theta-16\theta\frac{f_1}{f_2}).$$

This suggests that one can obtain a good approximate corrected estimate of π with:

$$\frac{\pi_{raw}}{(1-4P_{rs}-9\pi_{raw}-16\pi_{raw}\frac{f_1}{f_2})}$$
,

where π_{raw} is simply the observed fraction of sites that are called heterozygote. To estimate f_1/f_2 proceed as follows.

Let $F_1(c)$ be the probability of coverage c when one allele has dropped out, and let $F_2(c)$ be the probability of coverage c, when drop-out does not occur. Then assuming that the distribution of coverage of one allele is binomially distributed when conditioned on the total coverage at a site, and assuming that the coverage of the non-dropped-out allele has the same distribution as the coverage of a single allele at a non-drop-out site, it follows that

$$F_1(c) = \sum_{k \geq c} F_2(k)\,\mathrm{Bin}(c;k,0.5)$$

where Bin(c;k,0.5) is a binomial probability of c successes in k trials with the probability of success on each trial being 0.5. The sum should be for all k greater than or equal to c, but for simplicity we limit consideration to k <= 250. If we assume the threshold for reliable genotype calls is 20, then f_1 is the sum of F_1(c) for c going from 20 to 100. The quantity F_2(c) summed over values of c from 20 to 100 is f_2. Thus we have:

$$\frac{f_1}{f_2} = \frac{\sum_{c=20}^{100} \sum_{k \geq c}^{250} F_2(k)\,\mathrm{Bin}(c;k,0.5)}{\sum_{c=20}^{100} F_2(c)}$$

The quantities F_2(c) can be estimated very accurately from the observed coverage distribution using all sites. In practice it is easier to consider the coverage distributions conditional on coverage between 20 and 250. We thus estimate f_1/f_2 by:

$$\left\langle \frac{f_1}{f_2} \right\rangle = \frac{\sum_{c=20}^{100} \sum_{k \geq c}^{250} F_2^*(k)\,\mathrm{Bin}(c;k,0.5)}{\sum_{c=20}^{250} F_2^*(c)}$$

where

$$F_2^*(c) = \frac{N_2(c)}{\sum_{c=20}^{250} N_2(c)}$$

and N_2(c) is the observed number of sites with coverage c. Thus F_2^*(c) is the observed distribution of coverage conditional on coverage between 20 and 250. (The observed distribution is a mixture of sites with and without drop-out, but because polymorphism levels are so low the above approach gives an accurate estimate of F_2(c)).

The estimated value of f_1/f_2 varied from sample to sample, with a range of 0.38 to 0.73, with a mean of approximately 0.5. Assuming $f_1/f_2 = 0.5$ and π equal to 0.001, and $P_{rs} = 4/256$, we calculate that the raw π is expected to be approximately 3.4% lower than the true π.

To confirm that our approximate theoretical results above were accurate we carried out computer simulations. We generated pairs of sequences under a Jukes-Cantor neutral model without recombination, with equal frequencies of the four bases. We assumed $\theta = 0.001$. In our simulated sequences we located the restriction fragments of length 68-72, and tabulated the number of heterozygous sites without drop out (assuming $f_1/f_2 = 0.5$.) We found that the effect of allelic dropout was slightly over 3%, confirming our calculations above.

We also carried out simulations to check that divergence between sequences is biased by approximately the same multiplicative factor as estimates of π. To do this, we simulated polymorphism and divergence data using "ms" (Hudson 2002) to generate samples from two populations that split 2,000 generations in the past, using mutation rates and population sizes that resulted in levels of diversity and divergence similar to what we observed in our samples. Each ms sample specifies the locations of polymorphisms and which copies carry the ancestral and which the derived allele. To convert this information to a nucleotide sequence, a random ancestral sequence was generated with equal base frequencies expected. The resulting descendant sequences were generated from this ancestral sequence and the ms output. We found that divergence was biased downward by approximately 2% due to dropout. This bias is somewhat smaller than we found for diversity estimates. Our estimates for split times using equation (4) is based on the assumption that both divergence and diversity are biased by the same multiplicative factor. The somewhat reduced bias of divergence is expected to result in our split time estimates being upwardly biased. To quantitatively assess this bias in split time estimates, we again simulated polymorphism data with ms (Hudson 2002). In this case, for each replicate we generated samples as above for 10,000 unlinked segments, each 20,000 base pairs long with no recombination. (Segments of this length almost always generate zero, one or two restriction fragments of the appropriate length.) Generating 10,000 such segments resulted in approximately one megabase of sequence from restriction fragments which corresponds to the amount of sequence we used to estimate nucleotide diversity but somewhat more than we used for divergence estimates. For each replicate we employed equation (4) to estimate the split time. When the true split time was 50,000 years, we found our mean estimate from 1,000 replicates was 55,700 years with a standard deviation of 18,000 years. Thus the split time estimates are upwardly biased by approximately 10%.

Maximum likelihood estimates of nucleotide diversity

To estimate the fraction of sites in the genome of an individual that are heterozygous, which we denote by p, we utilized a maximum likelihood approach applied to all sites in our data that have coverage greater than or equal to 10. An observed configuration (i

A_1, j A_2), meaning i copies of the A_1 allele, and j copies of A_2, can occur in three different ways:

1) the site is heterozygous in the individual and no drop-out occurs at this site
2) the site is heterozygous in the individual and drop-out occurs at this site
3) the site is homozygous.

In cases 2 and 3, if there is apparent variation in the sample (i.e. i and j are both non-zero), then the apparent variation is due to sequencing error. We assume when the site is heterozygous and no dropout occurs, that the distribution of the number of copies of each allele is binomially distributed with parameter 0.5, conditional on the total coverage (i+j) at the site. When the variation is due to sequencing error we assume that the number of copies of each allele is again binomially distributed but with parameter, ε, which can be thought of as the sequencing error rate. We assume this parameter, ε, is independent of which two alleles are present, but we incorporate a term ($g_p(A_1,A_2)$) expressing the relative likelihood of different allele pairs, at heterozygous sites and a different term ($g_m(A_1,A_2)$) for the relative likelihood of each allelic pair for variation due to sequencing errors. With these considerations and based on the expressions given in the previous sections, we obtain the following expression for the probability of a configuration (i A_1, j A_2) at a site:

$$\begin{aligned} P(i\,A_1, j\,A_2) = [&\pi(1-17\pi-4P_{rs})g_p(A_1,A_2)F_2^*(i+j)\,\mathrm{Bin}(i;i+j;0.5)(2-I_{i,j}) + \\ &\pi(32\pi+8P_{rs})g_m(A_1,A_2)F_1^*(i+j)\,\mathrm{Bin}(i;i+j;\varepsilon) + \\ &(1-19\pi-81\pi^2)g_m(A_1,A_2)F_2^*(i+j)\,\mathrm{Bin}(i;i+j;\varepsilon) + \\ &8\pi(1-10\pi)g_m(A_1,A_2)F_1^*(i+j)\,\mathrm{Bin}(i;i+j;\varepsilon)]/\mathrm{Prob}(\mathrm{coverage}=i+j) \end{aligned}$$

where π is the probability a site is heterozygous. (This is the quantity we want to estimate.) And where $g_p(A_1,A_2)$ is the probability that the two alleles at a heterozygous site are A_1 and A_2 , $g_m(A_1,A_2)$ is the probability that the two alleles at a site due to sequencing error are A_1 and A_2, $F^*_2(n)$ is the probability of coverage n at a site without drop-out, $F^*_1(n)$ is the probability of coverage n when there is drop-out, $I_{i,j}$ is an indicator variable equal to one when i equals j, and zero otherwise, and where Bin(i;n;p) is the binomial probability of i successes in n trials with probability of success on each trial of p. And finally where:

$$\text{Prob(coverage } = \text{ i + j)} = \pi(1\text{-}17\pi - 4P_{rs})F_2^*(i+j) + \pi(32\pi + 8P_{rs})F_1^*(i+j) + (1\text{-}9\pi + 81\pi^2)F_2^*(i+j) + 16\pi(1-10\pi)F_1^*(i+j),$$

which is proportional to the probability that a site has coverage i+j, and is obtained from the expressions given in the previous sections. Dividing by this quantity means that our likelihood expression is based on probabilities conditional on the observed coverages.

The estimation $F^*_2(n)$, and $F^*_1(n)$ are described in the previous section. To estimate $g_m(A_1,A_2)$ we consider all sites that have coverage 20 to 100 that have one or two copies of the rarer allele. These sites are with high probability, homozygous sites with sequencing error. The fraction of these sites for which A_1 and A_2 are the two alleles found at the site is our estimate of $g_m(A_1,A_2)$. Similarly, to estimate $g_p(A_1,A_2)$ we consider all sites that have coverage 20 to 100, and for which the frequency of the rarer allele is (i+j)/2 -2, (i+j)/2 -1 , or (i+2)/2 . These sites are with high probability heterozygous sites. The fraction of these sites for which A_1 and A_2 are the two alleles found at the site is our estimate of $g_p(A_1,A_2)$. We find that g_p and g_m differ from each other substantially, and hence that the identity of the two nucleotides at an apparently polymorphic nucleotide position provides significant information about whether the site is truly polymorphic or not. Finally, we estimate e using high coverage sites, in this case coverage 15 to 100, that have zero to three copies of one allele. We treat each of these “erroneous” copies as an independent error, and thus get an error rate for a given coverage, c, with:

$$\sum_{i=1}^{3} n_i i \Bigg/ \sum_{i=0}^{3} c n_i$$

where n_i is the number of sites that have i copies of the rare allele and c is the coverage being considered. So the numerator here is the total number of “erroneous” alleles, and the denominator is the total number of copies in the data at coverage c. We simply take the arithmetic mean of these estimators for coverage 15 to 100.

We take all the parameters on the right hand side of the equation for P(i A_1, j A_2) to be known without error (except π), and assume the probability of the configuration at each site to be independent of the other sites. To obtain the maximum likelihood of π we maximize the product of the P(i A_1, j A_2)’s over all sites with coverage greater than 10. This is done by simply evaluating the log of the product on a grid of discrete values of π.

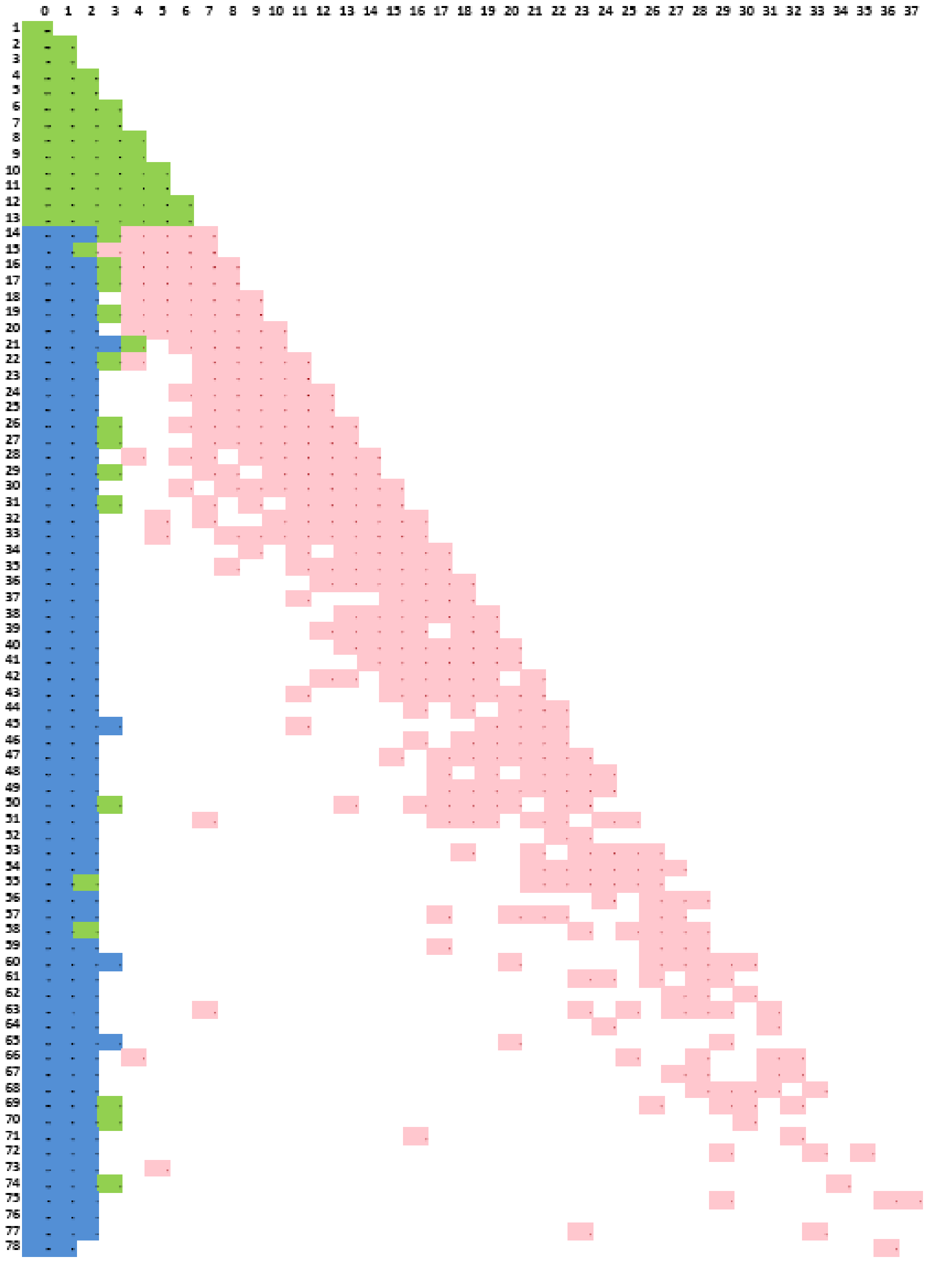
SUPPLEMENTARY FIGURE LEGENDS

Supplementary Figure 1. Example of a coverage matrix. The left column indicates the total coverage (up to coverage 78) and the top row indicates the minor allele coverage (up to coverage 37). Blue cells are called homozygous sites, red cells are called heterozygous sites, and green are uncalled sites.

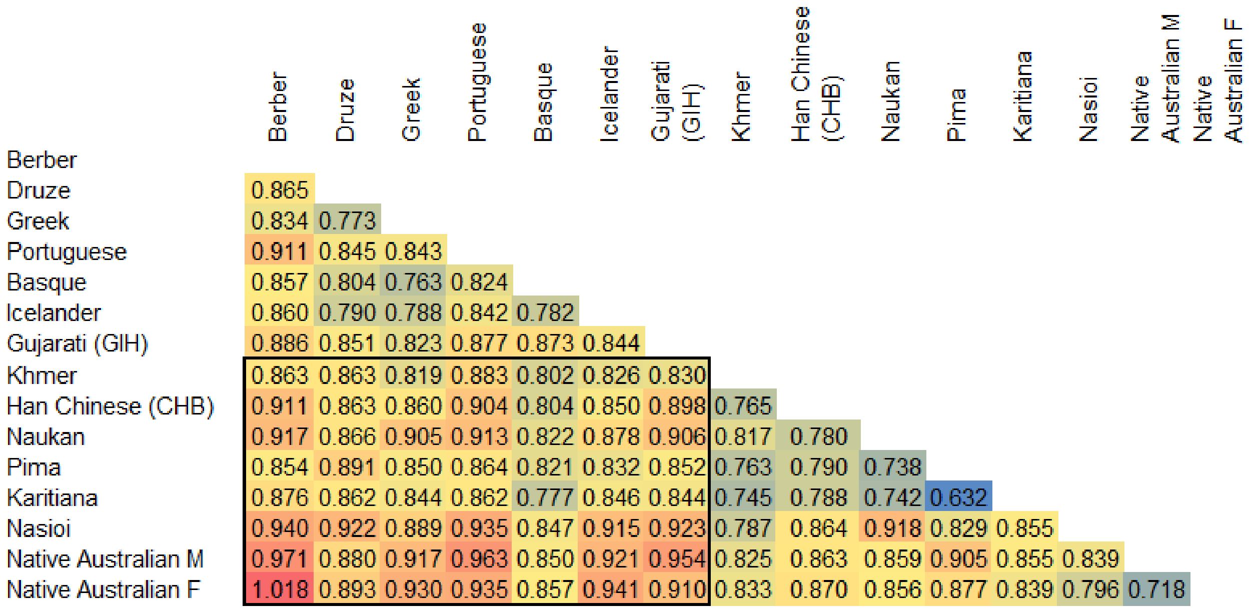
Supplementary Figure 2. Heatmap of the pairwise divergence between non-African samples. Divergence estimates between European samples and East Asians, Oceanic and Native American samples are demarked by the box.

Supplementary Figure 3. Structure plot generated using the resequencing data for the 20 samples included in the study, together with genotyping data for the 51 populations in the Human Genome Diversity Panel, 4 HapMap Phase 3 populations (LWK, MKK, GIH, TSI), and 4 additional populations (Naukan Yupik, Chukchi, Native Australian, Khoisan, Amhara, Oromo).

Supplementary Figure 1.

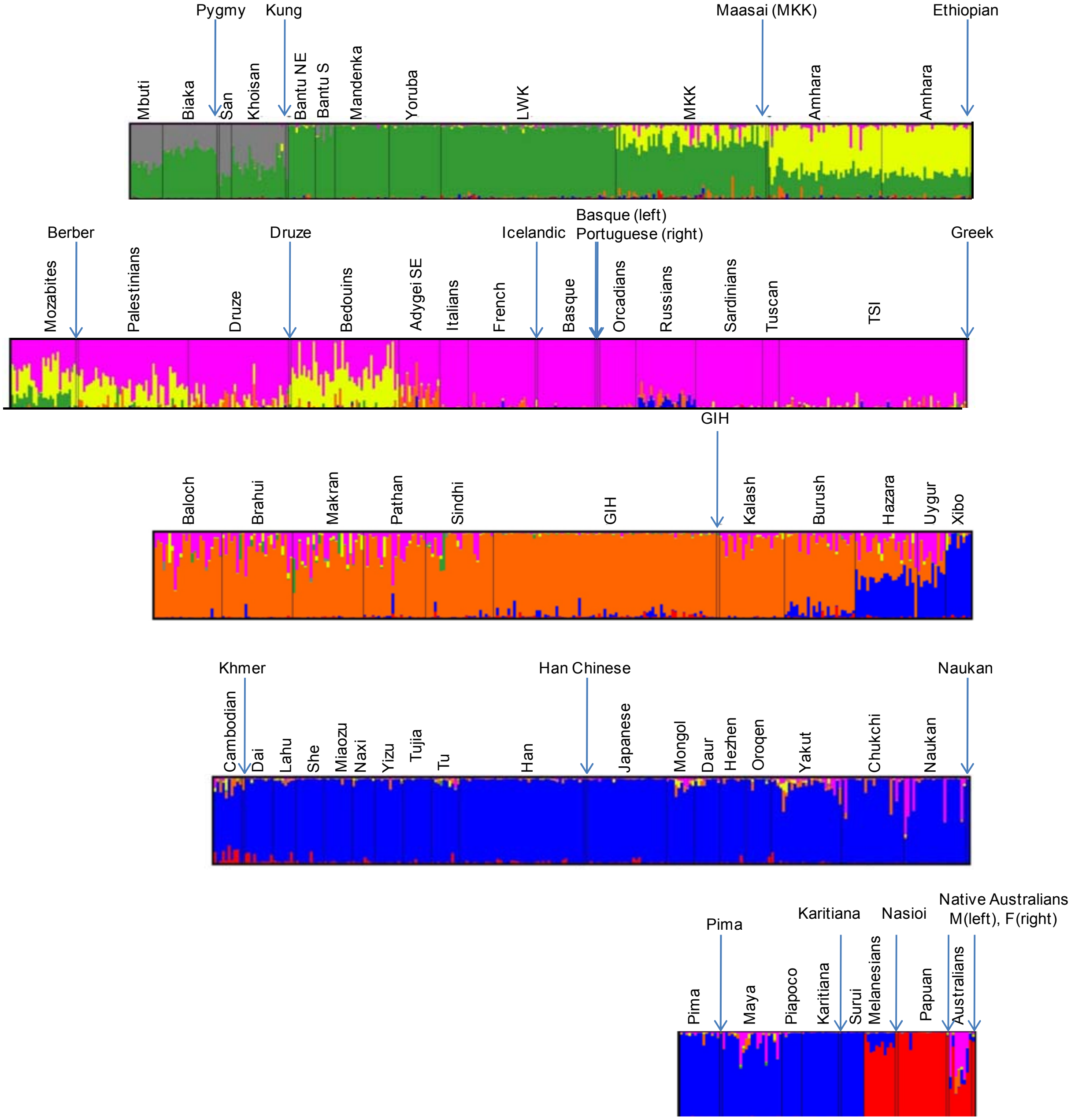


Supplementary Figure 2.



	Berber	Druze	Greek	Portuguese	Basque	Icelander	Gujarati (GIH)	Khmer	Han Chinese (CHB)	Naukan	Pima	Karitiana	Nasioi	Native Australian M	Native Australian F
Berber															
Druze	0.865														
Greek	0.834	0.773													
Portuguese	0.911	0.845	0.843												
Basque	0.857	0.804	0.763	0.824											
Icelander	0.860	0.790	0.788	0.842	0.782										
Gujarati (GIH)	0.886	0.851	0.823	0.877	0.873	0.844									
Khmer	0.863	0.863	0.819	0.883	0.802	0.826	0.830								
Han Chinese (CHB)	0.911	0.863	0.860	0.904	0.804	0.850	0.898	0.765							
Naukan	0.917	0.866	0.905	0.913	0.822	0.878	0.906	0.817	0.780						
Pima	0.854	0.891	0.850	0.864	0.821	0.832	0.852	0.763	0.790	0.738					
Karitiana	0.876	0.862	0.844	0.862	0.777	0.846	0.844	0.745	0.788	0.742	0.632				
Nasioi	0.940	0.922	0.889	0.935	0.847	0.915	0.923	0.787	0.864	0.918	0.829	0.855			
Native Australian M	0.971	0.880	0.917	0.963	0.850	0.921	0.954	0.825	0.863	0.859	0.905	0.855	0.839		
Native Australian F	1.018	0.893	0.930	0.935	0.857	0.941	0.910	0.833	0.870	0.856	0.877	0.839	0.796	0.718	

Supplementary Figure 3.



SUPPLEMENTARY TABLES

Supplementary table 1. False negative rate calculated by comparison with HapMap heterozygous sites, when a 5% binomial filter is applied to call heterozygous sites in the re-sequencing data set.

	Han Chinese (CHB)	Maasai (MKK)	Gujarati (GIH)
False negative rate	6.3%	9.4%	7.7%

Supplementary table 2. Number of heterozygous sites called on the X chromosome of male individuals after position and base quality filters were applied.

Sample	X error rate (x 10^{-3})
Basque	4.78
Druze	6.64
Ethiopian	6.78
Greek	8.17
Gujarati (GIH)	5.78
Han Chinese (CHB)	9.09
Icelander	9.20
Karitiana	2.67
Khmer	2.34
Kung	5.31
Maasai (MKK)	6.51
Mbuti Pygmy	4.94
Nasioi	9.20
Native Australian M	7.85
Naukan	6.78
Pima	5.68
Portuguese	9.41
Average	6.54

Supplementary table 3. Comparison of the maximum likelihood estimates of nucleotide diversity to previously published estimates

Sample	π (x 10^{-3}) [Wall et al (2008)]	π (x 10^{-3})
Basque	0.87	0.72
Biaka/Mbuti Pygmy	1.21	0.99
Han Chinese/CHB	0.81	0.67
Melanesian/Nasioi	0.78	0.65
San/Kung	1.26	1.00

Supplementary table 4. Nucleotide diversity from sites at any distance from genic regions estimated with the maximum likelihood approach (MLE π) compared to the method of moments estimates (π_m).

Sample	MLE π (x 10^{-3})	π_m(x 10^{-3})
Maasai (MKK)	1.04	0.97
Kung	1.00	1.00
Mbuti Pygmy	0.99	0.98
Ethiopian	0.92	0.93
Berber	0.84	0.81
Druze	0.81	0.81
Gujarati (GIH)	0.79	0.72
Portuguese	0.79	0.79
Khmer	0.74	0.70
Greek	0.73	0.73
Basque	0.72	0.73
Icelander	0.72	0.68
Native Australian M	0.69	0.67
Native Australian F	0.67	0.67
Han Chinese (CHB)	0.67	0.66
Nasioi	0.65	0.64
Pima	0.64	0.67
Naukan	0.63	0.59
Karitiana	0.58	0.53

REFERENCES

Cann, H.M., C. de Toma, L. Cazes, M.F. Legrand, V. Morel, L. Piouffre, J. Bodmer, W.F. Bodmer, B. Bonne-Tamir, A. Cambon-Thomsen, Z. Chen, J. Chu, C. Carcassi, L. Contu, R. Du, L. Excoffier, G.B. Ferrara, J.S. Friedlaender, H. Groot, D. Gurwitz, T. Jenkins, R.J. Herrera, X. Huang, J. Kidd, K.K. Kidd, A. Langaney, A.A. Lin, S.Q. Mehdi, P. Parham, A. Piazza, M.P. Pistillo, Y. Qian, Q. Shu, J. Xu, S. Zhu, J.L. Weber, H.T. Greely, M.W. Feldman, G. Thomas, J. Dausset, and L.L. Cavalli-Sforza. 2002. A human genome diversity cell line panel. *Science* **296:** 261-262.

Hancock, A.M., D.B. Witonsky, E. Ehler, G. Alkorta-Aranburu, C. Beall, A. Gebremedhin, R. Sukernik, G. Utermann, J. Pritchard, G. Coop, and A. Di Rienzo. 2010. Colloquium paper: human adaptations to diet, subsistence, and ecoregion are due to subtle shifts in allele frequency. *Proc Natl Acad Sci U S A* **107 Suppl 2:** 8924-8930.

Hudson, R.R. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* **18:** 337-338.

Li, J.Z., D.M. Absher, H. Tang, A.M. Southwick, A.M. Casto, S. Ramachandran, H.M. Cann, G.S. Barsh, M. Feldman, L.L. Cavalli-Sforza, and R.M. Myers. 2008. Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319:** 1100-1104.

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