SI Appendix

Balanced Polymorphisms

We have suggested previously that many of the selection events shared between populations may result from ongoing balanced selection (1, 2). Neutral theory predicts that it is extremely unlikely that two alleles will be maintained at high polymorphic frequency by chance. To identify possible balanced selection events, we identified selected clusters that are shared between the CEU and YRI populations. A scatter plot of the 509 identified clusters is shown. As noted (1), 80% of these selected alleles appear at least 2-fold older in the YRI population than in the CEU population. We suggest that the more rapid LDD for these alleles in African populations reflects the ancestral human coalescence under selective balance, whereas in European and Asian populations, ongoing selection "reset" the LD "clock" after bottlenecks resulting from founder effects or migrations. A category of genes that shows moderate overrepresentation in these inferred balanced selection events is pathogen-host interaction (for example AKAP10, IMMP2L, LRBA, LYG2, STAB1, and TCF12). Pathogen defense was likely under constant evolutionary pressure. Alleles in other overrepresented categories noted previously (1) include synaptic transmission (AKAP9, BSN, GLRA2, HTR1F, and RAPSN) and DNA repair (FANCC, RAD51C). Given our desire to analyze the rate of new selection during recent human evolutionary history, these 509 presumably "ancient" alleles were removed prior to calculating Fig. 1.

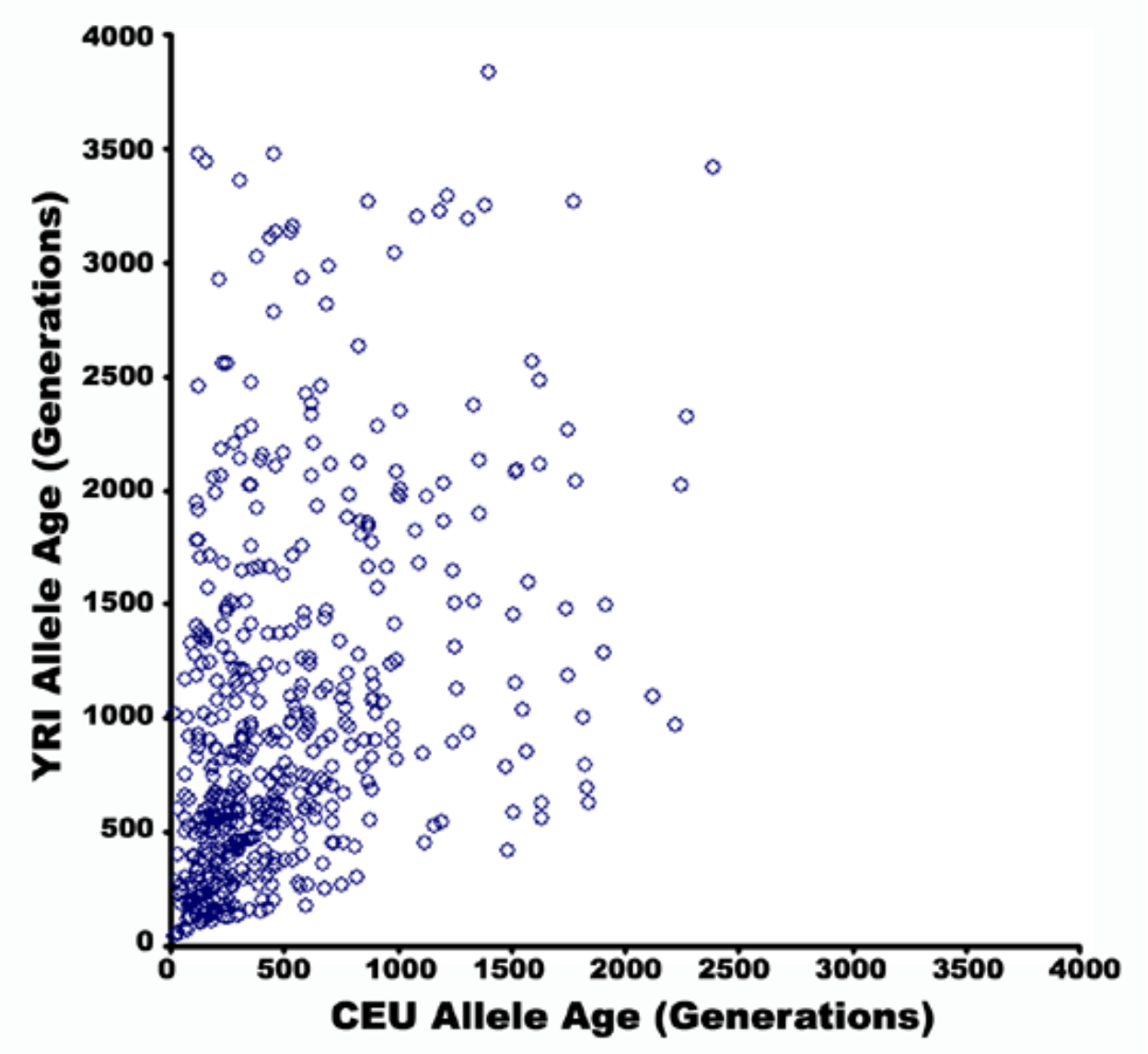

Fig. 4. Scatterplot of allele age in CEU versus YRI samples, for 509 clusters shared in both populations. Out of these 509, 404 are older in YRI than in CEU. The age discrepancy may be explained by the chance lack of long-term selective balances in early migrants from Africa, followed by subsequent dispersal of those old selected alleles into Eurasia.

Pseudohitchhiking

The empirical sum of site heterozygosities (ssh) for autosomal loci in human populations is between 4 × 10-4 (3) and 6 × 10-4 (4). Under neutrality, the mean sum of site heterozygosities (ssh) is expected to be 4Nu, where N is the effective population size, and u is the neutral mutation rate. The rate u has been estimated as approximately 10-9 per site per year for nonrepetitive nocoding DNA (5). This is generally in agreement with other estimates (6), indicating 1.23% divergence of human and chimpanzee genomes (7).

When adaptive substitutions are common, the mean ssh for a neutral locus is a function of the rate of adaptive substitution (ϱ), the rate of recombination (r) between the neutral and selected loci, the neutral mutation rate (u), and the population size (N); the relation is given by equation 11 in ref. 8:

$$\mathrm{ssh} = \frac{4Nu}{1 + 2N\rho y^2} \qquad [3]$$

In Eq. 3, ϱ is the rate of adaptive substitutions in the population at the selected site. Under our null hypothesis, this rate is simply the expected number of adaptive substitutions per genome per generation divided by 3 × 109 sites. For the YRI data, this rate is estimated as 4.4 × 10-9 per site per generation. The value y is the average frequency to which a selective sweep drives a partially linked neutral variant. It decreases as the rate of recombination between selected and neutral sites increases. Betancourt et al. (9) give an approximation for y in terms of the recombination rate, r, valid where r is very small compared with the selection coefficient s (i.e., r ≪ s):

$$y \approx 1 - \frac{r}{hs} \log 2N \qquad [4]$$

In Eq. 4, h is the degree of dominance, here assumed equal to unity. The effect of linkage on ssh is negligible where y is small. We estimated the effect of linked adaptive substitutions on the mean ssh across a region where y > 0.5. Assuming a recombination rate of 10-8 between adjacent sites, N = 104, and hs = 0.02, this encompasses a region of l = 100 kb on either side of the neutral site. The value y2 varies for each site in this region, so that Eq. 3 becomes:

$$\mathrm{ssh} = \frac{4Nu}{1 + 4Nl\rho \sum_{i=1}^{l} y_i^2} \qquad [5]$$

We calculated this for the parameter values listed above. These values provide a conservative estimate, because (i) selection is assumed to be slightly lower than the estimated value and (ii) the effect of linkage is ignored outside the region where $y > 0.5$.

Under neutrality, site heterozygosity should be independent of local recombination rate. But if adaptive substitutions are very common, then regions with low recombination should have reduced neutral variation compared to regions with high recombination. Recombination rates vary across the human genome between zero and 6 cM/Mb or higher, with the bulk of sites showing rates between 0.5 and 3 cM/Mb (10, 11). We found the expected relationship of recombination rate and ssh by finding the solution to Eq. 5, given applicable estimates of y from Eq. 4. We calculated this result at intervals of 0.1 cM/Mb across the range from 0.5 to 3.0, and

from this range of results calculated the expected relationship between local recombination rate and heterozygosity under the null hypothesis.

Demographic Model

We considered population growth during the last 80,000 years, roughly the time period across which the current data can detect selected variants (1). For times after 2500 years ago, we used size estimates for two populations, sub-Saharan Africa and Europe plus West Asia, taken from ref. 12. Europe and West Asia were combined in this demographic model because of the evidence for strong recent gene flow between the two areas (13), which would predict widely shared adaptive variants. For earlier times, we estimated sizes for the same populations as a proportion of global estimates (12, 14, 15), considering known climatic fluctuations and founding times for the different populations (16) (Fig. 3). We assumed exponential growth between point estimates of population numbers; this necessarily oversimplifies what must have been a fluctuating pattern of growth.

In an exponentially growing population, the fixation probability of selected dominant alleles is approximately $2(s + r)$, where r is the intrinsic rate of population growth (17). Through most of the past 40,000 years, growth was very slow, with $r \approx 0.0001$ or less per generation, fluctuating from time to time. The Neolithic intrinsic growth rate was as high as 0.01 or higher, substantially increasing the fixation probability of new selected variants. As above, both ν and $\bar{s}$ can be estimated by fitting Eq. 2 across all age intervals, in this case conditioned on the model of demographic history.

References

1. Wang ET, Kodama G, Baldi P, Moyzis RK (2006) *Proc Natl Acad Sci U S A* 103:135–140.

2. Wang ET, Moyzis RK (2007) *Mutation Res* 616:165–174.

3. Wang D, Fan J, Siao C, Berno A, Young P, Sapolsky R, Ghandour G, Perkins N, Winchester E, Spencer J, *et al*. (1998) *Science* 280:1077–1081.

4. Stephens JC, Schneider JA, Tanguay DA, Choi J, Acharya T, Stanley SE, Jiang R, Messer CJ, Chew A, Han JH, *et al*. (2001) *Science* 293:489–493.

5. Yi S, Ellsworth DL, Li WH (2002) *Mol Biol Evol* 19:2191–2198.

6. Wildman DE, Uddin M, Liu G, Grossman LI, Goodman M (2003) *Proc Natl Acad Sci U S A* 100:7181–7188.

7. The Chimpanzee Sequencing and Analysis Consortium (2005) *Nature* 437:69–87.

8. Gillespie JH (2000) *Genetics* 155:909–919.

9. Betancourt AJ, Kim Y, Orr HA (2004) *Genetics* 168:2261–2269.

10. Hellmann I, Ebersberger I, Ptak SE, Pääbo S, Przeworski M (2003) *Am J Hum Genet* 72:1527–1535.

11. Nachman MW (2001) *Trends Genet* 17:481–485.

12. Biraben JN (2003) *Population Sociétés* 394:1–4.

13. Richards M (2003) *Annu Rev Anthropol* 32:135–162.

14. Coale AJ (1974) *Sci Am* 231:40–52.

15. Weiss K (1984) *Hum Biol* 56:637–649.

16. Hawks JD (1999) *The Evolution of Human Population Size: A Synthesis of Fossil, Archaeological, and Genetic Data* Ph.D. thesis (University of Michigan, Ann Arbor).

17. Otto SP, Whitlock MC (1997) *Genetics*. 146:723–733.