Supplementary Material

Human population dispersal 'Out of Africa' estimated from linkage disequilibrium and allele frequencies of SNPs

Supplementary Figures 1-7

Supplementary Tables 1 -3

Supplementary File 1

Supplementary Figure 1 Discrepancy between T_F and T_{LD} divergence time estimates. There is a significant relationship between an increasing discrepancy between T_F and T_{LD} (measured as T_F/T_{LD}) and increasing T/N_e ratio (where T and N_e are both calculated from LD over 0.1cM) suggesting 'fixation bias' is at least partially responsible for the smaller T_{LD} divergence times. Comparisons involving the African-American population sample (red squares) provide an interesting empirical illustration of the potential differential impact of migration on T_F and T_{LD}. African-Americans are the result of recent admixture between highly divergent European and African parental populations. Comparisons involving the African-American sample have particularly discrepant T_F/T_{LD} ratios indicating migration may particularly bias T_{LD} estimates downwards. However, the Mexican population sample (green triangles), which is also the result of recent migration between European and Amerindian populations, does not show any unusual deviation. Other factors such as natural selection and error in the recombination map could also particularly bias T_{LD}.

Supplementary Figure 2 Comparison of T_F and T_{LD} divergence time estimates. (**A**) There is strong correlation (r^2~0.97) between T_F and T_{LD} over 105 pairwise population estimates excluding those from the recently admixed African-American and Mexican populations (marked by blue diamonds). However, absolute T_{LD} values are generally much less that T_F , which in turn seem to accord better with divergence times estimates from other loci and methods, as well as archaeological and fossil evidence (Stringer 2002). The slope of the line between the two estimators (with an intercept through 0) is 0.48 indicating that T_{LD} values are generally half those of T_F. T_F values are, however, quite similar to those estimated between the original HapMap2 African and European or Asian populations (~20-25 KYA) using a similar LD based approach (Sved et al. 2008). The downward discrepancy between T_{LD} relative to T_F is greatest for comparisons involving the admixed African-American

population and Eurasian populations (ASW, red squares). A similar discrepancy between admixed Mexicans and other populations is, however, not readily apparent (MEX, green triangles). (**B**) The discrepancy is small for closely related populations (T_F<100 generations). The slope of the indicated trend-line is ~0.82.

Supplementary Figure 3 Estimates of N_e t generations ago for each simulated single population scenario. N_e shown is the mean N_e at each t calculated from 100,000 independent replications for each scenario. Scenarios are named *t_r_g* following time of bottleneck (*t*), reduced population size at time of bottleneck (*r*) and rate of population growth following bottleneck (*g*), respectively.

Supplementary Figure 4 Principal Component Analysis (PCA) of 17 global populations. PC1 versus PC2 values derived from ~282K autosomal markers are plotted for 2765 individuals.

Supplementary Figure 5 LD patterns (r^2_{LD}) across recombination distances and populations. See **Supplementary Table 1** for population codes.

Supplementary Figure 6 SNP ascertainment bias. (**A**) LD patterns across physical distance categories from the fully ascertained ENCODE3 sequence data (eight x 100Kb regions), in a subset of 250 randomly chosen HapMap3 individuals (**B**) LD patterns across recombination distances using the genome wide SNPs in the same individuals. See **Supplementary Table 1** for population codes. Although the ENCODE3 data is provisional and the number of pairwise LD observation are relatively low (especially for distance classes >30kb), there is good correspondence between the qualitative patterns revealed in the fully

ascertained versus array genotyped samples. Specifically, the East Asian populations show stronger LD than Europeans as observed in the genome wide

SNP dataset. As we did not have genetic map information for the mainly newly discovered ENCODE3 SNPs we binned pairwise LD values by physical distance classes instead of by recombination distances. The correlation between r^2_{LD} values between the sequenced and genotyped data over distance classes up to 100kb/0.1cM is ~0.97.

Supplementary Figure 7 Diagram demonstrating simulated scenarios for populations undergoing multiple population splits. The three populations are shown with the original (ancestral) N_e, the times (in t generations ago) that population splits were simulated, and the extent of the bottleneck event at the time of the split. r is the proportional reduction in population size at the bottleneck.

References

Stringer, C. 2002. Modern human origins: progress and prospects. *Philos Trans R Soc Lond B Biol Sci* **357**(1420): 563-579.

Sved, J.A., McRae, A.F., and Visscher, P.M. 2008. Divergence between human populations estimated from linkage disequilibrium. *Am J Hum Genet* **83**(6): 737-743.