Supporting Text

Modeling. The regression model relates the Fraction Recombinant Chromosomes (FRC) on the *y*-axis to distance on the *y*-axis using the logistic equation

$$Y = F(X) = \frac{1}{2 + \beta e^{\lambda X}} + \varepsilon(0, \sigma^2), \quad (1)$$

where $\beta > 0$, $\lambda > 0$ for the right-side model and $\lambda < 0$ for the left-side model, with ε representing Gaussian additive noise with mean 0 and variance σ^2. For each locus or group of loci, the shape and offset parameters λ and β are estimated by maximum likelihood methods (maximum *a posteriori* methods are also possible using appropriate priors on λ and β). We derive a model for each primary locus, using the linkage disequilibrium (LD) decay observed for *G6PD* V202M as a starting point for analysis (1). Lambda is inversely proportional to the point of inflection for the sigmoid, β is directly related to the *y*-intercept, and σ represents the mean genome deviation from the model. The deviation term allows for both experimental error and deviations in local recombination rate (2). We initially set a detection threshold, using the LD decay (LDD) of *G6PD* V202M as our "model," at an average log likelihood (ALnLH) >2.6 SD (>99.5th percentile) from the genome average, or 0.61 for the Perlegen data set. The calculated genome-wide Perlegen ALnLH scores exhibit an average of 0.043, but with a standard deviation of 0.22. Hence, an ALnLH of 0.61 represents a highly unusual genetic architecture. While we need to set a threshold high enough to exclude spurious positive results (below), it cannot be so stringent that minor deviations from the model are excluded. A deviation ±0.10 FRC was chosen as a conservative estimate of fluctuation

that can still adhere to the model, yet detect inferred selection in alleles that exhibit LD patterns "shallower" or "broader" than *G6PD*. This stringent ALnLH cut-off excludes some regions that have LD patterns outside the ±0.10 cut-off. We therefore adjusted our probability function to tolerate negative differences less than the expected *G6PD* curve (i.e., "younger" alleles) but still account for positive FRC deviations. While this adjusted probability also detects the extensive LD found in other unusual genomic regions such as inversions, inferred inversions account for <0.2% of the identified regions.

Automation. From the processed Perlegen or International Haplotype Map (HapMap) data set, we ask whether at least one allele of each site has similar LDD to the model (>2.6 SD from the population mean) while the alternative allele is <1 SD from the genome average. Using Eq. **1**, the log likelihood of any allele *D* according to the *G6PD* V202M model *M* is given by,

$$Ln(P(D \mid M)) \propto \sum_i LnP(Y_i^{observed} \mid F(X_i)) = -\frac{N}{2} Ln(2\pi\sigma^2) + \sum_{i=1} \frac{-(Y_i^{observed} - F(X_i))^2}{2\sigma^2} \quad . (2)$$

Here, *N* is the number of neighboring single-nucleotide polymorphisms (SNPs) within a ±500 kb window binned based on the major and minor alleles of the primary locus under consideration. *X* is the vector containing the distance (in bp) of each ith neighbor to the target SNP (e.g. X_i is the distance of SNP *i* from this locus). $F(X_i)$ is the expected frequency according to the model. All values of *X* are weighted equally.

EASE Analysis. We find the nearest gene within a 100 kb radius from dbSNP (build 123). Overrepresentation analysis is performed using the EXPRESSION ANALYSIS SYSTEMATIC EXPLORER (EASE) package (3). EASE uses a robust version of the Fisher Exact Probability Test. It provides a probability for obtaining *x* number of genes in

category *y* in our list as compared to obtaining randomly the same number of genes in category *y* from the whole human genome.

Population Simulations.

To determine whether other population parameters could influence the LDD test for selection, we used the HapMap European ancestry (CEU) population (4). We chose this population for analysis as a typical population in which recent bottlenecks/admixtures have occurred. The CEU chromosome 7 ALnLH values are shown in Fig 6*A*. We generated a randomized chromosome 7 data set from this population by redistributing each SNP independently in a random uniform way and recomputed the ALnLH scores (details in Fig. 6*B*). The above process was repeated 1000 times. Typical scores from a single simulation are presented in Fig. 6*B*. Pairwise LD measurements in D′ statistics were obtained by randomly sampling 150 SNP pairs at distances of 5, 20, 60, or 160 kb. To simulate an extreme bottleneck, original HapMap genotypes of 5 unrelated individuals, retaining all currently observed LD blocks on chromosome 7, were chosen (details in Fig. 6*D*).

In addition to selection, we asked whether there are other mechanisms that could produce these unusual long-range genetic architectures (Figs. 3 and 4). First, one can ask if the threshold is unambiguously different from random. Given that many individual SNPs have ALnLH values calculated, and that the Perlegen and HapMap SNPs were chosen because of high heterozygosity, there is a concern that high ALnLH values could be achieved by chance. Random permutation of the actual Perlegen or HapMap data, then, is an effective way to determine whether a positive (>2.6 SD) score could be obtained by chance. Using such a simulated data set, the probability of reaching above an ALnLH of 0.2 (1 SD) is $P < 0.0007$ and above 0.71 (>2.6 SD) is $P = 0.0$ (Fig. 6*B*). We conclude, as expected, that chance alone cannot produce an ALnLH value of >2.6 SD (Figs. 3 and 4).

The randomized data set (Fig 6*B*) is also a reasonable model for expected human population structure, the "null" model in many population simulations. High-heterozygosity alleles are assumed to be present prior to the major coalescence of humans 50,000-100,000 years ago, and, in the absence of selection, exhibit little LD at distances of >5 kb (5, 6). The computed D′ for our randomized data set (≈0.20 at all distances of >5 kb) is identical to that predicted for a widely used coalescent simulation of human populations, where D' = 0.5 is <5 kb (5, 6). In this model, the human population is assumed to have constant effective size N = 10,000 until 5,000 generations before the present, followed by exponential expansion to 5 billion. This coalescent model has been widely used to predict the inferred LD between common SNPs in the absence of selection or bottlenecks (5, 6). Indeed, it was used as the basis (and rationale) for the Perlegen and HapMap projects (4, 7), since the low LD predicted by this model was experimentally observed in African ancestry populations, while observed LD was greater in European ancestry populations (D' = 0.5 at 60 kb in ref. 5 and the current HapMap data set (4); see Fig. 6*A*). The more extensive LD observed in European ancestry populations was interpreted as the result of an extreme bottleneck, representing an inbreeding coefficient of at least F = 0.2, corresponding to an effective population size of as small as 50 individuals for 20 generations (5).

Given that the random data set is a reasonable approximation of the high-heterozygosity SNP distribution expected for an ancient (50,000-100,000 years) population, one can use it to simulate various population structures that could lead to more extensive LD. In particular, one can estimate if the chosen threshold (Fig. 3) is high enough to eliminate detection of other potential sources of LD, such as population bottlenecks and/or admixture. Two different simulations were conducted to test these possibilities.

First, 162 haplotypes from the randomized HapMap CEU dataset were "infused" with 18 copies of a single haplotype, representing a contribution of 10% (Fig. 6*C*). This extreme

admixture/bottleneck model represents a population of 90 individuals in which 20% of the chromosomes, however, come from a single individual (10% from each homolog). It simulates the effect of both small population size and disproportionate contribution of a particular haplotype on the calculated ALnLH statistic. "Recombinations" were randomly generated from these haplotypes (once per chromosome arm), genotypes were generated in each generation randomly (for 500 generations, or approximately 10,000 "years"), and ALnLH values were calculated. Even for this extreme model, values of ALnLH only exceeded 0.2 (>1 SD) every 1.5 Mb (on average) during the first 10 generations and decreased to 7 Mb (on average) during the remaining 500 simulated generations, corresponding to decay of the original infused haplotype to average segments <100 kb (Fig. 6*C* and 7). More importantly, although there were rare cases where ALnLH reached 0.6 ($P < 0.0001$) during the simulation, the genetic architectures of these "admixed" SNPs were different than the inferred selected alleles (Fig. 8). As expected, while extreme bottlenecks/admixture can produce occasional "blocks" of LD with high ALnLH values, the random nature of the "overlaps" produced by this mechanism exhibit little dependence of LD Decay with distance. An ALnLH of >0.71 (>2.6 SD, Fig. 3) was never observed in this extreme simulation model of population bottlenecks/admixture. Therefore, bottlenecks/admixture of a less extreme (and more likely) size for human populations will also never produce an ALnLH >2.6 SD from the average.

A second simulation was conducted of a bottleneck of only 5 individuals, all of whom contain the currently observed selected LD blocks. Again, this simulation was constructed as an extreme test, so that more likely bottleneck sizes could be evaluated. If evidence for obtaining an ALnLH >2.6 SD for this simulation cannot be obtained, it cannot be obtained simulating less extreme bottlenecks. For this simulation, haplotypes were assembled from actual European ancestry (CEU) HapMap genotypes rather than the randomized genotypes (Fig. 6*D*). Again, recombinations were randomly generated, and genotypes and ALnLH values calculated for a total of 500 generations (10,000 years). While this model simulates a more extreme bottleneck than any likely to exist in European ancestry populations (5, 6), it further tests the decay of the observed unusual

genetic architectures in the absence of selection. As expected, in this simulation observed ALnLH values >0.71 (>2.6 SD) rapidly decay to a level 1/10th that initially observed in actual CEU HapMap data (Fig. 6A) by generation 40 (P = 0.0017; Fig. 6*D*). After 400 generations, essentially no ALnLH values >0.71 were found (P = 0.00006; Fig. 6*D*). We conclude that no plausible bottleneck in European ancestry populations (5, 6) can account for the observed genetic architectures. We can further conclude, as expected, that without ongoing selection, the observed LD blocks rapidly decay in length.

1. Sabeti, P. C., Reich, D. E., Higgins, J. M., Levine, H. Z., Richter, D. J., Schaffner, S. F., Gabriel, S. B., Platko, J. V., Patterson, N. J., McDonald, G. J., Ackerman, H. C., Campbell, S. J., Altshuler, D., Cooper, R., Kwiatkowski, D., Ward, R. & Lander, E. S. (2002) *Nature* **419,** 832-7.
2. Kong, A., Gudbjartsson, D. F., Sainz, J., Jonsdottir, G. M., Gudjonsson, S. A., Richardsson, B., Sigurdardottir, S., Barnard, J., Hallbeck, B., Masson, G., Shlien, A., Palsson, S. T., Frigge, M. L., Thorgeirsson, T. E., Gulcher, J. R. & Stefansson, K. (2002) *Nat Genet* **31,** 241-7.
3. Hosack, D. A., Dennis, G., Jr., Sherman, B. T., Lane, H. C. & Lempicki, R. A. (2003) *Genome Biol* **4,** R70.
4. The International HapMap Consortium (2005) *Nature* **437,** 1299-320.
5. Reich, D. E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P. C., Richter, D. J., Lavery, T., Kouyoumjian, R., Farhadian, S. F., Ward, R. & Lander, E. S. (2001) *Nature* **411,** 199-204.
6. Kruglyak, L. (1999) *Nat Genet* **22,** 139-44.
7. Hinds, D. A., Stuve, L. L., Nilsen, G. B., Halperin, E., Eskin, E., Ballinger, D. G., Frazer, K. A. & Cox, D. R. (2005) *Science* **307,** 1072-1079.