Since

$$\int_{t_0=0}^{\infty} e^{-t_0(1+i(\mu+\rho))}(1-e^{-t_0\rho})e^{-t(1+\mu+\rho)}dt_0 = \frac{\rho e^{-t(1+\mu+\rho)}}{(1+i(\mu+\rho))(1+\rho+i(\mu+\rho))},$$

we can let

$$K_i = \frac{\rho}{(1+i(\mu+\rho))(1+\rho+i(\mu+\rho))}$$

and conclude that

$$B_i(L) = B_{i+1}(L+1)$$

for all $i > 1$, whereas

$$B_1(L) = \sum_{i=1}^{L-1} K_i B_i(L-1).$$

7 Empirical validation using genome sequence data

To measure the accuracy of our predicted $p_L(\text{IBS})$ values, we found the lengths of all maximal ROHs in the eleven human genome sequences referenced in Table 13. The bases were re-called in a consistent fashion with the intent to make the quality good enough for population genetic analysis; out of a total of 33,686,389,482 base pairs, 9,743,948,741 (28.9%) were marked unreliable due to unreliable read mapping, proximity to indels, or other other attributes that made them suspect (see Base Calling Methods appendix), and we deleted these bases before proceeding. Our call sets for all sequences are available for download at ftp://ftp.sanger.ac.uk/pub/rd/humanSequences.

In addition to counting the number $N_{\text{ROH}}(L)$ of ROHs in each genome that are between $(L-1000)$ and L bases long (for L divisibly by 1000), we counted the number $N_{\text{ROH}}(L)_{0.10}$ of L-base regions that appear homozygous when we detect a tenth of all hets. Specifically, we generated *thinned ROHs* whose endpoints are the mutations with positions congruent to zero mod ten relative to the 5' end of the chromosome, referring to these endpoints as *observed hets* as opposed

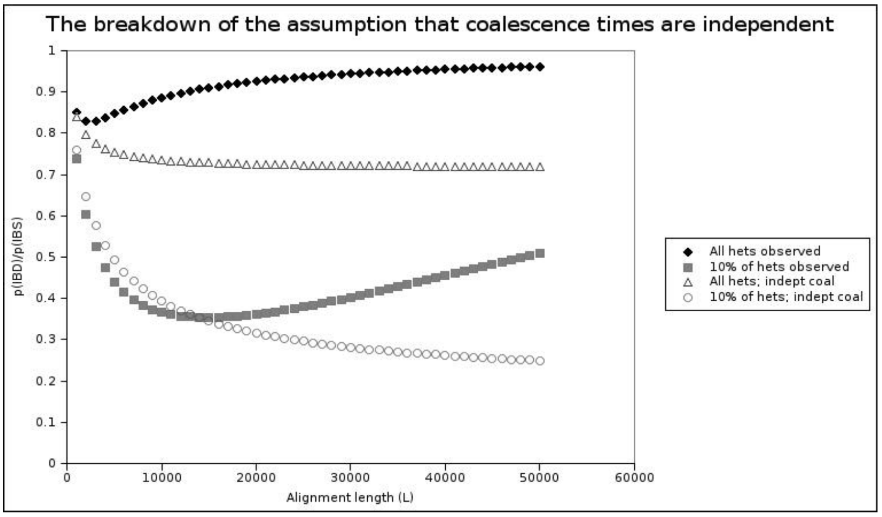

Figure 12: The two solid plots record the exact probability of IBD given IBS (assuming the SMC), the black plot with IBS required at every base in the alignment, and the grey plot where markers allow for only 10% of mutations to be detected. The empty triangles and empty circles record how that probability changes when we disregard linkage between non-IBD markers. The two curves almost never agree when complete sequences are used, in concordance with the fact that earlier methods do not claim to be accurate for such dense marker data.

Sequence Names	Origins
COLO-829-BL	Northern/Western European Ancestry [5]
NA12878, NA12891, NA12892	Northern/Western European Ancestry [1]
NA18507	Yoruba, Nigeria [7]
NA18506, NA18508, NA19239, NA19240	Yoruba, Nigeria (unpublished)
SJK	Korean [4]
YH	Chinese [60]

Figure 13: The eleven genomes used in our analysis

to *hidden hets*. We predict that

$$\frac{N_{\mathrm{ROH}}(L)}{N_{\mathrm{ROH}}(L)_{0.10}} = \frac{10 p_{L\max}(\mathrm{IBS})}{p_{L\max}(\mathrm{IBS})_{0.10}}, \tag{23}$$

adding the factor of 10 to account for the fact that there are ten times as many true ROHs as thinned ROHs (most of the excess ones being short).

Even though we take care to use genome data with a very low error rate, false positive hets (on the order of 1 per 10^5 bases) will present a significant problem for our analysis. We will be estimating the abundance of ROHs up to 10^7 bases long, and there is an overwhelming chance that their homozygosity will be broken up by false positives.

To correct for the breakup of ROHs by false positives, we estimate the false positive frequency f and multiply the measured value of $N_{\mathrm{ROH}}(L)/N_{\mathrm{ROH}}(L)_{0.10}$ by $(1-f/10)^L/(1-f)^L$, reasoning that $(1-f)^L$ is the probability that an L-base ROHs will be broken up by a false positive het. We choose $f = 1.5 \times 10^{-5}$ because $N_{\mathrm{ROH}}(L)/N_{\mathrm{ROH}}(L)_{0.10}$ tends toward $(1-f)^L/(1-f/10)^L$ in each genome as L gets large, while the ratio of thinned to true ROHs should tend toward 1.

The eleven plots of $N_{\mathrm{ROH}}(L)/N_{\mathrm{ROH}}(L)_{0.10}$ versus L cluster clearly by ethnicity (see Figures 14, 15, 16), and we account for the differences by finding effective population size histories that fit $10p_{L\max}(\mathrm{IBS})/p_{L\max}(\mathrm{IBS})_{0.10}$ well in the data from each ethnic group. We also experiment with varying the mutation rate μ, motivated by the fact that the 1000 Genomes consortium recently

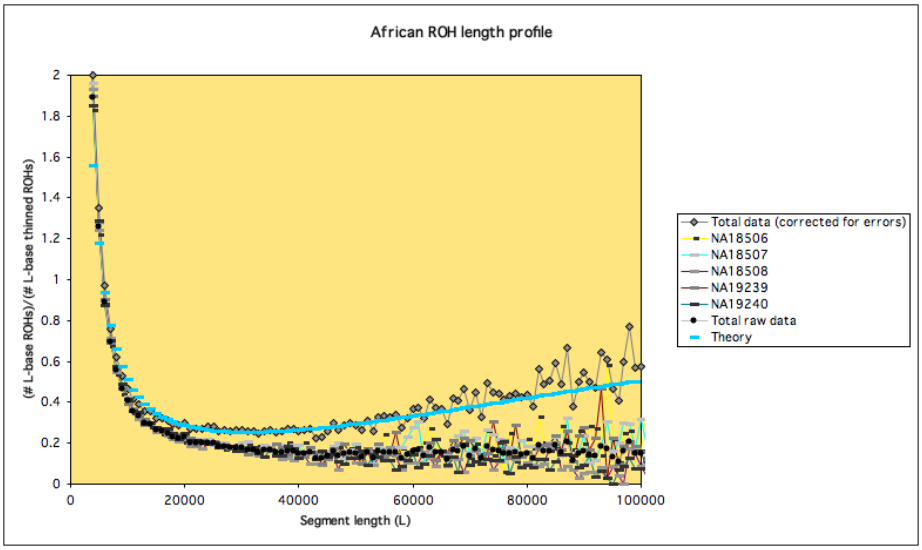

Figure 14: **A. Regions of homozygosity in African genome data** Here, we separately plot $N_{\mathrm{ROH}}(L)/N_{\mathrm{ROH}}(L)_{0.10}$ for each of five African genomes, then average this function across the genomes and correct it for 1.5×10^{-5} false positives per base. In blue is the theory plot $\frac{10p_{L\max}(\mathrm{IBS})}{p_{L\max}(\mathrm{IBS})_{0.10}}$ for a population of constant effective size $N = 14,000$ and a mutation rate of $m = 1.6 \times 10^{-8}$ per base per generation (one of many histories that minimize the sum of square distances from the data points to the predicted curve).

estimated μ to be 1×10^{-8} per base per generation [1] rather than 2.5×10^{-8}.

The measured $N_{\mathrm{ROH}}(L)/N_{\mathrm{ROH}}(L)_{0.10}$ ratios behave noisily for $L > 100,000$, likely because there are few such ROHs in the genome and each one is more likely than a short ROH to include recombination hotspots or other sites where the theory in this paper breaks down. Therefore, we define the best fit population history to be the one that minimizes the sum of squares distance from the predicted $N_{\mathrm{ROH}}(L)/N_{\mathrm{ROH}}(L)_{0.10}$ values to the measured $N_{\mathrm{ROH}}(L)/N_{\mathrm{ROH}}(L)_{0.10}$ values, the sum taken over L ranging from 10,000 to 100,000.

Let $T_{\mu,H}(L)$ denote the theory plot of $10p_{L\max}(\mathrm{IBS})/p_{L\max}(\mathrm{IBS})_{0.10}$ that is obtained a function of the mutation rate μ and the piecewise-constant popula-

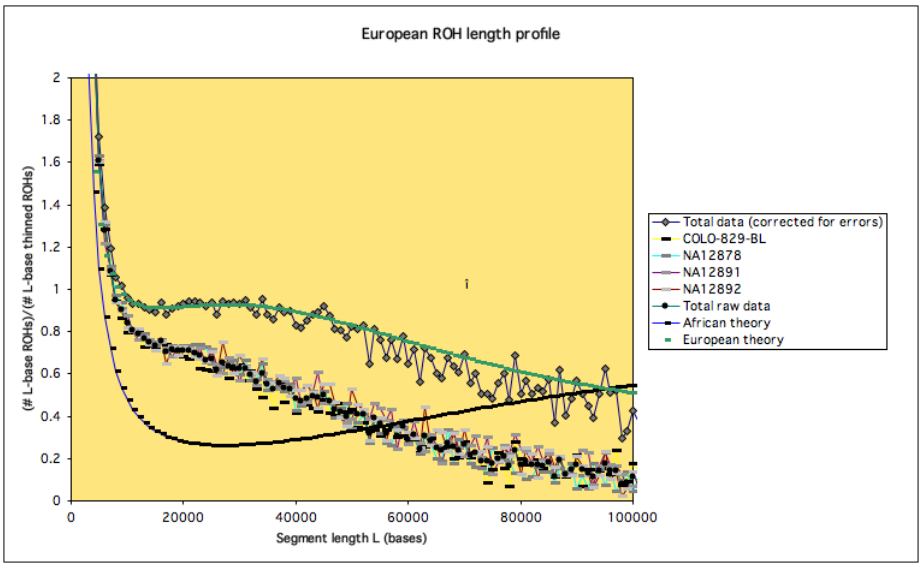

Figure 15: **B. Regions of homozygosity in European genome data** In green is the following single-bottleneck history, with mutation rate $m = 2.5 \times 10^{-8}$: $N = 11,900$, time ranging from 0 to 1240 generations ago (g.a.); $N = 4,530$, $1,240 - 1,770$ g.a.; $N = 15,000 \geq 1,770$ g.a. The African constant population size theory is included in black, for reference.

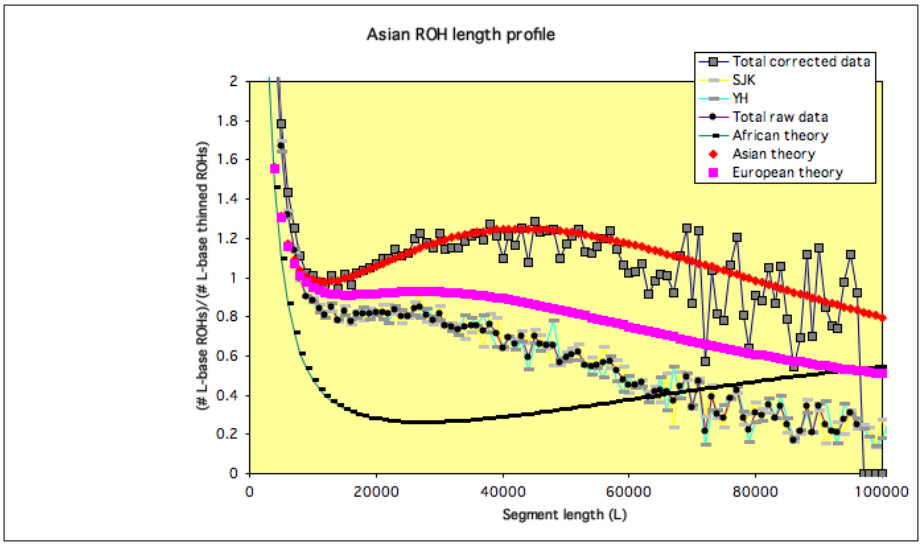

Figure 16: **C. Regions of homozygosity in Asian genome data** The single-bottleneck history shown in red, again assuming $m = 2.5 \times 10^{-8}$, is the following: $N = 8,670$, $0 - 1,380$ g.a.; $N = 1790$, $1,380 - 1,530$ g.a.; $N = 15,000 \geq 1,530$ g.a.. African and European theory plots are included for reference.

tion history H. Likewise, let $D_G(L)$ denote the set of data points $\{N_{\mathrm{ROH}}(10^3 \cdot i)/N_{\mathrm{ROH}}(10^3 \cdot i)_{0.10}\}_{i=1}^{10^3}$ that is obtained by counting all of the ROHs and thinned ROHs in some set of genomes G and correcting for 1.5×10^{-5} false positive hets per base. We measure the goodness of fit between the parameters (μ, H) and the data set G by calculating the sum of squares fit

$$SS(\mu, H, G, L) = \sum_{i=1}^{900} \left(T_{\mu,H}(10^3 \cdot (10+i)) - D_G(10^3 \cdot (10+i))\right)^2.$$

It remains to define a threshold for $SS(\mu, H, G, L)$ below which (μ, H) is deemed a good fit for G. Since $T_{\mu,H}(L)$ is not a straight line, we cannot perform a goodness-of-fit linear regression. We find it logical, instead, to define a threshold that depends on the noisiness of the curve $D_G(L)$, letting

$$N(G, L) = \sum_{i=1}^{(L-1)/1000} \left(D_G(10^3 \cdot (10+i+1)) - D_G(10^3 \cdot (10+i))\right)^2$$

denote the sum of squared distances between adjacent points of $D_G(L)$. If $T_{\mu,H}(L) = \frac{1}{2}(D_G(L) - D_G(L+1))$, making $T_{\mu,H}(L)$ a smoothed version of the data set $D_G(L)$, then

$$SS(\mu, H, G, L) = \frac{1}{4}N(G, L),$$

In each data plot $D_G(L)$, the left portion of the graph is much less noisy than the right portion and therefore provides more information about the mutation rate and population history. In the European genomes, for example, there is so little noise in the data set $D(G)|_{L<34000}$ that

$$\frac{1}{4}N(G_{\mathrm{European}}, 34000) < 0.0094,$$

while

$$\frac{1}{4}N(G_{\mathrm{European}}, 90000) > 0.26.$$

For each of the genome groups $G_{\mathrm{African}}, G_{\mathrm{European}}$, and G_{Asian}, we define L_{short} to be the largest L satisfying $\frac{1}{4}N(G, L-1) < 0.01$ and define L_{long} to be the longest $L \geq 1000$ satisfying $\frac{1}{4}N(G, L-1) < 0.5$ (specific values of L_{short} and L_{long} are recorded in Table 17). We then say that (μ, H) is a good fit for G if

$$SS(\mu, H, G, L_{\mathrm{short}}) < 0.01$$

Ethnicity	L_{short}	L_{long}
African	53000	90000
European	35000	89000
Asian	21000	60000

Figure 17: Thresholds for low noise ($N(G, L_{\text{short}} - 1) < 0.01$) and medium noise ($N(G, L_{\text{long}} - 1) < 0.2$) in the $D_G(L)$ ROH data sets. Our Asian data set, which contains half as many genomes as the others, appears commensurately noisier.

and

$$SS(\mu, H, G, L_{\text{long}}) < 0.5.$$

We searched for good parameter fits using a Monte Carlo Markov chain approach, beginning with a search of constant population size histories. As expected, the Africans are the only group for which we find good constant population size histories. Such histories fall within a narrow parameter space, namely $13,000 \leq N \leq 15,000$ and $1.55 \times 10^{-8} < m < 1.7 \times 10^{-8}$.

When we allow for a single population expansion or contraction, we find a large variety of histories that fit the African data well, though we still find no fits for the European or Asian data. These good African histories are all expansions when $m = 2.5 \times 10^{-8}$, all contractions when $m = 1 \times 10^{-8}$, and close to constant for $m = 1.75 \times 10^{-8}$, with a population size change in either the very recent past or the very distant past (see Figures 18, 19, and 20).

To find good theory fits for the European and Asian data, it was necessary to invoke a population bottleneck. To speed up our MCMC search, we fixed the mutation rate $m = 2.5 \times 10^{-8}$ and the ancestral population size $N_3 = 15,000$. This left two variable time parameters and two variable size parameters, enough to generate many optimal histories to fit both sets of non-African data (see Figures 21 and 22). In both sets of good histories, recent good-fit bottlenecks are shallower than ancient good-fit bottlenecks. The modern effective population size is lower, on average, in the Asian histories. This fits with the fact that the

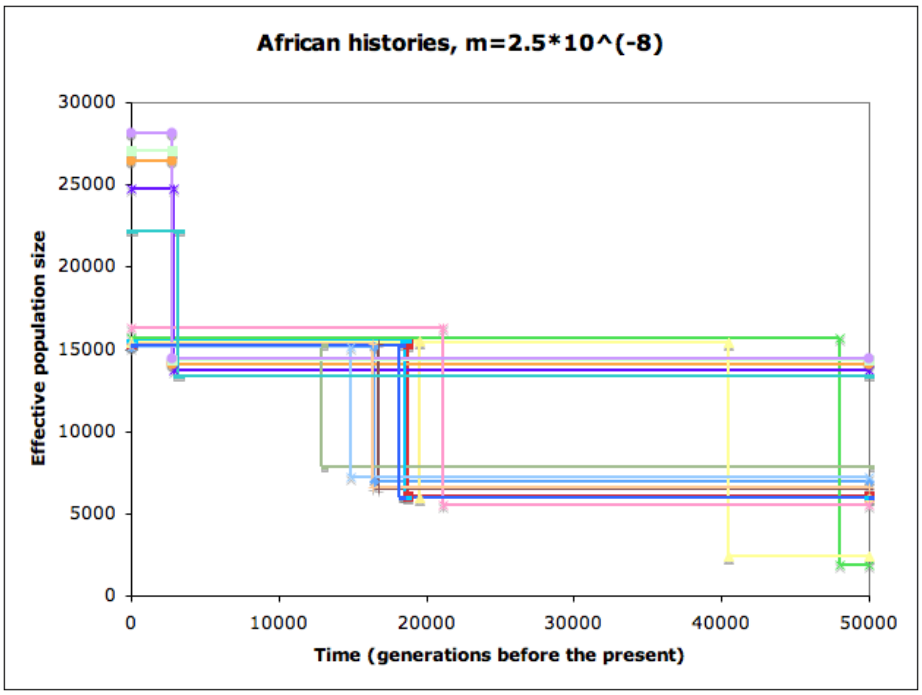

Figure 18: Some population histories satisfying our good fit criterion for African genome data assuming $m = 2.5 \times 10^{-8}$ mutations per base per generation

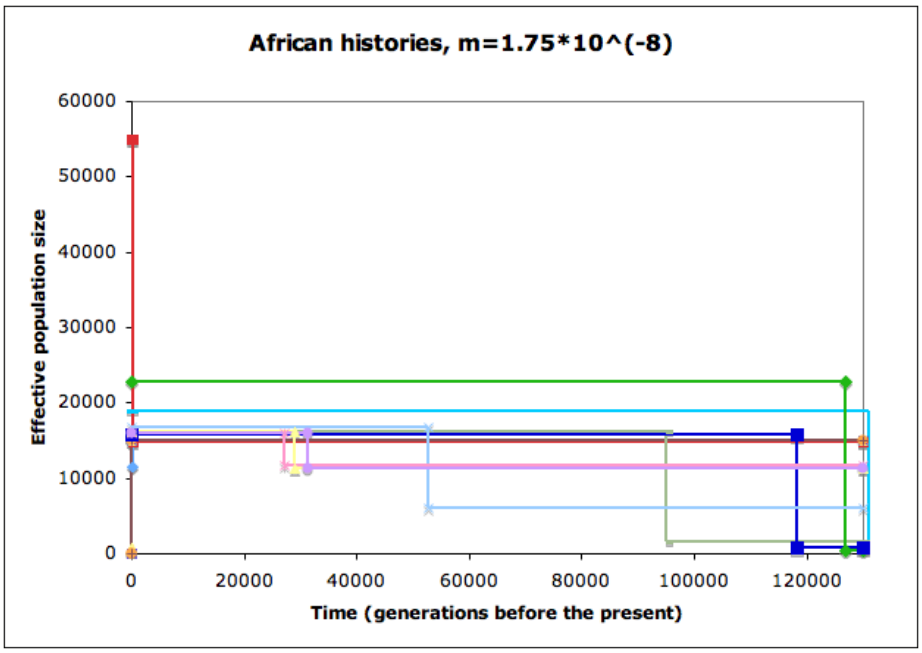

Figure 19: More good fit YRI histories; mutation rate $m = 1.75 \times 10^{-8}$ per base per generation

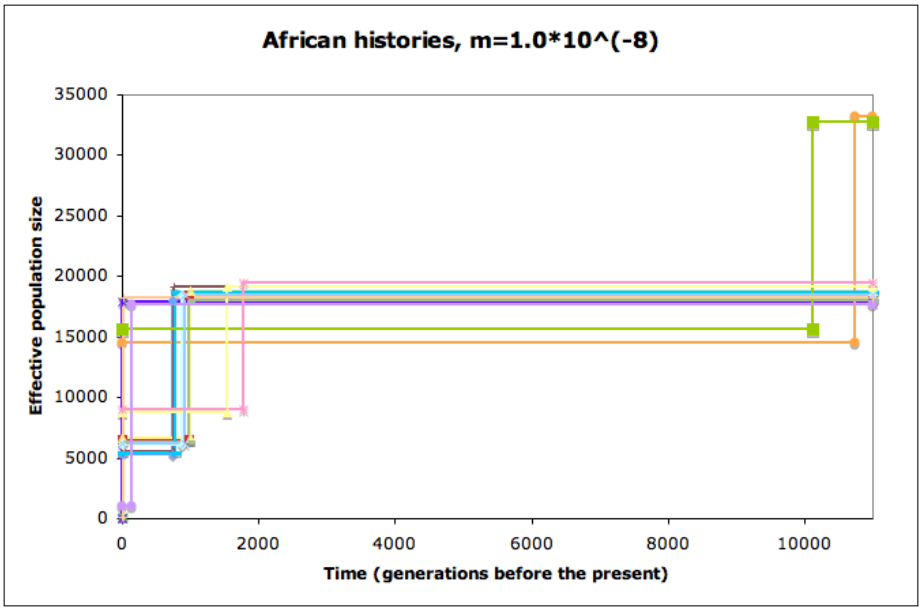

Figure 20: More good fit YRI histories, mutation rate $m = 1.0 \times 10^{-8}$ per base per generation

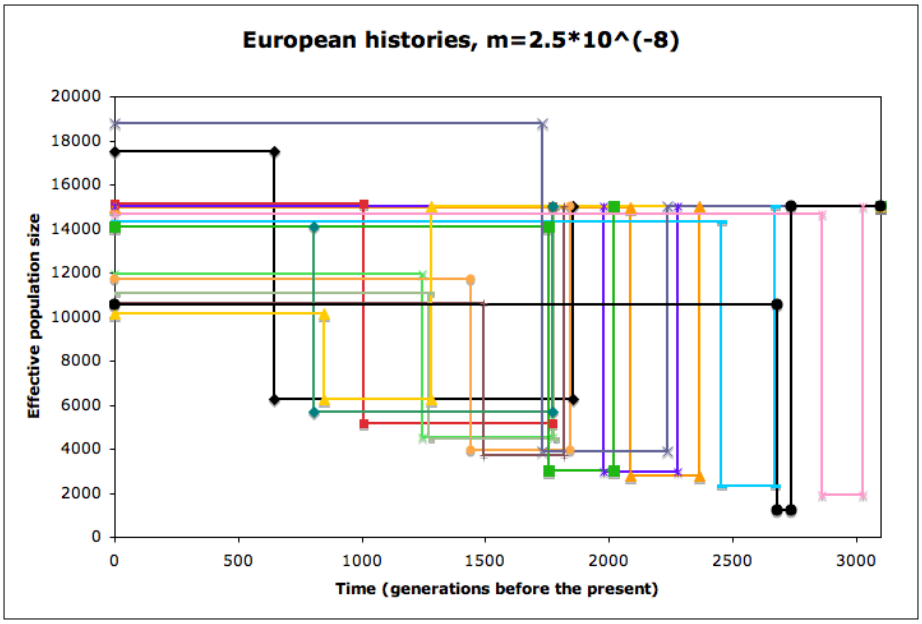

Figure 21: A population bottleneck is required to fit the CEU ROH data–either a shallow, recent bottleneck or a deeper, more ancient bottleneck.

Asian HapMap allele frequency spectrum shows more evidence of genetic drift than the European HapMap allele frequency spectrum [30].

It remains an open problem to mathematically describe the set of histories that fit the distribution of ROHs in each ethnic group. However, in showing that such histories exist, we achieve our aim of predicting the length distribution of ROHs in real genome data and validating the theory that we use to compute IBD probabilities.

8 Discussion

In this paper, we attempt to very precisely model patterns of linkage disequilibrium in genetic data, capturing its decay over long regions of the genome instead of assuming that certain blocks or pairs of loci assort independently. Rather than adding a new LD model to the myriad that exist already, we work