An MCMC Algorithm for Haplotype Assembly from Whole-genome Sequence Data: Supplementary Figures

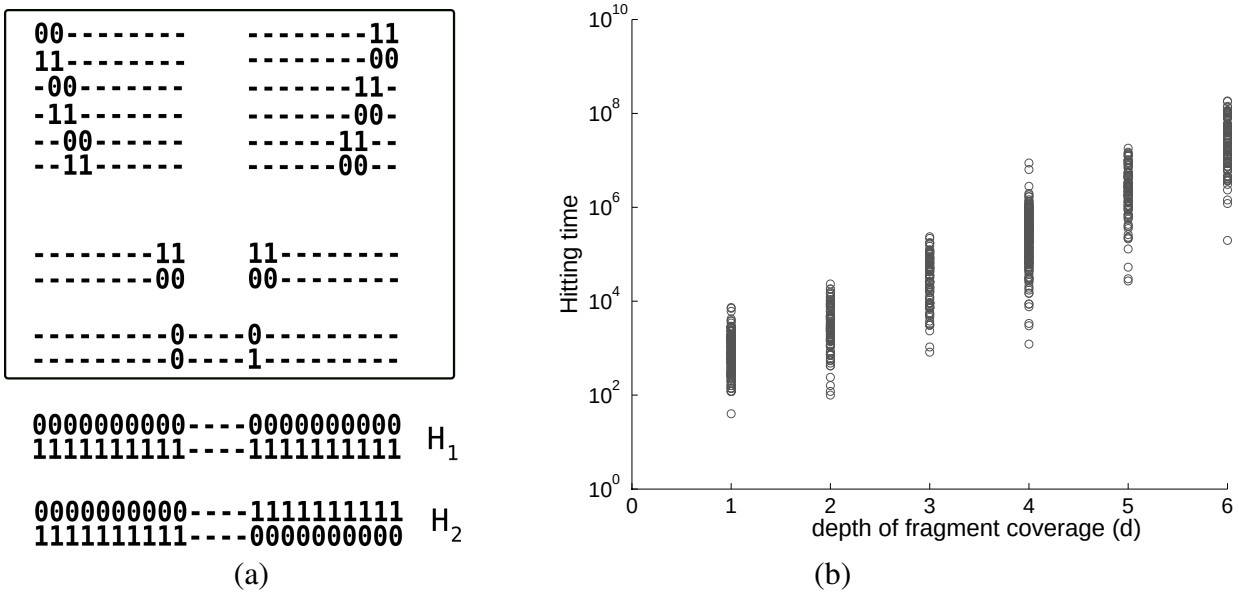

Figure S1: a) Example of a fragment matrix $X(n, d)$ for which two haplotypes H_1 and H_2 have equal likelihood. The matrix has n columns with each pair of adjacent columns (except the pair $(n/2, n/2 + 1)$) covered by d fragments ($d = 2$ and $n = 20$ as shown). b) Plot of hitting time (number of steps taken by the Markov chain $\mathcal{M}(\Gamma_1)$ to go from H_1 to H_2) as a function of the depth of fragment coverage (d). For each d value, the Markov chain was run 100 times (each value represented as a circle) with $\hat{q} = 0.05$.

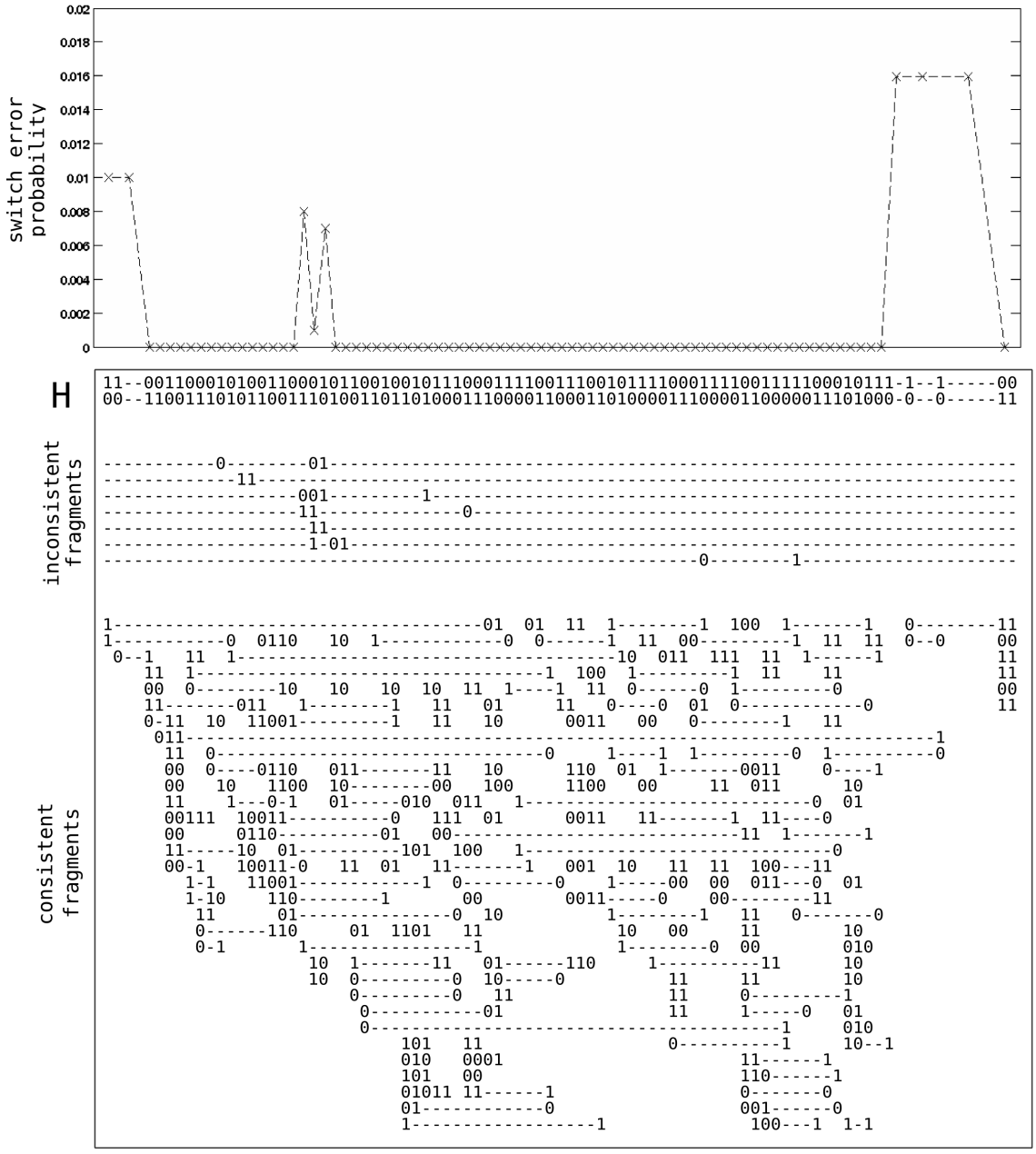

Figure S 2: An example of a fragment matrix for a haplotype segment from chromosome 22 of HuRef and switch error probabilities computed using samples from the MCMC algorithm. Shown in the lower part of the figure, is a fragment matrix and a haplotype pair H. "Inconsistent fragments" (one on each line) correspond to fragments that are inconsistent with H. "Consistent fragments" (multiple independent fragments on each line, two independent fragments separated by whitespace) that perfectly match one of the two haplotypes in H. This example illustrates two common features of the HuRef data relevant for haplotype assembly: (i) the "gapped" nature of the fragment matrix, i.e. the presence of links between non-adjacent variants, (ii) haplotypes do not always link all variants that they span. Shown above the haplotype pair, is a plot of the switch error probabilities for each pair of adjacent columns in the haplotype pair H computed using the MCMC algorithm.

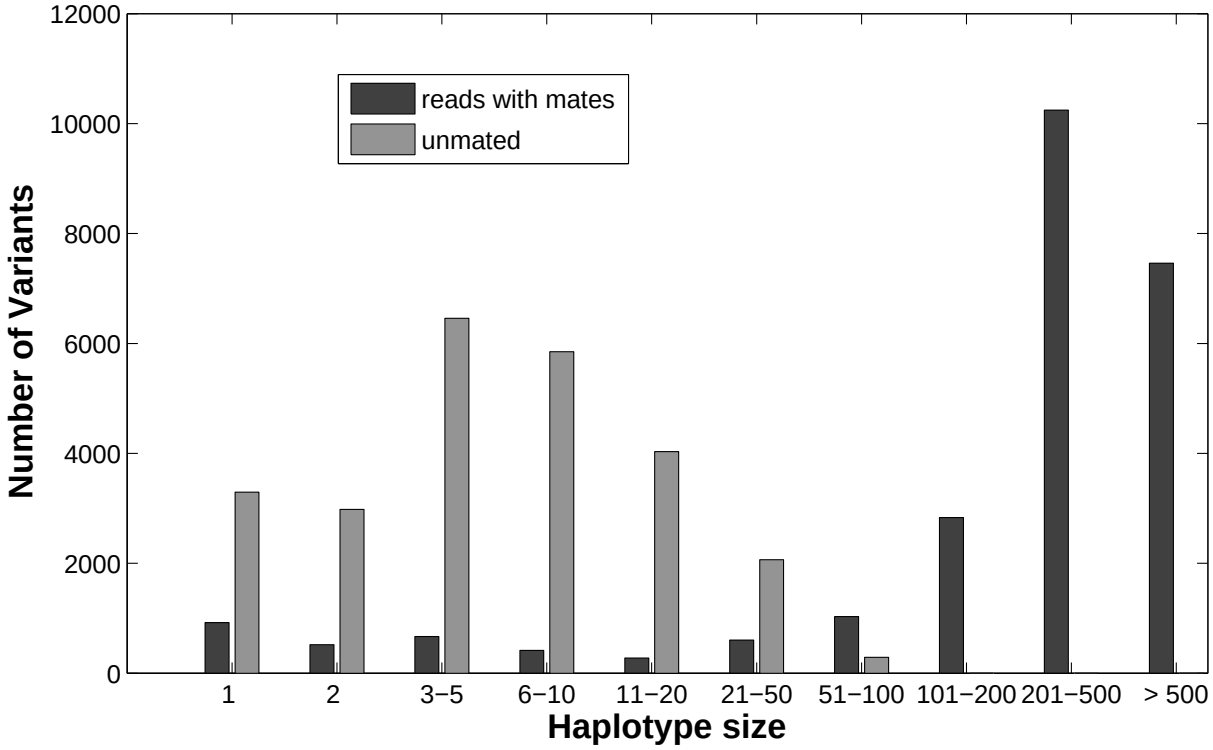

Figure S3: Distribution of the number of variants among haplotypes of different sizes (haplotype size is measured as the number of variants linked together in a haplotype) for chromosome 22 of HuRef. The y-axis is the aggregated number of variants that are part of haplotypes of a certain size. Haplotype size '1' corresponds to isolated variants not connected to any other variant. The 'reads with mates' distribution corresponds to the complete fragment matrix. The 'unmated' distribution is obtained by splitting mate pairs into separate fragments.

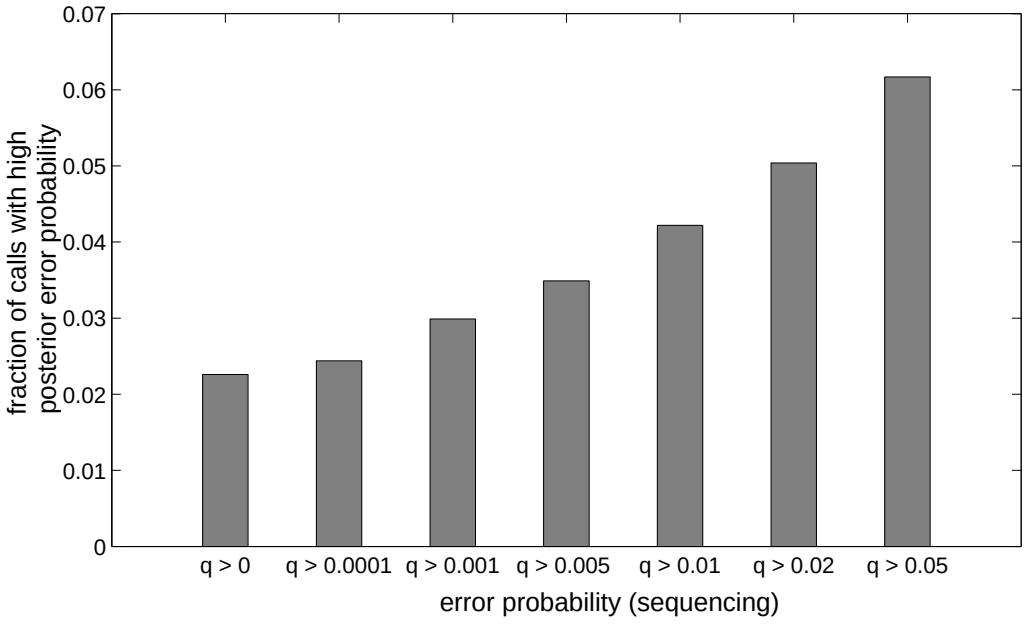

Figure S4: Fraction of variant calls with high posterior error probability (≥ 0.5) for different values of the error probability q (derived from the sequencing quality values) for chromosome 22 of HuRef.