Effect of Genetic Divergence in Identifying Ancestral Origin using HAPAA

Andreas Sundquist*, Eugene Fratkin*, Chuong B. Do, Serafim Batzoglou

Department of Computer Science, Stanford University, Stanford, CA 94305, USA

{asundqui, fratkin, chuongdo, serafim}@cs.stanford.edu

Figures

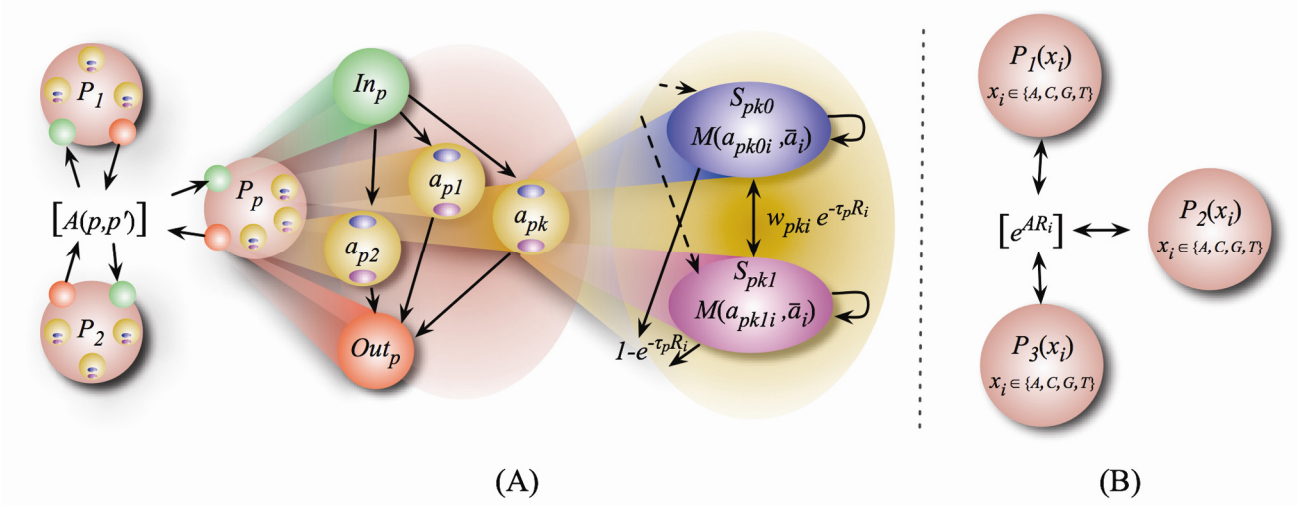

Figure 1: (A) *Hierarchical HMM state diagram for HAPAA.* On the left, inter- and intra-population transitions occur with probabilities governed by matrix $A(p, p')$. In the middle, each population P_p has a similar structure: entry state In_p transitions with uniform probability to a diploid model individual a_{pk}, then to exit state Out_p. On the right, in a_{pk} we transition into one of two states representing the haplotypes S_{pk0} and S_{pk1} of model individual a_{pk} with equal probability. Each haplotype emits its alleles a_{pkhi} via a mutation/error probability distribution $M(a_{pkhi}, \bar{a}_i)$. Haplotypes transition to each other with probability proportional to the phase switch error w_{pki}, and transition out of the diploid sample with probability governed by genetic distance to the next locus R_i and the population-specific recombination rate parameter τ_p. (B) *HMM state diagram for previous methods.* Each state represents a population and emits alleles according to frequency estimates for the populations, and admixture transition probabilities depend on the degree of admixture expected and other learned parameters. By construction, these methods assume a greater degree of independence between adjacent loci.

**These authors contributed equally to the work.*

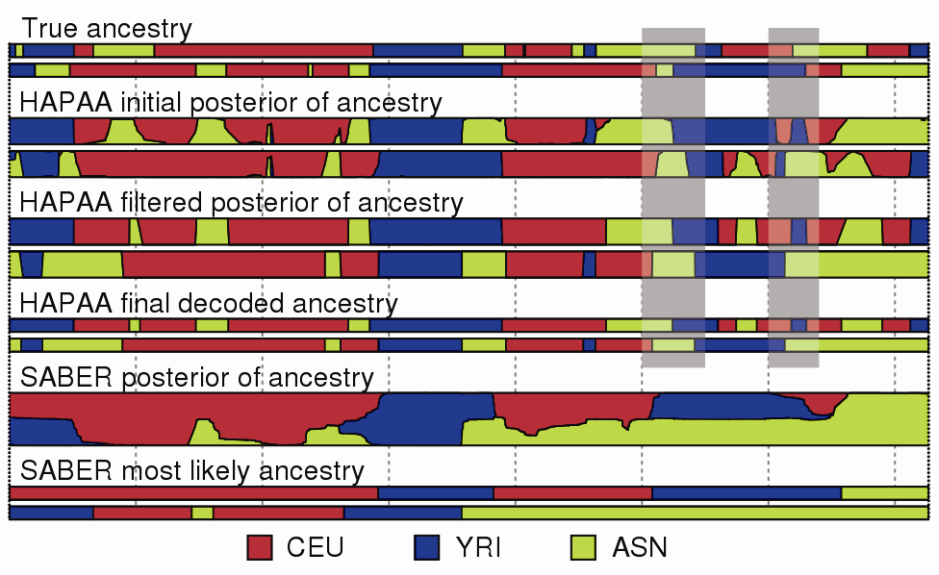

Figure 2: *Example inference on chromosome 22 of an individual admixed between three HapMap populations.* The top two tracks represent the true ancestries, followed by three stages of HAPAA processing, and finally posterior probabilities and Viterbi decoding by SABER. The gray bars highlight two locations with correctly inferred ancestry but with phase switching errors between the haplotypes.

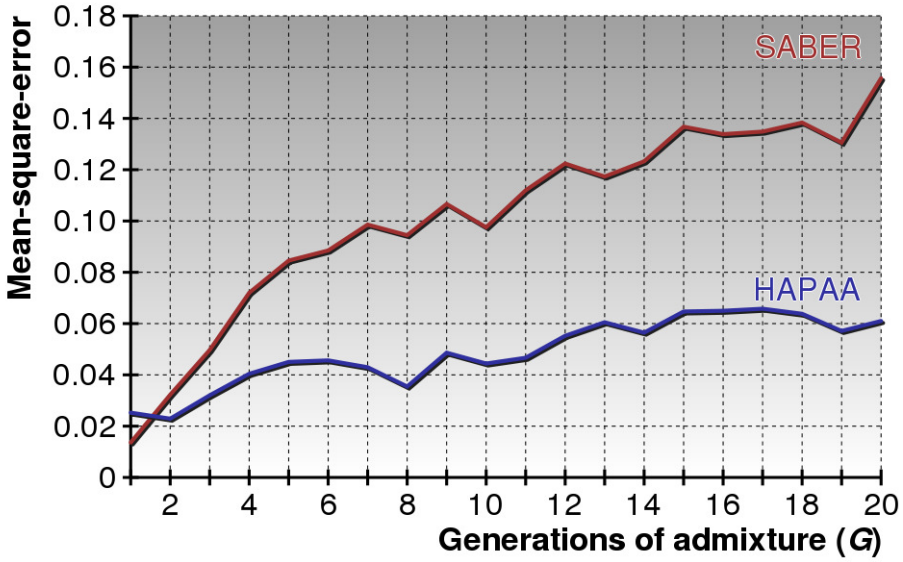

Figure 3: *Performance comparison between HAPAA and SABER.* We measured the mean-square-error of the inferred posterior probability of population ancestry on chromosome 22 for a varying number of generations of admixture. Tests were constructed by simulating admixture over G generations from 2^G individuals selected randomly from three HapMap populations.

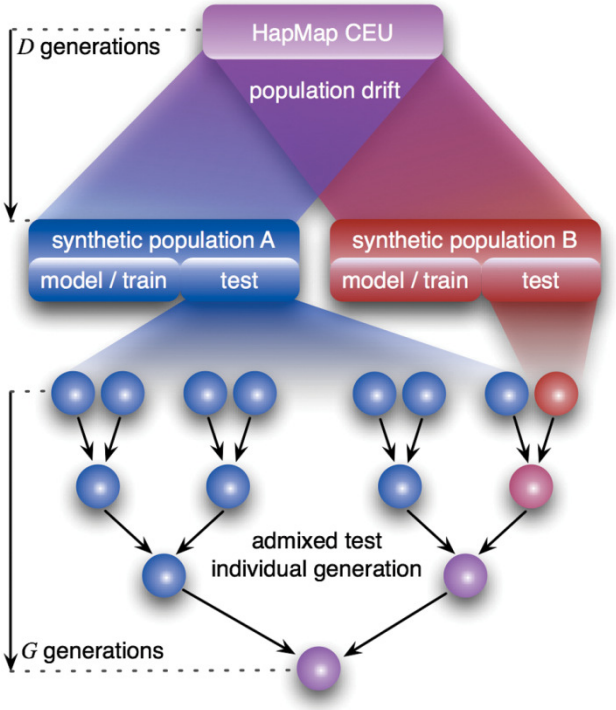

Figure 4: *Methodology for studying the effect of genetic divergence on ancestry inference.* We simulate pairs of randomly mating populations of fixed size 5,000 derived from the HapMap CEU population over D generations. We construct training and test individuals derived over G generations of admixture from $2^G - 1$ ancestors from one population and one ancestor from the other (minor) population.

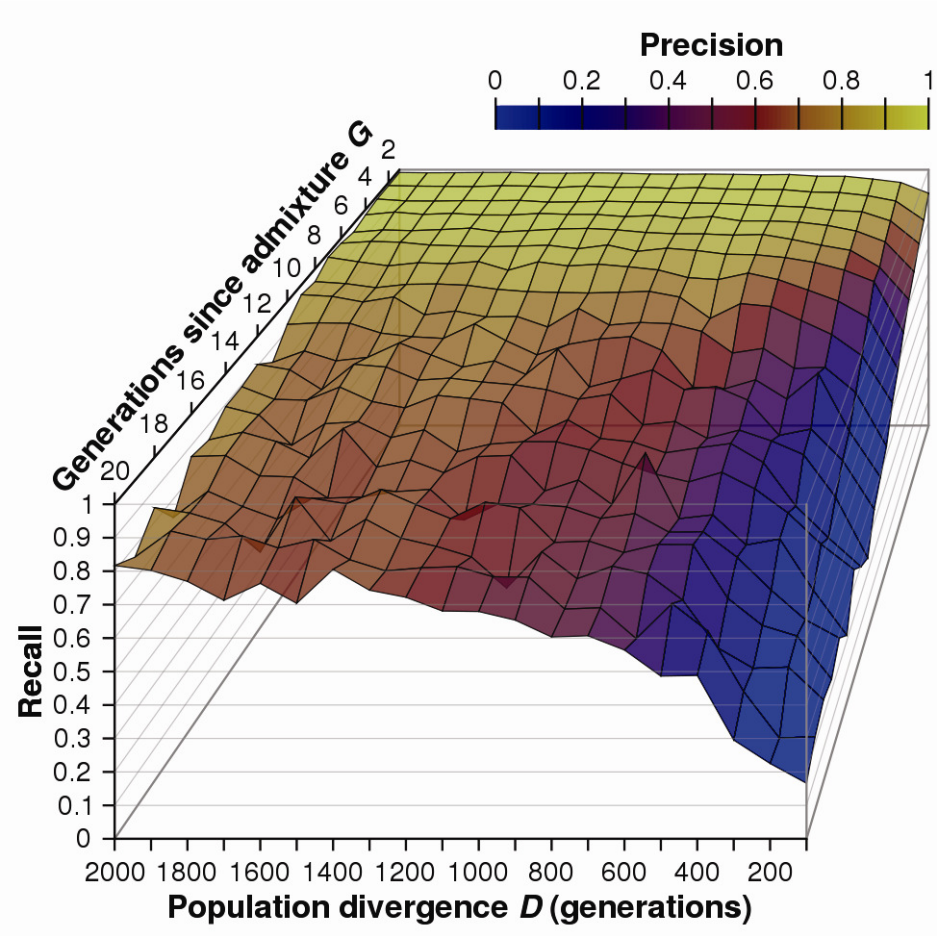

Figure 5: *Recall and precision of detecting minor population.* We simulated twenty pairs of populations separated by $D \in \{100,200,\ldots,2000\}$ generations of drift on the whole genome of Illumina 550K loci. For each D we constructed test individuals that were derived over $G \in \{1,2,\ldots,20\}$ generations of admixture from $2^G - 1$ ancestors from one population and one ancestor from the other (minor) population. Conditioned on the existence of at least one haploblock derived from the minor population, we measure the ability of HAPAA to identify these loci.

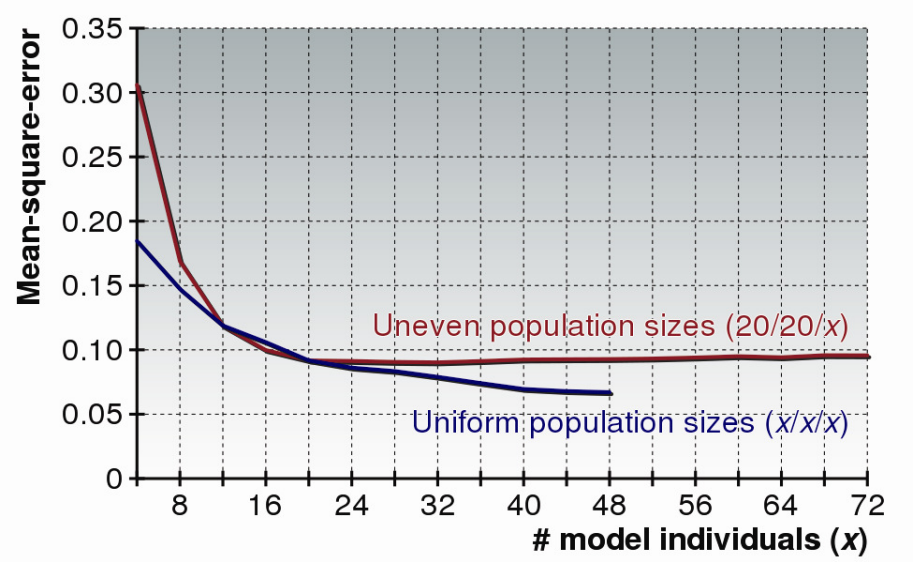

Figure 6: *Performance of HAPAA when varying the number of model individuals.* We created models with a varying number of individuals derived from three populations in the HapMap dataset within chromosome 22. For the *Uniform population sizes* we randomly picked $x \in \{4,8, \ldots ,48\}$ individuals to model each population, while for the *Uneven population sizes* we picked 20 individuals from CEU and YRI and $x \in \{4,8, \ldots ,72\}$ individuals from the ASN population. We benchmarked the mean-square-error performance of HAPAA on 1,000 test individuals admixed over $G \in \{1,2, \ldots ,20\}$ generations from the three populations.

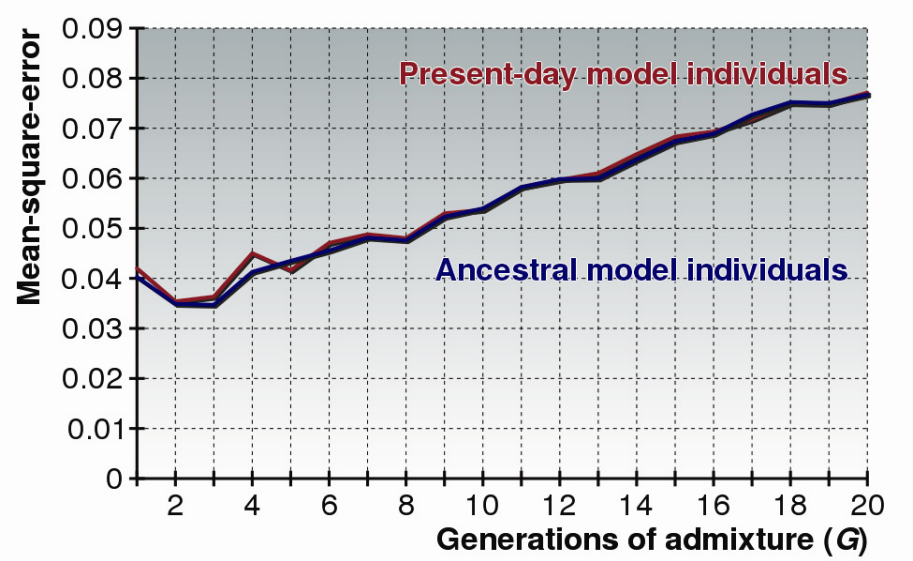

Figure 7: *Performance of HAPAA when using ancestral versus present-day model individuals.* For each $G \in \{1,2, \ldots ,20\}$ we constructed (1) a set of unrelated ancestral individuals by randomly selecting 45 from each HapMap population, and (2) 45 unrelated present-day model individuals for each population. Present-day individuals were generated over G generations of random mating using HapMap samples as templates. We tested on chromosome 22 on individuals admixed from the three populations over G generations using the mean-square-error metric.