Improved Bayesian inference of hybrids using genome sequences

Sneha Chakraborty[1,2] and Bruce Rannala[2]

[1]Department of Ecology and Evolutionary Biology, University of California, Los Angeles.

[2]Department of Evolution and Ecology, University of California, Davis.

Contributing authors: chakraborty@g.ucla.edu; brannala@ucdavis.edu;

ABSTRACT

A Bayesian hybrid inference method is developed which infers hybrids and backcrosses across two generations using sampled genomes from two populations. The method improves on that of Chakraborty and Rannala (2023) by accounting for uncertainty of population haplotype frequencies and correctly marginalizing over haplotypes while still modeling linkage and recombination across the genome. In analyses of simulated data the new method produced posterior probabilities nearly identical to the method of Chakraborty and Rannala (2023) when sample sizes were large. For small sample sizes, posterior probabilities produced by the new method tended to be lower as expected since it accounts for additional uncertainties of population haplotype frequencies. Statistical performance of the new method as measured by the ROC (Receiver Operation Characteristic) curve, appears equivalent to that of Chakraborty and Rannala (2023). The new method is applied to three recently published datasets for populations of kiwifruit (genus *Actinidia*), plateau fence lizard (*Sceloporus tristichus*) and puma (*Puma concolor*).

Keywords: hybrid inference; Bayesian inference; posterior predictive distribution; recombination; whole genome sequence data.

INTRODUCTION

The problem of identifying recent hybrids using genomic data is fundamental to many areas of biology such as conservation genetics (Fitzpatrick *et al.*, 2015), population genetics (Abbott *et al.*, 2016), speciation theory (Barton, 2001), and the study of invasive species (McGaughran *et al.*, 2023). Genomic datasets potentially provide powerful information for identifying hybrids but require a realistic model of hybrid transmission genetics to obtain accurate likelihoods and posterior probabilities. The Bayesian hybrid inference algorithm previously developed by Chakraborty and Rannala (2023) treats point estimates of population haplotype frequencies derived from the posterior mean of a sample as proxies for the unknown population haplotype frequencies and ignores uncertainties of these estimates in calculating likelihoods and posterior probabilities of hybrid genealogical classes. The theory developed in Chakraborty and Rannala (2023) was an improvement over an existing composite likelihood method, NewHybrids (Anderson and Thompson, 2002) in that it accounted for linkage between SNPs. We extend the method of Chakraborty and Rannala (2023) to account for the uncertainty of population haplotype frequencies by analytical integration under a conjugate prior. This produces posterior probabilities of individual assignments to hybrid genealogical classes conditional on a finite sample of individuals from each source population. Chakraborty and Rannala (2023) also used a method to marginalize over haplotypes (when 2 or more recombinations occur) that was biologically unrealistic. Essentially, they treated each contiguous sub-haplotype as if it were independently sampled from the population, which is not the case; the sub-haplotype (including all segments whether contiguous or not) should be sampled jointly. We also present here a new more realistic algorithm for correctly marginalizing over haplotypes. We use simulations to explore the effects of population sample sizes on assignment probabilities and compare the posterior assignments with those obtained using the Chakraborty and Rannala (2023) method. We also use the new method to analyze example empirical datasets for kiwifruit, lizards and puma.

THEORY

We consider a scenario where a sampled individual may be admixed between two populations A and B. Samples of pure individuals are available from populations A and B which we refer to as the *reference samples*. Here we develop the theory for a single chromosome; this is extended later by adding an additional subscript to indicate the particular chromosome being considered. Let $\mathbf{f}^k = \{f^k(h)\}$ be the haplotype frequencies in population $k \in \{A, B\}$, where $f^k(h)$ is the frequency of the h-th distinct haplotype. Let N_k be the number of diploid individuals in a phased reference sample from population k. It is assumed that H distinct haplotypes exist, each occurring in both populations. The set of distinct haplotypes compatible with genotypes observed in all sampled individuals (from both populations) provides an estimate of H. With no prior information, we assign equal prior probability density to the haplotype frequencies in each population (A or B), so the frequencies are $\mathbf{f}^k \sim$

Dirichlet$(1/H)$. The prior probability density of haplotype frequencies (i.e., before collecting reference samples) in population k is

$$p(\mathbf{f}^k) = \prod_{h=1}^{H} \frac{f^k(h)^{(1/H)-1}}{\Gamma(1/H)}. \tag{1}$$

Let $\mathbf{n} = \{\mathbf{n}_A, \mathbf{n}_B\}$ where $\mathbf{n}_k = \{n_{1k}, \ldots, n_{Hk}\}$ and n_{hk} is the observed number of copies of the hth distinct haplotype in a reference sample from population k. The probability mass of $\mathbf{n}_k$ conditioned on the haplotype frequencies follows a Multinomial distribution given by

$$\Pr(\mathbf{n}_k|\mathbf{f}^k) = \binom{2N_k}{n_{1k}, \ldots, n_{Hk}} \prod_{h=1}^{H} f^k(h)^{n_{hk}}, \tag{2}$$

where $2N_k = \sum_{h=1}^{H} n_{hk}$ is the total number of haplotypes sampled in population k. The posterior density of the haplotype frequencies, conditioned on the haplotypes observed in a reference sample from population k, follows a Dirichlet distribution, since the Dirichlet is a conjugate prior to a Multinomial distribution. The posterior probability density of haplotype frequencies is

$$p(\mathbf{f^k}|\mathbf{n}_k) = \Gamma(\theta_k) \prod_{h=1}^{H} \frac{f^k(h)^{\theta_k a_{hk} - 1}}{\Gamma(\theta_k a_{hk})}, \tag{3}$$

where, $\theta_k = 1 + 2N_k$ and $a_{hk} = \frac{n_{hk}+1/H}{\theta_k}$. Note that $\sum_{h=1}^{H} a_{hk} = 1$.

Posterior Predictive Distribution

To calculate the likelihoods of the individual diplotypes, accounting for uncertainties of haplotype frequencies, we require the *posterior predictive distribution*. The posterior predictive distribution of the observed haplotype counts, $\tilde{\mathbf{n}} = \{\tilde{n}_1, \ldots, \tilde{n}_H\}$ from a sampled individual diplotype, conditioned on the reference sample haplotype counts $\mathbf{n}_k$ from population k is

$$\Pr(\tilde{\mathbf{n}}|\mathbf{n}_k) = \frac{\Gamma(\tilde{\theta})\Gamma(\theta_k)}{\Gamma(\theta_k + \tilde{\theta} - 1)} \prod_{h=1}^{H} \left[\frac{\Gamma(\tilde{n}_h + n_{hk} + 1/H)}{\Gamma(\tilde{n}_h + 1)\Gamma(n_{hk} + 1/H)} \right]. \tag{4}$$

The log probability is

$$\begin{aligned} \log \Pr(\tilde{\mathbf{n}}|\mathbf{n}_k) = & \log \Gamma(\theta_k) - \sum_{h=1}^{H} \log \Gamma(n_{hk} + 1/H) \\ & + \log \Gamma(\tilde{\theta}) - \sum_{h=1}^{H} \log \Gamma(\tilde{n}_h + 1) \\ & - \log \Gamma(\theta_k + \tilde{\theta} - 1) + \sum_{h=1}^{H} \log \Gamma(n_{hk} + \tilde{n}_h + 1/H) \end{aligned} \tag{5}$$

where, $\tilde{\theta} = 1 + \sum_{h=1}^{H} \tilde{n}_h$.

Data and Parameters

Following Chakraborty and Rannala (2023), we consider a diploid individual with K chromosomes. Chromosome i contains L_i loci with phased biallelic single-nucleotide polymorphisms. The maternally (M) and paternally (P) inherited chromosomes are defined in the form of matrices:

$$\mathbf{x}^M = \{x_{ij}^M\},$$
$$\mathbf{x}^P = \{x_{ij}^P\},$$

where $\mathbf{x}_i^M$ and $\mathbf{x}_i^P$ are haplotypes for maternally and paternally inherited copies of chromosome i and $x_{ij}^M \in \{0,1\}$ (or, $x_{ij}^P \in \{0,1\}$) is the allele (coded as 0,1) present at the jth SNP locus on the maternally (M) (or, paternally (P)) inherited copy of chromosome i. The diplotype is the pair of haplotypes on homologous chromosomes, denoted as $\mathbf{x} = \{\mathbf{x}^M, \mathbf{x}^P\}$.

For chromosome i, we assume that H distinct haplotypes exist, each occurring in both populations A and B. The reference population haplotypes are defined in terms of $H \times L_i$ matrices:

$$O_A^{(i)} = O_B^{(i)} = \{o_{hj}^{(i)}\}, \quad (\text{for } h = 1, 2, \ldots, H \text{ and } j = 1, 2, \ldots, L_i)$$

where $\mathbf{o}_h^{(i)}$ is the h-th haplotype of chromosome i and $o_{hj}^{(i)} \in \{0,1\}$ is the allele (coded as 0,1) present at the j-th SNP locus on the chromosome i.

Likelihoods for Genealogical Classes

Two populations hybridizing over two generations will allow an individual to be classified into one of 6 distinct genealogical classes based on the number and arrangement of founder population originations. The 4 founders in this framework are considered to have pure ancestry from a reference population and the individual to be classified is at the base of the pedigree (see Figure 1 of Chakraborty and Rannala, 2023). The 6 genealogical classes are as follows: **a** and **d** are purebreds; **b** and **e** are backcrosses; **c** is an F1 hybrid and **f** is an F2 hybrid. We use the term genealogical class and model interchangeably. Here, we present formulas for calculating the likelihood of an individual's diplotype, $\mathbf{x} = \{\mathbf{x}^M, \mathbf{x}^P\}$ under each of the 6 possible genealogical classes. Let $G = \{a, b, c, d, e, f\}$ denote the set of all genealogical classes where $g \in G$. We define the indicator function for the i-th chromosome,

$$I(\mathbf{x}_i) = \begin{cases} 1 & \text{if } \mathbf{x}_i^M \neq \mathbf{x}_i^P \\ 0 & \text{if } \mathbf{x}_i^M = \mathbf{x}_i^P \end{cases} \tag{6}$$

where equality denotes identical haplotypes and inequality differences between haplotypes at one or more positions. We define a bijective function $\boldsymbol{\phi}(\mathbf{x}_i) : \mathbf{x}_i \mapsto \tilde{\mathbf{n}}$ such that $\boldsymbol{\phi}(\mathbf{x}_i) = [\phi_1(\mathbf{x}_i), \ldots, \phi_H(\mathbf{x}_i)]$ where,

$$\phi_h(\mathbf{x}_i) = \sum_{c \in \{M,P\}} I_h(\mathbf{x}_i^c), \qquad \forall\, h = 1 \ldots H \tag{7}$$

$$I_h(\mathbf{x}_i^c) = \begin{cases} 1 & \text{if } \mathbf{x}_i^c = \mathbf{o}_h^{(i)} \\ 0, & \text{otherwise} \end{cases} \qquad \text{and} \qquad \sum_{h=1}^{H} \phi_h(\mathbf{x}_i) = 2. \tag{8}$$

Similarly, we define a unit vector function $\boldsymbol{\phi}^c(\mathbf{x}_i) = [\phi_1^c(\mathbf{x}_i), \ldots, \phi_H^c(\mathbf{x}_i)]$, where $\phi_h^c(\mathbf{x}_i) = I_h(\mathbf{x}_i^c)$. Below we define the log-probability of the likelihood of the diplotype for any chromosome under each of the 6 genealogical classes. Note that the functions U and Q, which appear below and determine haplotype probabilities involving recombination, are quite complex and are therefore defined in separate appendices (Appendix A and Appendix B respectively). The definition of $Q(\mathbf{z}|d_i, r)$ is the same as that of (Chakraborty and Rannala, 2023) but for ease of reference is restated in Appendix B.

Model (a): Purebred population B

Both chromosomes ($\mathbf{x}_i^M$ and $\mathbf{x}_i^P$) come from Population B. The log-likelihood of the i-th diplotype for an individual is equivalent to the log-probability of the predicted counts of the two haplotypes as given by

$$\begin{aligned} \log P(\mathbf{x}_i|g = a, \mathbf{n}) = \log\Gamma(\theta_B) - \sum_{h=1}^{H} \log\Gamma(n_{hB} + 1/H) \\ + I(\mathbf{x}_i)\log 2 - \log\Gamma(\theta_B + 2) + \sum_{h=1}^{H} \log\Gamma(\phi_h(\mathbf{x}_i) + n_{hB} + 1/H) \end{aligned} \tag{9}$$

Model (b): Backcross with Population A

One chromosome comes from Population A and the other chromosome is a recombinant.

$$\begin{aligned} \log P(\mathbf{x}_i|g = b, \mathbf{n}) = I(\mathbf{x}_i)\log\left[\Pr(\boldsymbol{\phi}^M(\mathbf{x}_i)|\mathbf{n}_A)U(\mathbf{x}_i^P) + \Pr(\boldsymbol{\phi}^P(\mathbf{x}_i)|\mathbf{n}_A)U(\mathbf{x}_i^M)\right] \\ + [1 - I(\mathbf{x}_i)]\log\left[\Pr(\boldsymbol{\phi}^M(\mathbf{x}_i)|\mathbf{n}_A)U(\mathbf{x}_i^P)\right] \end{aligned} \tag{10}$$

where, $U(\mathbf{x}_i^c) = \sum_z U(\mathbf{x}_i^c|\mathbf{z}).Q(\mathbf{z}|d_i, r)$, and

$$\begin{aligned}\log \Pr(\boldsymbol{\phi}^c(\mathbf{x}_i)|\mathbf{n}_A) &= \log\Gamma(\theta_A) - \sum_{h=1}^{H} \log\Gamma(n_{hA} + 1/H) \\ &\quad - \log\Gamma(1+\theta_A) + \sum_{h=1}^{H} \log\Gamma(\phi_h^c(\mathbf{x}_i) + n_{hA} + 1/H), \qquad c \in \{M, P\}\end{aligned} \tag{11}$$

Model (c): F1 hybrid

One chromosome comes from Population A and the other from Population B

$$\begin{aligned}\log P(\mathbf{x}_i|g = c, \mathbf{n}) = I(\mathbf{x}_i) \log \Big[&\Pr(\boldsymbol{\phi}^M(\mathbf{x}_i)|\mathbf{n}_A)\Pr(\boldsymbol{\phi}^P(\mathbf{x}_i)|\mathbf{n}_B) \\ &+ \Pr(\boldsymbol{\phi}^P(\mathbf{x}_i)|\mathbf{n}_A)\Pr(\boldsymbol{\phi}^M(\mathbf{x}_i)|\mathbf{n}_B)\Big] \\ + [1 - I(\mathbf{x}_i)] \log &\Big[\Pr(\boldsymbol{\phi}^M(\mathbf{x}_i)|\mathbf{n}_A)\Pr(\boldsymbol{\phi}^P(\mathbf{x}_i)|\mathbf{n}_B)\Big]\end{aligned} \tag{12}$$

where,

$$\begin{aligned}\log \Pr(\boldsymbol{\phi}^c(\mathbf{x}_i)|\mathbf{n}_k) &= \log\Gamma(\theta_k) - \sum_{h=1}^{H} \log\Gamma(n_{hk} + 1/H) \\ &\quad - \log\Gamma(1+\theta_k) + \sum_{h=1}^{H} \log\Gamma(\phi_h^c(\mathbf{x}_i) + n_{hk} + 1/H),\end{aligned} \tag{13}$$

$c \in \{M, P\}, \quad k \in \{A, B\}$

Model (d): Purebred population A

Both chromosomes (x^M and x^P) come from Population A. The log-probability of the predicted counts of the two haplotypes is given by

$$\begin{aligned}\log P(\mathbf{x}_i|g = d, \mathbf{n}) &= \log\Gamma(\theta_A) - \sum_{h=1}^{H} \log\Gamma(n_{hA} + 1/H) \\ &\quad + I(\mathbf{x}_i)\log 2 - \log\Gamma(\theta_A + 2) + \sum_{h=1}^{H} \log\Gamma(\phi_h(\mathbf{x}_i) + n_{hA} + 1/H)\end{aligned} \tag{14}$$

Model (e): Backcross with Population B

One chromosome comes from Population B and the other chromosome is a recombinant.

$$\begin{aligned}\log P(\mathbf{x}_i|g=e,\mathbf{n}) = I(\mathbf{x}_i)\log\left[\Pr(\boldsymbol{\phi}^M(\mathbf{x}_i)|\mathbf{n}_B)U(\mathbf{x}_i^P)+\Pr(\boldsymbol{\phi}^P(\mathbf{x}_i)|\mathbf{n}_B)U(\mathbf{x}_i^M)\right]\\ +[1-I(\mathbf{x}_i)]\log\left[\Pr(\boldsymbol{\phi}^M(\mathbf{x}_i)|\mathbf{n}_B)U(\mathbf{x}_i^P)\right]\end{aligned} \tag{15}$$

where, $U(\mathbf{x}_i^c)=\sum_z U(\mathbf{x}_i^c|\mathbf{z}).Q(\mathbf{z}|d_i,r)$, and

$$\begin{aligned}\log\Pr(\boldsymbol{\phi}^c(\mathbf{x}_i)|\mathbf{n}_B) = \log\Gamma(\theta_B)-\sum_{h=1}^{H}\log\Gamma(n_{hB}+1/H)\\ -\log\Gamma(1+\theta_B)+\sum_{h=1}^{H}\log\Gamma(\phi_h^c(\mathbf{x}_i)+n_{hB}+1/H), \qquad c\in\{M,P\}\end{aligned} \tag{16}$$

Model (f): F2 hybrid

Both chromosomes are recombinants.

$$\log P(\mathbf{x}_i|g=f,\mathbf{n}) = I(\mathbf{x}_i)\log(2)+\log\left[U(\mathbf{x}_i^M)U(\mathbf{x}_i^P)\right] \tag{17}$$

where, $U(\mathbf{x}_i^c)=\sum_z U(\mathbf{x}_i^c|\mathbf{z}).Q(\mathbf{z}|d_i,r), \quad c\in\{M,P\}$. Therefore the likelihood of a diplotype, $\mathbf{x}=\{\mathbf{x^M},\mathbf{x^P}\}$ for an individual with K chromosomes under the g-th genealogical class is given by,

$$P(\mathbf{x}|G=g,\mathbf{n})=\prod_{i=1}^{K}P(\mathbf{x}_i|G=g,\mathbf{n}). \tag{18}$$

Posterior Probability of Genealogical Classes

The posterior probability that an individual with diplotype $\mathbf{x}$ belongs to the g-th genealogical class is

$$P(G=g|\mathbf{x},\mathbf{n})=\frac{\pi_g\times P(\mathbf{x}|G=g,\mathbf{n})}{\sum_{g\in G}\pi_i\times P(\mathbf{x}|G=g,\mathbf{n})}. \tag{19}$$

where, an individual belongs to genealogical class g with prior probability π_g with $g\in G$

SIMULATION STUDY

We used simulated data to compare the Mongrail program (Chakraborty and Rannala, 2023) with our new method implemented as Mongrail 2.0 which relaxes the assumption

of known population haplotype frequencies and correctly calculates marginal haplotype probabilities under recombination. We use a subset of the simulated datasets from Chakraborty and Rannala (2023), fixing the following combination of parameters: $K = 20, L = 10, c = 0.1, \alpha = 1$. We examined the relative performance of the two methods under different values of R (1cM or 50cM) and h (5 or 15). See Supplementary Article section S1.1 for details. Chakraborty and Rannala (2023) used a Dirichlet distribution to generate a diverse set of haplotype frequency distributions which were reused in the current study. Diplotypes for hybrid individuals were simulated based on the simulated haplotypes and their corresponding frequencies. For Mongrail 2.0, we use these previously generated frequencies as parameters for a Multinomial distribution to generate samples of haplotypes from the reference populations. We considered multinomial sample sizes of $N = 10, 100, 1000$ to generate different reference population datasets. For each simulated individual we either used Mongrail posterior probabilities computed in our previous paper (using population frequencies) or computed posterior probabilities using the posterior mean from the sample counts. We applied our new method to the same set of simulated individuals using the simulated reference samples to compute the posterior probabilities. We use the following three metrics to study the statistical performance of Mongrail and Mongrail 2.0:

1. Influence of sample size on posterior probabilities
2. AUC-ROC curve analysis

Influence of sample size on posterior probabilities

Here we compare the posterior probabilities previously computed using Mongrail and known population haplotype frequencies (Chakraborty and Rannala, 2023) with those of Mongrail 2.0 using different multinomial sample sizes ($N = 10, 100, 1000$). Only the stacked bar plots for 100 individuals from each of the six genealogical classes are shown as other results are essentially similar. Figure 1 shows that, irrespective of population sample size (N), both methods assign higher posterior probabilities to the true genealogical class of an individual for purebreds and F1 hybrids. The posterior probabilities are lower and distributed across more genealogical classes for Mongrail 2.0 when $N = 10$ and $N = 100$, reflecting the additional uncertainty when haplotypes sampling from the reference populations is properly accounted for. This effect is particularly notable for backcrosses and F2 hybrids.

Figure 2 shows the results for backcross and F2 hybrid individuals as the sample size increases from $N = 10$ to 1000 (moving top to bottom, first three rows). The magnitude of posterior probabilities produced by Mongrail 2.0 for the correct genealogical class (backcrosses and F2 hybrids) increases. The performance of Mongrail 2.0 under a multinomial sample size of $N = 1000$ (third row) is very similar to Mongrail used with known frequencies (last row). For a small sample size of $N = 10$ (first row), Mongrail 2.0 assigns more posterior probability to alternative genealogical classes as expected. It is evident that as the model complexity increases (from figure 1 to figure 2), the difference between the two methods Mongrail and Mongrail 2.0 is more pronounced. However, with a sample size of $N = 1000$, the haplotype count data

simulated for use with Mongrail 2.0 is close to the expected counts under the true haplotype frequency distribution as used by Mongrail. Thus the methods are comparable for $N = 1000$.

AUC-ROC curve analysis

Here we examine the performance of the two methods to classify the different genealogical classes for different classification thresholds using the AUC-ROC curve analysis. The greater the area under the ROC curve (AUC) the better the performance. We compare the AUC values of Mongrail (posterior probabilities computed using the posterior mean from the sample counts) with Mongrail 2.0 under different multinomial sample sizes ($N = 10, 100$ or 1000), recombination frequency ($R = 1$cM or 50cM) and distinct haplotype counts per chromosome for each population ($h = 5$ or 15). We are interested in the interaction of the different parameters (N, R and h) and their effect on classification power of the two methods. We plot the AUC values against increasing values of N as multi-line plots, where line colors represent different methods (Mongrail 2.0 and Mongrail) and line types (solid or dashed) different recombination frequencies ($R = 1$cM or 50cM). The results are shown in figure 3 where the top and bottom panels represent the cases for $h = 15$ and $h = 5$, respectively.

Figure 3 shows that irrespective of the method and any values of R or h, the AUC increases as the sample size increases from $N = 10$ to 1000. The performance of the two methods improves as recombination frequency increases from $R = 1$ cM to 50 cM (solid to dashed line) as expected. Performance of the alternate methods is almost indistinguishable across all six genealogical classes, especially for sample sizes $N = 100$ and 1000, regardless of values for R or h. We also observe that for $N = 10$ and $N = 100$, AUC values are higher for $h = 15$ relative to $h = 5$ across all genealogical classes. The difference is negligible for $N = 1000$ and more pronounced for $N = 10$. This suggests that with a small sample size there is limited information about population haplotype frequencies (and predicted haplotype counts) when the number of distinct haplotypes increases. This uncertainty affects both methods; Mongrail 2.0 is influenced through the distribution of predicted haplotype counts) and Mongrail is affected because the posterior mean is based on the sample counts. The AUC values in Figure 3 summarize the combined impact of the parameters h, R and N on the power of the two methods to detect genealogical classes. For further details on individual parameter effects see the individual ROC curves (Figures S1-S6) in the Supplementary Article section S1.2.

EMPIRICAL ANALYSIS

To assess the performance of the new method and the Mongrail 2.0 program when applied to empirical datasets for a range of organisms we analyzed three published hybridization datasets: kiwifruit (Yu *et al.*, 2023), plateau fence lizards (Leaché *et al.*, 2025) and Florida pathers (Aguilar-Gómez *et al.*, 2025).

Kiwifruit

We analyzed a dataset consisting of two deeply divergent kiwifruit species (*Actinidia eriantha* and *Actinidia hemsleyana*) and their hybrids using Mongrail 2.0. Actinidia, a dioecious variety of plant is mainly found in East and South Asia. *A. eriantha* is quite widespread unlike *A. hemsleyana* which occurs mainly in Eastern China. *Actinidia zhejiangensis* is thought to be a homoploid hybrid between *Actinidia eriantha* and *Actinidia hemsleyana*. This species is quite rare and is found in Jiangxi, Zhejiang and Fujian provinces in China. Yu *et al.* (2023) specifically chose these species to study the evolutionary consequence of homoploid speciation.

The dataset consists of 51 individuals (21 *A. hemsleyana*, 19 *A. eriantha* and 11 *A. zhejiangensis*). Whole genome resequencing data were assembled against *A. eriantha* genome and variant calling were performed to generate a Variant Call Format (VCF) data file. We used BCFtools to filter out missing sites and obtained a dataset with 5,345,755 SNPs on 29 chromosomes. We used Beagle 5.1 to phase the data for the two parental populations (*Actinidia eriantha* and *Actinidia hemsleyana*). See Supplementary Article section S2.1 for more details on recombination rate and markers. We used Mongrail 2.0 to calculate the posterior probabilities of genealogical classes for these data. In this framework *A. eriantha* is treated as population A and *A. hemsleyana* as population B. The genealogical classes are therefore as follows:

- Model a - *A. hemsleyana*
- Model b - Backcross with *A. eriantha*
- Model c - F1 hybrid (presumptive *A. zhejiangensis*)
- Model d - *A. eriantha*
- Model e - Backcross with *A. hemsleyana*
- Model f - F2 hybrid

Figure 4 shows the posterior probability distribution of 11 presumed *A. zhejiangensis* individuals. Mongrail 2.0 produces very high posterior probability for 10 out of 11 individuals. Based on a classification threshold of 0.95, 10 individuals were assigned to specific genealogical classes with posterior probability higher than 0.95. Seven were identified as F1 hybrids (Individuals: JNZJ-1, JNZJ-3, LQZJ-1, TSZJ-1, WCZJ-1, WCZJ-2, YP-1), two as F2 hybrids (Individuals: ASX-1, JNZJ-5) and one as backcrossed with *A. eriantha* (Individual: LSZJ-1). For direct comparison, we also applied Mongrail to this dataset (see Supplementary Section S3, Figures S7). Results were consistent across methods for all individuals.

Lizard

Next we analyzed plateau fence lizards (*Sceloporus tristichus*) sampled along a transect between Great Basin Grassland (north) and Great Basin Conifer Woodland (south) in eastern Arizona (USA) (collected in 2022 by Leaché *et al.* (2025)). The lizards found in grasslands (north) and juniper woodlands (south) are the two distinct parental populations. A hybrid zone is located in an ecotone where the vegetation transitions from grassland to juniper habitat between the parental populations. The sampling transect includes eight sites (Holbrook, Fivemile Wash, Washboard Wash, Woodruff,

Canoncito, Sevenmile Draw, Snowflake and Show Low) spanning a distance of 63.5 km where the hybrid zone is located primarily between the cities of Holbrook (north) and Show Low (south). We refer to parental grassland and juniper lizard populations as Holbrook and Show Low populations respectively. The remaining samples from six different sites are treated as potential hybrids. Leaché *et al.* (2025) collected these data to study the temporal introgression dynamics of the two populations sampled along the linear transect.

The dataset consists of 78 individuals (7 Holbrook, 11 Show Low and 60 transect samples). Among the 60 individuals sampled from the transect, 11 were from Fivemile Wash, 10 from Washboard Wash, 9 from Woodruff, 10 from Canoncito, 10 from Sevenmile Draw and 10 from Snowflake. Leaché *et al.* (2025) collected SNP data using double-digestion restriction site-associated DNA sequencing (ddRADseq) protocol, aligning the reads against a reference genome of *Sceloporus tristichus* (Bedoya and Leaché, 2021) and generated a Variant Call Format (VCF) data file. We used BCFtools (version 1.19-43) to filter out missing data, retaining only biallelic sites leaving 714 SNPs on 11 chromosomes. We used Beagle 5.1 to phase the data for the two parental populations (Holbrook and Show Low). See Supplementary Article section S2.2 for more details on recombination rate and markers. Mongrail 2.0 was used to calculate posterior probabilities of genealogical classes for individuals. In this framework Holbrook is treated as population A and Show Low as population B. The genealogical classes are therefore as follows:

- Model a - Show Low
- Model b - Backcross with Holbrook
- Model c - F1 hybrid
- Model d - Holbrook
- Model e - Backcross with Show Low
- Model f - F2 hybrid

Figure 5 shows the posterior probabilities of genealogical classes for each individual from the transect region. Most individual posterior probabilities are distributed primarily over three hybrid classes (F2 and backcrosses). Genealogical class **b** which is a backcross with Holbrook (color blue in Figure 5) has increased posterior probabilities in many individuals at the sites sampled from Fivemile Wash, Washboard Wash, Woodruff and Canoncito. By contrast, this pattern changes drastically for lizards sampled from Sevenmile Draw and Snowflake where the color red (genealogical class **a**: pure Show Low) and orange (genealogical class **e**: backcross with Show Low) is more dominant. Genealogical class assignment remains uncertain for most individuals. Only 6 out of 60 transect sampled individuals has a posterior probability higher than 0.95 for a specific genealogical class. Classification decisions based on a posterior probability threshold of 0.95 are shown in table 1. Three of 6 individuals are classified as F2 hybrids (Individuals: Five_5201, Wash_5192, Wood_5227), one as pure Holbrook (Individual: Wash_5253) and two as pure Show Low (Individuals: Nsno_5235, Snow_5264). The remaining 51 transect sampled individuals cannot be assigned with certainty to any of the genealogical classes. To allow direct comparison, we also applied Mongrail to this dataset (see Supplementary Section S3, Figures S8-S13). Results were

qualitatively similar for both methods, except for one individual (Nsno_5235, Figure S12).

Puma

Lastly we analyze a dataset comprising pumas (*Puma concolor*) found in southern Florida, also called Florida panthers. Florida panthers were once common in the US southeast but urban expansion and bounty hunting decimated the population, reducing their range to hardwood swamps and cypress prairies of south and central Florida. The species was listed as "endangered" in 1967 by the U.S. Fish and Wildlife Service under the Endangered Species Act. By the 1990s this small and isolated population of Florida panthers were suffering from reproductive failure and phenotypic defects due to inbreeding and less than 30 individuals remained in the wild. To rescue this population eight female panthers from Texas were released in southern Florida in 1995, five of which were successful in producing offspring. This introduction led to a four-fold increase in population size. Currently there are between 120 and 230 adult and subadult Florida panthers in the wild.

This conservation strategy for increasing the genetic diversity and fitness of a shrinking population is termed genetic rescue. Aguilar-Gómez *et al.* (2025) sequenced 29 post-rescue Florida panthers to study the long-term genomic consequences of genetic rescue. They also included previously published whole genome sequence data from two parental populations (Texas and pre-rescue Florida panthers) and two known F1 hybrids. The Texas population was represented by genome sequences collected from the five female Texas panthers originally introduced in 1995 (Ochoa *et al.*, 2019). The pre-rescue Florida panther population was represented by genome sequences of 2 individuals sampled by Ochoa *et al.* (2019) and 2 sampled by Saremi *et al.* (2019). The two known F1 hybrids were from Ochoa *et al.* (2019). We applied Mongrail 2.0 to analyze this dataset consisting of 5 Texas panthers, 4 pre-rescue Florida panthers and 31 post-rescue Florida panthers. Aguilar-Gómez *et al.* (2025) aligned these genomes against a reference genome for *Puma yagouaroundi* (an outgroup) and performed variant calling to generate a Variant Call Format (VCF) file. We filtered out missing data using BCFtools (version 1.19-43) and selected scaffolds larger than 20 Mb to produce a dataset with 1,107,855 biallelic SNPs across 34 scaffolds. We used Beagle 5.1 to phase the data for the two parental populations (Texas and pre-rescue Florida panthers). See Supplementary Article section S2.3 for more details on recombination rate and markers. Mongrail 2.0 was used to calculate posterior probabilities of the genealogical class for each post-rescue individual. In this framework, pumas from Texas are treated as population A and pre-rescue Florida panthers as population B. The genealogical classes are therefore as follows:

- Model a - pre-rescue Florida panther
- Model b - Backcross with Texas panther
- Model c - F1 hybrid
- Model d - Texas panther
- Model e - Backcross with pre-rescue Florida panther
- Model f - F2 hybrid

Figure 6 shows the posterior probability distribution for post-rescue Florida panthers. Most individuals are a mixture of either genealogical classes **e** and **f** (backcross with pre-rescue Florida panther and F2 hybrid) or **a** and **e** (pre-rescue Florida panther and backcross with pre-rescue Florida panther). Individuals AFP1 and AFP2 (known F1 hybrids) have posterior probabilities for genealogical class **c** (F1 hybrid) of 0.818 and 0.416, respectively, which are too low to warrant assignment. Based on a classification threshold of 0.95 only 7 of 31 individuals were assigned to specific genealogical classes. Table 2 describes the genealogical class assignments for these 7 post-rescue Florida panthers. Four of them are classified as F2 hybrids (Individuals: AFP10, AFP14, AFP24, AFP6), two as backcrosses with pre-rescue Florida panther (Individuals: AFP25, AFP3) and one as pre-rescue Florida panther (Individual: AFP29). For direct comparison, we applied Mongrail to this dataset (see Supplementary Section S3, Figures S14,S15). Results were largely consistent across methods, except for four individuals AFP18 (Figures S14), AFP11 (Figures S14), AFP29 (Figures S15) and AFP5 (Figures S15).

DISCUSSION

In studying the role of hybridization in any context (e.g., conservation genetics, speciation, etc) a necessary first step is to reliably identify hybrids. In our earlier paper (Chakraborty and Rannala, 2023) we developed a Bayesian method (implemented in the program Mongrail) to infer recent hybrids using genomic data. One assumption of Mongrail was that the population haplotype frequencies are known fixed parameters. For empirical datasets population haplotype frequencies are unknown; Chakraborty and Rannala (2023) estimated these using the posterior mean frequency (ignoring uncertainty of haplotype frequencies). Here we address this limitation and develop theory that allows for uncertainty of population haplotype frequencies by integrating over the posterior density using a conjugate prior. We further employed simulation analyses to compare the statistical performance of the two methods. We find that even relatively small samples can produce high posterior probabilities of assignments. The estimates obtained either by using posterior mean point estimates of haplotype frequencies or by integrating over the posterior density of frequencies conditional on the sample are highly similar unless the population sizes are very small, suggesting the previous approximation using estimated haplotype frequencies is fine in many cases. When comparing the new method (implemented in Mongrail 2.0) to the original Mongrail we observed greater uncertainty (represented by more probability associated with alternative genealogical classes) in the posterior probabilities (see Figure 1 and Figure 2). This outcome aligns with our expectation that a method integrating over the posterior density should exhibit more uncertainty than one treating haplotype frequencies as known. Moreover, the uncertainty in Mongrail 2.0 diminishes as the sample size increases, resulting in higher posterior probabilities for the correct genealogical classes. The fact that simulated haplotype counts analyzed using Mongrail 2.0 approach the expected counts under the true population haplotype frequency distribution is indicated by results for $N = 1000$, where the two methods produce essentially similar posterior probabilities.

The performance of both methods is impacted when the number of distinct haplotypes per chromosome (h) increases (from 5 to 15) for smaller sample sizes ($N = 10$) (see Figure 3). This is expected, as small sample sizes do not yield reliable estimates of haplotype count data when the number of multinomial categories is large. Additionally, as model complexity increases (purebred → F1 hybrid → backcross → F2 hybrid) the performance of both methods declines. Qualitatively the new method implemented in Mongrail 2.0 preserves the positive attributes of Mongrail. Specifically, the statistical power of Mongrail 2.0 increases with increasing recombination frequency (map-length of a chromosome) and increasing numbers of chromosomes.

This paper also presents a new algorithm to marginalize over haplotypes in order to calculate the probability of recombinant haplotypes; this is a corrected version of the algorithm described in our earlier paper (Chakraborty and Rannala, 2023). In our earlier approach, we only treated contiguous markers from the same population within a haplotype as jointly distributed, but in fact, all markers inherited from a given parental population should be treated as jointly distributed. This distinction is crucial, as it accurately reflects how meiosis and recombination generate recombinant haplotypes. Multiplying the marginal frequencies of contiguous markers from the same population incorrectly assumes that these segments are independent, contrary to biological reality. Instead contiguous markers derived from the same parental population are inherited together and using the joint frequency of co-transmitted markers correctly captures the ancestry of chromosomal segments formed by recombination. This revised algorithm performs the correct calculation with little additional computational expense. To illustrate this difference, we present a toy example in Supplementary Section S4, comparing the original Mongrail and updated Mongrail 2.0 algorithms. Both methods yield identical results when populations are in linkage equilibrium, which makes sense since the markers are effectively independent in that scenario. The difference between the two algorithms should be insignificant except in the case of large chromosomes that have experienced multiple recombination events. This is observed in Figure S16 (Supplementary Section S5) where Mongrail 2.0 places higher posterior probability on the true genealogical class (model **f**) by comparison with Mongrail for all individuals but one (i7). This holds true even for individuals that were difficult to resolve (i2, i3, i4, i6, i9). Posterior probabilities for individual i7 under genealogical class **f** match between methods to the second decimal place.

We explored Mongrail 2.0's performance on empirical datasets by applying it to analyze a range of data types comprising non-model organisms: kiwifruit, lizard and puma. Each dataset has unique features that present different challenges for hybrid inferences. The kiwifruit dataset is a genomic study investigating the evolutionary consequences of homoploid hybridization and testing for reduced fitness of hybrids. The lizard dataset includes individuals that were sampled along a linear transect to study the clinal nature of the hybrid zone and the relationship between physical distance and genealogical class. The puma genomic samples were analyzed to study the genomic impact of genetic rescue of Florida panthers through introduction of

Texas pumas and exemplify a recent hybridization event with known characteristics and source populations.

In the kiwifruit dataset Mongrail 2.0 classified 7 of 11 presumed *A. zhejiangensis* as F1 hybrids which supports the previous observation that F1 hybrids are more frequent in the hybrid zone (Yu *et al.*, 2023). We identified one individual (individual index: 7) as a backcross with *A. eriantha* (genealogical class **b**). A prior Newhybrids analysis identified the same individual (LSZJ-1) as a backcross with *A. eriantha* (Yu *et al.*, 2023). By contrast, we identified two individuals (individual index: 1, 5) as F2 hybrids that were previously identified as F1 hybrids (ASX-1 and JNZJ-5) (Yu *et al.*, 2023). Our results suggest that hybridization between *Actinidia eriantha* and *Actinidia hemsleyana* may not be a dead-end as previously assumed (Yu *et al.*, 2023). The presence of backcross and F2 hybrids contradicts the claim that F1 hybrids are infertile (Yu *et al.*, 2023). A possible reason for the lack of F2 individuals in the previous study may be Newhybrids tendency to inflate posterior probabilities of the F1 genealogical class (Chakraborty and Rannala, 2023).

Analysis of the lizard dataset using Mongrail 2.0 classified only 6 of 60 individuals with high posterior probability. Posterior probabilities are typically highly variable, often spread over F2 and backcrosses. There are several factors potentially affecting the method's ability to infer hybrids confidently. For example, Mongrail 2.0 is based on the assumption that the two populations have been interbreeding for 2 generations. Unclassified hybrids may be an outcome of backcrossing exceeding two generations. The small number of chromosomes analyzed (11 chromosomes) and small sample sizes for the parental populations might also have reduced the power. The power of Mongrail is known to increase with increasing number of chromosomes (Chakraborty and Rannala, 2023).

We also observed (Figure 5) a clear gradient in the proportion of assignment to genealogical classes as we move from north (Fivemile Wash) to south (Snowflake). This sharp transition south of Canoncito is expected as this site is located near the center of the hybrid zone. Mongrail 2.0 captures this spatial pattern which is a characteristic of clinal hybrid zone. The assignment probabilities for individuals from populations north of Canoncito suggests increased Holbrook ancestry as the prominent color is blue (genealogical class **b**: backcross with Holbrook). Whereas individuals from Sevenmile Draw and Snowflake have increased Show Low ancestry as the prominent colors are red (genealogical class **a**: Show Low) and orange (genealogical class **e**: backcross with Show Low). This clear transition in posterior probabilities of assignments is consistent with the findings of Leaché *et al.* (2025).

For the puma dataset, Mongrail 2.0 was able to infer genealogical classes with great certainty for 7 out of 31 post-rescue Florida panthers. Six of these were classified as F2 or backcrosses with pre-rescue Florida panther and one (individual index: 20) was classified as pre-rescue Florida panther. The remaining individuals have little Texas ancestry; the stacked bar plot (Figure 6) rarely exhibits genealogical classes **b** and **d**. These results provide an important insight into the genetic makeup of the post-rescue Florida panthers. They support the findings of Aguilar-Gómez *et al.* (2025) that genetic swamping due to the introduction is

unlikely to have occurred. Mongrail 2.0 did not produce high posterior probability for the two known F1 hybrids (individuals AFP1 and AFP2) assigning only 81.8% and 41.3% posterior probability to genealogical class **c** (F1 hybrid). These low posterior probabilities may be due to small sample sizes for the parental populations (5 Texas pumas and 4 pre-rescue Florida panthers).

To allow direct comparison, we applied both Mongrail and Mongrail 2.0 across all empirical datasets (see Supplementary Section S3, Figures S7-S15). Results were largely consistent across methods, except for one individual in the lizard dataset and four individuals in the puma dataset. Observed discrepancies in these cases may be due to small parental population sample sizes.

In summary, we have introduced an improved Bayesian algorithm for inferring hybrids and backcrosses across two generations using sampled genomes from two populations. This new algorithm, Mongrail 2.0, offers two major advancements over the original Mongrail (Chakraborty and Rannala, 2023). First, it relaxes the assumption of known population haplotype frequencies. Mongrail used the Multinomial-Dirichlet posterior mean from reference samples to estimate unknown population haplotype frequencies, disregarding uncertainty. In contrast, Mongrail 2.0 addresses this issue by using the *posterior predictive distribution* which integrates over uncertainties of population allele frequencies. Second, Mongrail 2.0 corrects the probability calculation of recombinant haplotypes by properly marginalizing over haplotypes. These two improvements come with little additional computational cost and simulations show that Mongrail 2.0 performs as effectively as Mongrail in generating high posterior probabilities for the correct genealogical classes when sample sizes are reasonably large, otherwise posterior probabilities are reduced reflecting the additional uncertainty due to uncertain population allele frequencies as expected. Mongrail 2.0 appears to be a statistically conservative method (having low flase positive rates), particularly when the reference population sample size is small, coupled with a high number of distinct haplotypes per chromosome. This is supported by empirical analyses, notably with the lizard and puma datasets, where Mongrail 2.0 faced challenges in inferring hybrids with high certainty. Despite these challenges, Mongrail 2.0 still incorporates linkage and recombination into its model, preserving both power and efficiency even with just 10 markers per chromosome. This speaks to its reliability, especially when the method produces high posterior probabilities. The diverse hybridizing non-model organism datasets used in this study demonstrate the broad applicability and utility of the new method.

SOFTWARE

The Open Source C program Mongrail 2.0 is available at https://github.com/mongrail/mongrail2.

DATA AVAILABILITY

Simulated datasets and scripts used for generating the simulations are available at https://github.com/Mongrail-2-0/simulations. The empirical datasets used in this paper can be obtained from the original authors upon request.

ACKNOWLEDGEMENTS

This work was supported by National Institutes of Health grant GM123306 and National Science Foundation grant DEB-1754254 to BR.

AUTHOR CONTRIBUTIONS

SC and BR co-developed the theory and co-wrote the manuscript. SC developed the Mongrail 2.0 software and performed the simulation study. SC conducted the analysis of the empirical datasets.

COMPETING INTERESTS

The authors declare no competing interests.

References

Abbott RJ, Barton NH, Good JM (2016). Genomics of hybridization and its evolutionary consequences. *Molecular Ecology* **25**: 2325–2332.

Aguilar-Gómez D, Yuan L, Zhang Y, Ochoa A, Culver M, Fitak RR, *et al.* (2025). Genetic rescue of florida panthers reduced homozygosity but did not swamp ancestral genotypes. *Proceedings of the National Academy of Sciences* **122**: e2410945122.

Anderson E, Thompson E (2002). A model-based method for identifying species hybrids using multilocus genetic data. *Genetics* **160**: 1217–1229.

Barton NH (2001). The role of hybridization in evolution. *Molecular Ecology* **10**: 551–568.

Bedoya AM, Leaché AD (2021). Characterization of a pericentric inversion in plateau fence lizards (Sceloporus tristichus): evidence from chromosome-scale genomes. *G3: Genes|Genomes|Genetics* **11**: jkab036.

Chakraborty S, Rannala B (2023). An efficient exact algorithm for identifying hybrids using population genomic sequences. *Genetics* **223**: iyad011.

Fitzpatrick BM, Ryan ME, Johnson JR, Corush J, Carter ET (2015). Hybridization and the species problem in conservation. *Current Zoology* **61**: 206–216.

Leaché AD, Davis HR, Singhal S (2025). Hybrid Zone Analysis Using Coalescent-Based Estimates of Introgression and Migration in Plateau Fence Lizards (Sceloporus tristichus). *Molecular Ecology* p. e17819.

McGaughran A, Dhami MK, Parvizi E, Vaughan AL, Gleeson DM, Hodgins KA, *et al.* (2023). Genomic tools in biological invasions: current state and future frontiers. *Genome Biology and Evolution* **16**: evad230.

Ochoa A, Onorato DP, Fitak RR, Roelke-Parker ME, Culver M (2019). De Novo assembly and annotation from parental and F1 Puma genomes of the Florida Panther Genetic Restoration Program. *G3: Genes|Genomes|Genetics* **9**: 3531–3536.

Saremi NF, Supple MA, Byrne A, Cahill JA, Coutinho LL, Dalén L, *et al.* (2019). Puma genomes from North and South America provide insights into the genomic consequences of inbreeding. *Nature Communications* **10**: 4769.

Yu X, Qin M, Qu M, Jiang Q, Guo S, Chen Z, *et al.* (2023). Genomic analyses reveal dead-end hybridization between two deeply divergent kiwifruit species rather than homoploid hybrid speciation. *The Plant Journal* **115**: 1528–1543.

Tables

Table 1 Assignment of 6 transect sampled plateau fence lizards

Individual Index	Sampling Location	Posterior Probability	Genetic Identification	Chosen Model
Five_5201	Fivemile Wash	0.986	F2	f
Wash_5192	Washboard Wash	0.95	F2	f
Wash_5253	Washboard Wash	0.95	Pure Holbrook	d
Wood_5227	Woodruff	0.989	F2	f
Nsno_5235	Sevenmile Draw	0.979	Pure Show Low	a
Snow_5264	Snowflake	0.984	Pure Show Low	a

Table 2 Assignment of 7 post-rescue Florida panthers

Individual Index	Posterior Probability	Genetic Identification	Chosen Model
AFP10	0.995	F2	f
AFP14	0.953	F2	f
AFP24	0.964	F2	f
AFP25	0.992	Backcross with pre-rescue Florida panther	e
AFP29	0.999	pre-rescue Florida panther	a
AFP3	0.965	Backcross with pre-rescue Florida panther	e
AFP6	0.988	F2	f

Figures

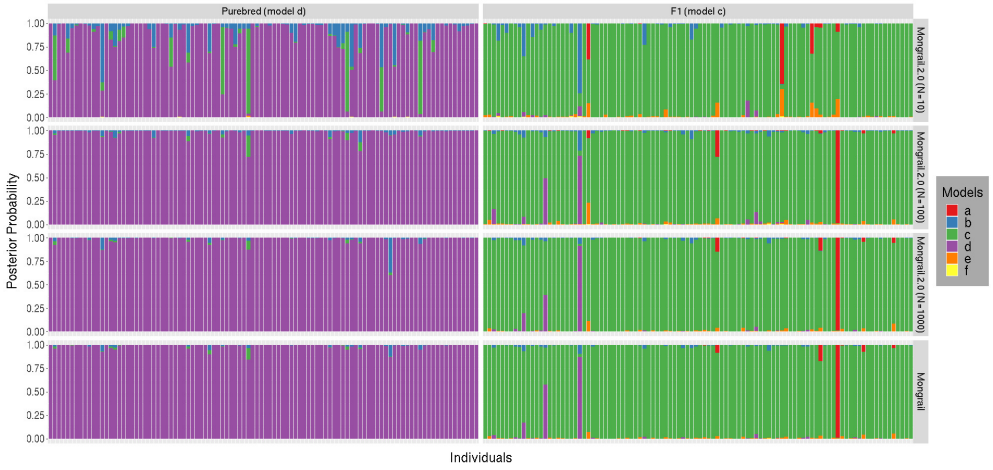

Fig. 1 Distribution of posterior probabilities for 100 individuals simulated under genealogical classes **d** (left column) and **c** (right column) using the following set of simulation parameters: $K = 20, L = 10, R = 50, h = 5, c = 0.1, \alpha = 1$. Posterior probabilities using Mongrail with known population haplotype frequencies is shown in bottom plot (4th row) and Mongrail 2.0 for different values of Multinomial sample counts $N = 10, 100, 1000$ in the first three rows respectively. The six genealogical classes are as follows: **a**-pure population B, **b**-backcross with population A, **c**-F1 hybrid, **d**-pure population A, **e**-backcross with population B, **f**-F2 hybrid.

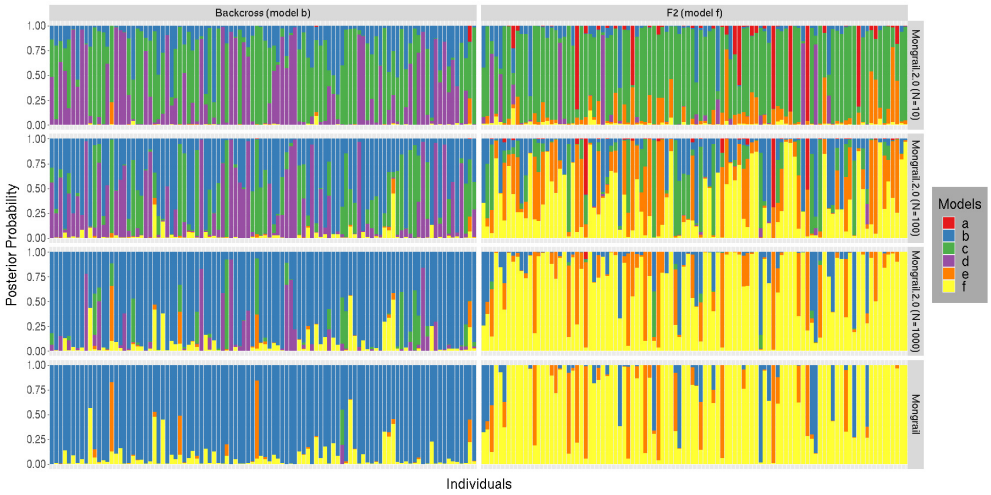

Fig. 2 Distribution of posterior probabilities for 100 individuals simulated under genealogical classes **b** (left column) and **f** (right column) using the following set of simulation parameters: $K = 20, L = 10, R = 50, h = 5, c = 0.1, \alpha = 1$. Posterior probabilities using Mongrail with known population haplotype frequencies is shown in bottom plot (4th row) and Mongrail 2.0 for different values of Multinomial sample counts $N = 10, 100, 1000$ in the first three rows respectively. The six genealogical classes are as follows: **a**-pure population B, **b**-backcross with population A, **c**-F1 hybrid, **d**-pure population A, **e**-backcross with population B, **f**-F2 hybrid.

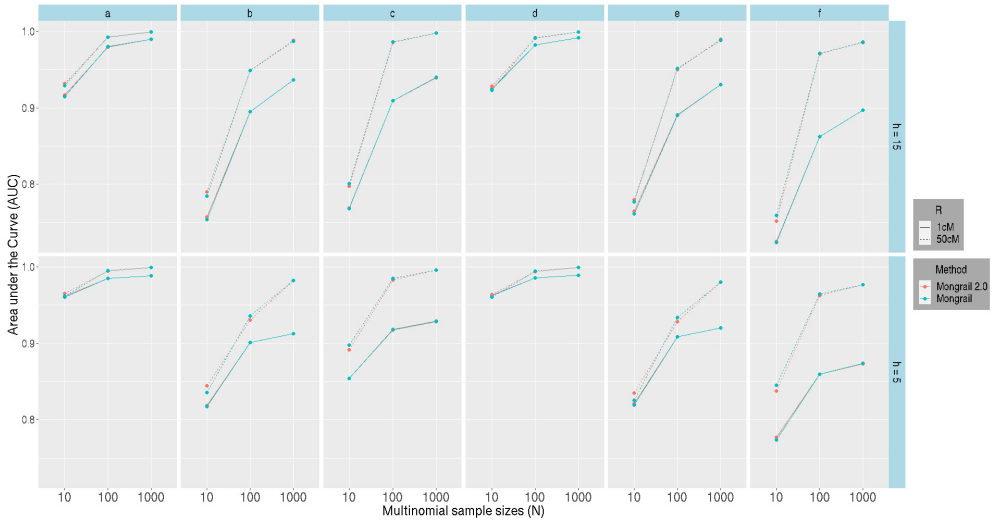

Fig. 3 AUC (Area under the curve) values (y-axis) plotted against different multinomial sample sizes N (x-axis) for Mongrail 2.0 (red line) and Mongrail (blue line). The plot is based on 10,000 individuals simulated using parameters: $K = 20, L = 10, c = 0.1, \alpha = 1$ and either expected recombination frequency of $R = 1$cM (solid linetype) or $R = 50$cM (dotted linetype). The first and second row corresponds to number of distinct haplotypes per chromosome with values $h = 15$ and 5 respectively. Results for the six genealogical classes (**a**-**f**) are shown from left to right in both rows. The 6 genealogical classes are as follows: **a**-pure population B, **b**-backcross with population A, **c**-F1 hybrid, **d**-pure population A, **e**-backcross with population B, **f**-F2 hybrid.

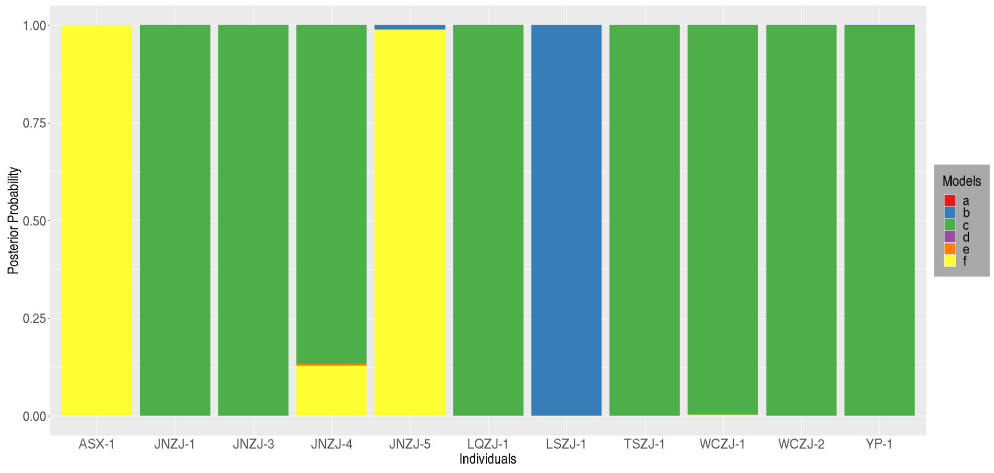

Fig. 4 Stacked bar plot showing the distribution of posterior probabilities for 11 presumed *A. zhejiangensis* individuals. The genealogical classes are : **a** - *A. hemsleyana*, **b** - Backcross with *A. eriantha*, **c** - F1 hybrid, **d** - *A. eriantha*, **e** - Backcross with *A. hemsleyana* and **f** - F2 hybrid.

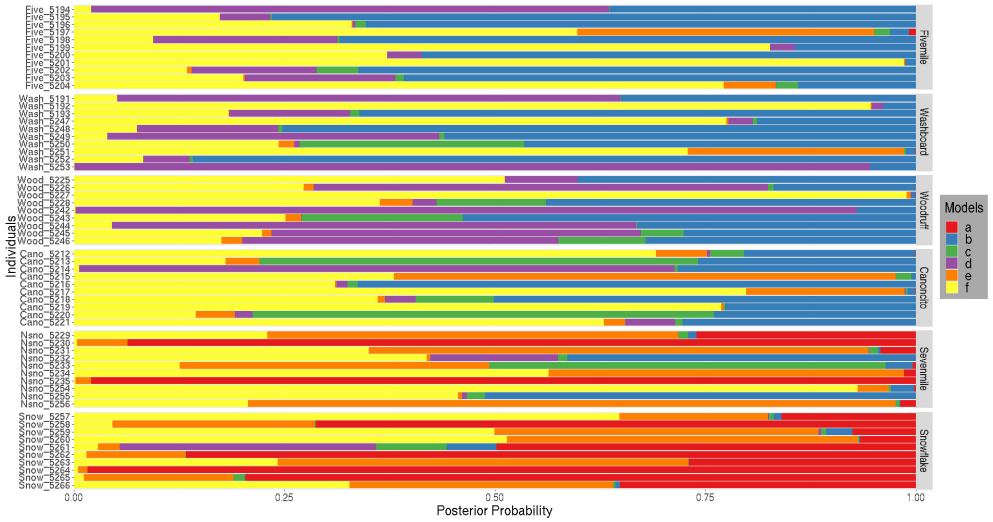

Fig. 5 Horizontal stacked bar plot showing the distribution of posterior probabilities for 60 plateau fence lizards (*Sceloporus tristichus*) sampled from the transect. Individuals are arranged according to the sampled sites along the transect. Panels correspond to the six sites - Fivemile Wash, Washboard Wash, Woodruff, Canoncito, Sevenmile Draw and Snowflake which are ordered latitudinally. The genealogical classes are : **a**-pure Show Low, **b**-backcross with Holbrook, **c**-F1 hybrid, **d**-pure Holbrook, **e**-backcross with Show Low, **f**-F2 hybrid.

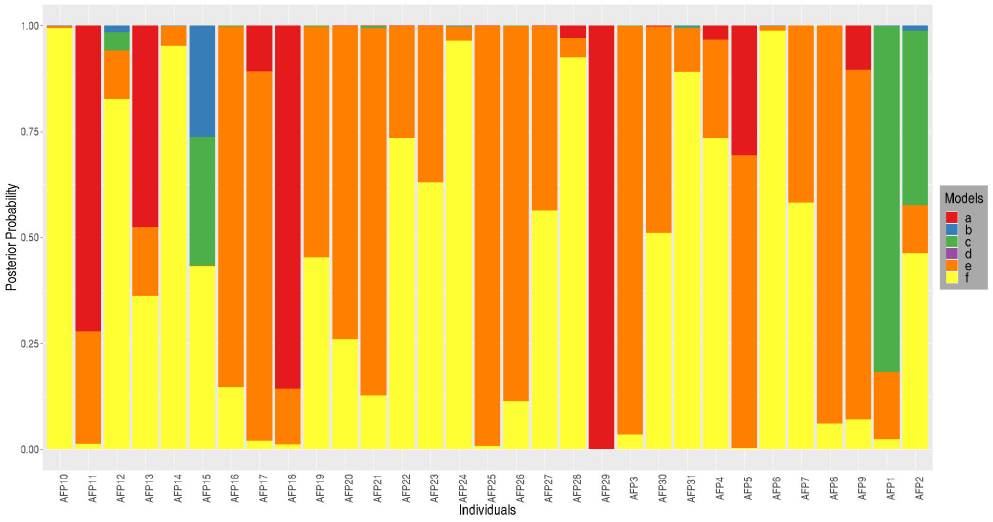

Fig. 6 Stacked bar plot showing the distribution of posterior probabilities for 31 post-rescue Florida panthers. The genealogical classes are : **a** - pre-rescue Florida panther, **b** - Backcross with Texas panther, **c** - F1 hybrid, **d** - Texas panther, **e** - Backcross with pre-rescue Florida panther and **f** - F2 hybrid.

Appendix A Recombinant haplotype probability

Following Chakraborty and Rannala (2023), an ancestry vector for a haplotype is a boolean array defined as $\mathbf{z} = \{z_j\}$ where $z_j \in \{0, 1\}$ with 0 or 1 indicating that a SNP locus at position j originates from population A or B, respectively. The probability of a recombinant haplotype $\mathbf{x}_i^c$ for chromosome i is given by

$$U(\mathbf{x}_i^c) = \sum_{\mathbf{z}} U(\mathbf{x}_i^c|\mathbf{z}).Q(\mathbf{z}|d_i, r), \quad c \in \{M, P\} \tag{A1}$$

The $Q(\mathbf{z}|d_i, r)$ term is defined in (Chakraborty and Rannala, 2023) and is restated in Appendix B. The term $U(\mathbf{x}_i^c|\mathbf{z})$ is defined below.

If a haplotype is entirely from either population A or B, the ancestry state $\mathbf{z}$ is a $L_i \times 1$ vector composed entirely of zeros or ones, respectively. The probability of the haplotype is

$$U(\mathbf{x}_i^c|\mathbf{z}) = \begin{cases} Pr(\boldsymbol{\phi}^c(\mathbf{x}_i)|\mathbf{n}_A), & \text{if } \mathbf{z} = [0, 0, \ldots 0] \\ Pr(\boldsymbol{\phi}^c(\mathbf{x}_i)|\mathbf{n}_B), & \text{if } \mathbf{z} = [1, 1, \ldots 1] \end{cases} \tag{A2}$$

For notational simplicity, we assume $L_i = L, \forall i = 1, 2, \ldots K$ (the argument extends to unequal L_i). If the haplotype $\mathbf{x}_i^c$ is recombinant between A and B, it can be divided into two sub-haplotypes, s_A^* and s_B^* based on ancestry state $\mathbf{z}$. A sub-haplotype from population k includes only alleles at markers derived from population k. Sub-haplotype s_k^* contains alleles of haplotype $\mathbf{x}_i^c$ whose markers belong to population $k \in \{A, B\}$. The probability of $\mathbf{x}_i^c$ given $\mathbf{z}$ is a product of sub-haplotype probabilities for s_A^* and s_B^*. The probability of a sub-haplotype s_k^* is calculated using the marginal counts of the sub-haplotype obtained by marginalizing over the reference sample haplotype counts ($\mathbf{n}_k$) from population k.

We previously defined matrices of population haplotypes O_A and O_B having dimensions $H \times L$, where each row is a distinct biallelic haplotype, H is the number of distinct haplotypes and L is the number of markers. Let I be an indexing set, $I = \{1, 2, \ldots L\}$. For a given ancestry state $\mathbf{z}$, let $I_k \subset I$ denote the set of indexes of markers with ancestry from population k. I_A and I_B satisy the conditions,

$$I_A \cup I_B = I \quad \text{and} \quad I_A \cap I_B = \phi. \tag{A3}$$

We define a $L \times 1$ unit vector $\mathbf{e}_j^{(L)} = \begin{bmatrix} e_{1j} & e_{2j} & \ldots & e_{Lj} \end{bmatrix}^{\mathrm{T}}$ such that for all $j' = 1, 2, \ldots, L$ and $j \in I_k$

$$e_{ij} = \begin{cases} 1 & \text{when } j' = j \\ 0 & \text{otherwise.} \end{cases} \tag{A4}$$

We define the matrix E_k of order $L \times |I_k|$ as

$$E_k = [\mathbf{e}_j^{(L)}] \quad \forall j \in I_k \tag{A5}$$

To extract sub-haplotypes whose ancestry belongs to population k, we post-multiply E_k with O_k

$$O_k E_k = S_k, \tag{A6}$$

where S_k is of order $H \times |I_k|$. The rows of the resulting matrix S_k are sub-haplotypes. The set of distinct sub-haplotypes in matrices S_A and S_B, denoted by $\mathcal{A}$ and $\mathcal{B}$ respectively, and their corresponding counts (marginals) in the two populations can be found as follows.

Define a $H \times 1$ unit vector $\mathbf{e}_h^{(H)} = \begin{bmatrix} e_{1h} & e_{2h} & \dots & e_{Hh} \end{bmatrix}^{\mathrm{T}}$ such that for all $h, h' = 1, 2, \dots, H$

$$e_{h'h} = \begin{cases} 1 & \text{when } h' = h \\ 0 & \text{otherwise.} \end{cases} \tag{A7}$$

Each of the matrices S_k can be expressed as a column of row vectors

$$S_k = \begin{bmatrix} s_{(k)1}^{\mathrm{T}} \\ s_{(k)2}^{\mathrm{T}} \\ \vdots \\ s_{(k)H}^{\mathrm{T}} \end{bmatrix} \qquad \text{where } s_{(k)h}^{\mathrm{T}} = \begin{bmatrix} s_{(k)h1} & s_{(k)h2} & \dots & s_{(k)h|I_k|} \end{bmatrix}. \tag{A8}$$

To get the h-th row of S_k we perform the operation

$$(\mathbf{e}_h^{(H)})^{\mathrm{T}} S_k = s_{(k)h}^{\mathrm{T}} \tag{A9}$$

Let I_H be an indexing set, $I_H = \{1, 2, \dots, H\}$. For h-th row we define a linear transformation,

$$f(s_{(k)h}^{\mathrm{T}}) = \sum_{j=1}^{|I_k|} s_{(k)hj} 2^{|I_k|-j} \tag{A10}$$

For population A, this linear transformation f defines a partition $\mathcal{P}_\mathcal{A}$ on S_A. $\mathcal{P}_\mathcal{A}$ partitions set I_H into $|\mathcal{A}|$ disjoint subsets $A_1, A_2, \dots, A_{|\mathcal{A}|}$ where

$$f(s_{(A)u}^{\mathrm{T}}) = f(s_{(A)v}^{\mathrm{T}}), \tag{A11}$$

for any $u, v \in A_y$ satisfying the following three conditions:

i. $A_y \neq \phi \quad$ for all $y = 1, 2, \dots, |\mathcal{A}|$

ii. $I_H = \bigcup_{y=1}^{|\mathcal{A}|} A_y$

iii. $A_y \cap A_{y'} = \phi \quad \forall y \neq y' \quad$ where $y, y' = 1, 2, \dots, |\mathcal{A}|$

Now we update the corresponding counts of the distinct sub-haplotypes. Let us define the indexing set $\tilde{I}_\mathcal{A} = \{1, 2, \dots, |\mathcal{A}|\}$. We can describe a set by associating its element with members of an index set. We define a set N_A that contains the reference sample haplotype counts from population A.

$$N_A = \{n_{hA} | h \in I_H\}. \tag{A12}$$

Now based on partition $\mathcal{P}_\mathcal{A}$ applied on S_A, we define set $N'_A = \{n'_{yA} | y \in \tilde{I}_\mathcal{A}\}$ where $n'_{yA} = \sum_{u \in A_y} n_{uA}$.

Therefore we have $|\mathcal{A}|$ distinct sub-haplotypes (denoted by $\tilde{s}_y^A$ where $y = 1, 2, \dots |\mathcal{A}|$ and their corresponding counts denoted by n'_{yA}). Similarly, we can define a partition

$\mathcal{P}_{\mathcal{B}}$ on S_B which partitions set I_H into $|\mathcal{B}|$ disjoint subsets. Subsequently, we have $|\mathcal{B}|$ distinct sub-haplotypes (denoted by $\tilde{s}_y^B$ where $y = 1, 2, \ldots |\mathcal{B}|$ and their corresponding counts denoted by n'_{yB}).
For $k \in \{A, B\}$, the sub-haplotype s_k^* is obtained by post multiplying E_k by $\mathbf{x}_i^c$

$$s_k^* = \mathbf{x}_i^c E_k, \tag{A13}$$

where s_k^* is of order $1 \times |I_k|$. The probability of s_k^* is

$$\begin{aligned} \log P(s_k^*|\mathbf{n}'_k) = \log\Gamma(\theta'_k) - \sum_{y=1}^{|m|} \log\Gamma(n'_{yk} + 1/|m|) \\ - \log\Gamma(1+\theta'_k) + \sum_{y=1}^{|m|} \log\Gamma(I_y(s_k^*) + n'_{yk} + 1/|m|) \end{aligned} \tag{A14}$$

where,

$$I_y(s_k^*) = \begin{cases} 1 & \text{if } s_k^* = \tilde{s}_y^k \\ 0 & \text{otherwise} \end{cases}, \theta'_k = 1 + \sum_{y=1}^{|m|} n'_{yk} \quad \text{and} \quad m = \begin{cases} \mathcal{A} & \text{if } k = A \\ \mathcal{B} & \text{if } k = B \end{cases} \tag{A15}$$

This is equivalent to calculating the posterior predictive distribution of the observed count, s_k^*, conditioned on $\mathbf{n}'_k = [n'_{1k}, n'_{2k}, \ldots, n'_{|m|k}]$. The probability of $\mathbf{x}_i^c$ given $\mathbf{z}$ is

$$U(\mathbf{x}_i^c|\mathbf{z}) = \exp\left\{\log P(s_A^*|\mathbf{n}'_A) + \log P(s_B^*|\mathbf{n}'_B)\right\}. \tag{A16}$$

As an example let us consider a case with $L = 5$ markers and $H = 7$ distinct haplotypes. The 7 haplotypes are expressed as rows in the following 7×5 matrices:

$$O_A = O_B = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 \end{bmatrix}$$

Let the recombinant haplotype be $\mathbf{x}_i^c = \begin{bmatrix} 1 & 1 & 1 & 0 & 1 \end{bmatrix}$ and the ancestry state be $\mathbf{z} = \begin{bmatrix} 0 & 0 & 1 & 0 & 1 \end{bmatrix}$. The ancestry state $\mathbf{z}$, indicates that markers 1, 2 and 4 belong to population A and markers 3 and 5 belong to population B. The indexing sets are

$$I = \{1, 2, 3, 4, 5\}; \quad I_A = \{1, 2, 4\}; \quad I_B = \{3, 5\}$$

Thus,

$$|I_A| = 3; \quad |I_B| = 2$$

$$E_A = \begin{bmatrix} e_1 & e_2 & e_4 \end{bmatrix}; \quad E_B = \begin{bmatrix} e_3 & e_5 \end{bmatrix}$$

E_A is of order 5×3 and E_B is of order 5×2. The matrices of the sub-haplotypes are calculated as

$$S_A = O_A E_A = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix} \quad \text{and} \quad S_B = O_B E_B = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}$$

All rows of S_A and 3 rows of S_B are distinct, so the sets of distinct sub-haplotypes are

$$\mathcal{A} = \{000, 001, 010, 101, 100, 110, 111\}; \quad \mathcal{B} = \{00, 10, 01\}$$

The indexing set is

$$I_H = \{1, 2, \ldots 7\}.$$

Applying the transformation f on S_A, partitions the set I_H into $|\mathcal{A}| = 7$ disjoint subsets $A_y = \{y\}$ for all $y = 1, 2, \ldots 7$. Similarly, applying f on S_B partitions the set I_H into $|\mathcal{B}| = 3$ disjoint subsets

$$B_1 = \{1, 3\}; \quad B_2 = \{2, 5, 7\}; \quad B_3 = \{4, 6\}.$$

For example, B_3 contains indexes of the rows of S_B (4 and 6) that are identical since

$$f(s^{\mathrm{T}}_{(B)4}) = f(s^{\mathrm{T}}_{(B)6}) = 1$$

For the set of distinct sub-haplotypes $\mathcal{A}$ the marginal counts are

$$n'_{yA} = \sum_{u \in A_y} n_{uA} = n_{yA}, \quad y = 1, 2, \ldots 7$$

For $\mathcal{B}$, the marginal counts are $n'_{1B} = \sum_{u \in B_1} n_{uB} = n_{1B} + n_{3B}$; $n'_{2B} = \sum_{u \in B_2} n_{uB} = n_{2B} + n_{5B} + n_{7B}$ and $n'_{3B} = \sum_{u \in B_3} n_{uB} = n_{4B} + n_{6B}$. The distinct sub-haplotypes for population A and B are denoted by $\tilde{s}^A_y \quad (y = 1, 2, \ldots 7)$ and $\tilde{s}^B_y \quad (y = 1, 2, 3)$. Given the recombinant haplotype $\mathbf{x}^c_i = \begin{bmatrix} 1 & 1 & 1 & 0 & 1 \end{bmatrix}$ and the ancestry state $\mathbf{z} = \begin{bmatrix} 0 & 0 & 1 & 0 & 1 \end{bmatrix}$, the sub-haplotypes s^*_A and s^*_B are given by

$$s^*_A = \mathbf{x}^c_i E_A = \begin{bmatrix} 1 & 1 & 0 \end{bmatrix}$$

$$s^*_B = \mathbf{x}^c_i E_B = \begin{bmatrix} 1 & 1 \end{bmatrix}$$

We see that $s^*_A = \tilde{s}^A_6 = 110$ implies $I_6(s^*_A) = 1$ and $\theta'_A = 1 + \sum_{y=1}^{7} n'_{yA} = 1 + \sum_{y=1}^{7} n_{yA}$.

The log-probability of sub-haplotype s^*_A is given by,

$$\begin{aligned}\log P(s_A^*|\mathbf{n}'_A) &= \log\Gamma(\theta'_A) - \sum_{y=1}^{7}\log\Gamma(n'_{yA}+1/7)\\ &\quad - \log\Gamma(1+\theta'_A) + \sum_{y=1}^{7}\log\Gamma(I_y(s_A^*)+n'_{yA}+1/7)\\ &= \log\Gamma(\theta'_A) - \log\Gamma(1+\theta'_A) + \log\Gamma(1+n'_{6A}+1/7) - \log\Gamma(n'_{6A}+1/7)\end{aligned}$$

But $s_B^* = 11 \neq \tilde{s}_y^B$(for any $y = 1, 2, 3$) implies $I_y(s_B^*) = 0$ for all $y = 1, 2, 3$ and $\theta'_B = 1 + \sum_{y=1}^{3} n'_{yB} = 1 + \sum_{y=1}^{7} n_{yB}$. The log-probability of sub-haplotype s_B^* is given by

$$\begin{aligned}\log P(s_B^*|\mathbf{n}'_B) &= \log\Gamma(\theta'_B) - \sum_{y=1}^{3}\log\Gamma(n'_{yB}+1/3)\\ &\quad - \log\Gamma(1+\theta'_B) + \sum_{y=1}^{3}\log\Gamma(I_y(s_B^*)+n'_{yB}+1/3)\\ &= \log\Gamma(\theta'_B) - \log\Gamma(1+\theta'_B)\end{aligned}$$

Therefore,

$$U(\mathbf{x}_i^c|\mathbf{z}) = \exp\left\{\log P(s_A^*|\mathbf{n}'_A) + \log P(s_B^*|\mathbf{n}'_B)\right\}$$

Appendix B Derivation of $Q(\mathbf{z}|d_i, r)$

Before redefining $Q(\mathbf{z}|d_i, r)$, we first need to define several other terms that will be used in its definition. We already defined in the Theory section (Data and Parameters) that chromosome i contains L_i loci with phased biallelic single-nucleotide polymorphisms. Earlier in Appendix A we have also defined an ancestry vector for a haplotype, $\mathbf{z} = \{z_j\}$ where $z_j \in \{0, 1\}$ denotes the population ancestry state of the marker j (where 0 represents population A and 1 indicates the position originates from population B). For chromosome i, the physical distance between markers is defined by d_{ij}, where d_{ij} represents the distance between markers $j-1$ to j and d_{i1} is the distance from the 5' end of chromosome i to marker 1. We assume a uniform recombination rate r on chromosome i, measured in centiMorgans (cM) per unit of physical distance. Therefore the map distance between markers $j-1$ and j (for chromosome i) is given by $d_{ij} \times r$ (cM). Under the assumption of no interference (i.e., recombination events on different intervals are independent of each other) and a uniform recombination rate, recombinations can be modeled as a Poisson process along the chromosome. Thus, the number of recombinations in an interval of

length d_{ij} follows a Poisson distribution with a mean of rd_{ij}. Hence the probability of an even number of recombinations occuring in an interval of length d_{ij} is given by

$$\sum_{n=0}^{\infty} \frac{\mathrm{e}^{-rd_{ij}}(rd_{ij})^{2n}}{[2n]!} = \mathrm{e}^{-rd_{ij}}(\cosh[rd_{ij}] - 1) + \mathrm{e}^{-rd_{ij}}. \tag{B17}$$

Note that zero recombination (or no recombination) falls under the category of an even number of recombination events. An even number of recombinations results in no change to the ancestry state of the markers. In contrast, an odd number of recombinations between two markers changes the ancestry state of the marker to the right of an interval. Thus, the probability of ancestry change in the interval d_{ij} is given by

$$P(d_{ij}, r) = 1 - (\mathrm{e}^{-rd_{ij}}\{\cosh[rd_{ij}] - 1\} + \mathrm{e}^{-rd_{ij}}). \tag{B18}$$

Following the above result, the probability of a particular ancestry state $\mathbf{z}$ for L_i SNP loci is given by,

$$\begin{aligned} Q(\mathbf{z}|d_i, r) = & \left\{\frac{1}{2} \times P(d_{i1}, r)^{z_1} \times [1 - P(d_{i1}, r)]^{1-z_1} \times P^*\right\} + \\ & \left\{\frac{1}{2} \times P(d_{i1}, r)^{|z_1-1|} \times [1 - P(d_{i1}, r)]^{1-|z_1-1|} \times P^*\right\}, \end{aligned} \tag{B19}$$

where,

$$P^* = \prod_{l=2}^{L_i} \left\{P(d_{il}, r)^{|z_l - z_{l-1}|} \times [1 - P(d_{il}, r)]^{1-|z_l-z_{l-1}|}\right\}. \tag{B20}$$

We now explain the derivation of $Q(\mathbf{z}|d_i, r)$ for a particular ancestry state $\mathbf{z}$. The summation in equation B19 of the two terms enclosed in curly braces represents two mutually exclusive and exhaustive events: a chromosome can either be sampled from population A or from population B, each with probability 1/2. Assuming no interference as we move along the chromosome from left to right, transitions from one population ancestry state to another will be calculated as independent conditional probabilities. Consider the first term, where the chromosome is sampled from population A. If $z_1 = 0$, this indicates that the ancestry state did not change on interval d_{i1}, and the probability of no change is $[1 - P(d_{i1}, r)]$. Alternatively, if $z_1 = 1$, the ancestry state changes on interval d_{i1}, which has a probability of $P(d_{i1}, r)$. The second term, where the chromosome is sampled from population B, is derived similarly. For the remaining loci $(z_l; l > 1)$, probabilities are combined into the term P^* defined in equation B20, where $P(d_{il}, r)$ represents the probability of an ancestry state change $(z_l \neq z_{l-1})$, and $[1 - P(d_{il}, r)]$ represents the probability of no change $(z_l = z_{l-1}; l \neq 1)$.