

UC Davis

UC Davis Electronic Theses and Dissertations

Title

An Efficient Exact Algorithm for Identifying Hybrids Using Population Genomic Sequences

Permalink

<https://escholarship.org/uc/item/1m53s12k>

Author

Chakraborty, Sneha

Publication Date

2023

Peer reviewed|Thesis/dissertation

An Efficient Exact Algorithm for Identifying Hybrids Using Population Genomic Sequences

By

SNEHA CHAKRABORTY
DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Biostatistics

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Bruce H. Rannala, Chair

Brian R. Moore

Ziheng Yang

Committee in Charge

2023

Copyright© Sneha Chakraborty, 2023. All rights reserved.

In the loving memory of my grandma, Pratima Banerjee (1951-2023)

Contents

Abstract	v
Acknowledgments	vii
Chapter 1. Introduction	1
Chapter 2. Theory	5
2.1. Genealogical class	5
2.2. Model and likelihood	6
2.3. Allelic state of markers	11
2.4. Likelihood	13
2.5. Likelihoods for genealogical classes	13
2.6. Bayesian inference	15
2.7. Estimation of population haplotype frequencies	16
2.8. Inference for unphased individuals	17
Chapter 3. Simulation Methods	19
3.1. Comprehensive simulation	19
3.2. Coalescent simulation	24
Chapter 4. Simulation Results	26
4.1. Comprehensive simulation	26
4.2. Coalescent simulation	34
Chapter 5. Empirical Analysis of Spotted and Barred Owl Hybridization	38
5.1. Background	38
5.2. Materials and methods	39
5.3. Results	43

Chapter 6. Discussion	50
6.1. Data availability	54
6.2. Funding	55
Appendix A. Simulation	56
A.1. Comprehensive simulation: Methods	56
A.2. Simulating diplotypes for markers	58
A.3. Coalescent simulation: Simulating diplotypes	62
Appendix B. Empirical Dataset	64
B.1. Mongrail: Analysis of putative hybrids	64
B.2. Applying NewHybrids to the owl dataset	64
Bibliography	69

Abstract

The identification of individuals that have a recent hybrid ancestry (between populations or species) has been a goal of naturalists for centuries. Since the 1960s, codominant genetic markers have been used with statistical and computational methods to identify F1 hybrids and back crosses. Existing hybrid inference methods assume that alleles at different loci undergo independent assortment (are unlinked or in population linkage equilibrium). Genomic datasets include thousands of markers that are located on the same chromosome and are in population linkage disequilibrium which violate this assumption. Existing methods may therefore be viewed as composite likelihoods when applied to genomic datasets and their performance in identifying hybrid ancestry (which is a model-choice problem) is unknown. Here we develop a new program Mongrail that implements a full-likelihood Bayesian hybrid inference method that explicitly models linkage and recombination, generating the posterior probability of different F1 or F2 hybrid, or backcross, genealogical classes. We use simulations to compare the statistical performance of Mongrail with that of an existing composite likelihood method (NewHybrids) and apply the method to analyze genome sequence data for hybridizing species of barred and spotted owls.

Chapter 1 reviews the different types of hybrid inference methods present in literature from the 1960s till present. The review traces the gradual development of the inference methods with advancement in sequencing technologies. We discuss how the assumption of independence among loci (applied by most existing methods) adversely affects the analysis of current genomic datasets.

In Chapter 2 we propose a hybridization model based on diplotypes under a two-generational pedigree when we consider two sympatric diploid populations. We present a novel way of calculating the exact likelihood of the SNP data under a model with recombination by using the knowledge of physical distance between the markers. Our method requires phased data and population haplotype frequencies to be known when calculating the likelihoods. But we also present some alternatives when these quantities are unknown. We use a point estimate to estimate the haplotype frequencies for the two populations using individuals who are unlikely to be hybrids. And for hybrid individual without any phase information we calculate the likelihood by integrating over all compatible diplotypes.

Chapter 3 describes the two major simulation study designs that were used to compare the two inference methods (NewHybrids and Mongrail) on datasets where the markers were linked. The first design (Comprehensive Simulation) involved generating a diverse set of haplotype frequency distributions. Whereas the second design (Coalescent Simulation) used a structured coalescent model with recombination, thus allowing the statistical performance of the two methods to be evaluated under biologically realistic conditions. Chapter 4 presents an exhaustive summary of the simulation results for both the study designs. We find that in general, Mongrail is more effective in distinguishing hybrids and backcrosses compared to NewHybrids under both the simulation study designs. One of the most noteworthy findings from the simulation study is that the number of chromosomes and the map-length of the chromosome, contribute more to power (to infer the correct genealogical class) than the number of markers. This outcome is extremely advantageous since it is more computationally challenging to increase the number of markers than increasing the number of chromosomes or map-length.

Chapter 5 presents the application of our method to a genomic dataset on spotted owls, barred owls and their hybrids. We give a brief background on the dataset and describe the methods we employed to analyze the dataset. Mongrail was able to infer the genealogical classes for all putative hybrids with high posterior probability.

Finally Chapter 6 briefly summarizes the findings from the simulation study designs and the empirical analysis. We conclude with a discussion about the strengths and weaknesses of our method and future research directions.

Acknowledgments

The journey of getting a PhD was anything but smooth sailing. But as the saying goes “*The wind and the waves are always on the side of the ablest navigator*”, this ship would not have made it to the shore without my advisor Bruce Rannala at the helm. He has been such a wonderful guide. I thank Bruce for his mentorship and extreme patience as I slowly transitioned to population genetics, an area of research completely new to me. His passion for research and drive for excellence is truly inspiring. I thank him for his brilliant inputs and invaluable insights that added so many layers to this research problem. He has been very supportive during some of my tough times. He was very generous with his time and guidance. Our weekly brainstorming sessions were not confined within the walls of research. We had fun discussions about sailboats, music, movies and basically anything and everything. Interaction with Bruce has been the most rewarding experience at UC Davis and he has helped me become a better researcher.

I sincerely thank my committee members Ziheng Yang and Brian Moore for their continuous support and meticulous feedback that brought some huge improvements to my thesis over the years. I am very grateful to the staff members in the Statistics and Evolution & Ecology department for helping me throughout my graduate life at UC Davis.

It has been such an honour to study Statistical Genetics under the tutelage of late Professor Saurabh Ghosh at ISI, Kolkata. I shall be forever indebted to him for kindling my research interest in the field of genetics. I would like to express my sincere gratitude to my excellent teacher Subhasis Ray Choudhury who helped me fall in love with Mathematics and in the process changed my life forever. It is impossible to thank all my teachers during my days as an undergraduate and masters student but I am very thankful to each and every one of them.

I think it takes a village to finish a PhD, so my journey would be incomplete without thanking my friends. It has been great to have friends like Samayita Bhattacharjee, Anubhab Gupta, Subhadip Dey, Sayantani Sarkar, Abhijnan Chattopadhyay, Anna Nagel, Maxime Pouokam, Paromita Dubey, Abhishek Roy and Indrajit Jana. I have had such wonderful memories with my friends who stood by me through thick and thin and made Davis a home away from home.

Last but not the least, I would like to thank my family especially my parents for their unconditional love, support and endless sacrifices. I owe everything to them.

CHAPTER 1

Introduction

For hundreds of years naturally occurring hybrids between species have been identified based on their morphology, often intermediate by comparison with the pure parental species. During the twentieth century, other diagnostics such as karyotypes, blood groups and isozymes were increasingly used to identify hybrids. However, such diagnoses are inherently subjective and confounded by the presence of backcrosses. Codominant genetic markers potentially offer a more objective form of data and have been increasingly used to identify hybrid individuals in natural populations ever since the development of allozyme markers in the 1960s [23, 29]. Genetic distance-based “hybrid indexes” were proposed in the 1970s, for example, to assign to individuals from known hybrid zones degrees of hybrid ancestry using allozyme allele frequencies of source populations [25]. Applications of allozyme markers to identify population or species hybrids in animals and plants flourished during the 1980s [17, 26, 28], and microsatellites [37] or AFLPs [10] were commonly used during the 1990s and 2000s. The applications were often species conservation or, in the case of fish, stock management [40]. During the 1980s, diagnostic criteria were proposed to identify hybrids based on explicit considerations of Mendelian inheritance. Some of these [5] rely on fixed allelic differences between pairs of populations (or species). If loci with fixed differences exist, F1 hybrids will be heterozygous for all such loci and (if markers are unlinked) the expected fraction of individuals that can be identified as F1 hybrids versus F2 backcrosses can be predicted as a function of the number of marker loci [5]. Diagnostic statistics have also been developed to identify F1 or F2 hybrids and specific backcrosses (F1 x population 1, F1 x population 2) if one or more alleles are exclusive to each population but not necessarily fixed [31].

Fixed differences (or exclusive alleles) do not always exist between populations (or species) and (even if they do) it is impossible to be certain an allele is fixed (or exclusive) without exhaustive sampling. Another class of hybrid inference methods thus relax the requirement for fixed allelic differences, instead relying on differences in allele frequencies between populations (or species) and

calculating probabilities of multilocus genotypes in F1 and F2 hybrids versus pure individuals [8]. Some methods for estimating individual proportions of admixture between populations [33, 34, 36], or migrant ancestry [44] also potentially identify F1 hybrids. In particular, [2] developed a powerful Bayesian method for distinguishing between hybrids and backcrosses based on multinomial sampling theory that has been widely used. In this paper, we focus on extending the hybrid inference method of [2] to accommodate genomic data using multilocus genotypes, or haplotypes, without a requirement for fixed or exclusive alleles.

1.0.1. Assumptions of hybrid inference methods. Model-based methods for identifying hybrids assume random mating (Hardy-Weinberg equilibrium) within populations (species) and some form of independence among loci. The independence requirement (among loci) is usually stated as an assumption of either unlinked loci [5, 31, 34], or (linked or unlinked) loci that are in population linkage equilibrium [2, 8, 36]. The distinction can be important because markers on the same chromosome may be in linkage equilibrium, whereas none are unlinked.

The statistical assumption of independence among loci implies that the joint probability of alleles at multiple loci, calculated under a model with linkage, should be equal to the probability calculated as a product (across loci) of the marginal probabilities. This is generally stated without proof and thus axioms of the statistical methods are implicit. The more general requirement of statistical independence of alleles among loci seems to underlie both types of assumptions (unlinked loci or linkage equilibrium) in most methods. For large genomic datasets, many loci will be linked and in population linkage disequilibrium so that the assumptions of existing methods will be violated. Thus, alternative methods that relax these assumptions are desirable. A method that relaxes the independence assumption must formulate the problem in terms of linked markers on chromosomes and haplotype frequencies in populations, rather than multilocus genotypes and allele frequencies.

1.0.2. Full versus composite likelihood. If a likelihood or Bayesian statistical method is applied that assumes unlinked loci, when the loci are actually linked, the method becomes a composite likelihood. Composite likelihoods can produce efficient point estimators, with estimates converging to the true parameter value with increased sample size [42]. However, the problem of

identifying hybrid class is a model choice problem not a point estimation problem – the performance of composite likelihood methods for model choice is poorly understood [43]. A full-likelihood method that explicitly models recombination, thus allowing linked loci, would eliminate the need for composite likelihood approximations and the potential problems of existing estimators. This paper aims to develop a full-likelihood method that explicitly models recombination during the formation of hybrids.

With the advent of genomic datasets of potentially millions of markers linked on chromosomes for which only a few recombination events are expected to occur per meiosis, the independence assumptions of composite likelihood methods will inevitably be violated. However, existing simulation studies examining the statistical performance of hybrid identification methods explicitly assume that markers are unlinked [41]. The effects of linkage on the performance of these composite likelihood methods is thus unknown. In particular, when thousands of markers are used so that the assumptions of the methods are strongly violated the effects of model violations could be extreme. Although the simulation method described in [45] includes a model of linkage, it has not yet been used to compare different hybrid inference methods. This paper develops a simulator that includes both linkage and recombination during the formation of population hybrids, allowing the statistical performance to be evaluated under a more realistic model for both existing (composite likelihood) hybrid inference methods and the new full-likelihood methods developed in this paper.

1.0.3. Hybrid inference using genomic data with linked markers. There are two major factors accounting for the requirement, ubiquitous among existing hybrid inference methods, of independent assortment of alleles among loci: (1) the positions of first-generation genetic markers were usually unknown (a linkage or a physical map was not available) for most organisms; (2) explicitly modeling linkage and recombination is complex and computationally demanding. Advances in genome sequencing are rapidly altering this first factor, providing inexpensive physical maps of millions of SNP markers for virtually any species, and statistical methods and computer speed are rapidly converging to reduce or eliminate the second limiting factor (computational complexity).

In this paper, we present a full-likelihood Bayesian method for identifying hybrids and backcrosses using biallelic SNP loci available from population genomic samples. The method explicitly models recombination by using a physical map of the markers. To maximize the efficiency of the

method we implement many of the computations as low-level bit operations in the C programming environment. To examine the effects of linkage on composite likelihood methods we conducted a simulation study of the performance of the [2] composite likelihood method, when used with linked markers, by comparison with our full-likelihood method which incorporates linkage. As an application of our method we analyze a real data set consisting of spotted owls, barred owls and their hybrids [15,21]. This dataset includes an improved spotted owl genome assembly and 51 high coverage whole-genome sequences [15].

CHAPTER 2

Theory

Most methods using genetic markers to identify population hybrids of a diploid species have assumed that genotypes at different loci undergo independent assortment (e.g., are unlinked or in linkage equilibrium). In particular, the widely used method of [2] incorporates this assumption – when markers are actually linked their method becomes a composite likelihood approximation. Here, we extend the [2] model to allow genomic sequence data comprised of linked SNPs to be jointly analyzed using an exact full-likelihood approach. We consider two sympatric diploid populations, labelled A and B, and assume they were initially isolated but have been interbreeding for the last n generations. We follow [2] who considered all the combinations of hybrid ancestries that can result with n generations of interbreeding between two populations but consider explicit results only for the case of a recent population hybridization event ($n \leq 2$ generations). Here, we consider the diploid genome sequence for a single individual and how this may be used to infer the hybrid status of the individual. In the absence of close relatedness between sampled individuals the hybrid status of multiple individuals may be inferred independently.

2.1. Genealogical class

A non-inbred pedigree of n generations describing the ancestors of a single individual includes 2^n founders, so in our two generation case there are $2^2 = 4$ founders. We consider the founders to be purebred. We can identify 6 distinct classes of such two-generational pedigrees by considering the number and arrangement of founders originating from a specified population (population A, for example). Following [31] we refer to these as genealogical classes (see Figure 2.1), where genealogical classes **a** and **d** are purebreds; **b** and **e** are backcrosses; **c** is a F1 hybrid and **f** is a F2 hybrid. We use the term genealogical class and model interchangeably. Our approach differs from [31] in that we consider diplotypes rather than marker genotypes. The diplotype is the pair of haplotypes on homologous chromosomes.

If the map distances between markers and the population haplotype frequencies are known, then under a random mating process (within populations) the probability of the observed marker data (likelihood) is completely specified for any given genealogical class. The focus of this paper will be on developing and implementing an efficient algorithm for calculating this likelihood, thus allowing inference of the genealogical class of an individual using linked SNP marker data.

2.2. Model and likelihood

Here we describe the likelihood calculation for a set of multiple biallelic SNP markers located on the two copies (maternal and paternal) of a single chromosome from an individual of a diploid sexual species that undergoes recombination. Because different chromosomes segregate independently during meiosis, the likelihood for markers on multiple chromosomes is simply a product across the likelihoods for the individual chromosomes. Following [2] the objective of the inference will be to infer the hybridization history of an individual (genealogical class) which is essentially a model choice problem. We will use posterior model probabilities to evaluate the support for different genealogical classes. Here, we present the data, model parameters, and formulas needed for calculating the likelihood under each of the 6 possible genealogical classes (see Figure 2.1) for 2 populations.

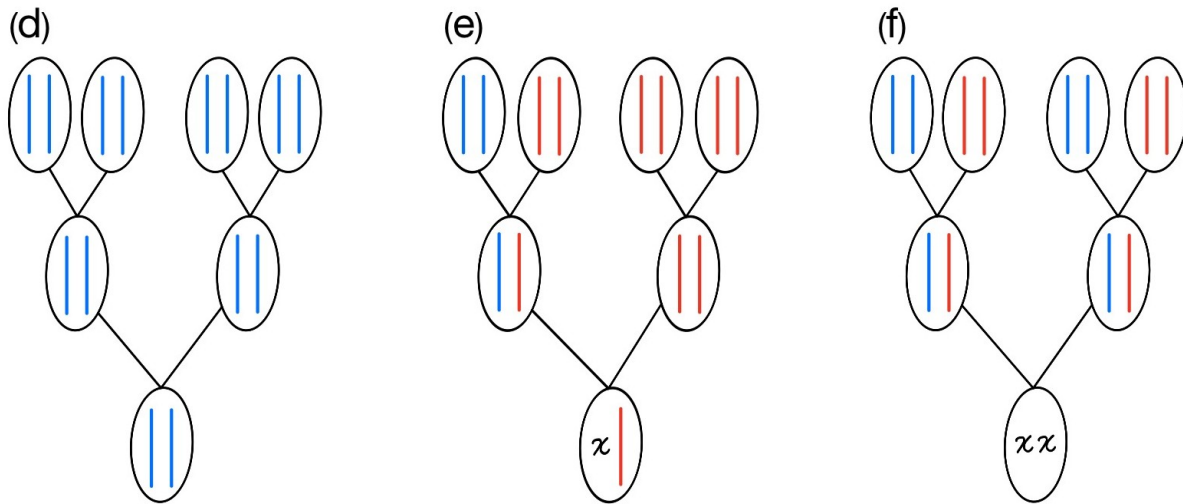
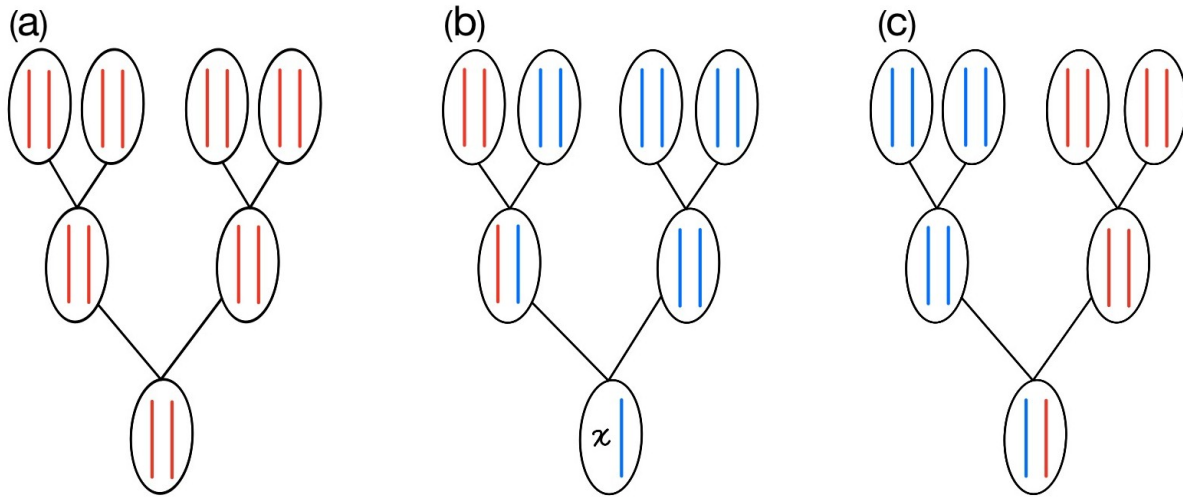
2.2.1. Data and parameters. Consider a sample of K chromosomes from a diploid individual. Chromosome i contains L_i loci with phased biallelic single-nucleotide polymorphisms. We represent the maternally (M) and paternally (P) inherited copies of the chromosomes as matrices,



$$\begin{aligned}\mathbf{x}^M &= \{x_{ij}^M\}, \\ \mathbf{x}^P &= \{x_{ij}^P\},\end{aligned}$$

where $x_{ij}^M \in \{0, 1\}$ is the allele (coded as 0,1) present at the j th SNP locus on the maternally inherited copy of chromosome i , etc. The complete data for an individual are then $\mathbf{x} = \{\mathbf{x}^M, \mathbf{x}^P\}$. We define the physical distances between markers on chromosomes as

$$\mathbf{d} = \{d_{ij}\},$$

6



 Chromosome originating from population A
 Chromosome originating from population B

χ Chromosome is mixture of haplotypes from population A and B

FIGURE 2.1. Pedigrees of relationships among founders for $n = 2$ generations. Circles represent diploid individuals and the pair of lines within each individual represent the two chromosomes; the blue line denotes the chromosome originating from population A and the red line represents the chromosome from population B. Founders are individuals at the top of the pedigree. An individual's genealogical class (a-f), is defined by the population origins among founders.

where d_{ij} is the distance on chromosome i from marker $j - 1$ to j and d_{i1} is the distance from the 5' end of chromosome i to marker 1.

We define the recombination rates on the intervals between markers as

$$\mathbf{r} = \{r_{ij}\},$$

where r_{ij} is the recombination rate on chromosome i for the interval between markers $j - 1$ and j in units of centiMorgans (cM) per unit of physical distance. For example, if physical distance is measured in megabases (Mb) the units would be cM/Mb. Recall that 1 cM = 1 % recombination per meiosis. The map distance for the interval $[j - 1, j]$ of chromosome i is defined by the product $d_{ij} \times r_{ij}$ and is measured in units of cM (percent recombination per meiosis).

We consider a model with hybridization between 2 populations A and B. The ancestry matrix for each chromosome specifies the population origin of each marker locus,

$$\begin{aligned} \mathbf{z}^M &= \{z_{ij}^M\}, \\ \mathbf{z}^P &= \{z_{ij}^P\}, \end{aligned}$$

where populations A and B are denoted by 0 and 1, respectively, and $z_{ij}^M = 0$ specifies that marker j of the maternally inherited copy of chromosome i originates from a founder chromosome that was in population A, and so on. The complete ancestry matrix for an individual is then $\mathbf{z} = \{\mathbf{z}^M, \mathbf{z}^P\}$.

Population haplotype frequencies are also needed to calculate likelihoods. We assume that the population is randomly mating so that diplotype probabilities can be calculated directly from haplotype frequencies for non-hybrid segments of chromosomes. We define $f^A(x_i)$ to be the frequency of haplotype x_i on chromosome i in population A and $f^B(x_i)$ its frequency in population B. In this paper, we treat population haplotype frequencies as known when calculating likelihoods. In empirical analyses, haplotype frequencies are estimated using individuals who are unlikely to be hybrids (for example, individuals sampled outside of a hybrid zone). For the individual being tested for potential hybrid status we integrate over the unknown haplotype phase taking account of uncertainty (see section Inference for unphased individuals).

2.2.2. Population ancestry states of markers. For the case of 2 populations and 2 generations of hybridization the only non-trivial chromosome ancestry probability to calculate is that for a chromosome that is a recombinant between two pure chromosomes, one from population A and the other from population B. We consider autosomes so that it does not matter which specific chromosome is maternally, or paternally, inherited. We assume no interference (independence of recombination events on different intervals) allowing the transitions from one population ancestry state to another along the chromosome from left to right (see Figure 2.2) to be calculated as independent conditional probabilities. Here, we present the probability calculation for chromosome i with L_i linked loci. We omit the subscript i from the population ancestry vector z for simplicity.

The population origin of a SNP locus to the right of an interval changes whenever there is an odd number of recombinations (1,3,5, etc) on the interval. We denote a distinct population state (or, ancestry state) as

$$\mathbf{z} = \{z_j\}; j = 1, \dots, L_i,$$

where $z_j \in \{0, 1\}$ is the population origin of the marker j (where 0 represents population A and 1 population B).

There are 2^{L_i} possible distinct population states for L_i loci, for example:

$$\underbrace{0 \dots 0}_{L_i}, \underbrace{1 \dots 1}_{L_i}, \underbrace{0 \dots 0}_{L_i-1} \underbrace{1}_1, \underbrace{0 \dots 0}_{L_i-2} \underbrace{11}_2, \dots$$

For simplicity, here we treat the rate of recombination as uniform ($r_{ij} = r$) on chromosome i , although the implementation allows recombination rates to vary as specified by the user. Given the assumption of no interference, recombinations occur as a Poisson process along the chromosome, such that the number of recombinations in an interval of length d_{ij} is Poisson distributed with a mean of rd_{ij} . Accordingly, the probability that an even number of recombinations occur on an interval of length d_{ij} is

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

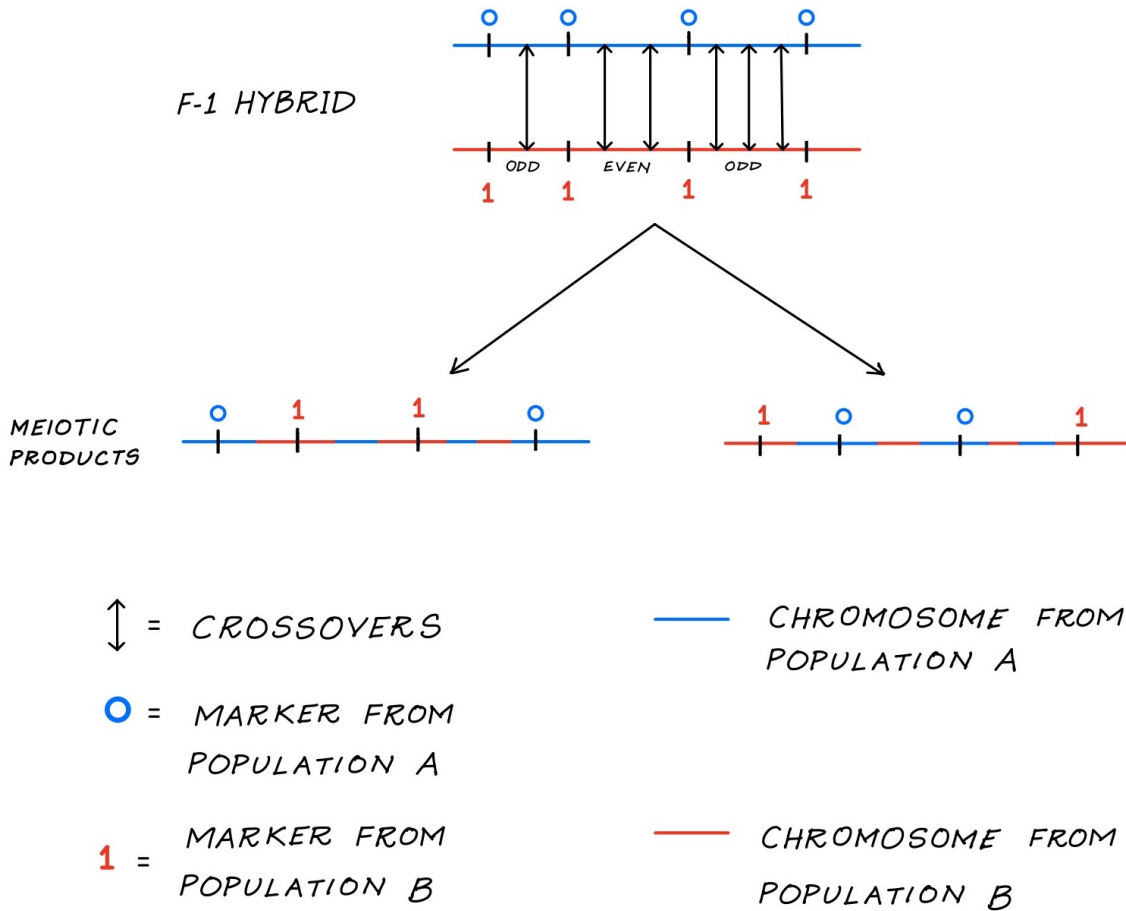


FIGURE 2.2. Population origin of $L = 4$ markers as a result of recombination (or, crossovers) between two pure chromosomes (a F1 hybrid): one from population A (blue) and other from population B (red). The two possible gametes (recombinant haplotype) produced at the end of meiosis is shown with the population origin labelled on the markers. An odd number of recombinations between two markers changes the population origin of the marker to the right of an interval whereas an even number of recombinations results in no change in the population state of the markers.

Thus the probability that there is a change of population on the interval between the $(j - 1)$ -th marker and j -th marker is

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

where the last term in parentheses on the right is the probability of no recombinations (which also results in no change of population ancestry). Using this result we can calculate the probability of a particular ancestry state \mathbf{z} for L_i SNP loci,

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

where

$$P^* = \prod_{k=2}^{L_i} \left\{ P(d_{ik}, r)^{|z_k - z_{k-1}|} \times [1 - P(d_{ik}, r)]^{1-|z_k - z_{k-1}|} \right\}.$$

Here we explain the derivation of $Q(\mathbf{z}|d_i, r)$ for a particular ancestry state \mathbf{z} . A chromosome can either be sampled from population A (with probability 1/2) or from population B (with probability 1/2). This explains the summation of the two (mutually exclusive and exhaustive events) represented by the terms in curly braces. Considering the first term, given that the chromosome was sampled from population A, if $z_1 = 0$ no change of population state occurred on interval d_{i1} . The probability of no change is $[1 - P(d_{i1}, r)]$. Otherwise, if $z_1 = 1$ the population state changed on interval d_{i1} . The probability a change occurs is $P(d_{i1}, r)$. The derivation for a chromosome sampled from population B (the second term) is similar. For the remaining loci ($z_k; k > 1$) the probability is $P(d_{ik}, r)$ if the ancestry state z_k changed from z_{k-1} , otherwise the probability is $[1 - P(d_{ik}, r)]$ if there was no change ($z_k = z_{k-1}; k \neq 1$), and these probabilities are combined into the term P^* .

2.3. Allelic state of markers

We now consider the probabilities of the SNP alleles observed conditional on the ancestry states of the markers. Let x_i be the SNP haplotype for a (maternally or paternally inherited) chromosome i and let \mathbf{z} be the population ancestry vector (again dropping the chromosome i subscript for simplicity). The probability of the SNP alleles on the chromosome for the two trivial cases (either the entire chromosome comes from population A or from population B) are given by

$$f(x_i|\mathbf{z}) = \begin{cases} f^A(x_i), & \text{if } z_j = 0 \quad \forall j = 1, \dots, L_i \\ f^B(x_i), & \text{if } z_j = 1 \quad \forall j = 1, \dots, L_i \end{cases}$$

Now consider the non-trivial case (i.e., the chromosome is a recombinant between two pure chromosomes, one from population A and the other from B). First, we define the marginal frequency of alleles on sub haplotypes $\{x_{iu}, x_{i(u+1)}, x_{i(u+2)}, \dots, x_{iv}\}$ from the u -th marker to the v -th marker where $1 \leq u < v \leq L_i$ in population A (or, B) be $f^A(x_{i[u,v]})$ or, $f^B(x_{i[u,v]})$ respectively. Then, for any ancestry state \mathbf{z} we consider each segment of consecutive 0's or 1's, counting the number of segments and the lengths of each segment (in units of number of markers rather than physical distance). Consider the case of c segments with lengths j_1, j_2, \dots, j_c . The calculation differs for even versus odd numbers of segments.

Case I: Even number of segments $c = 2n$ and $z_1 = 0$ (i.e., the starting state is population A),

$$\begin{aligned}
f(x_i|\mathbf{z}) &= f^A(x_{i[1,j_1]}) \\
&\times f^B(x_{i[j_1+1,j_1+j_2]}) \\
&\times f^A(x_{i[j_1+j_2+1,j_1+j_2+j_3]}) \\
&\vdots \\
&\times f^B(x_{i[j_1+j_2+\dots+j_{c-1}+1,j_1+j_2+\dots+j_c]})
\end{aligned}$$

The terms are thus an alternating sequence of marginal frequencies from populations A and B and there are n marginal frequency terms from each. If $z_1 = 1$ the first term instead starts with population B and the last term is from population A.

Case II: Odd number of segments $c = 2n + 1$ and $z_1 = 0$ (i.e., the starting state is population A),

$$\begin{aligned}
f(x_i|\mathbf{z}) &= f^A(x_{i[1,j_1]}) \\
&\times f^B(x_{i[j_1+1,j_1+j_2]}) \\
&\times f^A(x_{i[j_1+j_2+1,j_1+j_2+j_3]}) \\
&\vdots \\
&\times f^B(x_{i[j_1+j_2+\dots+j_{c-2}+1,j_1+j_2+\dots+j_{c-1}]})) \\
&\times f^A(x_{i[j_1+j_2+\dots+j_{c-1}+1,j_1+j_2+\dots+j_c]})
\end{aligned}$$

In this case ($z_1 = 0$), there are $(n + 1)$ marginal frequency terms from population A and n from B. If $z_1 = 1$, the first and last terms instead come from population B so that $(n + 1)$ marginal frequency terms are from population B and n from A.

2.4. Likelihood

For the maternally inherited chromosome i , the probability of a L_i -loci hybrid SNP haplotype (i.e., a recombinant SNP haplotype), x_i^M , unconditioned on the ancestry state of the markers is given by

$$\begin{aligned}
L(x_i^M|d_i, r) &= \sum_z L(x_i^M, \mathbf{z}|d_i, r) \\
&= \sum_z f(x_i^M|\mathbf{z}) \cdot Q(\mathbf{z}|d_i, r)
\end{aligned}$$

where the sum is over the set of all possible ancestry states at L_i markers (there are 2^{L_i} distinct combinations). Because the number of distinct combinations grows with L_i , it is currently computationally practical to analyze 15 or fewer loci per chromosome.

2.5. Likelihoods for genealogical classes

Here, we present formulas for calculating the likelihood of a diplotype, $\mathbf{x} = \{\mathbf{x}^M, \mathbf{x}^P\}$ for an individual under each of the 6 possible genealogical classes (see Figure 2.1). We use the term

genealogical class and model interchangeably. We introduce the variable $G : G = g$ denotes that the individual belongs to the genealogical class g . We define the indicator function,

$$I = \begin{cases} 1 & \text{if } x_i^M \neq x_i^P \\ 0 & \text{if } x_i^M = x_i^P \end{cases}$$

Model (a): Both chromosomes (x_i^M and x_i^P) come from Population B (i.e., $z_j = 1 \quad \forall j = 1, \dots, L_i$)

$$L(x^M, x^P | G = 1, \mathbf{d}, r) = \prod_{i=1}^K \{2^I \cdot f^B(x_i^M) \cdot f^B(x_i^P)\},$$

Model (b): One chromosome comes from Population A and the other chromosome is a recombinant.

$$\begin{aligned} L(x^M, x^P | G = 2, \mathbf{d}, r) = & \\ & \prod_{i=1}^K \left(\left[f^A(x_i^M) \left\{ \sum_z f(x_i^P | \mathbf{z}) \cdot Q(\mathbf{z} | d_i, r) \right\} + \right. \right. \\ & \left. \left. f^A(x_i^P) \left\{ \sum_z f(x_i^M | \mathbf{z}) \cdot Q(\mathbf{z} | d_i, r) \right\} \right]^I \right. \\ & \left. \times \left[f^A(x_i^M) \left\{ \sum_z f(x_i^P | \mathbf{z}) \cdot Q(\mathbf{z} | d_i, r) \right\} \right]^{1-I} \right), \end{aligned}$$

Model (c): One chromosome comes from Population A and the other from Population B

$$\begin{aligned} L(x^M, x^P | G = 3, \mathbf{d}, r) & \\ & = \prod_{i=1}^K \left[\left\{ f^A(x_i^M) \cdot f^B(x_i^P) + f^A(x_i^P) \cdot f^B(x_i^M) \right\}^I \right. \\ & \left. \times \left\{ f^A(x_i^M) \cdot f^B(x_i^P) \right\}^{1-I} \right], \end{aligned}$$

Model (d): Both chromosomes (x_i^M and x_i^P) come from Population A (i.e., $z_j = 0 \quad \forall j = 1, \dots, L_i$)

$$L(x^M, x^P | G = 4, \mathbf{d}, r) = \prod_{i=1}^K \{2^I \cdot f^A(x_i^M) \cdot f^A(x_i^P)\},$$

Model (e): One chromosome comes from Population B and the other chromosome is a recombinant.

$$L(x^M, x^P | G = 5, \mathbf{d}, r) = \prod_{i=1}^K \left(\left[f^B(x_i^M) \left\{ \sum_z f(x_i^P | \mathbf{z}) \cdot Q(\mathbf{z} | d_i, r) \right\} + f^B(x_i^P) \left\{ \sum_z f(x_i^M | \mathbf{z}) \cdot Q(\mathbf{z} | d_i, r) \right\} \right]^I \times \left[f^B(x_i^M) \left\{ \sum_z f(x_i^P | \mathbf{z}) \cdot Q(\mathbf{z} | d_i, r) \right\} \right]^{1-I} \right),$$

Model (f): Both chromosomes are recombinants.

$$L(x^M, x^P | G = 6, \mathbf{d}, r) = \prod_{i=1}^K \left[2^I \cdot \left\{ \sum_z f(x_i^M | \mathbf{z}) \cdot Q(\mathbf{z} | d_i, r) \right\} \cdot \left\{ \sum_z f(x_i^P | \mathbf{z}) \cdot Q(\mathbf{z} | d_i, r) \right\} \right]$$

Likelihoods for 3 genealogical classes (a,d and c) shown in Figure 2.1 are trivial to calculate because recombination has no effect on the population origins of linked sites. For models (a) and (d) the sampled individual's chromosomes are either entirely A or entirely B. For model (c) the individual is a first-generation hybrid so that one chromosome is entirely A and the other entirely B. The probability of the data for these genealogical classes can therefore be calculated directly by assuming random mating and applying the usual Hardy Weinberg formulas using the population haplotype frequencies in place of allele frequencies. Models (b), (e) and (f) all require calculation of the probabilities of one or more recombinant haplotypes (indicated by an x in Figure 2.1).

2.6. Bayesian inference

A Bayesian approach to hybrid inference requires a prior distribution for the genealogical classes. An individual belongs to genealogical classe g with prior probability π_g with $g = 1, 2, \dots, 6$ and $\sum_{g=1}^6 \pi_g = 1$. The posterior probability an individual belongs to the g th genealogical class is

$$P(G = g | x^M, x^P, \mathbf{d}, r) = \frac{\pi_g \times P(x^M, x^P | G = g, \mathbf{d}, r)}{\sum_{i=1}^6 \pi_i \times P(x^M, x^P | G = i, \mathbf{d}, r)}.$$

In general, a prior distribution on genealogical classes could incorporate factors such as differential fitnesses among classes, varying frequencies of mating encounters between individuals of different populations (or species), and so on. Lacking specific information on hybridization rates, one can use a discrete uniform prior $\pi_g = 1/6, \forall g = 1, \dots, 6$ which reduces the posterior likelihood to

$$P(G = g | x^M, x^P, \mathbf{d}, r) = \frac{P(x^M, x^P | G = g, \mathbf{d}, r)}{\sum_{i=1}^6 P(x^M, x^P | G = i, \mathbf{d}, r)}.$$

2.7. Estimation of population haplotype frequencies

The likelihood theory above treats population haplotype frequencies as known fixed parameters. For most empirical datasets population haplotype frequencies are unknown; we thus estimate them using posterior probability densities. To estimate frequencies “purebred” individuals (those least likely to be hybrids; possibly sampled outside a hybrid zone) are identified from each of the two populations and analyzed separately. Let $\mathbf{f}^i = \{f^i(h)\}$ be a vector of the haplotype frequencies in population i , where $f^i(h)$ is the frequency of the h th distinct haplotype in i and $i \in \{A, B\}$. Let N_i be the number of diploid individuals sampled from population i . 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 and B). Results are formulated for population A (B is equivalent). The prior probability density of haplotype frequencies (i.e., before sampling) is

$$\Pr(\mathbf{f}^A) = \prod_{h=1}^H \frac{\{f^A(h)\}^{(1/H)-1}}{\Gamma(1/H)}.$$

Thus, $\mathbf{f}^A \sim \text{Dirichlet}(1/H)$. Let the vector $\mathbf{n}_A = \{n_{1A}, \dots, n_{HA}\}$, where n_{hA} is the observed number of copies of the h th distinct haplotype in a sample from population A. The probability density of \mathbf{n}_A conditioned on the haplotype frequencies follows a Multinomial distribution given by

$$\Pr(\mathbf{n}_A | \mathbf{f}^A) = \binom{\sum_{h=1}^H n_{hA}}{n_{1A}, \dots, n_{HA}} \prod_{h=1}^H \{f^A(h)\}^{n_{hA}},$$

where

$$\sum_{h=1}^H n_{hA} = 2N_A$$

is the total number of haplotypes observed in population A. The posterior density of the haplotype frequencies, conditioned on the haplotypes observed in a sample from population A, follows a Dirichlet distribution, since the Dirichlet is a conjugate prior to a Multinomial distribution. The posterior probability density of haplotype frequencies is

$$\Pr(\mathbf{f}^A | \mathbf{n}_A) = \Gamma(\theta_A) \prod_{h=1}^H \frac{\{f^A(h)\}^{\theta_A a_{hA} - 1}}{\Gamma(\theta_A a_{hA})},$$

where,

$$\theta_A = (1 + \sum_{h=1}^H n_{hA}) = (1 + 2N_A)$$

and

$$a_{hA} = \frac{n_{hA} + 1/H}{1 + \sum_{h=1}^H n_{hA}} = \frac{n_{hA} + 1/H}{1 + 2N_A}.$$

Note that $\sum_{h=1}^H a_{hA} = 1$. The posterior mean is used to estimate \mathbf{f}^A and is given by

$$(2.1) \quad \mathbb{E}(\mathbf{f}^A | \mathbf{n}_A) = \frac{\theta_A a_{hA}}{\sum_{h=1}^H (\theta_A a_{hA})} = a_{hA}.$$

For simplicity, we ignore uncertainties of haplotype population frequencies when inferring hybrid classes, using the posterior mean as a proxy for the true frequency. Uncertainties could be accounted for by instead integrating over the posterior density rather than using the posterior mean.

2.8. Inference for unphased individuals

The above calculations require phased haplotypes for all individuals. Although improved genomic sequencing technologies provide more experimental information about phase than in the past, complete phase information is not always available. For the “pure” individuals one can obtain phase by applying existing population-based haplotype phasing methods [6, 7, 38] to each population separately. However, with unphased putative hybrid individuals using either pure population for phasing would be incorrect, and thus could potentially lead to biased inferences, unless

the individual turns out to be a non-hybrid. Therefore, instead of trying to estimate the phase of a hybrid individual we opted to integrate over all possible haplotype phase states.

Let $\mathbf{X} = \{X_{ij}\}$ be the matrix of genotype data of a hybrid individual, where X_{ij} is the genotype at the j -th SNP locus for chromosome i . Given the unphased multilocus genotype data X_i for chromosome i , let C_i denote its set of compatible diplotypes. Thus for chromosome i , given the g -th genealogical class, the likelihood of the unphased hybrid individual is

$$(2.2) \quad P(X_i | G = g, \mathbf{d}, r) = \sum_{\{x_i^M, x_i^P\} \in C_i} P(x_i^M, x_i^P | G = g, \mathbf{d}, r)$$

Here we are summing over all compatible diplotypes to obtain the marginal likelihood, taking into account the uncertainty in phasing. For a sample of K chromosomes, the likelihood of the unphased hybrid is

$$(2.3) \quad \prod_{i=1}^K P(X_i | G = g, \mathbf{d}, r)$$

The chromosome probabilities are multiplied because the different chromosomes undergo independent assortment during meiosis. These equations were implemented in a new inference program named Mongrail.

CHAPTER 3

Simulation Methods

Two different simulation study designs were used to evaluate the statistical performance of our new inference method Mongrail versus NewHybrids. The first design involved generating a diverse set of haplotype frequency distributions, ensured by using either a symmetrical or non-symmetrical Dirichlet distribution. This allowed a comprehensive comparison of performance over a broad range of conditions. The second design used a structured coalescent model with recombination, with haplotypes generated under a neutral Wright-Fisher model from each of two populations (A and B) connected by migration, allowing the statistical performance of the inference method to be evaluated under biologically realistic conditions.

3.1. Comprehensive simulation

We simulated diplotypes for individuals that were uniformly assigned to one of the 6 genealogical classes. The chromosomes of parents were randomly assigned haplotypes according to the population haplotype frequencies in populations A and B, which must first be specified. We first describe the procedure used to simulate haplotype marker configurations and their corresponding frequencies (for two populations A and B) for each chromosome. We use simple procedures designed to mimic population genetic processes rather than explicitly simulating a population genetic model as it allowed more direct control over levels of variation in the populations. For simplicity, in all simulations we fixed the length of each chromosome to be 240 Mb and the recombination rate to be 1.2 cM/Mb respectively. The simulation experiment used a factorial design allowing the performance of the methods to be assessed for many combinations of parameters. The parameters (factors) and their values were as follows:

- (1) Number of chromosomes: $K = 1, 2, 5, 10, 20$
- (2) Number of loci per chromosome: $L = 1, 5, 10$

- (3) Expected recombination frequency (in cM): $R = 1, 25, 50$ between the first and the last locus
- (4) Number of distinct haplotype sequences per chromosome for each population: $h = 5, 10, 15$
- (5) Allelic configurations of haplotypes, generated by simulating the switches between allele states (see description below). Switch rates used: $c = 0.1, 1, L/2$
- (6) Haplotype frequencies, following a Dirichlet distribution (symmetrical with parameters $\alpha = 1, 5$, or non-symmetrical with parameters $w = 5, 20$) (see description below)

To simulate haplotype configurations we mimic recombination by using a “switching process” (that flips the adjacent marker state) operating along the chromosome. The switch rate on a particular interval is $p = c/L$ where p is the probability of a switch from 0 to 1 (or 1 to 0). To simulate haplotype frequencies, we used either a symmetrical or non-symmetrical Dirichlet distribution with parameters $\alpha_1, \alpha_2, \dots, \alpha_h$. For the symmetric Dirichlet distribution we set $\alpha_i = \alpha$ for $i = 1, 2, \dots, h$ and consider two cases: $\alpha = 1$ or $\alpha = 5$.

For the non-symmetric Dirichlet distribution we use:

$$\alpha_i = \begin{cases} 0.7 \times w, & i = 1 \\ 0.3 \times \frac{w}{h-1}, & i > 1 \end{cases}$$

and again consider two cases: $w = 5, 20$. These combinations produce a diverse set of distributions (see Appendix A.1). The above combinatorial design produced 1100 simulation combinations in total (see below). For each, we simulated from 10,000 to 100,000 genealogical classes using a discrete uniform probability distribution on the six distinct genealogical classes. As noted above, given a genealogical class, the diplotype of an individual is generated by simulating a pair of chromosomes (whether one or both chromosomes are pure or recombinant depends on the genealogical class). The simulator is available as an option in our program.

With six parameters (factors) and the number of corresponding levels for each (5 for K , 3 for L , 3 for R , 3 for h , 3 for c , and 4 for α and ω) the number of simulation combinations (when $L > 1$) is $C_{L>1} = 5 \times 2 \times 3 \times 3 \times 3 \times 4 = 1080$. If $L = 1$, for each possible chromosome number (K) four ways exist to generate population haplotypes using Beta distributions (univariate cases of the Dirichlet distribution) with parameters α_1 and α_2 :

- (1) $\alpha_1 = \alpha_2 = 1$
- (2) $\alpha_1 = \alpha_2 = 5$
- (3) $\alpha_1 = 0.7 \times 5, \alpha_2 = 0.3 \times 5$
- (4) $\alpha_1 = 0.7 \times 20, \alpha_2 = 0.3 \times 20$

Thus, the number of combinations (for $L = 1$) is $C_{L=1} = 5 \times 4 = 20$ and the total number of simulation combinations is $C_{L>1} + C_{L=1} = 1080 + 20 = 1100$.

We analyzed the statistical performance of Mongrail and NewHybrids [2] for all the simulation combinations by computing for each simulated individual the posterior probabilities of belonging to each of the six genealogical classes. NewHybrids analyses multilocus genotypes and population allele frequencies, rather than haplotypes and haplotype frequencies. The genotypes are completely specified given the haplotypes. In our analyses, we ignore uncertainty of haplotype (or allele) frequencies, treating them as known. In this case, we do not require MCMC sampling to compute the posterior probability of an individual belonging to each of the six different hybrid categories under the NewHybrids model, thus we implemented a version of NewHybrid without MCMC sampling to calculate posterior probabilities with known population allele frequencies. Our goal is to compare the two methods using the same assumptions. Avoiding MCMC allows a large-scale comparison without having to worry about whether NewHybrids MCMC analyses have converged. Uncertainty of haplotype or allele frequencies will be an additional source of variance for both estimators. It is straightforward to incorporate individual haplotype and population haplotype frequency inference into Mongrail and thus account for this source of uncertainty. For both methods, we considered a discrete uniform prior on the six distinct genealogical classes.

We performed a comparative analysis of statistical performance between Mongrail and NewHybrids using five different performance metrics:

- (1) Accuracy of posterior probabilities
- (2) Power to identify genealogical classes
- (3) Posterior distribution of genealogical classes
- (4) ROC curve analysis of power versus Type I error
- (5) Sensitivity to biological and experimental parameters

We describe the procedures used to evaluate the methods for each of these criteria below.

3.1.1. Accuracy of posterior probabilities. The aim of this analysis was to verify that the posterior probabilities of genealogical classes for individuals provide consistent and unbiased estimates of the frequencies at which individuals belong to the classes. The underlying true genealogical class is known for a simulated individual and a Bayesian method should produce posterior probabilities for genealogical classes (models) that correspond to the frequency at which that genealogical class is the one under which the individual was simulated. To evaluate this, individual posterior probabilities for the g th genealogical class were binned into 10 intervals each of length 0.1. The proportion of individuals in each bin for which the true genealogical class was g was calculated and plotted against the mid-point of the posterior probability for the interval. The method is performing well if, in each interval with midpoint posterior probability p , the frequency of individuals binned into that interval for whom the true genealogical class is g is close to p . This will produce a straight line along the diagonal of the plot. This expected relationship holds for all the six genealogical classes. For example, 95% of individuals that are placed in bin (0.9, 1.0) for genealogical class g should have true genealogical class g while the remaining 5% will have a true genealogical class that belongs to one of the other five alternatives.

3.1.2. Power to identify genealogical classes. The aim of this analysis was to examine how often a method produces a high posterior probability for the true genealogical class used to simulate an individual. If the posterior probability is accurate, a greater proportion of high posterior probability outcomes will indicate greater power. The posterior probability assigned to an individual for each of the six genealogical classes was plotted as a stacked bar plot. Different colors are used to illustrate the different segments in the bar. Each colored segment represents the relative contribution of one of the six genealogical class posterior probabilities for that individual. One can conclude the method is performing well for individuals simulated under the g th genealogical class if the color assigned to the g th class is the dominant one in the plot. If the bar has many colors with uniform representation there is low support for any particular model and thus low power.

3.1.3. Posterior distribution of genealogical classes. The aim of this analysis was to examine the posterior probability of the true genealogical class for individuals simulated under that genealogical class. If a method is powerful most individuals should have a posterior probability

concentrated near posterior probability 1 for the true genealogical class (a high frequency of individuals at the right of the graph). If a method is accurate few individuals should have a posterior probability concentrated near posterior probability 0 for the true genealogical class (a low frequency of individuals at the left of the graph).

3.1.4. ROC curve analysis of power versus Type I error. The aim of this analysis was to examine the power of the methods to detect genealogical classes relative to Type I error for different classification thresholds. We plot the ROC (Receiver Operating Characteristics) curve for all six genealogical classes. The ROC curve for the g th genealogical class is created by plotting the true positive rate (equivalent to power, or *sensitivity*) against the false positive rate (equivalent to Type I error, or $1 - \textit{specificity}$). The true positive rate is defined as the proportion of individuals simulated under the g th genealogical class and classified as belonging to that class. The false positive rate is defined as the proportion of individuals simulated under another class but classified as belonging to the g th class. The ROC curve measures the performance of the methods for classifying individuals into genealogical classes as a function of varied classification thresholds (based on posterior probabilities). For each of the six genealogical classes we overlay the ROC curves of Mongrail and NewHybrids to allow comparisons between the two methods.

3.1.5. Sensitivity to biological and experimental parameters. The aim of this analysis was to determine how sensitive the method is to changes of key biological and experimental parameters: number of chromosomes (K), number of loci (L), and recombination frequency (R). All of these factors are affected both by the biology of the organism under study and by the experimental design. We expect increased information with increasing values of K, L or R . However, we do not know a priori how large the effect will be on the posterior probabilities of genealogical classes. To reduce the state space for this analysis, we consider fixed values for other parameters that determine haplotype frequencies in populations – these other parameters are typically beyond the experimentalists control and are also expected to have less predictable effects on method performance. We used: $h = 5, c = 0.1$ and $\alpha = 1$. We plot the proportion of cases for which the posterior probability of belonging to the correct genealogical class is above a threshold value of 0.9 against the number of chromosomes analyzed using multi-line plots, where each line represents

different combinations of values for the remaining two parameters: number of loci (except $L \neq 1$) and recombination frequency (R).

3.2. Coalescent simulation

The program *ms* [24] was used to simulate samples from two populations evolving according to a neutral Wright-Fisher model under various demographic histories. The program employs a structured coalescent model with recombination. Haplotype marker samples were simulated from the two populations (A and B) and used to estimate corresponding population haplotype frequencies. The diploid effective population size was $N_0 = 10,000$ for each population. We simulated 100 sampled chromosomes, each 1 Mb in length, for each population. The simulator is haploid so this corresponds to a sample of 50 diploid individuals from each population. We simulated 20 chromosomes for each diploid individual. The per-generation recombination probability over the entire chromosome was fixed to $r = 0.01$, thus the recombination rate was 1 cM/Mb. The population-scaled recombination rate parameter was then $\rho = 4N_0 \times r = 40,000 \times 0.01 = 400$. We assumed a symmetrical island model with $M = 4N_0m$, where m is the fraction of each population made up of new migrants each generation. We chose five different values $M = 0.1, 0.25, 1, 10, 100$ to study the effects of migration on the performance of the two methods in identifying genealogical classes.

We used *ms* to generate gene trees representing the history of the sampled chromosomes. The *seq-gen* program [35] was then used to simulate sequences on gene trees under a Jukes-Cantor mutation model. The population-scaled per-site mutation rate, θ , was assumed to be $\theta = 4N_0\mu = 0.00004$, where $\mu = 10^{-9}$ is the mutation rate per generation. The program *snp-sites* [32] was used to extract SNPs from simulated sequences into a Variant Call Format (VCF) file. Subsequent processing and manipulations, such as extracting biallelic SNPs into separate VCF files for the two populations, were performed using BCFtools [11]. We chose a subset of $L = 10$ markers for each analysis such that the markers were approximately equidistant to each other and spanned the chromosome. Since we specified a recombination rate of 1 cM/Mb, the physical distance between the first and last marker was 1 Mb (the length of simulated chromosomes). Having obtained haplotype marker configurations for samples from both populations, unknown population haplotype frequencies were estimated using the Multinomial-Dirichlet posterior mean (see equation 2.1).

For each of five different values of M , we simulated 1000 individuals for each of the four genealogical classes: purebred (model **d**), F1 (model **c**) hybrid, backcross (model **b**) and F2 (model **f**) hybrid. For brevity, we used only one backcross model (**b**) and one purebred model (**d**). Simulating a diploid individual is equivalent to generating a diplotype (a pair of haplotypes). Haplotypes required to form these individuals (belonging to any of the four genealogical classes) were simulated simultaneously with the population sample haplotypes that were generated under the structured coalescent process. The description of the procedure to generate diploypes assigned to one of the four genealogical classes can be found in Appendix A.3.

For each simulated individual we computed posterior probabilities under each of the six genealogical classes using either Mongrail or NewHybrids. We performed a comparative analysis of statistical performance between Mongrail and NewHybrids using two different performance metrics:

- (1) Power to identify genealogical classes
- (2) ROC curve analysis of power versus Type I error

CHAPTER 4

Simulation Results

4.1. Comprehensive simulation

It is impossible to present an exhaustive summary of the properties of the methods when applied to all 1100 combinations of simulation parameters considered in this paper. Instead, we give a general description of the most obvious patterns observed when applying each type of analysis to all the datasets and then provide specific examples for a subset of the combinations for which these patterns were most apparent. All the simulated datasets and scripts to perform the analyses are available at <https://github.com/mongrail/simulations>.

4.1.1. Accuracy of posterior probabilities. The general pattern observed across datasets when the analysis was done using Mongrail was that the average posterior probabilities matched the proportion of individuals correctly assigned to the specified genealogical class. However, the posterior probabilities and proportions typically did not match one another when using NewHybrids. The one exception was the case of a single locus per chromosome – in that case the assumptions of NewHybrids (independent assortment of alleles across loci) were satisfied and the posterior probabilities appeared correct. In other cases, posterior probabilities obtained using NewHybrids were higher than the proportion correctly assigned when posterior probabilities were high and were lower than the proportion correctly assigned when posterior probabilities were low.

As an example, we generated 100,000 individuals under the set of simulation parameters: $K = 20$, $L = 10$, $R = 50$, $h = 10$, $c = 0.1$, $\alpha = 1$. The Mongrail and NewHybrids programs were used to produce posterior probabilities for each distinct genealogical class for each individual. The posterior probabilities were binned into intervals as described in Section 3.1.1. The results are shown in Figure 4.1. The proportion of individuals having the correct model is plotted against the midpoint posterior probability of the interval they were binned into. The results for Mongrail, shown in red, indicate a precise linear relationship with points lying very near the identity line; this is expected

if the posterior probabilities are accurate. The results for NewHybrids, shown in blue, display an irregular curve with posterior probabilities higher than the proportion of correct genealogical classes when posterior probabilities are greater than about 40% and posterior probabilities lower than the proportion correct when posterior probabilities are lower than about 40%. Moreover, Mongrail performs well across all the genealogical classes (compare the 6 panels of Figure 4.1 from left to right). The performance of the NewHybrids method appears non-uniform across the models, it seems to perform worse for models representing hybrids (**c**, **f**) and backcrosses (**b**, **e**) compared to the pure parental ones (**a**, **d**). Thus, even with a high recombination rate the NewHybrids composite likelihood method is too liberal. Since we are primarily interested in high posterior probabilities, it is a serious problem if estimates of high posterior probabilities are overconfident (the observed pattern fits our expectation that since NewHybrids is a composite likelihood method it will tend to underestimate the uncertainty). See <https://github.com/mongrail/simulations> for additional examples of this behavior obtained using other combinations of parameters.

4.1.2. Power to identify genealogical classes. The general pattern observed across simulated datasets was that Mongrail typically placed higher posterior probability on the true genealogical class of an individual than did NewHybrids. NewHybrids always had more uniform probabilities among models but, although still worse, was closer in performance to Mongrail for individuals that were pure A or B (models **a** and **d**). In the case of 1 locus per chromosome, Mongrail and NewHybrids generate identical plots for posterior probabilities. This is expected because in this case NewHybrids is not a composite likelihood (all loci are unlinked).

Figure 4.2 shows the results for the particular simulation combination $K = 20$, $L = 10$, $R = 50$, $h = 5$, $c = 0.1$, $\alpha = 1$. Due to space constraint we show the stacked bar plots for only 100 individuals from each of the six genealogical classes. The Mongrail method has high power in distinguishing the pure individuals (models **a**, **d**) and the F1 hybrids (model **c**)(top graph, panels labelled **a**, **d** and **c** in Figure 4.2). The posterior probability that Mongrail assigns to the correct genealogical class is for most individuals greater than 0.95. Even for the F2 hybrids (model **f**) and backcrosses (models **b**, **e**) support for the correct genealogical class increases substantially from the prior (uniform) to the posterior. When the posterior probability of belonging to the correct model **b** (backcross with pure population A) is less than 0.9, the remaining posterior probability is mostly

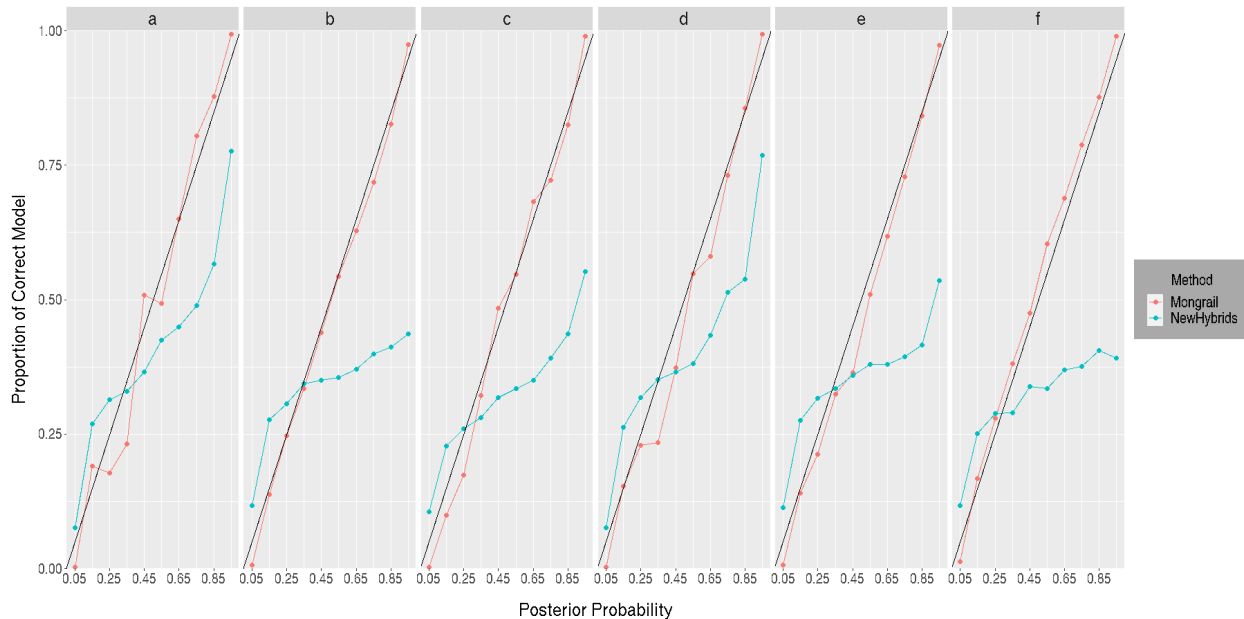


FIGURE 4.1. The proportion of correct assignments to genealogical classes (y-axis) for individuals binned according to the posterior probability of the genealogical class plotted against the midpoint posterior probability of each binning interval (x-axis). The 6 panels represent the 6 different genealogical classes. Results for posterior probabilities obtained using Mongrail are plotted in red and results for posterior probabilities obtained using NewHybrids are plotted in blue. If posterior probabilities match proportions the points should fall on a linear line with slope of unity (shown in black). Results are based on simulated data for 100,000 individuals using the simulation parameters: $K = 20$, $L = 10$, $R = 50$, $h = 10$, $c = 0.1$, $\alpha = 1$. 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.

assigned to the other hybrid and backcross categories. Similar patterns are observed for models **e** and **f** as we see a lot more variation (other colours which do not correspond to the true model).

Analyzing the same individuals using the NewHybrids method (bottom versus top graph) the posterior probabilities appear much more uniform across the genealogical classes with no particular genealogical class being well supported (this is particularly true for individuals that are hybrids or backcrosses; panels **b**, **c**, **e** and **f**). These results suggest that the NewHybrids method may often produce poorly resolved genealogical classes for individuals when used with linked marker data. This is particularly the case when individuals are F2 hybrids or backcrosses.

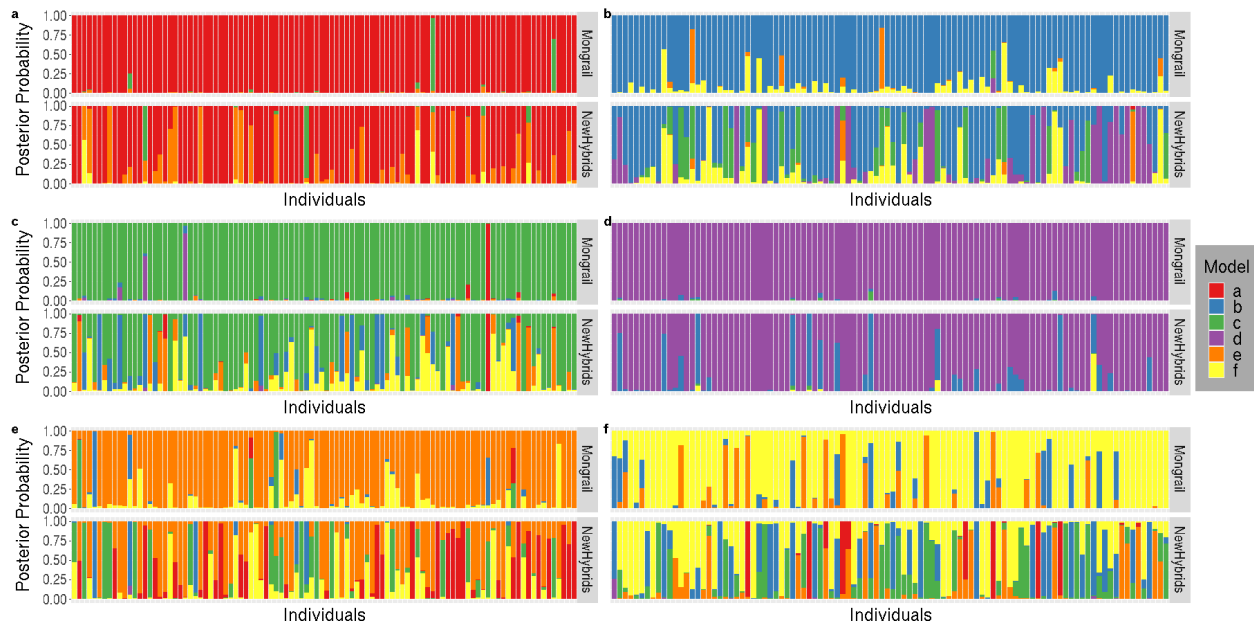


FIGURE 4.2. Distributions of posterior probabilities for random subsets of 100 individuals simulated under each of the 6 genealogical classes (plots labelled a-f). Posterior probabilities for Mongrail are shown in the top plot and for NewHybrids in the bottom plot. The posterior probabilities for different genealogical classes are represented by segments of different colors. The proportion of the stacked bar plot comprised of a particular color indicates the posterior probability of the model corresponding to that color. The following simulation parameters were used: $K = 20$, $L = 10$, $R = 50$, $h = 5$, $c = 0.1$, $\alpha = 1$. 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.

4.1.3. Posterior distribution of genealogical classes. This analysis examines the distribution of the posterior probability of a genealogical class when it is the true genealogical class for an individual (see description above). The general pattern across simulated datasets was that Mongrail tends to assign very high posterior probabilities only to the true model and very low posterior probabilities only to incorrect models. When the data are less informative the probabilities tend to be more uniform, resembling the prior. NewHybrids, on the other hand, less frequently assigns high probability to the true model and frequently assigns very low probabilities to the true model. These patterns are exemplified in Figure 4.3. To create this figure we generated 100,000 individuals under the following set of simulation parameters: $K = 20$, $L = 10$, $h = 10$, $c = 0.1$, and $\alpha = 1$ with recombination frequencies on the interval of both $R = 1\text{cM}$ and $R = 50\text{cM}$. The first row of the

figure shows results for Mongrail with a relatively uninformative dataset ($R = 1\text{cM}$). Even with low information, high posterior probabilities are frequently obtained for the true model when it is pure (model **a** or **d**) and for the other cases the probabilities are distributed quite uniformly and mostly intermediate. With $R = 50\text{cM}$ (second row of figure), high posterior probabilities occur much more commonly for the true model when it is any of the 6 models.

The results for the NewHybrids method are shown in rows 3 and 4 of Figure 4.3. Since NewHybrids assumes unlinked markers we expect the size of the region (frequency of recombination) will have no effect on the frequency distribution of posterior probabilities (the two rows with either $R = 1\text{cM}$ or $R = 50\text{cM}$ are identical). High posterior probabilities are obtained for the true model only when it is a pure population model (model **a** or **d**). For all the other models the most frequent outcome is a very low posterior probability for the true model. Thus, NewHybrids has very low power to infer the true model when it is not a pure population model and will often exclude the true model, assigning very low probability to it.

4.1.4. ROC curve. Here we examine the relative trade-off between power and type I error for each method using ROC curves (see description above). A method that has high power and low error should produce a curve that increases steeply and plateaus at a value approaching 1. The greater the area beneath the curve the better the performance. In general, Mongrail produces an ROC curve that strictly lies above the curve produced using NewHybrids when $L > 1$. As an example, we simulated 10,000 individuals using the combinations of parameters: $K = 20$, $L = 5$, $h = 5$, $c = 0.1$, $\alpha = 1$, and either $R = 1$ or $R = 10$. The results are shown in Figure 4.4.

The Mongrail inference method outperforms NewHybrids across all the six genealogical classes. In the case of a recombination frequency of 1cM , the difference between the two methods is least pronounced with either pure population A (model **d**) or B (model **a**) ancestry. As the recombination frequency increases from 1cM to 50cM , the ROC curve for the Mongrail method improves, approaching the top left corner, whereas the curve for NewHybrids is virtually unchanged. The NewHybrids method assumes unlinked markers and thus we expect its ROC curve to be unaffected by recombination (compare the blue curves on the top row to the bottom one).

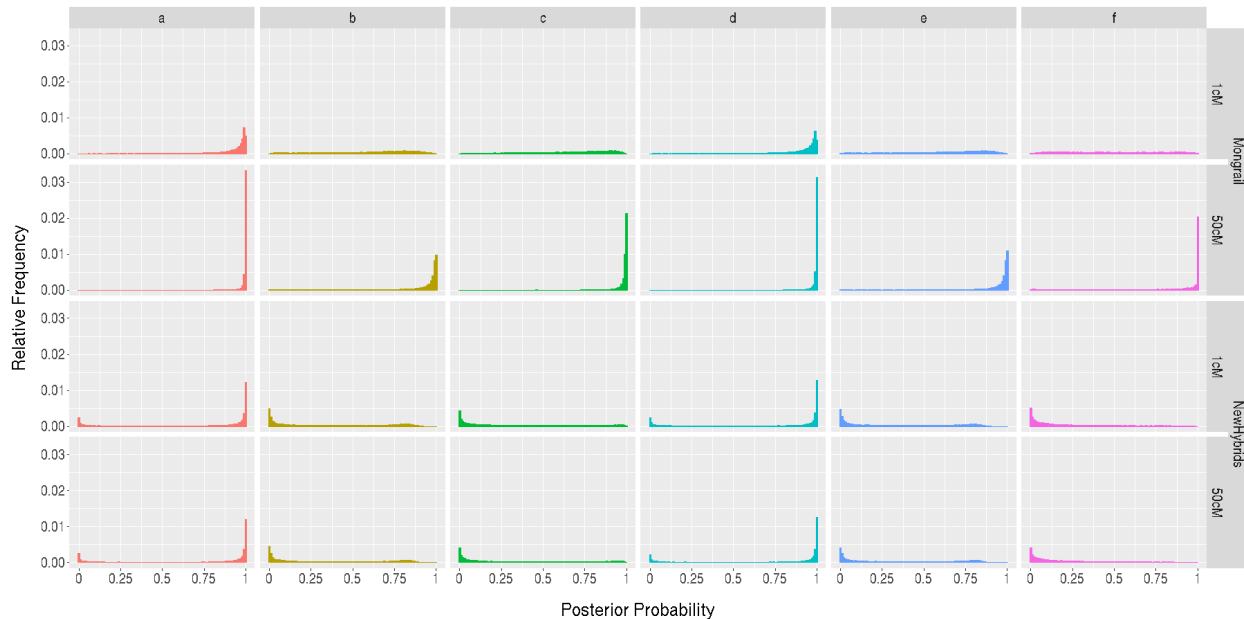


FIGURE 4.3. Histogram showing the relative frequency of the posterior probabilities obtained for the true model when analyzing simulated data using either the Mongrail (top two rows) or NewHybrids (bottom two rows) programs. The results are shown for data simulated using true models a-f (6 plots from left to right in each row). The plot is based on 100,000 individuals simulated using the following set of simulation parameters: $K = 20$, $L = 10$, $h = 10$, $c = 0.1$, and $\alpha = 1$ with recombination frequencies on the interval of both $R = 1\text{cM}$ and $R = 50\text{cM}$. 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.

4.1.5. Sensitivity to biological and experimental parameters. Here we examine the relative influence of 3 parameters: number of chromosomes, number of loci, and recombination frequency on the power of the methods to identify genealogical classes (using a posterior probability of 0.9 as a threshold for classifying individuals). Other parameters were held constant (see above). Figure 4.5 shows the proportion of cases with posterior probability for the true genealogical class greater than 0.9 (y-axis) as a function of the number of chromosomes (x-axis) for different combinations of number of loci and expected recombination rate for the Mongrail method (top row) and for NewHybrids (bottom row). Some interesting patterns emerge from the multi-line plot for Mongrail (Figure 4.5). Each coloured line in Figure 4.5 shows that as the number of chromosomes considered increases, the proportion of cases where the posterior probability is above 0.9 increases as well. The increase is particularly large when the number of chromosomes increases from 5 to 20. This is

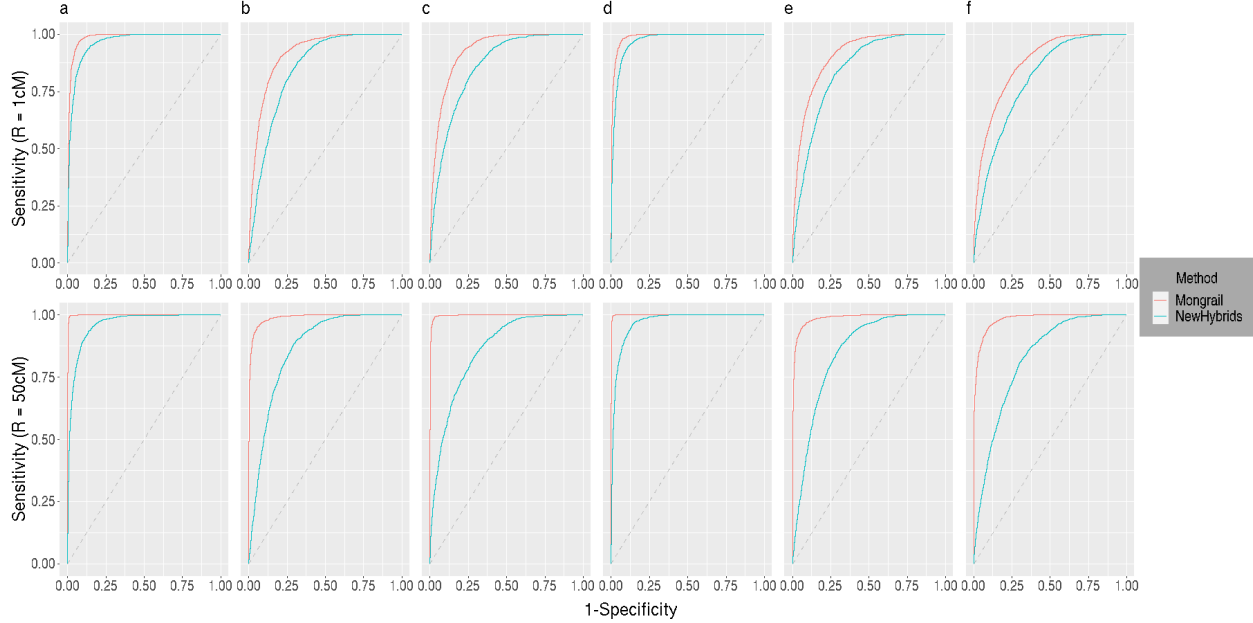


FIGURE 4.4. Receiver Operator Characteristics (ROC) curves for Mongrail (red line) and NewHybrids (blue line). The plot is based on 10,000 individuals simulated using parameters: $K = 20$, $L = 5$, $h = 5$, $c = 0.1$, and $\alpha = 1$. The top row uses a smaller region ($R = 1\text{cM}$) with a lower number of expected recombinations and the bottom row uses a larger region ($R = 50\text{cM}$) with a higher number of expected recombinations. 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.

true across all genealogical classes. However, as observed for other metrics the genealogical classes **a,c** or **d** receive higher posterior probabilities for the correct class as compared to the other three genealogical classes, likely because the data are more informative in these cases. Another striking trend is apparent across all the genealogical classes, the proportion above 0.9 changes little with an increase in number of markers but steadily increases with increasing recombination frequency from 1cM to 50 cM. In summary, increasing the number of chromosomes or recombination frequency has a large effect on the posterior probabilities (the former having greatest effect) whereas increasing the number of loci has little effect. With only one or two generations of mating one expects few recombination events, even on large intervals and so a small number of markers are sufficient to capture the available information from the data. Because chromosomes undergo independent assortment adding additional chromosomes has a much greater effect on power. Similarly, increasing

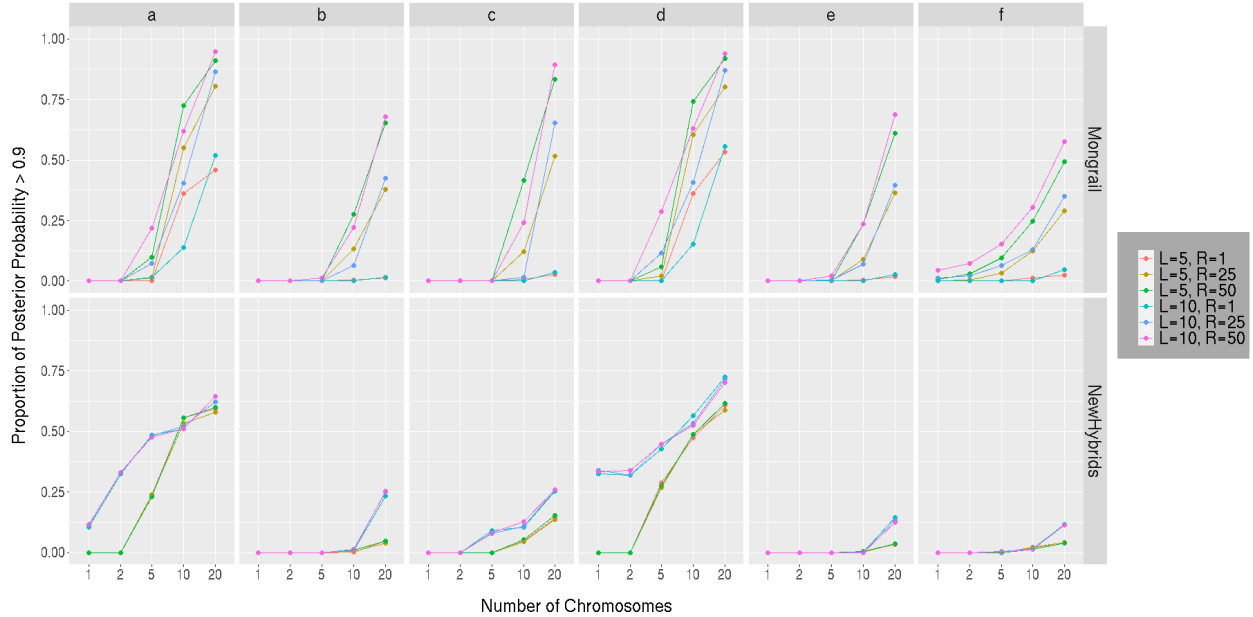


FIGURE 4.5. Proportion of simulated individuals with posterior probabilities for the correct genealogical class greater than 0.9 (y-axis) versus the number of chromosomes sampled per individual (x-axis). Different colored curves are plotted for 6 different combinations of numbers of loci and expected recombination rates (in cM). The results for Mongrail are shown in the top row and for NewHybrids in the bottom row. Results for the 6 genealogical classes that individuals were simulated under are given from left to right in each row. 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.

the size of a region of chromosome (even with a fixed total number of markers) increases the chances of observing recombination events and also increases power.

We analyze the same simulated dataset using the NewHybrids method (bottom row of Figure 4.5) for a side-by-side comparison of the effects of the three factors on the two competing methods. As expected, we see that the proportions for $R = 1, 25$ and 50 merge, regardless of the number of loci. Because NewHybrids is a composite likelihood method, we expect that recombination frequency should have little or no effect on the posterior probabilities and this is indeed the case. We find that increasing number of chromosomes increases the proportion of cases for which the posterior probability is greater than 0.9 but the changes are pronounced only for the pure (**a**, **d**) and F1 hybrid (**c**) individuals. In fact, an increase in the number of loci (from 5 to 10) increases

the proportion only for these three classes. Only when the number of chromosomes is $K = 20$, is there is a difference in the proportions (at $L = 5, 10$) for the remaining genealogical classes (**b,e,f**).

4.2. Coalescent simulation

We performed simulation analyses for $M = 0.1, 0.25, 1, 10, 100$ and four genealogical classes: purebred (model **d**), F1 (model **c**) hybrid, backcross (model **b**) and F2 (model **f**) hybrid. For brevity, we describe the general pattern observed across the analyses and present details only for a subset of specific illustrative cases. Results for the complete set of simulated datasets and scripts to perform the analyses are available at <https://github.com/mongrail/simulations>.

4.2.1. Power to identify genealogical classes. The general pattern observed across simulated datasets for genealogical classes purebred (model **d**) and F1 hybrids (model **c**) with $M \leq 1$ was that both Mongrail and NewHybrids performed well in identifying true genealogical classes. Both methods produced high posterior probabilities for the true genealogical class with neither obviously superior. However, when considering all four genealogical classes (models **d,c,b,f**) with $M > 1$, performance of Mongrail in identifying the true genealogical class is typically better than NewHybrids, although the power to distinguish correct genealogical classes diminishes for both methods as M increases.

Figure 4.6 shows results for purebreds (model **d**, first column) and F1 hybrids (model **c**, second column). Figure 4.7 shows results for backcrosses (model **b**, first column) and F2 hybrids (model **f**, second column). Both figures show results for a range from relatively low to high values of migration ($M = 0.25, 1, 10$). The stacked bar plots for only 100 representative individuals from each of these four genealogical classes are shown due to space constraints. For all four genealogical classes (Figure 4.6 and 4.7) when $M \leq 1$ (first four rows), Mongrail (first and third row) and Newhybrids (second and fourth row) perform similarly well, both placing high posterior probability on the true genealogical classes. For most individuals, the posterior probability assigned to the true genealogical class by both methods is greater than 0.9. More probability associated with incorrect models (colors not corresponding to the true model) is evident for backcrosses and F2 hybrids in Figure 4.7 compared to the pure and F1 genealogical classes (Figure 4.6). For $M > 1$ (last two rows), both methods show more uncertainty across all four genealogical classes but Mongrail

appears to perform better on average. The other general pattern observed is that both methods struggle to resolve the correct genealogical class as model complexity increases (from purebred to F1 hybrid in Figure 4.6 or from backcross to F2 hybrid in Figure 4.7) especially when the migration rate is high ($M = 10$). Difficulty increases with an increase in model complexity (models **d,c,b,f**) across all values of M . This is evident from the more uniform distribution of posterior probabilities across genealogical classes. For extremely high migration between the two populations, NewHybrids performs less well in resolving genealogical classes by comparison with Mongrail. This is especially true for F1 or F2 hybrids or backcrosses.

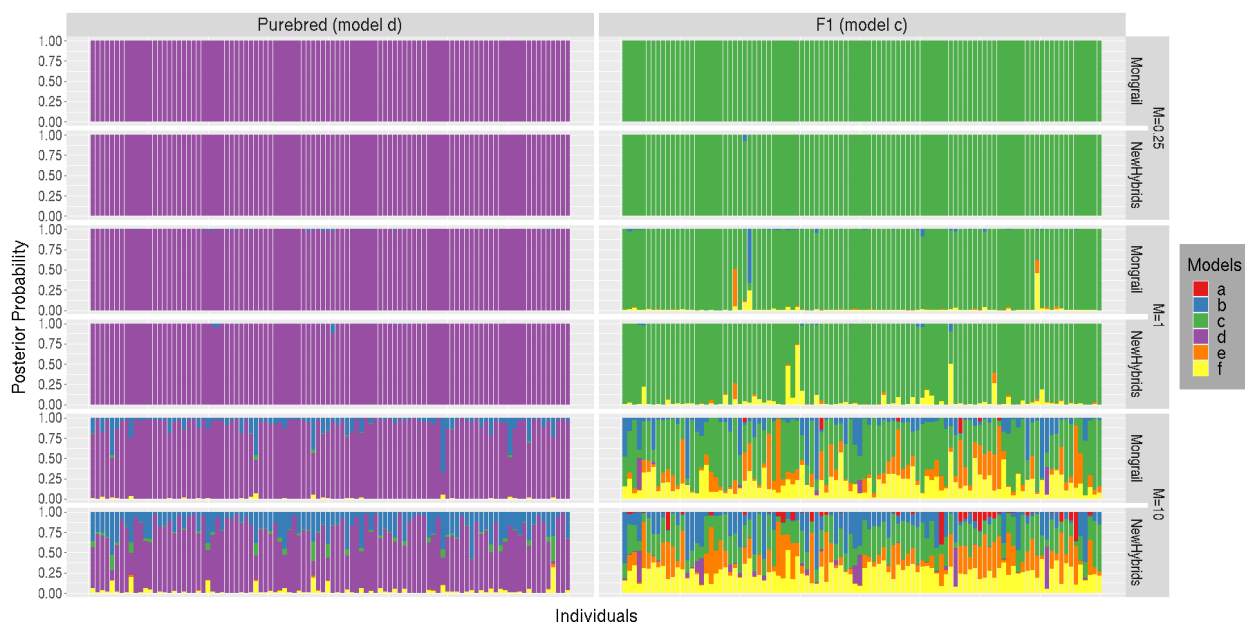


FIGURE 4.6. Distributions of posterior probabilities for random subsets of 100 individuals simulated (through a structured coalescent process) under each of the 2 genealogical classes (plots labelled d and c) for three values of migration parameter $M = 0.25$ (1st, 2nd rows), 1 (3rd, 4th rows), 10 (5th, 6th rows). For each value of M , posterior probabilities for Mongrail are shown in the top plot and for NewHybrids in the bottom plot. The posterior probabilities for different genealogical classes are represented by segments of different colors. The proportion of the stacked bar plot comprised of a particular color indicates the posterior probability of the model corresponding to that color. 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.

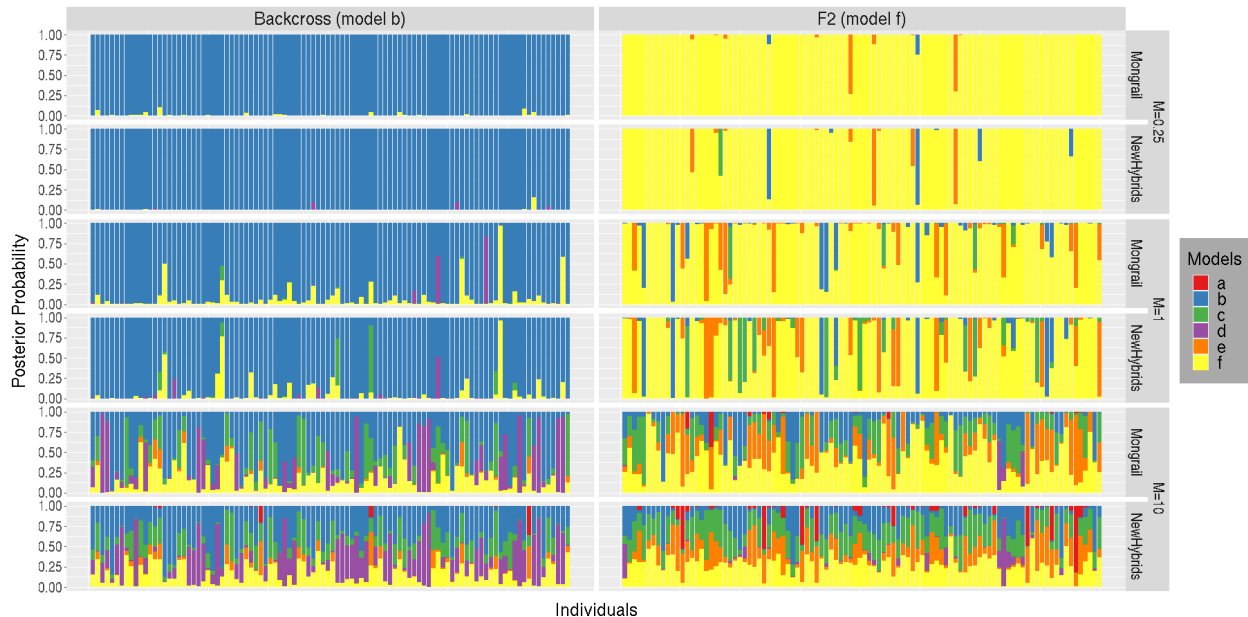


FIGURE 4.7. Distributions of posterior probabilities for random subsets of 100 individuals simulated (through a structured coalescent process) under each of the 2 genealogical classes (plots labelled b and f) for three values of migration parameter $M = 0.25$ (1st, 2nd rows), 1 (3rd, 4th rows), 10 (5th, 6th rows). For each value of M , posterior probabilities for Mongrail are shown in the top plot and for NewHybrids in the bottom plot. The posterior probabilities for different genealogical classes are represented by segments of different colors. The proportion of the stacked bar plot comprised of a particular color indicates the posterior probability of the model corresponding to that color. 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.

4.2.2. ROC curve. The relative trade-off between power and type I error was analyzed for each method using ROC curves for the datasets simulated under a structured coalescent process. The greater the area beneath the curve, the better the performance. In general, for all $M \geq 1$, Mongrail produces an ROC curve that lies strictly above the curve produced by NewHybrids across all four genealogical classes (models **d,c,b,f**). We present the results for 1000 individuals simulated under each of the four genealogical classes for $M = 0.25, 1, 10$ in Figure 4.8.

Mongrail outperforms NewHybrids across all four genealogical classes when $M \geq 1$. The performance difference between the two methods increases with model complexity, with $M(\geq 1)$ fixed. As M increases, the difference between the two methods gets more pronounced for all the four genealogical classes. When $M < 1$ for the purebred (model **d**) and F1 hybrid (model **c**), the

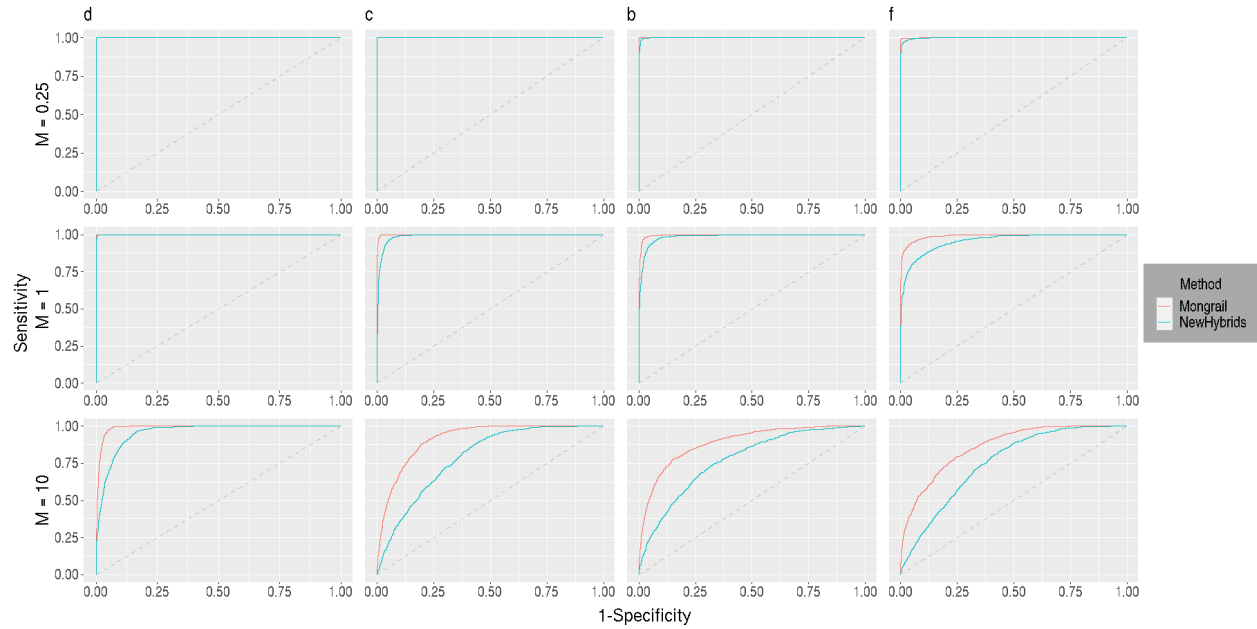


FIGURE 4.8. Receiver Operator Characteristics (ROC) curves for Mongrail (red line) and NewHybrids (blue line). The plot is based on 1000 individuals simulated under each of the four genealogical classes (**d,c,b,f**) through a structured coalescent process. The first, second and third row corresponds to migration parameter with values $M = 0.25, 1$ and 10 respectively. Results for the six genealogical classes (**d,c,b,f**) are shown from left to right in all three 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.

two curves overlap with no visible difference between methods. When $M < 1$ for the backcross (model **b**) and F2 hybrid (model **f**), although the Mongrail curve lies above the NewHybrids curve, the difference is negligible.

Empirical Analysis of Spotted and Barred Owl Hybridization

5.1. Background

Spotted Owls (SO) are native to the forests of western Northern America, mainly the Pacific Northwest, California and Mexico [3, 12, 13]. Barred Owls (BO) are native to eastern North America but have expanded their range to the west coast of North America thus encroaching on the territory of the endangered spotted owl [14, 18, 30], whose population is already in decline due to habitat loss caused by logging and wildfires [9, 16, 39]. There are three recognized subspecies of the spotted owl ranging in distribution from British Columbia to Mexico: Northern Spotted Owl (NSO), California Spotted Owl (CSO) and Mexican Spotted Owl (MSO). The NSO and MSO have been listed as “threatened” under the Endangered Species Act since the early 1990s by the US Fish and Wildlife Service (USFWS). As the two species can hybridize (sympatric populations of spotted and barred owls exist from British Columbia to southern California) [19, 20, 27], frequent hybridization may threaten the genetic integrity of the spotted owls.

The study by [15] is the largest genomic study conducted on spotted owls, barred owls and their hybrids. They obtained sequences from spotted and barred owls sampled outside and across their hybrid zone in western Northern America. The sampling locations of all of the owls included in the study are presented in Table 1. For County level information see [15] (Supplementary Table S5). [15] improved upon a previously generated SO genome assembly [22] (using data from 10x genomics and Bionano Genomics) and generated high-coverage (mean $31.70\times$, ± 6.51) whole genome sequence data from 51 owl samples consisting of 11 spotted owls, 25 barred owls, 2 known hybrids (identified in [21]) and 13 potential hybrids. The 51 owl samples included a female SO sample named Sequoia [22] used for constructing the new and more contiguous reference genome assembly.

TABLE 1. Table showing the sampling locations (states in North America) of the 51 owl samples consisting of 11 spotted owls, 25 barred owls, 2 known hybrids and 13 potential hybrids. The purebred samples (SO and BO) are further categorized into their recognized sub-species.

Species Category	State	#individuals	Total #individuals
Northern Spotted Owl	California	5	8
	Oregon	2	
	Washington	1	
California Spotted Owl	California	3	3
Eastern Barred Owl	Kentucky	2	12
	Ohio	2	
	New York	3	
	Massachusetts	3	
	New Jersey	1	
	Indiana	1	
Western Barred Owl	California	13	13
Putative Hybrid	California	7	13
	Oregon	5	
	Washington	1	
Known Hybrid	California	1	2
	Oregon	1	

5.2. Materials and methods

We used Mongrail to analyze the 51 owl samples, using the filtered Variant Call Format (VCF) file available at <https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?analysis=SRZ190173>. The dataset includes 17,385,299 biallelic single nucleotide polymorphisms (SNPs) across 82 large autosomal scaffolds and 8,543,351 of these had high-confidence genotype calls ($GQ \geq 40$) in all individuals. We restricted our analyses to the 15 largest autosomal scaffolds and filtered out sites with any missing data. Here we treat each scaffold as a chromosome. For each scaffold, we extracted the SO (10 individuals, excluding Sequoia) and BO (25 individuals) populations into two separate VCF files. Processing and manipulation of the high-throughput sequencing data was performed using BCFtools [11]. The unknown population haplotype frequencies (for SO and BO) were first estimated for each scaffold. Phased haplotype information is needed to estimate frequencies but the data were not phased for all SNPs. BEAGLE version 5.1 [7] was used to phase each population separately for each scaffold. We chose 10 markers from each scaffold for the analysis. The distribution of markers across scaffolds varied (see Section 5.2.1, 5.2.2). The

frequencies of phased haplotypes for the two populations were estimated using the Multinomial-Dirichlet posterior mean (see equation 2.1). In our model framework, BO is treated as Population A and SO as Population B. Thus the pure and hybrid classification of this owl dataset with its equivalent genealogical class based on our model:

- Model a - Spotted Owl (SO)
- Model b - Backcross with Barred Owl (F1 x BO)
- Model c - F1 hybrid
- Model d - Barred Owl (BO)
- Model e - Backcross with Spotted Owl (F1 x SO)
- Model f - F2 hybrid

The putative hybrids have an unusual plumage pattern making them difficult to distinguish from the western BO based purely on morphology. The main aim of this empirical analysis is to examine whether Mongrail can successfully place the samples into different genealogical classes, especially the hybrids (see Appendix B.1). We also examine the effect of certain biological and experimental parameters on the power to identify genealogical classes, suggesting optimal ways to choose parameters (number of chromosomes or scaffolds and region size or expected recombination frequency/map length).

We perform four main analyses (bash scripts used to perform the analyses are available at Scripts) to examine the owl dataset which are as follows:

- (1) Spatial variation of model posterior probabilities across scaffolds
- (2) Effect of successive scaffold inclusion for varied map length
- (3) Sensitivity of results to assumed recombination rate
- (4) Assignment of 15 hybrids (13 putative and 2 known)

The procedures used to perform these four analyses are described below.

5.2.1. Spatial variation of model posterior probabilities. The purpose of this analysis is to study the variation in posterior probabilities of genealogical classes for owl samples across the genome by using different sets of $L = 10$ loci from a particular scaffold region of variable map size ($R = 1.5\text{cM}$ and $R = 50\text{cM}$). This was implemented using a sliding window approach over

an entire scaffold starting from the 5' end and moving towards 3'. A window implies a set of L markers each of length R , where the markers are almost equidistant to each other. We say “almost” since the theoretically equidistant marker positions are not all present in the actual dataset. The recombination rates for SO and BO are not available, so following [21] we assumed a recombination rate of 1.5 cM/Mb (the average recombination rate of zebra finch). At $R = 50\text{cM}$ (a physical length of 33.33 Mb), we chose 100 initiating markers to create windows evenly spaced across the scaffold. Then for every window the posterior probability assigned to an owl sample for each of the six genealogical classes was plotted as a stacked bar plot. Different colors are used to illustrate the different segments in the bar. Each colored segment represents the relative contribution of one of the six genealogical class posterior probabilities for that owl sample. Using the same 100 initiating markers but with $R = 1.5\text{cM}$ (a physical length of 1 Mb) we created similar stacked bar plots. The two stacked bar plots are compared to examine the effects of varying map length on posterior probabilities. Within each plot we examine the consistency of the posterior probabilities across the entire scaffold.

5.2.2. Effect of successive scaffold inclusion for varied map length. The purpose of this analysis is to examine the variation in posterior probabilities of genealogical classes for owl samples as we increase the number of scaffolds for different levels of recombination frequency. For each sample we considered $L = 10$ markers from all 15 scaffolds where the scaffolds are arranged in a decreasing order of length. We study the cumulative effect of adding the scaffolds successively starting from the largest scaffold (Super-Scaffold_7) of size 72.11 Mb to the smallest scaffold (Super-Scaffold_47) of size 21.02 Mb on the posterior probabilities. A recombination rate of 1.5 cM/Mb is assumed for the same reasons as mentioned earlier. We perform the analysis under two cases. In the first case, we use a constant recombination frequency (or, map length) of $R = 1.5\text{cM}$ (a physical length of 1Mb) for every scaffold. For the second one, we vary the recombination frequency over each scaffold in a way such that its equivalent physical length is approximately equal (rounded off to a whole number) to the length of the scaffold. Say, for the largest scaffold which is of length 72.11 Mb we use a map length of 108cM which is equivalent to a physical length of 72 Mb. This design that incorporates almost the entire scaffold size as the map length has the advantage of retaining maximum information. Henceforth we refer to this second case as “maximally informative”. Now

we need to choose a set of $L = 10$ markers from each scaffold. Here we adopt the window approach (as described in the previous section) for each scaffold and arbitrarily select the middlemost window under this “maximally informative” case. For every owl sample we plot the posterior probability for the g th genealogical class against the cumulative number of scaffolds using multi-line plots, where each line represents a particular genealogical class. From the simulation analysis we found that increasing the number of chromosomes results in an increase in information. So we examine this multi-line plot to see whether a particular genealogical class is favored over others as the number of scaffolds increases. Ideally the g th genealogical class is said to be “preferred” over the rest, if the posterior probability for the g th genealogical class approaches one while the rest of the lines concentrates around zero as the number of scaffold increases. Using the same initiating markers as chosen earlier for each scaffold but with a constant map length of $R = 1.5\text{cM}$ we produce similar multi-line plots. The multi-line plots for the two cases of recombination frequency are compared to study the effects of different map length on posterior probabilities.

5.2.3. Sensitivity of results to assumed recombination rate. In the previous two analyses we studied the behaviour of the posterior probabilities of genealogical classes under a fixed recombination rate of 1.5 cM/Mb . This value has been extrapolated from another species, the zebra finch (*Taeniopygia guttata*) as the recombination rate for owls was not available. This raises the question whether the choice of recombination rate affects the inference of genealogical classes. We conducted a sensitivity analysis to study the behaviour of the posterior probabilities of the “preferred” model (inferred in the previous analyses) for different values of the recombination rate when successively increasing the number of scaffolds. We chose three different levels of recombination rate (in cM/Mb): $r = 0.5, 1.5$ and 5 . For each individual we plot the posterior probability of the “preferred” genealogical class against the cumulative number of scaffolds using multi-line plots, where each plot represents a particular value of recombination rate r . The map length used for scaffolds is similar to the “maximally informative” case described above.

5.2.4. Assignment of 15 hybrids (13 putative and 2 known). This analysis aimed to compare the genealogical classifications obtained from Mongrail with previous classifications of the same individuals. We used all 15 scaffolds and assumed a recombination rate of 1.5cM/Mb (with

a “maximally informative” region size) to construct a table of the posterior probabilities of the “preferred” model for each of the 15 hybrids (along with the primary and genetic identification information of [15]). The purpose of this table was to examine whether posterior probabilities of the “preferred” model are high enough to make an assignment call for each of the 15 hybrid owl samples using a pre-specified threshold posterior probability (for example, 0.99). Given that an individual can be classified, we examined whether our inference matched the previous genetic identification [15].

5.3. Results

All analyses were performed using the 50 owl samples described in Table 1 excluding Sequoia. For brevity, detailed results are presented for only 5 individuals. The 5 owls were chosen to be representative of the five observed categories genetically identified by [15]. The categories along with the sample names are presented in Table 2.

TABLE 2. Details for five individuals (out of the 50 owls) chosen for detailed analysis. Primary and genetic identifications are from [15].

Primary identification	Genetic identification	Sample names
spotted owl (SO)	SO	ZRH625
barred owl (BO)	BO	ZRHG101
putative hybrid	BO	CYWC009
known hybrid	backcross (F1 × BO)	ZRH607
known hybrid	F1 hybrid (F1)	ZRH962

Phased haplotypes were not available for these owls. Putative “pure” individuals ZRH625 (SO) and ZRHG101 (BO) were phased using BEAGLE version 5.1 [7]. For the remaining putative hybrid, or backcross, individuals inferences were averaged over the probability distribution of possible phase resolutions using equations 2.2 and 2.3.

5.3.1. Spatial variation of model posterior probabilities. In this analysis we examine how posterior probabilities vary across a scaffold. We also examine the effect of the size of a window (map length) on the distribution of posterior probabilities. Figure 5.1 shows the distribution of posterior probabilities across the largest scaffold (Super-Scaffold_7) of length 72.11 Mb. The distribution of the posterior probability on the six genealogical classes (denoted by each bar) is

highly consistent across the entire scaffold for all the samples. Increasing map length from $R = 1.5$ cM (left) to $R = 50$ cM (right) increases the posterior probabilities for the proposed genealogical class for the pure individuals and the F1 hybrid. The pattern in the putative hybrid (row 3) and backcross (row 4) is less clear. For these two owl samples there is a slight increase in the posterior probability for the pure barred owl genealogical class (model **d**) but also increased support for F1 hybrid (model **c**) in some regions. Such patterns might suggest more ancient hybridization not included in our model. There is also more variation in the posterior probabilities for the larger window size along the scaffold, likely because the markers are spread over a larger region thus increasing the chances of recombination. The stacked bar plots for the putative hybrid and the backcross individuals look very similar suggesting that distinguishing the hybrids may be quite difficult when using a small region of a single chromosome. Distinguishing the genealogical classes of these two hybrids may not be possible based on a single scaffold.

5.3.2. Effect of successive scaffold inclusion for varied map length. This analysis evaluates the cumulative effect of the number of scaffolds and of map length on the posterior probability for each genealogical class. Another aim is to examine whether the genealogical class with highest posterior probability for an individual analyzed using Mongrail matches the genetically identified category of [15] for that individual.

To examine the cumulative effect of adding scaffolds we used the 15 largest scaffolds, which were sequentially added in descending order by size. In figure 5.2, as we move from left (Super-Scaffold_7) to right (Super-Scaffold_47) on the x-axis, the number of scaffolds denoted by K increases from 1 to 15. The general pattern observed is that the posterior probability for genealogical classes **a**, **d** and **c** increases monotonically (asymptotically approaching 1) as the number of scaffolds is increased for the pure spotted owl (ZRH625), pure barred owl (ZRHG101) and F1 hybrid (ZRH962). Across the three samples, in the “maximally informative” case, the posterior probability of the “preferred” model exceeds 0.9 when $K = 3$ and exceeds 0.99 when $K = 5$. The “maximally informative” case (right) occurs when we choose map lengths (for each scaffold) such that the equivalent physical length is near the length of the scaffold. With a smaller region (lower expected recombination frequency) of $R = 1.5$ cM (left), the posterior probability of the genealogical class **a** (or, **d**) exceeds 0.9 when $K = 4$ and exceeds 0.99 when $K = 7$ for spotted owl (or, barred owl), respectively. For

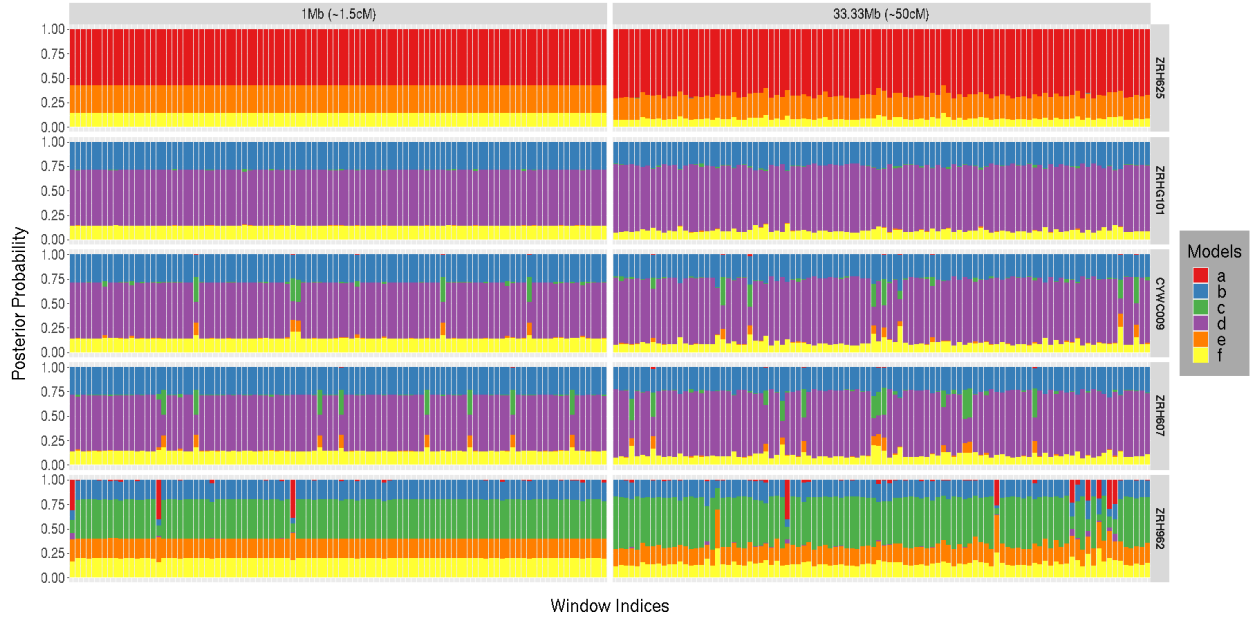


FIGURE 5.1. Distribution of posterior probabilities is constructed for 5 owl samples (ZRH625, ZRHG101, CYWC009, ZRH607 and ZRH962) for 100 windows (set of $L = 10$ markers), where the initiating markers are evenly spaced across the largest scaffold Super-Scaffold_7 (size 72.11Mb). The right column uses a larger region (33.33Mb equivalent to $R = 50\text{cM}$) of the scaffold and the left column uses the same initiating markers but of smaller length (1Mb equivalent to $R = 1\text{cM}$). Six different colors are used to represent the 6 different genealogical classes. The proportion of the stacked bar plot comprised of a particular color indicates the posterior probability of the model corresponding to that color. The primary and genetic identification information about the 5 owl samples are provided alongside the sample identifiers. A recombination rate of $r = 1.5\text{cM}/\text{Mb}$ is assumed for this plot. 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. In our model framework, BO is treated as population A and SO as population B.

the F1 hybrid at $R = 1.5\text{cM}$ the posterior probability of genealogical class **c** exceeds 0.9 when $K = 5$ and exceeds 0.99 when $K = 9$. These results suggests that either increasing the number of scaffolds or increasing the within-scaffold region size increases information, but additional scaffolds have a greater effect. The smaller the region (lower the expected recombination frequency) the higher the number of scaffolds (K) needed for the posterior probability of the “preferred” model to approach 1.

For the putative hybrid, even in the “maximally informative” case, more scaffolds are needed by comparison with the pure barred owl for the genealogical class **d** to become the “preferred” one.

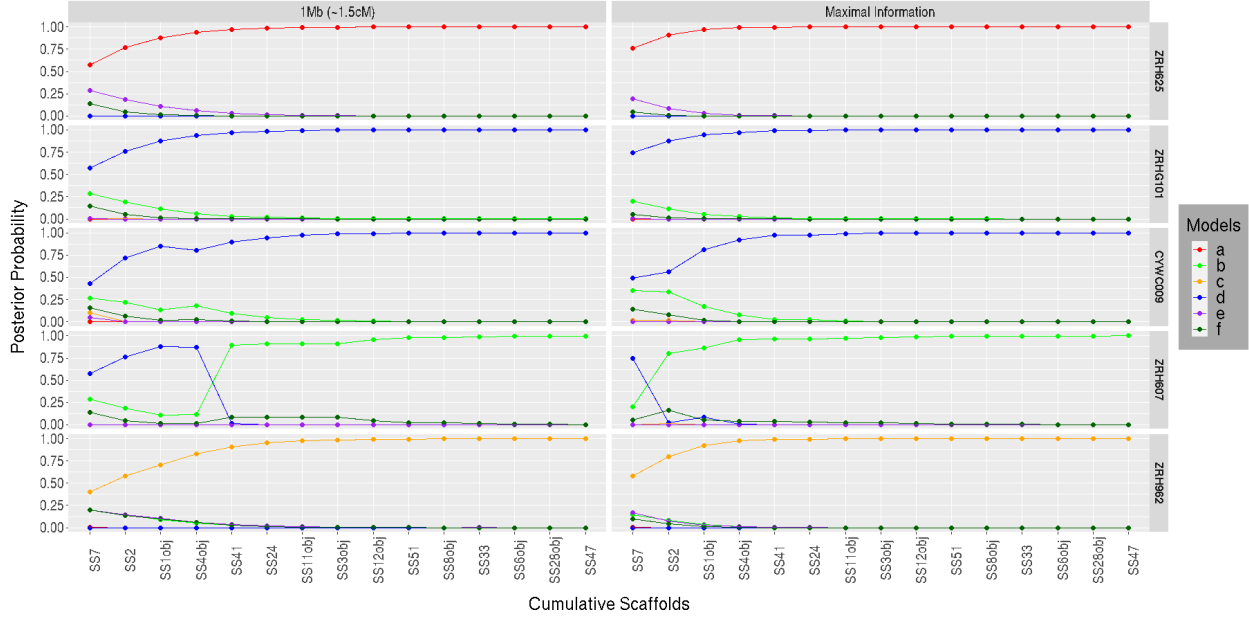


FIGURE 5.2. Posterior probability plotted against the cumulative number of scaffolds for 5 owl samples (ZRH625, ZRHG101, CYWC009, ZRH607 and ZRH962). Different colored lines are plotted for 6 different genealogical classes. Moving along the x-axis, the scaffolds are arranged from the largest to smallest and the number of scaffolds (K) increases from 1 to 15. The right column referred to as the “maximally informative” case considers very large map length (its equivalent physical length is almost equal to the length of the scaffold) for every scaffold and a set of $L = 10$ markers are chosen from the middle of each scaffold where the markers are almost equidistant from each other. The left column considers a smaller region (1Mb equivalent to $R = 1.5\text{cM}$) for each scaffold with $L = 10$ markers with the same initiating markers chosen previously (for the “maximally informative” case). The primary and genetic identification information about the 5 owl samples are provided alongside the sample identifiers. A recombination rate of $r = 1.5\text{cM}/\text{Mb}$ is assumed for this plot. 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. In our model framework, BO is treated as population A and SO as population B.

The posterior probability of genealogical class **d** exceeds 0.9 when $K = 5$ and exceeds 0.99 when $K = 8$. A similar posterior probability pattern is observed for $R = 1.5\text{cM}$ with an increasing number of scaffolds. Though the “preferred” model is the same (**d**) for both the samples (pure barred owl and putative hybrid) a striking contrast is observed between posterior probability plots for the two individuals. This difference may suggest why the putative hybrid individual was difficult to differentiate from a pure barred owl based solely on morphology (it had an unusual plumage pattern).

It is possible that this individual is descended through a backcross greater than 2 generations ago. Distinctive patterns of posterior probability are observed in the putative backcross individual for both choices of region size (expected recombination frequency). Two genealogical classes **b** and **d** seem to be competing for relative support, with **d** preferred initially. As the number of scaffolds is increased a point is reached where the trend of the two line plots is reversed and genealogical class **b** is increasingly favoured, ultimately emerging as the “preferred” model. The main difference between the scaffold addition plot for the two different region sizes ($R = 1.5\text{cM}$ and “maximally informative”) is that the transition to the asymptotically preferred model takes place with fewer scaffolds in the “maximally informative” case (the transition occurs as $K = 1$ increases to $K = 2$ in the “maximally informative” case versus as $K = 4$ increases to $K = 5$ in the $R = 1.5\text{cM}$ case). The posterior probability for genealogical class **b** exceeds 0.99 when $K = 10$ in the “maximally informative” case but more scaffolds ($K = 13$) are needed for the posterior probability to exceed 0.99 when the map length is $R = 1.5\text{cM}$.

In conclusion, the “preferred” genealogical class for each owl sample matches the previously genetically identified class indicating that Mongrail is successful in inferring genealogical classes irrespective of the choice of region size ($R = 1.5\text{cM}$ or “maximally informative”). The larger “maximally informative” region size ultimately supported the same genealogical classes as the smaller region but converged to a high posterior probability with fewer scaffolds. This suggests that a researcher has some flexibility to design genomics experiments with different numbers of scaffolds and region sizes according to the limitations imposed by their budget or study organism.

5.3.3. Sensitivity of results to assumed recombination rate. The Mongrail method requires that recombination rate be known (in units of cM/Mb). For the Spotted and Barred Owl species direct estimates of recombination rates across the genome (from pedigree analysis for example) are unavailable. Instead, we used a recombination rate of $1.5\text{cM}/\text{Mb}$ based on the average recombination rate of the zebra finch. Here we examine the sensitivity of the results generated by Mongrail to assumptions about the recombination rate. Specifically, we examined the influence of different values of recombination rate on the inferred genealogical class. Figure 5.3 shows the posterior probability for the “preferred” model as a function of the number of scaffolds (K) at different values of recombination rate for each owl sample. A model is “preferred” if its posterior

probability approaches one as the value of K increases. The general pattern observed across all the owl samples is that the “preferred” models for the spotted owl, barred owl, putative hybrid, backcross and F1 hybrid when all scaffolds are used are, respectively, **a** (Row 1), **d** (Row 2), **d** (Row 3), **b** (Row 4) and **c** (Row 5) regardless of the recombination rate that was used. Thus, for the owl dataset, varying the recombination rate over a broad range does not change the genealogical classes inferred by Mongrail. The same genealogical class is likely to be identified even if the recombination rate is badly mis-specified and our extrapolation of the recombination rate from another species is unlikely to be misleading for this dataset.

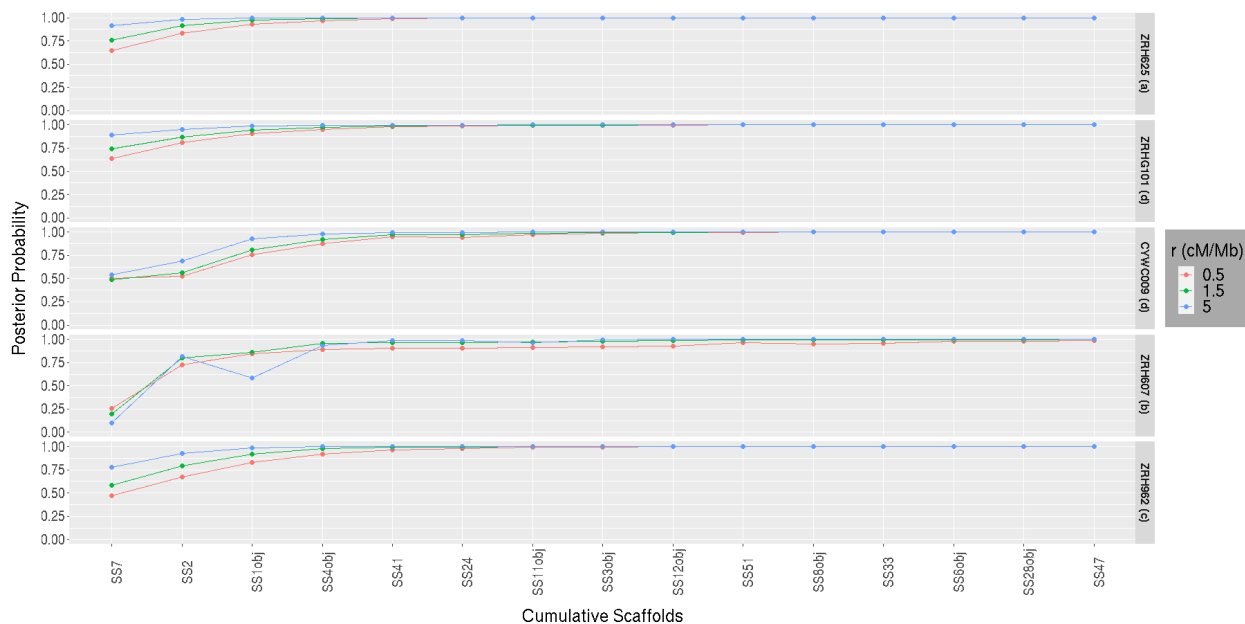


FIGURE 5.3. Posterior probability of the “preferred” model is plotted against the cumulative number of scaffolds for 5 owl samples (ZRH625, ZRHG101, CYWC009, ZRH607 and ZRH962) at different levels of recombination rate ($r = 0.5, 1.5, 5\text{cM/Mb}$). Different colored plots denote different values of recombination rate. A genealogical class (or, model) is said to be “preferred” if its posterior probability approaches 1 while the posterior probability for the rest concentrates around 0 with large number of scaffolds. The “preferred” model is shown in parenthesis placed beside the sample names. Moving along the x-axis, the scaffolds are arranged from the largest to smallest and as we move from left to right the number of scaffolds (K) increases from 1 to 15. A set of $L = 10$ markers are chosen from the middle of each scaffold where the map length for every scaffold is selected in such a way that its equivalent physical length is almost equal to the length of the scaffold. The markers are almost equidistant from each other. The primary and genetic identification information about the 5 owl samples are provided alongside the sample identifiers.

5.3.4. Assignment of 15 hybrids (13 putative and 2 known). Here we examine whether the Mongrail method produced posterior probabilities for “preferred” models high enough to allow classification of a hybrid sample. Table 3 shows that if 0.99 is chosen as the threshold value to make a decision, we were able to make an assignment call for all hybrids except one (TLW519). The posterior support for the one unassigned hybrid sample TLW519 is still very high (0.96) although below the threshold (0.99) and it too would be assigned if a lower threshold (such as 0.95) were used. Based on our model framework, genealogical class **d**, **c** and **b** refers to a pure barred owl, a F1 hybrid and a backcross with barred owl respectively. We see that the inferred genealogical class for each sample matches the prior genetic classification suggesting that Mongrail provides reasonable results that align with previous conclusions (analysis with NewHybrids produces similar results, see Appendix: Table B1).

TABLE 3. Table showing the assignment call for all 15 hybrid owl samples . The primary and genetic identification were provided by [15]. Assuming a recombination rate of 1.5cM/Mb and considering the “maximally informative” case (the map length for each scaffold is chosen in such a way that its equivalent physical length is almost equal to the length of the scaffold) for all 15 scaffolds, the “preferred” model along with its posterior probability is presented in the table. Based on our model framework, genealogical class **d**, **c** and **b** refers to a pure barred owl, a F1 hybrid and a backcross (with barred owl) respectively.

Sample Names	Primary Identification	Genetic Identification	Posterior Probability	Preferred Model
TLW521	Unknown	Barred Owl	1	d
TLW532	Unknown	Barred Owl	1	d
AFRD90	Unknown	Barred Owl	1	d
CYWC009	Unknown	Barred Owl	1	d
1957-00137	Unknown	F1	1	c
1957-00240	Unknown	F1	1	c
1957-00243	Unknown	F1	1	c
LCW1363	Unknown	F1	1	c
LCW1383	Unknown	F1	1	c
ZRH600	Unknown	F1	1	c
ZRH610	Unknown	F1	1	c
ZRH962	Known Hybrid	F1	1	c
TLW519	Unknown	Backcross	0.961	b
TLW528	Unknown	Backcross	1	b
ZRH607	Known Hybrid	Backcross	0.999	b

CHAPTER 6

Discussion

In this article we developed a full-likelihood Bayesian method to identify species hybrids, explicitly modeling recombination, using population genome sequence data (multiple biallelic SNP markers). We compute the posterior probability of an individual belonging to different genealogical classes (pure population, F1 or F2 hybrid, or backcross). Hybridization is an important evolutionary process influencing biodiversity and impacting diversity generating processes such as speciation [1]. Mongrail should thus be particularly useful to researchers in the fields of conservation biology or population management. In particular, identifying hybrids is a first step in studying various consequences of natural hybridization such as hybrid inviability, introgression between species, and so on [4]. The method developed here offers an improvement over the hybrid inference method of [2] for the analysis of population genomic data by explicitly modeling linked markers. Rapid progress in sequencing technologies has produced datasets with hundreds of thousands of linked loci, clearly violating this assumption. Treating linked loci as independent reduces NewHybrids to a composite likelihood method, potentially leading to overconfidence and/or inaccuracy in identifying hybrids.

We considered a simple model of recombination which closely mimics the population genetic process. This allowed us to build a full-likelihood method with linked markers without the need to use MCMC, reducing the burden of heavy computational cost and the risk of non-convergence. We compared the performance of Mongrail and NewHybrids by performing two extensive simulation studies. The first type, a comprehensive simulation study, aimed at investigating the two inference methods under a broad range of haplotype frequency distributions. The second type, based on a structured coalescent model was performed to generate biologically realistic linkage disequilibrium patterns among haplotypes from the source populations. The simulator generates a pair of haplotypes (a diplotype) for an individual from two populations assuming linkage and recombination in the formation of hybrids. Simulated data were analyzed to compute the posterior probabilities

of genealogical classes using both Mongrail and NewHybrids. This allowed us to evaluate their relative statistical performance in distinguishing the different genealogical classes.

The comprehensive simulation study demonstrated the improved performance of Mongrail over NewHybrids using varying levels of biological and experimental parameters such as number of chromosomes, number of loci, expected recombination frequency (in cM) etc. The simulation studies indicated that Mongrail was often able to correctly identify the genealogical classes with high certainty, whereas NewHybrids more often failed to infer the correct genealogical class, especially in the case of hybrids or backcrosses. Even with high values of expected recombination frequency, NewHybrids often overestimated the posterior probabilities (as expected in a composite likelihood). With only two generations of mating few recombinations are expected over the entire chromosome, violating the assumption of the [2] method that markers are unlinked.

The comprehensive simulation study also suggested that for, both methods, increasing the number of chromosomes has a large effect on the power to infer the correct genealogical class, whereas increasing the number of markers has little effect. Increasing the map-length of the chromosome also greatly increases power by increasing the expected number of recombinations. One outcome of this is that relatively few markers provide sufficient power, reducing the potential computational cost incurred due to a large number of linked loci. Currently, our program implementing this algorithm (Mongrail) allows only 10 markers per chromosome, although the statistical model allows for an arbitrary number of markers. This limitation of the current program does not negatively impact our empirical analysis since we can utilize the power of increased numbers of chromosomes. Many diploid species have sufficient numbers of chromosomes to render the method powerful. The computational complexity increases only linearly in the number of chromosomes.

The coalescent simulation study evaluated the performance of the two methods in inferring individual genealogical classes with varying levels of migration between populations. We considered a wide range of migration, $M \in \{0.1, 0.25, 1, 10, 100\}$. Applying Wright's approximate formula for expected F_{ST} (fixation index), $F_{ST} = 1/(4N_0m + 1) = 1/(M + 1)$, these values correspond to $F_{ST} \in \{0.91, 0.8, 0.5, 0.091, 0.0091\}$. Low values of migration translate to a high value of F_{ST} , indicating greater genetic differentiation between the two populations and vice versa. This simulation study suggested both methods perform well inferring purebreds for low values of migration

($M \leq 1$ equivalent to $F_{ST} \geq 0.5$). But with an increase in migration, Mongrail performs better in distinguishing the genealogical classes on average compared to Newhybrids. As model complexity increases (from F1 hybrid to backcross to F2 hybrid) the performance of both methods declines. In general, Mongrail is more effective in distinguishing hybrids and backcrosses compared to Newhybrids under the simulation conditions we examined, even when genetic differentiation between populations is low.

We applied Mongrail to a previously published whole genome sequence dataset consisting of spotted owls, barred owls and their hybrids. Mongrail was able to infer genealogical classes for all the putative hybrids as well as the purebreds with high posterior probability. This was true despite using only 10 markers from each of 15 largest autosomal scaffolds. The markers were spread evenly across the entire length of scaffold to attain maximum information from the data. No prior information were available on the hybridization rates or differential fitnesses among the six genealogical classes, thus a discrete uniform prior on the classes was used. We also assumed a uniform recombination rate on each scaffold – as no prior information was available for the owl dataset – and extrapolated the specific value from another species, the zebra finch. Mongrail is able to accommodate variable recombination rates if such information is available. The genealogical classifications did not change when other much higher (or lower) rates were used, suggesting that for many species a rough estimate of the recombination rate should be sufficient for use of Mongrail.

The model underlying Mongrail assumes random mating (Hardy-Weinberg equilibrium), and the method requires phased diplotypes and specified haplotype frequencies for the two populations. Many genomic datasets are comprised of unphased genotype data and the population haplotype frequencies are usually unknown. In our study, we phased the pure individuals using BEAGLE version 5.1 [7] on each population separately. With advances in sequencing technologies complete phased data may be common in future, thus eliminating the need for phase inference. Population haplotype frequencies were estimated using the Multinomial-Dirichlet posterior mean of the reference samples. An alternative would be to allow for uncertainties by integrating over the posterior density. An advantage of our approach is that we calculate the likelihood of hybrid individuals without assuming phase, by integrating over all compatible haplotypes thus taking into account the uncertainties.

A current limitation of Mongrail, is that it only allows biallelic SNP loci. The vast majority of SNP loci in most species are biallelic, so plenty of biallelic SNPs are available for use rendering this constraint largely unimportant. Nonetheless, the method can be easily extended to multiple alleles by redefining the bit operations of the program. An assumption of the current method is that hybridization between the two species (populations) occurred within the last $n = 2$ generations. Therefore the extent to which many generations of backcrossing ($n > 2$) affects our inference method is unknown. Since we suspect that individuals resulting from many generations of unidirectional backcrossing may resemble purebreds genetically this potentially limits the scope of our method. However, information probably dissipates quickly with additional generations of hybridization and the number of possible models quickly increases – there could also be identifiability issues. We suggest that it is sensible to focus exclusively on identifying recent hybrid ancestry until further theoretical studies confirm our ability to infer more distant ancestries. The approaches developed here could be extended to allow more generations of hybridization but computational expense will increase dramatically.

Finally, though Mongrail does not require any fixed differences (or, exclusive alleles) between the two populations, high levels of genetic differentiation increase the power of the method to identify hybrids. In particular, for the empirical data analysis, we found the SO genomes had little polymorphism and were very distinct from the much more variable BO genomes. The exceptionally low genetic diversity of SO genomes is likely due to a recent population decline. This may explain the high posterior probabilities we obtained as resulting from strong genetic differentiation between the two species.

Mongrail currently requires knowledge of haplotype frequencies in source populations. If this information is unavailable we need “pure” individuals (no recent hybrid ancestry) to be present in the sample in order to estimate these parameters. This implies that the two pure populations should be clearly distinguishable. One can be fairly confident of choosing individuals who are likely to be pure if sampled from two non-overlapping geographical regions. This condition is automatically ensured for most allopatric populations. It is currently difficult to apply Mongrail to individuals sampled from a sympatric region when the two species (populations) cannot be separated or to a population with clinal variation of haplotype frequencies. To address this issue a possible extension

of our approach could be to jointly infer population haplotype frequencies and genealogical classes. This would require developing a Markov Chain Monte Carlo method similar to the one presented in [2] (which jointly infers allele frequencies and genotype frequency classes). If pure individuals can be distinguished, it is recommended that at least 10 individuals are sampled from each population to produce sensible estimates of population haplotype frequencies. An important question, beyond the scope of the current study, concerns the effect of population sample size on errors of estimates of population haplotype frequencies (and resulting genealogical classifications).

Analyzing larger number of markers per chromosome has little effect in increasing the power to infer genealogical classes, yet an increased number of markers requires more haplotype frequencies to be estimated. Thus, there appears to be a trade-off between information gained from additional markers and cost incurred by additional parameter estimates.

In conclusion the method presented in this paper has the power to infer hybrids using linked genetic data without the requirement for any fixed or exclusive alleles to be present between two diploid populations. Extensive simulations show the potentially adverse effects of applying the widely used program NewHybrids (which assumes unlinked loci) to genomic data composed of large numbers of linked markers. The fact that the number of chromosomes, and the size of the intervals, contribute more to power than the number of markers allows an exact likelihood approach to be developed that is powerful without an excessive computational burden. Due to the analytical nature of the theory and consequent absence of simulation-based methodologies (such as MCMC) from the inference procedure the method is computationally efficient so that most of the runs finish within a few minutes. Rapid advances in sequencing technologies and bioinformatics tools, along with decreasing costs of genome sequencing and assembly, will increase the availability of genomic datasets for hybridizing non-model organisms. Therefore efficient statistical methods for identifying hybrids, such as Mongrail, that account for linkage will be increasingly needed in conservation biology and related disciplines.

6.1. Data availability

Simulated datasets and scripts for generating simulations are available at <https://github.com/mongrail/simulations>. Scripts for analyzing the empirical dataset are available at <https://github.com/mongrail/simulations>.

[//github.com/mongrail/Scripts](https://github.com/mongrail/Scripts). The Open Source C program Mongrail, implementing the algorithms presented in this paper, is available at <https://github.com/mongrail>. The owl dataset analyzed in this paper is publicly available at <https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?analysis=SRZ190173>.

6.2. Funding

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

APPENDIX A

Simulation

A.1. Comprehensive simulation: Methods

Here, we describe the details of the simulation procedures used in the paper and of the statistical methods used to summarize the results in making a comparison of the statistical performance of Mongrail and NewHybrids.

The simulation experiment used a factorial design allowing the performance of the methods to be assessed for many combinations of parameters. The parameters (factors) and their values were as follows:

- (1) Number of chromosomes: $K = 1, 2, 5, 10, 20$
- (2) Number of loci per chromosome: $L = 1, 5, 10$
- (3) Expected recombination frequency (in cM): $R = 1, 25, 50$ between the first and the last locus
- (4) Number of distinct haplotype sequences per chromosome for each population: $h = 5, 10, 15$
- (5) Allelic configurations of haplotypes, generated by simulating the switches between allele states (see Section A.1.1). Switch rates used: $c = 0.1, 1, L/2$
- (6) Haplotype frequencies, following a Dirichlet distribution (symmetrical with parameters $\alpha = 1, 5$, or non-symmetrical with parameters $w = 5, 20$) (see Section A.1.2)

For simplicity, in all simulations we fixed the length of each chromosome (D) to be 240 Mb and the recombination rate (r) to be 1.2 cM/Mb respectively.

The R scripts for simulating haplotype configurations (Section A.1.1), simulating haplotype frequencies (Section A.1.2) and simulating marker positions (Section A.1.3) are available at Scripts.

A.1.1. Simulating haplotype configurations. The parameters for the simulation study were generated using R version 3.6.3. To simulate haplotype configurations we mimic recombination by using a “switching process” (that flips the adjacent marker state) operating along the

chromosome. The switch rate on a particular interval is $p = c/L$ (when $L \neq 1$) where p is the probability of a switch from 0 to 1 (or 1 to 0).

The allele state for the first marker is simulated randomly from a Bernoulli(1/2) distribution. Given the value of L and c , the allele states for the rest of the $(L - 1)$ markers are simulated from a Bernoulli(p) distribution (following the description in the previous paragraph). We repeat the process above until we obtain h distinct haplotypes. These simulations were performed using the *rbinom* function in R.

When $L = 1$, there are only two possible haplotypes 0 and 1 (the two allele states possible for a single marker). Thus for $L = 1$, the number of haplotypes $h = 2$.

We use the same haplotype configurations for both the populations (A and B) for each of the K chromosomes.

A.1.2. Simulating haplotype frequencies. Given the h distinct haplotypes generated in section A.1.1, we simulate their corresponding population frequencies. Let \mathbf{f}_k^A and \mathbf{f}_k^B be the haplotype frequencies in population A and B respectively for chromosome k for $k = 1, 2, \dots, K$. We perform these simulations using either a symmetrical or non-symmetrical Dirichlet distribution. The simulations were generated using the *rdirichlet* function in R. We simulate h frequencies from a Dirichlet Distribution with parameters $\alpha_1, \alpha_2, \dots, \alpha_h$.

For the symmetric Dirichlet distribution we set $\alpha_i = \alpha$ for $i = 1, 2, \dots, h$ and consider two cases: $\alpha = 1$ or $\alpha = 5$.

For the non-symmetric Dirichlet distribution we use:

$$\alpha_i = \begin{cases} 0.7 \times w, & i = 1 \\ 0.3 \times \frac{w}{h-1}, & i > 1, \end{cases}$$

and again consider two cases: $w = 5$ or $w = 20$.

When $L = 1$, there are only 2 haplotypes, and we simulate $h = 2$ frequencies from Beta distributions (univariate cases of the Dirichlet distribution) with parameters α_1 and α_2 :

$$(1) \alpha_1 = \alpha_2 = 1$$

$$(2) \alpha_1 = \alpha_2 = 5$$

$$(3) \alpha_1 = 0.7 \times 5, \alpha_2 = 0.3 \times 5$$

$$(4) \alpha_1 = 0.7 \times 20, \alpha_2 = 0.3 \times 20$$

We repeat this process for each of the K chromosomes.

A.1.3. Assigning marker positions. Given the value of $L(\neq 1)$, R and r , we calculate the distance between the first and last marker in units of base pair (bp) which is

$$\frac{R \times (10^6)}{r}.$$

We consider all L markers to be equidistant from each other. Therefore the inter-marker distance (in units of bp) is given by

$$\frac{R \times (10^6)}{r \times (L - 1)}.$$

For simplicity, we consider the position of the first marker (in bp) as $\frac{R \times (10^6)}{r}$ and obtain the positions of the rest of the $L - 1$ markers (in bp) following the description above. For $L = 1$, we arbitrarily chose 120000bp as the marker position. We use the same marker positions for both the populations (A and B) for each of the K chromosomes.

A.2. Simulating diplotypes for markers

All simulations in this section were done in the C programming environment. This simulation environment is available as an option in our program Mongrail. Given that the haplotypes and their corresponding frequencies (along with the marker positions) have been generated for the two populations A and B, generating a diploid individual is equivalent to generating a diplotype (a pair of haplotypes). Generating a diplotype from any of the models: **a** (pure population B), **c** (F1 hybrid) or **d** (pure population A) is straightforward. For these models, each of the two chromosomes derives entirely from one population (A or B). To simulate a chromosome, a sample of size 1 from a multinomial distribution with h types was simulated, where h is the number of haplotypes, and the multinomial proportions \mathbf{f}^j are the haplotype frequencies in the source population, $j \in \{A, B\}$. For this purpose we used the function *gsl_ran_multinomial* from the GNU scientific library. For model **a** (or, **d**) both the chromosomes are simulated independently from population B (or, A) respectively.

In case of model **c**, one chromosome is simulated from population A and the other from population B.

For the other models, one (**b**, **e**) or both (**f**) the chromosomes are recombinant (a mixture between two pure parental chromosomes, one from population A and the other from B). Recombinant chromosomes were simulated in two steps:

Step I

Simulate two pure parental chromosomes (one from population A and the other from B) which is equivalent to simulating from model **c** (described above). Denote the two simulated pure chromosomes as C_A (Population A) and C_B (Population B) respectively. For example, let's say we generated the following two chromosomes for $L = 10$ markers:

$$C_A \equiv 1100011010$$

$$C_B \equiv 0111001001$$

Here the 0 or 1 indicate the allele present at each marker since we consider phased biallelic SNP markers. Therefore C_A (or, C_B) is a binary string of size L . We shall carry on with this example in the following steps.

Step II

Simulate recombinations between the two pure chromosomes to produce a pair of recombinant chromosomes. This is achieved as follows:

- (a) **Simulate the number of recombinations:** We simulate the number of recombinations (n_r) over the length of the chromosome. Assuming the rate of recombination over the chromosome is uniform, and recombination events are independent and never occur simultaneously, n_r follows a Poisson distribution with rate parameter λ (the expected number of recombination events over the entire chromosome). Given the rate of recombination (r) in units of cM/Mb and the length of the chromosome (say, D) in units of Mb, the map distance of the chromosome in units of

centiMorgans (cM) is $D \times r$ and $\lambda = (D \times r)/100$. We used the function *gsl_ran_poisson* from GNU Scientific Library to generate n_r .

- (b) **Simulate the positions of recombinations:** The positions of recombinations on the chromosome, conditional on the number of recombinations (n_r), are simulated from a continuous uniform distribution over the length of the chromosome (in cM). For example, if $n_r = 3$, we might generate a single crossover event between each of the following pair of marker positions: (3, 4), (5, 6) and (7, 8). We used the function *gsl_rng_uniform_pos* from GNU Scientific Library to generate the recombination positions.
- (c) **Obtain the population origin of markers:** The population origin (or, ancestry state) of the markers was obtained conditional on the positions of recombinations. The population origin of a SNP marker to the right of an interval changes whenever there is an odd number of recombinations (similarly, an even number of recombinations results in no change). We move from left i.e., 5' end of the chromosome to the right towards the 3' end of the chromosome. Based on the example developed so far, two haplotype ancestry states are produced for $L = 10$ markers:

$$AAABBAABBB \quad \text{and} \quad BBBAABBAAA$$

- (d) **Obtain the allelic state (0/1) of the markers:** The allele type (0 or 1) for each marker is derived from C_A or C_B depending on whether the ancestry state is A or B respectively for the marker under consideration. For the particular C_A and C_B we considered at the beginning and based on the ancestry state *AAABBAABBB*, the alleles at the markers from position 1 to 3 and from positions 6 to 7 should come from population A. This requires that the alleles at these positions should match those of C_A (indicated by the underbraces) and the alleles for the rest of the markers match those of chromosome C_B :

$$C_A \equiv \underbrace{110} \underbrace{00} \underbrace{11} \underbrace{010}$$

$$C_B \equiv 011 \underbrace{10} \underbrace{01} \underbrace{001}$$

Therefore, given the ancestry state $AAABBAABBB$, the allele states of the markers for one recombinant haplotype are

$$r_{H_1} \equiv 1101011001.$$

Similarly, given the ancestry state is $BBBAABBAAA$, the allele states of the markers for the other haplotype are

$$r_{H_2} \equiv 0110001010.$$

Thus, there are two r_{H_1} and r_{H_2} recombinant haplotypes. We used bit operations in C to perform this step.

- (e) **Simulate a recombinant chromosome:** The two recombinant chromosomes (r_{H_1} and r_{H_2}) are equally likely to occur. So we choose one of them from a Bernoulli(1/2) distribution (say, getting r_{H_1} is defined as success). We used the function `gsl_ran_bernoulli` from GNU Scientific Library to generate a recombinant chromosome.

A.2.1. Distribution of Linkage Disequilibrium (LD). We obtained the linkage disequilibrium coefficient (r^2) under a particular set of simulation combinations (cases where haplotype frequencies were simulated using symmetric Dirichlet Distribution). We only consider the linkage disequilibrium coefficient (r^2) between the first and last marker. For brevity we plot the distribution of r^2 values for four specific simulation combinations (Figure A1):

- (1) $L = 5, h = 5, \alpha = 5, c = 1$
- (2) $L = 10, h = 15, \alpha = 1, c = 0.1$
- (3) $L = 5, h = 15, \alpha = 5, c = L/2$
- (4) $L = 5, h = 15, \alpha = 5, c = 0.1$

We find that under our comprehensive simulation setup, the r^2 values range from really low values (close to 0) to very high values (close to 1). Therefore this shows that the comprehensive simulation design does not explicitly produce high LD values. It produces a broad range of LD values ranging from 0 to 1.

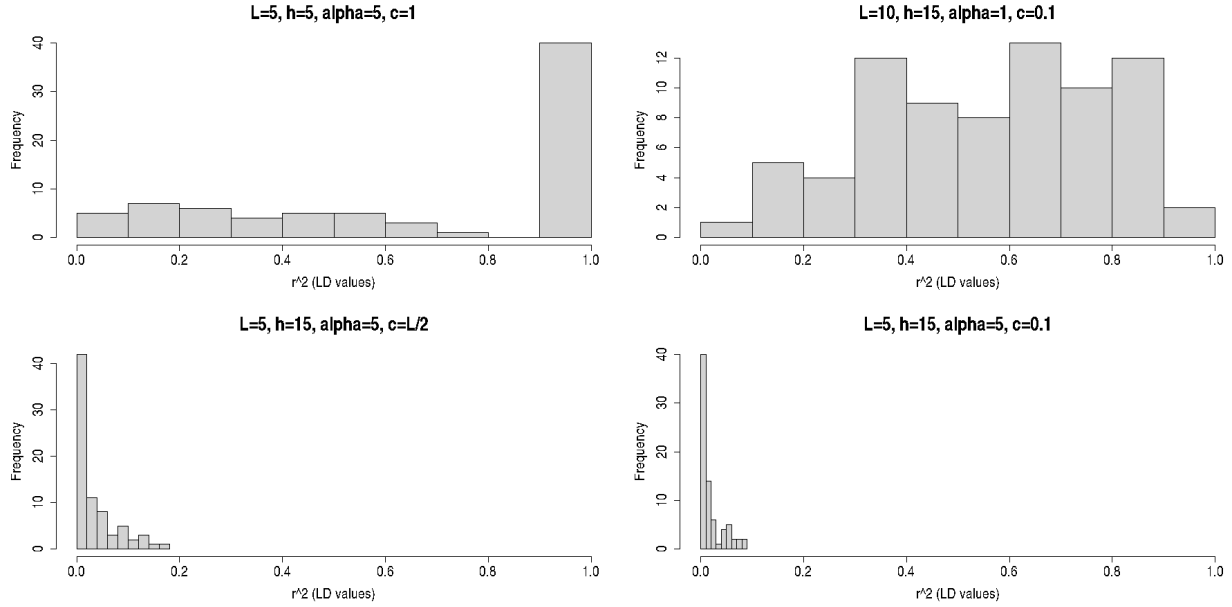


FIGURE A1. Histogram showing the frequency distribution of linkage disequilibrium coefficient (r^2) under four specific simulation combinations: (a) $L = 5, h = 5, \alpha = 5, c = 1$ (top-left), (b) $L = 10, h = 15, \alpha = 1, c = 0.1$ (top-right), (c) $L = 5, h = 15, \alpha = 5, c = L/2$ (bottom-left), (d) $L = 5, h = 15, \alpha = 5, c = 0.1$ (bottom-right).

A.3. Coalescent simulation: Simulating diplotypes

Here we describe the procedure for generating diplotypes under different genealogical classes given a sample of chromosomes from two populations simulated under the structured coalescent model. Generating a diplotype which is a purebred (model **d**) or F1 (model **c**) is straightforward. Under model **d** both the chromosomes arise entirely from population B. To simulate a purebred diploid individual from population B, we simulate 2 additional chromosomes from population B in the coalescent simulation; the two chromosomes from the purebred diplotype. In an F1 hybrid, one chromosome arises from population A and the other from population B. To simulate an F1 hybrid, we simulate 1 additional chromosome from each of the two populations (A and B) during the coalescent simulation, the two chromosomes form the F1 diplotype. For a backcross (model **b**) one chromosome arises entirely from population A and the other is a recombinant between a pair of chromosomes, one from population A and the other from B. Therefore we simulate 3 additional chromosomes (2 from population A and 1 from population B) and follow Step II of Section A.2 to produce a recombinant chromosome (using one of the chromosomes from population A and

another from B). The A-B recombinant and A chromosomes together form the backcross diplotype. For an F2 hybrid (model **f**) both chromosomes are recombinant between populations A and B. To simulate an F2 hybrid 4 additional chromosomes are simulated (2 from population A and 2 from population B) during the coalescent simulation and each pair (one chromosome from A and one from B) undergo recombination according to Step II of Section A.2. The two A-B recombinant chromosomes together form the F2 diplotype.

A.3.1. Summary of number of population haplotypes generated. We computed the mean and standard deviation of the number of distinct haplotypes observed in populations A and B under the coalescent simulation. These results are presented for all values of $M = 0.1, 0.25, 1, 10, 100$ in Table A1. As the value of M increases, the mean number of unique

TABLE A1. The mean and standard deviation (S.D.) of the number of distinct 10-locus haplotypes generated under each coalescent simulation scenario for two populations A and B.

Migration Rate (M)	Population A		Population B	
	Mean	S.D.	Mean	S.D.
0.1	9.17277	5.7704	9.1582	5.74928
0.25	13.3307	7.23048	13.3239	7.25902
1	17.9721	8.62542	17.9387	8.62992
10	21.3009	8.84554	21.2143	8.78405
100	22.6983	8.89778	22.6209	8.85955

haplotypes increases in both simulated populations. This is expected as an increase in migration introduces additional shared haplotypes into both populations. Since the coalescent simulation was carried out under a symmetric island model the two populations are equivalent and thus the mean and standard deviation are nearly identical between the two populations.

APPENDIX B

Empirical Dataset

B.1. Mongrail: Analysis of putative hybrids

Marker genotypes were filtered for each scaffold using positions chosen such that they satisfied physical distances specified for the different analyses (either a sliding window analysis or a maximally informative distribution) using BCFtools, and reformatted for input to Mongrail using Awk and Sed. A C program was used to enumerate all compatible diplotypes for putative hybrid individuals for each choice of markers, which formed the input for Mongrail. For each scaffold and genealogical class, the total likelihood is obtained as a sum of the likelihood calculated for each compatible diplotype. When multiple scaffolds are analyzed, likelihoods are multiplied across scaffolds, treating scaffolds as equivalent to independent chromosomes. Posterior probabilities are obtained from the normalized likelihoods (using a uniform prior on genealogical classes). Results were plotted in R using Tidyverse, RColorBrewer and reshape.

B.2. Applying NewHybrids to the owl dataset

B.2.1. Assignment of 15 hybrids (13 putative and 2 known). We applied NewHybrids to the owl dataset to obtain genealogical classifications of the putative hybrids. We used the same set of 10 markers from the 15 scaffolds that were considered for Mongrail (the case where the assumed recombination rate was 1.5cM/Mb under a “maximally informative” region size). Therefore we used 150 markers in total to construct a table (see Table B1) of the posterior probabilities of the “preferred” model for each of the 15 hybrids (along with the primary and genetic identification information of [15]).

Based on the NewHybrids model framework, genealogical class d, c and b refers to a pure barred owl, a F1 hybrid and a backcross with barred owl respectively. The results are similar to those obtained by applying Mongrail to this owl dataset (see Table 3). The preferred model has

Sample Names	Primary Identification	Genetic Identification	Posterior Probability	Preferred Model
TLW521	Unknown	Barred Owl	1	d
TLW532	Unknown	Barred Owl	1	d
AFRD90	Unknown	Barred Owl	1	d
CYWC009	Unknown	Barred Owl	1	d
1957-00137	Unknown	F1	1	c
1957-00240	Unknown	F1	1	c
1957-00243	Unknown	F1	1	c
LCW1363	Unknown	F1	1	c
LCW1383	Unknown	F1	1	c
ZRH600	Unknown	F1	1	c
ZRH610	Unknown	F1	1	c
ZRH962	Known Hybrid	F1	1	c
TLW519	Unknown	Backcross	0.99912	b
TLW528	Unknown	Backcross	1	b
ZRH607	Known Hybrid	Backcross	1	b

TABLE B1. Table showing the assignment call for all 15 hybrid owl samples using NewHybrids.

posterior probability 1 for all samples suggesting that both Mongrail and NewHybrids similarly provide reasonable results that align with previous conclusions.

B.2.2. Comparison of posterior probability distributions. In this analysis we compare the posterior probabilities under the two methods (Mongrail and NewHybrids) for 5 putative hybrids. These owls were chosen to be representative of the three observed categories genetically identified by [15]. The categories along with the sample names are presented in Table B2.

Primary identification	Genetic identification	Sample names
putative hybrid	BO	CYWC009
known hybrid	F1 hybrid (F1)	ZRH962
putative hybrid	backcross (F1 × BO)	TLW519
putative hybrid	backcross (F1 × BO)	TLW528
known hybrid	backcross (F1 × BO)	ZRH607

TABLE B2. Details for 5 individuals chosen for detailed analysis. Primary and genetic identifications are from [15].

We chose 10 markers from the largest scaffold (Super-Scaffold_7 of length 72.11 Mb) and consider two different map lengths. We considered $R = 1.5$ cM and the “maximally informative” case described in section 5.2.2.

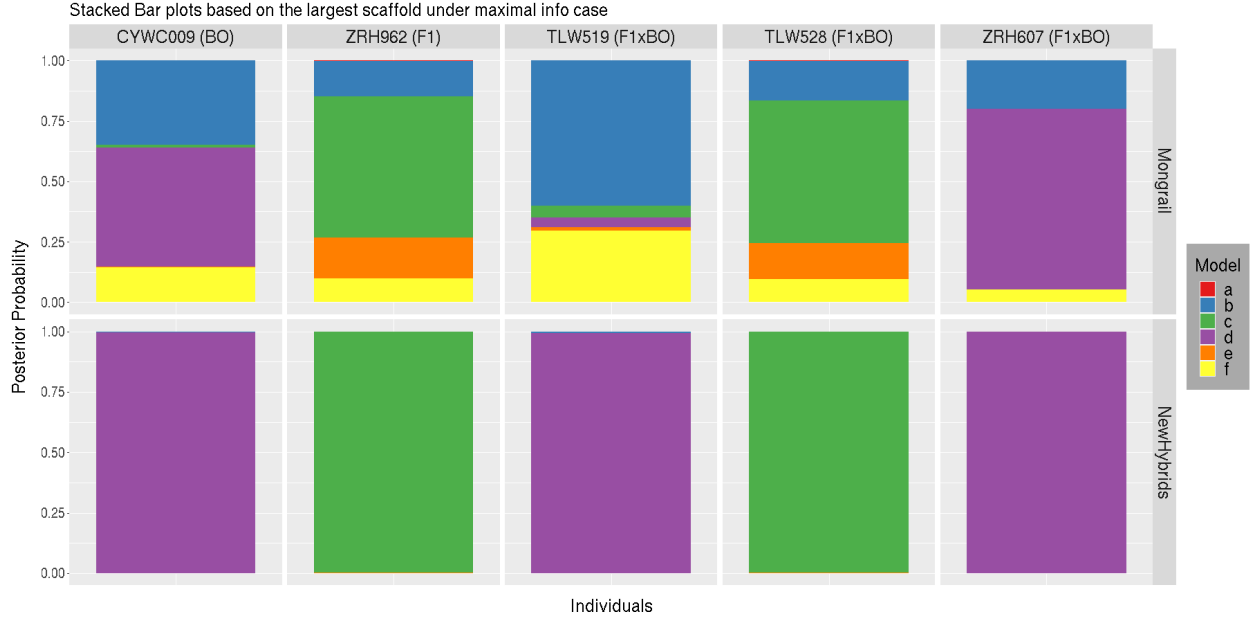


FIGURE B1. Distribution of posterior probabilities is constructed for 5 owl samples (CYWC009, ZRH962, TLW519, TLW528, ZRH607). A set of $L = 10$ markers are chosen from the largest scaffold Super-Scaffold_7 (size 72.11 Mb) according to the “maximally informative” case. For each individual, posterior probabilities for Mongrail are shown in the top plot and for NewHybrids in the bottom plot. The posterior probabilities for different genealogical classes are represented by segments of different colors. The proportion of the stacked bar plot comprised of a particular color indicates the posterior probability of the model corresponding to that color. 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. In our model framework, BO is treated as population A and SO as population B.

There is a greater distribution of posterior probabilities among genealogical classes under Mongrail by comparison with NewHybrids (compare top and bottom rows in Figures B1 and B2). Even with 10 markers, NewHybrids tends to produce extremely high posterior probabilities for specific genealogical classes for both small (Figure B2) and large (Figure B1) intervals. Thus NewHybrids seems to be overconfident even when information is quite low. In one case (individual TLW519) the posterior probability for genealogical class d is nearly one in the “maximally informative” case (Figure B1), whereas for a smaller interval ($R = 1.5$ cM) the posterior probability for genealogical class d drops to 0.02 and the posterior probability for genealogical class c increases to 0.83 (Figure B2). In both cases the genealogical class with the highest posterior probability differs

from the preferred genealogical class. Mongrail appears more conservative, never assigning a posterior probability to a non-preferred model of more than 0.30 for individual TLW519.

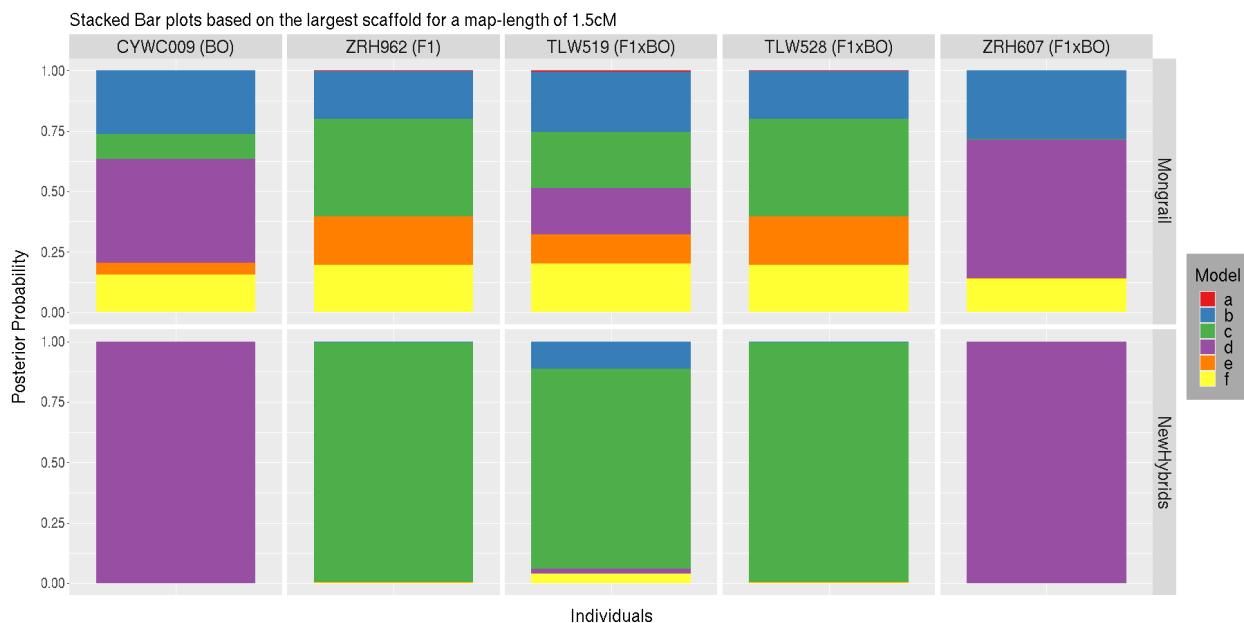


FIGURE B2. Distribution of posterior probabilities is constructed for 5 owl samples (CYWC009, ZRH962, TLW519, TLW528, ZRH607) for a set of $L = 10$ markers, where the markers are evenly spaced across the largest scaffold Super-Scaffold_7 (size 72.11 Mb) with a constant map length of $R = 1.5\text{cM}$. The initiating marker is same as the first marker in the “maximally informative” case. For each individual, posterior probabilities for Mongrail are shown in the top plot and for NewHybrids in the bottom plot. The posterior probabilities for different genealogical classes are represented by segments of different colors. The proportion of the stacked bar plot comprised of a particular color indicates the posterior probability of the model corresponding to that color. 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. In our model framework, BO is treated as population A and SO as population B.

B.2.3. NewHybrids analysis using unlinked SNPs with fixed differences. To

explore the effects of fixed differences between these two species on hybridization inference with NewHybrids, we applied NewHybrids to the same set of 5 putative hybrids analyzed previously (See Table B2) but specifically chose loci with fixed differences. We randomly chose a single marker from among all the markers with fixed differences for each of the 15 scaffolds. As shown in Figure B3, when loci are fixed for alternate alleles NewHybrids produces extremely high posterior

probabilities (greater than 0.998) for specific genealogical classes, even when using only 15 markers. For the individuals analyzed, the “preferred” genealogical class matches the prior genetic classification. Caution is needed when considering such results, since without exhaustive population sampling it is always uncertain whether “fixed” differences are truly fixed or instead an artifact of genotyping error or failure to sample an allele.

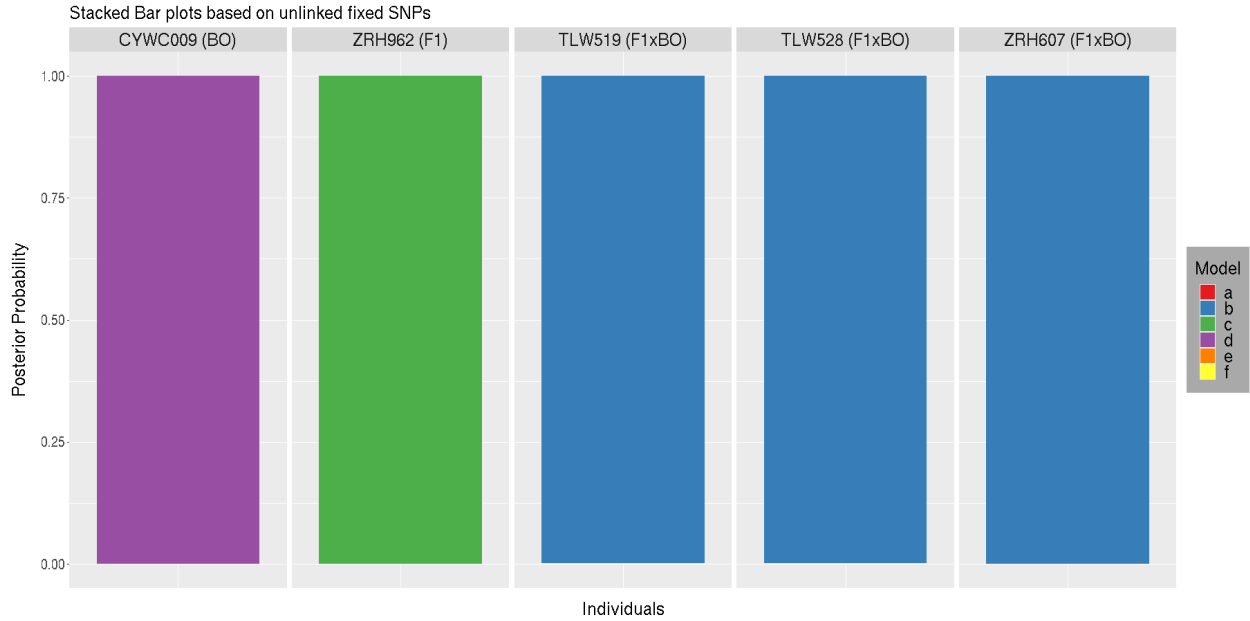


FIGURE B3. Distribution of posterior probabilities for 5 owl samples (CYWC009, ZRH962, TLW519, TLW528, ZRH607) obtained by applying NewHybrids to a set of 15 fixed SNPs, where each SNP is chosen randomly from each of the 15 scaffolds. The posterior probabilities for different genealogical classes are represented by segments of different colors. The proportion of the stacked bar plot comprised of a particular color indicates the posterior probability of the model corresponding to that color. 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. In our model framework, BO is treated as population A and SO as population B.

Bibliography

- [1] R. ABBOTT, D. ALBACH, S. ANSELL, J. W. ARNTZEN, S. J. BAIRD, N. BIERNE, J. BOUGHMAN, A. BRELSFORD, C. A. BUERKLE, R. BUGGS, ET AL., *Hybridization and speciation*, *Journal of Evolutionary Biology*, 26 (2013), pp. 229–246.
- [2] E. ANDERSON AND E. THOMPSON, *A model-based method for identifying species hybrids using multilocus genetic data*, *Genetics*, 160 (2002), pp. 1217–1229.
- [3] R. G. ANTHONY, E. D. FORSMAN, A. B. FRANKLIN, D. R. ANDERSON, K. P. BURNHAM, G. C. WHITE, C. J. SCHWARZ, J. D. NICHOLS, J. E. HINES, G. S. OLSON, ET AL., *Status and trends in demography of northern spotted owls, 1985–2003*, *Wildlife Monographs*, 163 (2006), pp. 1–48.
- [4] M. L. ARNOLD, *Natural hybridization as an evolutionary process*, *Annual Review of Ecology and Systematics*, 23 (1992), pp. 237–261.
- [5] J. C. AVISE AND M. J. VAN DEN AVYLE, *Genetic analysis of reproduction of hybrid white bass x striped bass in the savannah river*, *Transactions of the American Fisheries Society*, 113 (1984), pp. 563–570.
- [6] B. L. BROWNING, X. TIAN, Y. ZHOU, AND S. R. BROWNING, *Fast two-stage phasing of large-scale sequence data*, *The American Journal of Human Genetics*, 108 (2021), pp. 1880–1890.
- [7] S. R. BROWNING AND B. L. BROWNING, *Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering*, *The American Journal of Human Genetics*, 81 (2007), pp. 1084–1097.
- [8] D. E. CAMPTON AND F. M. UTTER, *Natural hybridization between steelhead trout (*salmo gairdneri*) and coastal cutthroat trout (*salmo clarki clarki*) in two puget sound streams*, *Canadian Journal of Fisheries and Aquatic Sciences*, 42 (1985), pp. 110–119.
- [9] D. A. CLARK, R. G. ANTHONY, AND L. S. ANDREWS, *Relationship between wildfire, salvage logging, and occupancy of nesting territories by northern spotted owls*, *The Journal of Wildlife Management*, 77 (2013), pp. 672–688.
- [10] L. CONGIU, I. DUPANLOUP, T. PATARNELLO, F. FONTANA, R. ROSSI, G. ARLATI, AND L. ZANE, *Identification of interspecific hybrids by amplified fragment length polymorphism: the case of sturgeon*, *Molecular Ecology*, 10 (2001), pp. 2355–2359.

- [11] P. DANECEK, J. K. BONFIELD, J. LIDDLE, J. MARSHALL, V. OHAN, M. O. POLLARD, A. WHITWHAM, T. KEANE, S. A. MCCARTHY, R. M. DAVIES, ET AL., *Twelve years of SAMtools and BCFtools*, Gigascience, 10 (2021), p. giab008.
- [12] R. J. DAVIS, B. HOLLEN, J. HOBSON, J. E. GOWER, AND D. KEENUM, *Northwest forest plan—the first 20 years (1994–2013): status and trends of northern spotted owl habitats*, Gen. Tech. Rep. PNW-GTR-929. Portland, OR: US Department of Agriculture, Forest Service, Pacific Northwest Research Station. 54 p., 929 (2016).
- [13] E. D. FORSMAN, E. C. MESLOW, AND H. M. WIGHT, *Distribution and biology of the spotted owl in oregon*, Wildlife Monographs, (1984), pp. 3–64.
- [14] A. B. FRANKLIN, K. M. DUGGER, D. B. LESMEISTER, R. J. DAVIS, J. D. WIENS, G. C. WHITE, J. D. NICHOLS, J. E. HINES, C. B. YACKULIC, C. J. SCHWARZ, ET AL., *Range-wide declines of northern spotted owl populations in the pacific northwest: A meta-analysis*, Biological Conservation, 259 (2021), p. 109168.
- [15] N. T. FUJITO, Z. R. HANNA, M. LEVY-SAKIN, R. C. BOWIE, P.-Y. KWOK, J. P. DUMBACHER, AND J. D. WALL, *Genomic variation and recent population histories of spotted (*strix occidentalis*) and barred (*strix varia*) owls*, Genome Biology and Evolution, 13 (2021), p. evab066.
- [16] J. L. GANEY, H. Y. WAN, S. A. CUSHMAN, AND C. D. VOJTA, *Conflicting perspectives on spotted owls, wildfire, and forest restoration*, Fire Ecology, 13 (2017), pp. 146–165.
- [17] W. S. GRANT, G. B. MILNER, P. KRASNOWSKI, AND F. M. UTTER, *Use of biochemical genetic variants for identification of sockeye salmon (*oncorhynchus nerka*) stocks in cook inlet, alaska*, Canadian Journal of Fisheries and Aquatic Sciences, 37 (1980), pp. 1236–1247.
- [18] R. GUTIÉRREZ, M. CODY, S. COURTNEY, AND A. B. FRANKLIN, *The invasion of barred owls and its potential effect on the spotted owl: a conservation conundrum*, Biological Invasions, 9 (2007), pp. 181–196.
- [19] S. M. HAIG, T. D. MULLINS, E. D. FORSMAN, P. W. TRAIL, AND L. WENNERBERG, *Genetic identification of spotted owls, barred owls, and their hybrids: legal implications of hybrid identity*, Conservation Biology, 18 (2004), pp. 1347–1357.
- [20] T. E. HAMER, E. D. FORSMAN, A. FUCHS, AND M. WALTERS, *Hybridization between barred and spotted owls*, The Auk, (1994), pp. 487–492.
- [21] Z. R. HANNA, J. P. DUMBACHER, R. C. BOWIE, J. B. HENDERSON, AND J. D. WALL, *Whole-genome analysis of introgression between the spotted owl and barred owl (*strix occidentalis* and *strix varia*, respectively; aves: Strigidae) in western north america*, G3: Genes, Genomes, Genetics, 8 (2018), pp. 3945–3952.
- [22] Z. R. HANNA, J. B. HENDERSON, J. D. WALL, C. A. EMERLING, J. FUCHS, C. RUNCKEL, D. P. MINDELL, R. C. BOWIE, J. L. DERISI, AND J. P. DUMBACHER, *Northern spotted owl (*strix occidentalis caurina*) genome: divergence with the barred owl (*strix varia*) and characterization of light-associated genes*, Genome Biology and Evolution, 9 (2017), pp. 2522–2545.

- [23] H. HARRIS, *Genetics of enzyme polymorphisms in man*, Proceedings of the Royal Society of London. Series B. Biological Sciences, 164 (1966), pp. 298–310.
- [24] R. R. HUDSON, *Generating samples under a wright–fisher neutral model of genetic variation*, Bioinformatics, 18 (2002), pp. 337–338.
- [25] G. W. HUNT AND R. K. SELANDER, *Biochemical genetics of hybridisation in european house mice*, Heredity, 31 (1973), pp. 11–33.
- [26] R. JOLY AND W. ADAMS, *Allozyme analysis of pitch× loblolly pine hybrids produced by supplemental mass-pollination*, Forest Science, 29 (1983), pp. 423–432.
- [27] E. G. KELLY AND E. D. FORSMAN, *Recent records of hybridization between barred owls (*Strix varia*) and northern spotted owls (*S. occidentalis caurina*)*, The Auk, 121 (2004), pp. 806–810.
- [28] T. LAMB AND J. C. AVISE, *Directional introgression of mitochondrial DNA in a hybrid population of tree frogs: the influence of mating behavior*, Proceedings of the National Academy of Sciences, 83 (1986), pp. 2526–2530.
- [29] R. C. LEWONTIN AND J. L. HUBBY, *A molecular approach to the study of genic heterozygosity in natural populations. ii. amount of variation and degree of heterozygosity in natural populations of drosophila pseudoobscura*, Genetics, 54 (1966), p. 595.
- [30] L. L. LONG AND J. D. WOLFE, *Review of the effects of barred owls on spotted owls*, The Journal of Wildlife Management, 83 (2019), pp. 1281–1296.
- [31] J. NASON AND N. ELLSTRAND, *Estimating the frequencies of genetically distinct classes of individuals in hybridized populations*, Journal of Heredity, 84 (1993), pp. 1–12.
- [32] A. J. PAGE, B. TAYLOR, A. J. DELANEY, J. SOARES, T. SEEMANN, J. A. KEANE, AND S. R. HARRIS, *Snpsites: rapid efficient extraction of snps from multi-fasta alignments*, biorxiv, (2016), p. 038190.
- [33] S. PIRY, A. ALAPETITE, J.-M. CORNUET, D. PAETKAU, L. BAUDOUIN, AND A. ESTOUP, *GeneClass2: a software for genetic assignment and first-generation migrant detection*, Journal of Heredity, 95 (2004), pp. 536–539.
- [34] J. K. PRITCHARD, M. STEPHENS, AND P. DONNELLY, *Inference of population structure using multilocus genotype data*, Genetics, 155 (2000), pp. 945–959.
- [35] A. RAMBAUT AND N. C. GRASS, *Seq-gen: an application for the monte carlo simulation of dna sequence evolution along phylogenetic trees*, Bioinformatics, 13 (1997), pp. 235–238.
- [36] B. RANNALA AND J. L. MOUNTAIN, *Detecting immigration by using multilocus genotypes*, Proceedings of the National Academy of Sciences, 94 (1997), pp. 9197–9201.
- [37] M. S. ROY, E. GEFFEN, D. SMITH, E. A. OSTRANDER, AND R. K. WAYNE, *Patterns of differentiation and hybridization in north american wolflike canids, revealed by analysis of microsatellite loci.*, Molecular Biology and Evolution, 11 (1994), pp. 553–570.

- [38] M. STEPHENS, N. J. SMITH, AND P. DONNELLY, *A new statistical method for haplotype reconstruction from population data*, *The American Journal of Human Genetics*, 68 (2001), pp. 978–989.
- [39] D. J. TEMPEL, H. A. KRAMER, G. M. JONES, R. GUTIÉRREZ, S. C. SAWYER, A. KOLTUNOV, M. SLATON, R. TANNER, B. K. HOBART, AND M. Z. PEERY, *Population decline in california spotted owls near their southern range boundary*, *The Journal of Wildlife Management*, 86 (2022), p. e22168.
- [40] F. UTTER AND N. RYMAN, *Genetic markers and mixed stock fisheries*, *Fisheries*, 18 (1993), pp. 11–21.
- [41] J.-P. VÄHÄ AND C. R. PRIMMER, *Efficiency of model-based bayesian methods for detecting hybrid individuals under different hybridization scenarios and with different numbers of loci*, *Molecular Ecology*, 15 (2006), pp. 63–72.
- [42] C. VARIN, N. REID, AND D. FIRTH, *An overview of composite likelihood methods*, *Statistica Sinica*, (2011), pp. 5–42.
- [43] C. VARIN AND P. VIDONI, *A note on composite likelihood inference and model selection*, *Biometrika*, 92 (2005), pp. 519–528.
- [44] G. A. WILSON AND B. RANNALA, *Bayesian inference of recent migration rates using multilocus genotypes*, *Genetics*, 163 (2003), pp. 1177–1191.
- [45] B. F. WRINGE, E. C. ANDERSON, N. W. JEFFERY, R. R. STANLEY, AND I. R. BRADBURY, *Development and evaluation of SNP panels for the detection of hybridization between wild and escaped Atlantic salmon (*Salmo salar*) in the western Atlantic*, *Canadian Journal of Fisheries and Aquatic Sciences*, 76 (2019), pp. 695–704.