# Population assignment from genotype likelihoods for low-coverage whole-genome sequencing data

Matthew DeSaix[1], Marina Rodriguez[1], Kristen Ruegg[1], and Eric Anderson[2]

[1]Colorado State University
[2]Southwest Fisheries Science Center

June 2, 2023

**Abstract**

Low-coverage whole genome sequencing (WGS) is increasingly used for the study of evolution and ecology in both model and non-model organisms; however, effective application of low-coverage WGS data requires the implementation of probabilistic frameworks to account for the uncertainties in genotype likelihood data. Here, we present a probabilistic framework for using genotype likelihood data for standard population assignment applications. Additionally, we derive the Fisher information for allele frequency from genotype likelihood data and use that to describe a novel metric, the effective sample size, which figures heavily in assignment accuracy. We make these developments available for application through WGSassign, an open-source software package that is computationally efficient for working with whole genome data. Using simulated and empirical data sets, we demonstrate the behavior of our assignment method across a range of population structures, sample sizes, and read depths. Through these results, we show that WGSassign can provide highly accurate assignment, even for samples with low average read depths ($< 0.01$X) and among weakly differentiated populations. Our simulation results highlight the importance of equalizing the effective sample sizes among source populations in order to achieve accurate population assignment with low-coverage WGS data. We further provide study design recommendations for population-assignment studies and discuss the broad utility of effective sample size for studies using low-coverage WGS data.

# Population assignment from genotype likelihoods for low-coverage whole-genome sequencing data

Matthew G. DeSaix[*,§], Marina D. Rodriguez[*], Kristen C. Ruegg[*], Eric C. Anderson[†,‡,*]

[*]Department of Biology, Colorado State University, Fort Collins, Colorado, USA. [†]Fisheries Ecology Division, Southwest Fisheries Science Center, National Marine Fisheries Service, NOAA, Santa Cruz, California, USA. [‡]Department of Fisheries, Wildlife, and Conservation Biology, Colorado State University, Fort Collins, Colorado, USA.

*Keywords*: Fisher information, genetic stock identification, next-generation sequencing, population genomics, statistical genetics

[§]*Corresponding Author*: mgdesaix@gmail.com

*Running Title*: Population assignment from genotype likelihoods

## Abstract

Low-coverage whole genome sequencing (WGS) is increasingly used for the study of evolution and ecology in both model and non-model organisms; however, effective application of low-coverage WGS data requires the implementation of probabilistic frameworks to account for the uncertainties in genotype likelihood data. Here, we present a probabilistic framework for using genotype likelihood data for standard population assignment applications. Additionally, we derive the Fisher information for allele frequency from genotype likelihood data and use that to describe a novel metric, the *effective sample size*, which figures heavily in assignment accuracy. We make these developments available for application through WGSassign, an open-source software package that is computationally efficient for working with whole genome data. Using simulated and empirical data sets, we demonstrate the behavior of our assignment method across a range of population structures, sample sizes, and read depths. Through these results, we show that WGSassign can provide highly accurate assignment, even for samples with low average read depths (< 0.01X) and among weakly differentiated populations. Our simulation results highlight the importance of equalizing the effective sample sizes among source populations in order to achieve accurate population assignment with low-coverage WGS data. We further provide study design recommendations for population-assignment studies and discuss the broad utility of effective sample size for studies using low-coverage WGS data.

## Introduction

In just a few years, next-generation sequencing (NGS) technologies have revolutionized the study of evolution and ecology in both model and non-model organisms, and have become established as standard tools in molecular ecology. In particular, whole genome sequencing (WGS) can provide sequence data from a large proportion of the genome and is increasing in use. While large-scale WGS projects can be prohibitively expensive at the necessary read depths for accurately calling individual genotypes, low-coverage WGS offers a cost-effective approach aimed at reducing the read depth per individual while retaining sufficient information for genomic analyses. However, since low-coverage WGS precludes the ability to call individual genotypes, probabilistic frameworks are used to account for the uncertainty in an individual's genotype (Nielsen *et al.* 2011; Buerkle & Gompert 2013). Extending common analyses in the field of molecular ecology to accommodate genotype uncertainty through the direct use of genotype likelihoods is a necessary advance for broadening the utility of low-coverage WGS.

The creation of probabilistic frameworks for allele frequency estimation, genotype calling, and single nucleotide polymorphism (SNP) calling have made low-coverage WGS practical for many applications (Nielsen *et al.* 2011, 2012; Kim *et al.* 2011). By first estimating the joint site frequency spectrum for individuals without calling individual genotypes, priors on allele frequency can improve the calling of individuals' genotypes and SNPs. Population genetics analyses have been further advanced through the development of methods that quantify genetic differentiation and investigate population structure with principal components analysis, while accounting for uncertain genotypes (Fumagalli *et al.* 2013). Similarly, accurate estimates of individual admixture proportions (Skotte *et al.* 2013) and pairwise relatedness (Korneliussen & Moltke 2015) can be obtained using genotype likelihoods. The widespread use of these methods is facilitated by software that is both user-friendly and computationally efficient (e.g. ANGSD (Korneliussen *et al.* 2014), ngsTools (Fumagalli *et al.* 2014), PCangsd (Meisner & Albrechtsen 2018)). However,

a fundamental analysis for molecular ecology yet to be developed for low-coverage WGS data is population assignment.

Population assignment methods are used to determine an individual's population of origin and have provided insight into ecological and evolutionary processes, such as dispersal, hybridization, and migration, as well as informed conservation and management decisions (Manel *et al.* 2005). The traditional assignment test uses an individual's multilocus genotype and the source populations' allele frequencies to calculate the likelihood of the genotype originating from each of the populations (Paetkau *et al.* 1995; Rannala & Mountain 1997). Using this framework, the recent increase in available markers (e.g., from RADseq approaches) has made possible highly accurate assignment of individuals among weakly differentiated populations by using subsets of informative loci for population structure (e.g. (DeSaix *et al.* 2019; Ruegg *et al.* 2014; Benestan *et al.* 2015)). The traditional assignment test is readily extended to analyses such as genetic stock identification (GSI), to determine the proportion of source populations in a mixture of individuals Smouse *et al.* (1990). To date, methods for performing assignment tests require known genotypes and have not been implemented to use genotype likelihoods.

Assignment tests are well suited for application with low-coverage WGS data, because they rely heavily on allele frequency estimates, for which a number of approaches are already developed. For accurate allele frequency estimation from low coverage WGS data, simulation studies have demonstrated that prioritizing larger sample sizes of individuals with lower sequencing depth is the most cost-effective strategy (Buerkle & Gompert 2013; Lou *et al.* 2021; Fumagalli 2013). Specific recommendations include aiming for individual sequencing depths of 1x (Buerkle & Gompert 2013) or having at least 10 individuals sequenced with a total per-population sequencing depth of at least 10x (Lou *et al.* 2021). The goal of these strategies is to maximize information for estimating allele frequencies given finite resources for sequencing depth and number of samples. Lower sequencing depth decreases the amount of information about population allele frequency, while using larger sample sizes increases the amount of information.

However, information is not directly quantified in these studies; rather comparison of known versus simulated allele frequencies were used to arrive at these general rules of thumb (Buerkle & Gompert 2013; Lou *et al.* 2021). The development of an information metric that accounts for read-depth variation across genotypes would provide a valuable method to quantify the thresholds of information needed for parameter estimation with low-coverage WGS data.

Here we present WGSassign, an open-source software package of population assignment tools for genotype likelihood data from low coverage WGS. The objectives of WGSassign are: 1) provide common assignment methods that use genotype likelihoods, instead of called genotypes, 2) evaluate the information available in low-read-depth sequencing data for allele frequency estimation, and 3) achieve computational efficiency for processing large numbers of samples with genome-wide data. WGSassign provides methods for individual assignment, estimation of mixture proportions, and leave-one-out cross-validation of samples of known origin. Additionally, it calculates a $z$-score metric that can indicate when samples originate from an unsampled source population. For the second objective, we calculate Fisher Information and determine the *effective sample size*—the number of samples with completely observed genotypes that would yield the same amount of statistical information for estimating allele frequency as the observed genotype likelihoods in a dataset. This calculation of effective sample size has broad utility for population genomics studies using low-coverage WGS.

We validate WGSassign and investigate its behavior with an extensive set of simulations and demonstrate its use on two empirical datasets. In the first, we apply WGSassign to weakly differentiated groups of yellow warblers (*Setophagia petechia*). In the second, we apply WGSassign to two well-differentiated Chinook salmon (*Oncorhynchus tshawytscha*) populations to demonstrate that when sufficient effective sample sizes of the source population are available, unknown individuals can be assigned accurately, even at extremely low read depths.

## Methods

WGSassign is written in Python 3 (`https://www.python.org/`) and requires the following modules: numpy (`https://numpy.org/`), cython (`https://cython.org/`), and scipy (`https://scipy.org/`). Detailed instructions for using WGSassign are available at `https://github.com/mgdesaix/WGSassign`.

*Population Assignment*

We assume that there are $K$ sampled source populations to which an individual can be assigned using data from $L$ biallelic loci in the genome. Let a diploid individual's genotype at locus $\ell$ ($1 \leq \ell \leq L$) be represented by $G_\ell \in \{0, 1, 2\}$, which counts the number of alleles matching the reference genome carried by the individual at locus $\ell$. Denote by $\theta_{k,\ell}$ the true—but typically unkown—frequency of the alternate allele at locus $\ell$ within source population $k$. Under the assumption of Hardy-Weinberg equilibrium, the probability of $G_\ell$, when the individual is from population $k$ is:

$$P(G_\ell|\theta_{k,\ell}) = \begin{cases} (1-\theta_{k,\ell})^2 & \text{if } G_\ell = 0 \\ 2(\theta_{k,\ell})(1-\theta_{k,l}) & \text{if } G_\ell = 1 \\ (\theta_{k,\ell})^2 & \text{if } G_\ell = 2. \end{cases} \quad (1)$$

With low-coverage sequencing data, $G_\ell$ is not observed with certainty. Rather, evidence about the unknown genotype is obtained from sequencing reads covering the locus. Let $R_\ell$ denote the sequencing read data from an individual at locus $\ell$. The evidence for the state of $G_\ell$ from the read data is summarized as the likelihood of the genotype given the read data, which is simply the probability of the read data given the genotype, considered as a function of the

genotype:

$$P(R_\ell|G_\ell) = \begin{cases} g_{\ell,0} & \text{for } G_\ell = 0 \\ g_{\ell,1} & \text{for } G_\ell = 1 \\ g_{\ell,2} & \text{for } G_\ell = 2. \end{cases} \tag{2}$$

Without loss of generality, we consider these likelihoods to be scaled so that they sum to one: $g_{\ell,0} + g_{\ell,1} + g_{\ell,2} = 1$. Such likelihoods are typically a function of the number of reads of each allele observed and the corresponding base quality scores, and they are computed during genotype calling by a variety of programs such as bcftools (Li *et al.* 2009; Li 2011), GATK (McKenna *et al.* 2010), and ANGSD (Korneliussen *et al.* 2014). An accessible review of the different models providing genotype likelihoods is found in (Lou *et al.* 2021).

To do population assignment from the read data of an individual (rather than from directly observed genotypes) requires, for each locus, $\ell$, the likelihood that the individual came from a source population $k$, say, given the individual's read data. This is simply the probability of the read data from the individual given that the individual came from source population $k$, with allele frequencies $\theta_{k,\ell}$. Thus, we require $P(R_\ell|\theta_{k,\ell})$, which can be calculated from (1) and (2) using the law of total probability:

$$\begin{aligned} P(R_\ell|\theta_{k,\ell}) &= \sum_{G_\ell=0}^{2} P(R_\ell|G_\ell)P(G_\ell|\theta_{k,\ell}) \\ &= g_{\ell,0}(1-\theta_{k,\ell})^2 + g_{\ell,1}2(\theta_{k,\ell})(1-\theta_{k,\ell}) + \\ &\quad g_{\ell,2}(\theta_{k,\ell})^2. \end{aligned} \tag{3}$$

If the $L$ loci in the genome are not in linkage disequilibrium (LD), and are hence independent of one another, within source populations, then the likelihood of source population $k$ given $R$, the read sequencing data across the entire genome, is simply the product over loci.

$$P(R|\theta_k) = \prod_{\ell=1}^{L} P(R_\ell|\theta_{k,\ell}), \tag{4}$$

where $\theta_k$ denotes the set of all $L$ allele frequencies in population $k$. Of course, with lcWGS some variants may be near one another and will then likely be in LD. In such a case (4) is not correct, but, rather, is a composite-likelihood approximation to the true likelihood (which is largely intractable). Composite likelihood estimators often produce unbiased results, but, because they do not take account of the dependence of different variables in the likelihood, they typically underestimate the uncertainty in the estimates (Larribe & Fearnhead 2011). We discuss this later. For each individual of unknown origin, this likelihood can be computed for each source population, $k$, and the relative values of those likelihoods gives the evidence that the individual came from each of the source populations. If the prior probability $\pi_k$ that an individual came from source population $k$ is available for $k \in \{1, \ldots, K\}$, then the likelihoods can be used to compute the posterior probability that the individual came from each of the source populations:

$$P(Z = k|R, \theta_1, \ldots, \theta_K, \pi_1, \ldots, \pi_K) = \frac{\pi_k P(R|\theta_k)}{\sum_{i=1}^{K} \pi_k P(R|\theta_k)}, \tag{5}$$

where $Z$ is a random variable indicating the origin of the individual.

In practice, the allele frequencies in each source population are not known with certainty. Accordingly, these frequencies must be estimated from sequencing read data from individuals known to be from the source populations (these are often referred to as "reference samples.") We estimate these by maximum likelihood. The probability of the read data, $R_\ell^{(i)}$, from the $i^{\text{th}}$ reference sample, given that it came from source population $k$, is, following (3),

$$\begin{aligned} &P(R_\ell^{(i)}|\theta_{k,\ell}) = \\ &g_{\ell,0}^{(i)}(1-\theta_{k,\ell})^2 + g_{\ell,1}^{(i)}2(\theta_{k,\ell})(1-\theta_{k,\ell}) + g_{\ell,2}^{(i)}(\theta_{k,\ell})^2, \end{aligned} \tag{6}$$

where the genotype likelihoods are now adorned with a superscript $^{(i)}$ to denote they are for the $i^{\text{th}}$ reference sample. Assuming the samples from source population $k$ are not related, the log-likelihood for $\theta_{k,\ell}$ given the read data from all $n_k$ reference samples from population $k$ is:

$$L(\theta_{k,\ell}) = \sum_{i=1}^{n_k} \log P(R_\ell^{(i)}|\theta_{k,\ell}) \tag{7}$$

In our implementation, we first use the Expectation-Maximization algorithm (Dempster *et al.* 1977) described in the supplement to Meisner & Albrechtsen (2018) to obtain the maximum likelihood estimates (MLEs) of the population allele frequencies, $\hat{\theta}_{k,\ell}$, from the reference samples. Then, when calculating $P(R|\theta_k)$ we substitute $\tilde{\theta}_{k,\ell}$ for $\theta_{k,\ell}$, calculated as follows:

$$\tilde{\theta}_{k,\ell} = \begin{cases} \hat{\theta}_{k,\ell} & \text{if } \hat{\theta}_{k,\ell} > 0 \\ \frac{1}{2(n_k+1)} & \text{if } \hat{\theta}_{k,\ell} = 0, \\ 1 - \frac{1}{2(n_k+1)} & \text{if } \hat{\theta}_{k,\ell} = 1, \end{cases} \tag{8}$$

where, again, $n_k$ is the number of reference samples from source population $k$. This provides a correction for cases in which the the allele exists in a source population, but was not detected in the reference samples from that population—effectively, it adds one more individual to the sample that carries one copy of the allele not previously seen in that reference population.

As should be clear from the preceding development, the accuracy of population assignment depends, at least in part, on the accuracy of the estimates of the allele frequencies from each source population. The following section develops theory (which is then implemented in WGSassign) that provides the user with a measure of allele frequency estimate accuracy, calculated from the genotype likelihoods in the reference samples, that takes account of both sample size and read depth.

*Fisher Information and Effective Sample Size*

[Figure 1 about here.]

The likelihood that an individual originated from a source population depends on the read data (summarized as a genotype likelihood) and also on the estimated allele frequencies of the source populations. In turn, the accuracy of the estimated allele frequency depends on the number of individuals in the reference sample from the source population and read depth of

those individuals (Buerkle & Gompert 2013; Lou *et al.* 2021; Fumagalli 2013). Fewer individuals sampled and lower sequencing depth will result in less information in the data regarding allele frequency.

As noted above, estimates of the allele frequencies are made by maximum likelihood using the sequencing data on the reference samples from each source population. Fisher information is a statistical metric that quantifies the amount of information in a sample for estimating an unknown, continuous parameter (Fisher 1922). It measures the curvature of the log-likelihood function, and is inversely related to the variance. In visual terms, a sharply peaked log-likelihood curve (i.e., one with greater curvature) for a parameter indicates greater certainty in the estimated parameter (and, also higher Fisher information) than a flatter log-likelihood function. Formally, the curvature is measured by the negative second derivative of the log-likelihood function. The *observed* Fisher information for allele frequency is that negative second derivative evaluated at the MLE

$$I_o(\theta_{k,\ell}) = -\frac{\partial^2 L(\theta_{k,\ell})}{\partial \theta_{k,\ell}^2}\bigg|_{\theta_{k,\ell}=\hat{\theta}_{k,\ell}}. \tag{9}$$

Appendix A shows how $I_o^{(i)}(\theta_{k,\ell})$, the observed Fisher information for $\theta_{k,\ell}$ in the reads from a single individual, $i$, is found to be:

$$\begin{aligned} I_o^{(i)}(\theta_{k,\ell}) = \Bigg[ & \frac{2(g_{\ell,0}^{(i)} + g_{\ell,2}^{(i)} - 2g_{\ell,1}^{(i)})}{g_{\ell,0}^{(i)}(1-\hat{\theta}_{k,\ell})^2 + g_{\ell,1}^{(i)} 2\hat{\theta}_{k,\ell}(1-\theta_{k,\ell}) + g_{\ell,2}^{(i)}\hat{\theta}_{k,\ell}^2} \\ & + \Bigg( \frac{2\hat{\theta}_{k,\ell}(g_{\ell,0}^{(i)} + g_{\ell,2}^{(i)} - 2g_{\ell,1}^{(i)}) + 2(g_{\ell,1}^{(i)} - g_{\ell,0}^{(i)})}{g_{\ell,0}^{(i)}(1-\hat{\theta}_{k,\ell})^2 + g_{\ell,1}^{(i)} 2\hat{\theta}_{k,\ell}(1-\hat{\theta}_{k,\ell}) + g_{i,2}^{(i)}\hat{\theta}_{k,\ell}^2} \Bigg)^2 \Bigg]. \end{aligned} \tag{10}$$

The observed Fisher information from all $n_k$ reference samples is then simply, $I_o(\theta_{k,\ell}) = \sum_{i=1}^{n_k} I_o^{(i)}(\theta_{k,\ell})$.

To derive $\tilde{n}_l$, our effective sample size metric for locus $\ell$, we compare this observed Fisher information to the *expected* Fisher information that would be obtained from $2\tilde{n}_\ell$ gene copies with allelic type directly observed (Appendix A) from a population in which the true allele frequency is $\hat{\theta}_{k,\ell}$:

$$I_e(\theta_{k,\ell}) = \frac{2\tilde{n}_\ell}{\hat{\theta}_{k,\ell}(1-\hat{\theta}_{k,\ell})}. \tag{11}$$

Equating $I_o(\theta_{k,\ell})$ to $I_e(\theta_{k,\ell})$ and solving for $\tilde{n}_\ell$ yields

$$\tilde{n}_\ell = \frac{1}{2} I_o(\theta_{k,\ell}) \times \hat{\theta}_{k,\ell}(1 - \hat{\theta}_{k,\ell}). \tag{12}$$

This is the number of diploid individuals with perfectly observed genotypes that provides the same information (and hence accuracy) for estimating $\theta_{k,\ell}$ as is available from the sequencing read data from the $n_k$ reference samples from source population $k$. We term $\tilde{n}_\ell$, calculated as above, the *effective sample size* of the read data from the reference samples of source population $k$ at locus $\ell$. In practice, to avoid issues of non-differentiability on the boundaries of the space (i.e., at $\theta = 0$ or $\theta = 1$) we calculate $\tilde{n}_\ell$ using $\tilde{\theta}_{k,\ell}$. The effective sample size for an individual is then derived by taking the mean of $\tilde{n}_l$ across all loci, $\tilde{n} = \frac{1}{L} \sum_{l=1}^{L} \tilde{n}_l$.

Fisher information and effective sample size calculated in this way are useful summaries for understanding the trade-offs between sequencing more individuals at lower depth versus fewer individuals at higher depth, at least as it pertains to accurately estimating allele frequencies. In the context of population assignment, the effective sample size, in particular, provides an accessible metric for how good (or bad) the source-population allele frequencies can be expected to be. As we will see later, Fisher information also provides a valuable way to standardize the effective sample size of the reference samples from each population—an important consideration when using WGSAssign. A useful statistic for accomplishing this is the individual-specific average effective size for individual $i$:

$$\tilde{n}^{(i)} = \frac{1}{L} \sum_{\ell=1}^{L} \frac{1}{2} I_o^{(i)}(\theta_{k,\ell}) \times \hat{\theta}_{k,\ell}(1 - \hat{\theta}_{k,\ell}), \tag{13}$$

where $I_o^{(i)}(\theta_{k,\ell})$ is the contribution to the observed Fisher information of the reads from individual $i$:

$$I_o^{(i)}(\theta_{k,\ell}) = -\frac{\partial^2 \log P(R_\ell^{(i)} | \theta_{k,\ell})}{\partial \theta_{k,\ell}^2}\bigg|_{\theta_{k,\ell} = \hat{\theta}_{k,\ell}}.$$

$\tilde{n}^{(i)}$ ranges between 0 and 1.

We also implement a $z$-score calculation for determining whether an individual's genotype is unlikely to have come from one of the $K$ source populations, but rather, from an unsampled population. The full derivation of the method is shown in Appendix B. In short, we determine the expected distribution of log probabilities of an individual's genotype likelihood data arising from a population (given the individual's allele counts across loci and the population's allele frequencies), using a central limit theorem approximation. The $z$-score is then calculated by subtracting the mean expected likelihood from the observed likelihood and dividing the difference by the standard deviation of the expected likelihoods. Given that the actual distribution of the $z$-score is likely to deviate from a standard normal distribution, we further standardize the observed $z$-score by the $z$-scores of the reference individuals from the source populations. Individuals truly from an assigned population are expected to have $z$-scores within several standard deviations of the normal distribution, while individuals from an unsampled but differentiated population are expected to have $z$-scores that fall below the expected range of a standard unit normal random variate.

*Simulations to illustrate the effective sample size*

We used the R programming language to run simulations that illustrate how Fisher information and effective sample size vary across a range of simulated read depths and true allele frequencies. Our simulations assumed a sample size of 100 diploid individuals and a single biallelic locus, with allelic types within individuals being independent of each other.

For each individual, we simulated read depth from a Poisson distribution with mean $D_{\text{ave}}$ and allelic types upon each read by sampling from the two gene copies within the individual with equal probability and switching the allelic type with probability 0.01 for each read to simulate sequencing errors. Genotype likelihoods from the reads were calculated according to the simulation model. We calculated the maximum likelihood estimate (MLE) for $\theta$ from the genotype data as the observed proportion of alleles, and for the sequencing read data, we used

the EM algorithm to compute the MLE. Using these estimates, we then computed the observed information from the genotypes and from the genotype likelihoods.

To determine the effective sample size, we calculated the expected information for observed genotypes, assuming the true value of $\theta$ was the MLE from genotype likelihoods and then used (12).

We ran these simulations across values of $D_{\text{ave}} \in \{0.1, 0.5, 1, 2, 3, 4, 5, 7, 10, 15, 20, 30, 50\}$ and values of $\theta \in \{0.01, 0.05, 0.10, \ldots, 0.90, 0.95, 0.99\}$, simulating 50 replicate samples for each combination.

*Genetic Simulations*

To demonstrate the efficacy of WGSAssign in performing population assignment for a range of samples, read depths and genetic differentiation among populations we simulated a series of genetic datasets using msprime (Kelleher *et al.* 2016). In the first simulation, we implemented two-population island models with an effective population size of 1000 individuals in each population. We simulated ancestry for a genomic sequence of $10^8$ bases with a recombination rate of $10^{-8}$ and a mutation rate of $10^{-7}$. To vary the genetic differentiation between populations, we varied the lineage migration rate parameter between 0.0005 and 0.05 in 20 equal increments. From both populations we sampled 10, 50, 100, or 500 individuals. Pairwise $F_{\text{ST}}$ was calculated between the two populations using the sampled individuals and the genetic variants were output in variant call format.

Genotype likelihoods were produced with vcfgl (`https://github.com/isinaltinkaya/vcfgl`) based on mean read depths of 0.1X, 0.5X, 1X, 5X, 10X, or 50X. For each of the 480 parameter combinations (10 migration rates, 4 sample sizes, and 6 read depths) we simulated 10 replicates, for a total of 2,400 simulated datasets. We used bcftools (Li *et al.* 2009; Li 2011) to remove any SNPs with a minor allele frequency less than 0.05. We converted the data to Beagle file format with custom scripts, and used these data as input into WGSAssign.

To determine the influence of sampling design (i.e. number of samples in a source population and their read depths), as well as amount of genetic differentiation, on assignment accuracy, we calculated the effective sample size and leave-one-out (LOO) assignment accuracy for each population. In WGSAssign, LOO is performed by iteratively removing an individual of known origin from its source population, calculating allele frequencies within the source populations using the remaining individuals, and then calculating the likelihood that the removed individuals originated from each of the different source populations. The LOO method is widely used to avoid the bias that arises from using training data that also includes data being tested. The assigned population was determined by maximum likelihood. We also measured the run time for the calculation of allele frequency and effective sample size, as well as the LOO calculation.

In the second simulation, we assessed the influence on assignment accuracy of using unequal effective sample sizes of source populations. In population assignment applications, unequal sample sizes in different populations will result in different levels of precision in the allele frequency estimation. We implemented two-population island models as in the previous simulation, but included all sample combinations of 10, 50, and 100 individuals for the two populations. We also used 10 equal increments of migration rates from 0.005 to 0.05, and simulated read depths of 1X, 5X, and 10X. We then filtered by a minor allele frequency of 0.05 and randomly selected 100,000 SNPs to be used for the effective sample size calculation and LOO assignment.

In the third simulation, we assessed the performance of the WGSAssign $z$-score metric for determining whether an individual of unknown origin being assigned to a population is actually from an unsampled population. We implemented a three-population stepping-stone model with 20, 60, or 110 individuals using msprime. Individuals had simulated mean read depths of 1X or 5X, and we customized vcfgl (`https://github.com/isinaltinkaya/vcfgl`) to output allele counts for the major and minor alleles. We used populations 1 and 2 in the stepping-stone model as reference populations and calculated the reference $z$-scores using WGSAssign from all but 10 the individuals in these two populations. We assigned 10 individuals from population 3 and

10 from population 2 to the reference populations (1 and 2) using WGSassign. We calculated the $z$-scores of these individuals' assignments to demonstrate the behavior of the $z$-score metric for correctly assigned individuals (i.e., the individuals from population 2 that were assigned to population 2) versus individuals from an unsampled population (i.e., the individuals from population 3 that were assigned to population 2).

*Application to Empirical Data*

We used WGSassign on data from yellow warblers to test its accuracy when applied to individuals from a species exhibiting isolation by distance (Bay *et al.* 2021; Gibbs *et al.* 2000). Previous work on yellow warblers has found weak differentiation between populations, with pairwise $F_{ST}$ values on the order of 0.01 or less (Gibbs *et al.* 2000). Blood samples from 105 individuals was collected via brachial venipuncture in the years 2020 and 2021. These served as reference samples from 3 populations—North, Central, and South—previously described in Bay *et al.* (2021) and Gibbs *et al.* (2000). We extracted DNA from blood using the manufacturer's protocol for Qiagen DNEasy Blood and Tissue Kits. Whole genome sequencing libraries were prepared following modifications of Illumina's Nextera Library Preparation protocol (Schweizer & DeSaix 2023) and sequenced on a HiSeq 4000 at Novogene Corporation Inc., with a target sequencing depth of 2X per individual.

Sequences were trimmed with TrimGalore version 0.6.5 (https://github.com/FelixKrueger/TrimGalore) and mapped to the NCBI yellow warbler reference genome (Sayers *et al.* 2022) (accession number JANCRA010000000) using the Burrows-Wheeler Aligner software version 0.7.17 (Li & Durbin 2009). After mapping, the resulting SAM files were sorted, converted to BAM files, and indexed using Samtools version 1.9 (Li *et al.* 2009). We used MarkDuplicates from GATK version 4.1.4.0 (McKenna *et al.* 2010) to mark read duplicates and clipped overlapping reads with the clipOverlap function from bamUtil (`https://genome.sph.umich.edu/wiki/BamUtil:_clipOverlap`). To reduce sequencing depth variation, we used the DownsampleSam function from GATK to down-

sample reads from BAM files with greater than 2X coverage, to 2X coverage. To identify genetic markers from low-coverage WGS data, we used stringent filtering options in ANGSD version 0.9.40 (Korneliussen *et al.* 2014). We retained reads with a mapping quality of at least 30 and base quality of at least 33. We retained SNPs that had read data in at least 50% of individuals and a minor allele frequency greater than 0.05. The filtered variants were output as genotype likelihoods and stored in a Beagle-formatted file.

We implemented principal components analysis (PCA) to ensure reference samples from each of our source populations actually showed geographic signatures of clustering in the PCA. Genetic differentiation among the breeding populations was calculated by creating site allele frequency files for each breeding population and calculating $F_{ST}$ in ANGSD (Korneliussen *et al.* 2014). In order to assess our ability to accurately assign individuals of unknown origin to breeding populations, we determined the accuracy of assignment of the known breeding origin individuals using WGSAssign's leave-one-out approach.

For the second empirical dataset, we applied WGSAssign to previously published data from Chinook salmon (Thompson *et al.* 2020) to assess its utility in situations with low to extremely low read depth and poor-quality DNA. For this scenario, we entertained the task of assigning Chinook salmon to either the Klamath River basin, or the Sacramento Basin. These populations are quite distinct, with pairwise $F_{ST}$ values between the basins on the order of 0.1. So, it should be quite easy to distinguish fish from the two basins. However, in whole genome sequencing data from Thompson *et al.* (2020) there were several fish from rivers in the Klamath basin collected from carcasses with low read depth. These fish were excluded from most analyses in Thompson *et al.* (2020) because they did not reliably cluster with other fish from their populations on a PCA; however we evaluate here if their basin of origin can be recovered using WGSAssign. Additionally, through downsampling of reads from the BAM files we investigate if average read depths as low as 0.001X in the sample being assigned can deliver accurate assignments.

We included fish from the closely related Feather River Spring, Feather River Fall, San Joaquin Fall, and Coleman Late Fall collections as members of the Sacramento River source population, while fish from the closely related Salmon River Fall and Spring and Trinity River Fall and Spring collections constitute samples from the Klamath River source population. With 64 fish in each source population, we removed the 12 fish from each that had the fewest sequencing reads to serve as our 24 "unknown" fish to be assigned to the populations. The remaining 52 in each population served as the reference samples.

The genotype likelihoods for the reference sample were in a VCF file produced by GATK. This was filtered using bcftools (Danecek *et al.* 2021) to retain only biallelic SNPs with a minor allele frequency > 0.05 which were missing data in fewer than 30% of the samples. Additionally, data from chromosome 28, which holds a region strongly differentiated between spring-run and fall-run Chinook salmon (Thompson *et al.* 2020) was excluded. These genotype likelihoods were stored in a Beagle-formatted file using a custom script.

The data for the test samples were extracted from BAM files. We used `samtools stats` (Li *et al.* 2009) to determine the average read depth in each BAM and used that number with `samtools view` to downsample each BAM five times with five separate seeds to average read depth levels of 0.001X, 0.005X, 0.01X, 0.05X, 0.1X, 0.5X, and 1.0X, when those read depths were lower than the full read depth of the file. Genotype likelihoods for the 24 individuals were then called with ANGSD v0.940 (Korneliussen *et al.* 2014) using the `-sites` options to call only the sites found in the Beagle-formatted file of the reference samples. After genotype likelihood estimation in the test samples, the Beagle file of reference samples was filtered to include only the sites output by ANGSD. The resulting Beagle files were then passed to WGSAssign to compute the likelihood of population origin for each of the test fish, and the results were plotted using R version 4.0 (R Core Team 2022).

## Results

*Effective Sample Size Simulations*

As expected, observed Fisher information for allele frequency from sequencing read data increases as the average sequencing depth increases, reaching a limit at the observed information from fully observed genotypes. The absolute value of the observed Fisher information varies widely over the different allele frequencies, however the relative values of information from genotypes and from sequencing reads varies less, and the effective sample size is largely consistent across the range of minor allele frequencies from 0.05 to 0.5, showing the effective sample size to be a useful metric. Fisher information and effective sample size are shown for three representative values of $\theta$ (0.05, 0.3, and 0.5) in Figure 1. The flattening of the curves for observed information from sequencing data as the average read depth increases indicates the diminishing returns of additional sequencing depth versus additional samples, for estimating allele frequencies that has been noted previously (Buerkle & Gompert 2013; Lou *et al.* 2021; Fumagalli 2013).

*Genetic Simulations*

In the first simulation, genetic differentiation between the sampled individuals from the two populations ranged from -0.003 - 0.13 $F_{ST}$. Across all read depths within each category of number of samples (10, 50, 100, 500), assignment accuracy increased with genetic differentiation, and generally high assignment accuracy was achieved even with low genetic differentiation (Figure 2). Accuracy above 90% was reached for all simulations within the 500 samples category with $F_{ST}$ > 0.004, 100 samples category with $F_{ST}$ > 0.006, 50 samples category with $F_{ST}$ > 0.015, and the 10 samples category with $F_{ST}$ > 0.043. When excluding simulations with populations with the lowest effective sample sizes (< 0.1 individuals), high assignment accuracy was reached for all simulations at $F_{ST}$ > 0.015 (Figure 2). Within each sample size category, increasing average read depth, and therefore effective sample size, resulted in higher assignment accuracy, especially when populations had weak genetic differentiation (Figure 2).

[Figure 2 about here.]

Runtime for the simultaneous calculation of Fisher information, effective sample size, and allele frequency for populations in WGSAssign was fast. With 2 populations and 100,000 loci being analyzed in parallel with 20 threads, runtime was less than 10 seconds for populations with 100 samples or less, and between 15 and 30 seconds for populations with 500 samples. Leave-one-out assignment requires population allele frequency to be recalculated for each individual in the population, and time required for that re-calculation increases linearly with sample size. Accordingly, runtime for LOO cross-validation is expected to increase quadratically with increasing number of samples per population, and we observe this: for 100 samples for the two populations at 1X mean individual read depth LOO assignment had a mean runtime of 51 seconds and for 500 samples run time was 1,743 seconds. Run times also increase with lower read depth due to the increase in iterations needed in the expectation-maximization algorithm for allele frequency calculation used from PCangsd (Meisner & Albrechtsen 2018).

When $F_{ST}$ is greater than 0.01, effective sample sizes as low as approximately 3 individuals achieve assignment accuracy of greater than 90% (Figure 3). Examining simulations with weak genetic differentiation ($0.005 < F_{ST} < 0.01$), shows that a minimum effective sample size of 10 individuals is needed for consistently high assignment accuracy (Figure 3). At the weakest genetic differentiation of $F_{ST} < 0.005$, consistently high assignment accuracy is not necessarily achieved across all simulations, but a minimum effective sample size of 100 individuals is needed for an assignment accuracy of greater than 80%.

[Figure 3 about here.]

*Assignment bias due to unequal sample sizes*

Our simulation results for unequal sample sizes demonstrate that high assignment bias occurs when populations have different numbers of samples (Figure 4). When populations have the

same number of samples, with the same average read depths, assignment accuracy overall increases with genetic differentiation and there is no evidence of bias, with one population having higher accuracy than another population. However, when populations have unequal sample sizes, individuals from the less-sampled population tend to be assigned to the more-sampled population, even when genetic differentiation is higher ($F_{ST} > 0.01$). This bias is exacerbated when effective sample size is lower (i.e. the populations have lower read depths).

[Figure 4 about here.]

*Determining an individual's origin from an unsampled population*

At higher genetic differentiation ($F_{ST} > 0.1$), samples can readily be identified as coming from an unsampled population using the $z$-score metric in WGSAssign (Figure 5. At such high differentiation, individuals from an unsampled population tend to have $z$-scores less than 3 compared to individuals correctly assigned to a population having $z$-scores in $(-3, 3)$, as expected of a standard unit normal. With weaker genetic differentiation ($F_{ST} < 0.1$), sample size and read depth have a more noticeable effect on the behavior of the $z$-score metric (Figure 5). Generally, higher source sample sizes and read depths allow individuals from unsampled populations to be distinctively identified from individuals that are truly from a source population.

[Figure 5 about here.]

*Application to Empirical Data*

Yellow warbler reference samples were accurately assigned to either the North, Central, or East populations using leave-one-out self-assignment. All 35 reference samples from both the North and East populations were assigned with 100% accuracy, and of the 35 birds from the Central population, 34 were correctly assigned.

Chinook salmon were accurately assigned to either the Sacramento or Klamath river basins

even at read depths as low as 0.001X (Figure 6). All 12 test samples from the Sacramento river were correctly assigned at all read depth levels, and, of the 12 Klamath test fish, 11 were correctly assigned at all read depth levels, while one was correctly assigned at all read depth levels except for one of the five replicates at read depth 0.001X. The four samples with lowest full read depth (the four at the bottom of Figure 6) have log-likelihood ratios that are noticeably smaller than those of the remaining 20 fish at all downsampled read depth levels, possibly indicating that, in addition to being samples with low depth, they might also have yielded very poor quality DNA.

[Figure 6 about here.]

## Discussion

Here, we present WGSAssign and demonstrate its utility for population assignment with low-coverage WGS data. Our results, from both simulated and empirical data, show that low-coverage WGS data can be used to achieve high assignment accuracy even among weakly differentiated populations ($F_{ST} < 0.01$). We show that balancing effective sample size among populations is essential for avoiding assignment bias due to variation in the precision of allele frequency estimation for different populations. Effective sample size can also be used to guide decisions in study design for choosing the number of samples and sequencing depth in a given population. The ability to perform population assignment on large numbers of individuals, cost-effectively sequenced at low-coverage across the whole genome, further expands the utility of low-coverage WGS for population and conservation genomics.

### *Performance of WGSAssign and implications for population-assignment studies*

Our implementation of WGSAssign allows users to perform population-assignment analyses from genotype likelihood data. Features of WGSAssign include standard and leave-one-out (LOO) population assignment, as well as calculations of effective sample sizes (of both individuals and populations) and a $z$-score metric for determining whether an individual is from an unsampled population. Importantly, as implemented, these analyses can be parallelized across loci, which allows for fast computation of data produced from low-coverage WGS, even for computationally intensive applications such as LOO assignment. Studies of wild populations are typically limited in the number of samples available for sequencing, where 50 may be a large number of samples for a given population. With such a sample size, leave-one-out assignment at a standard low-coverage read depth of 1X could be expected to have a runtime on the order of minutes for multiple populations and a million loci.

Implicit in standard population assignment tests is that there will always be a population with a maximum likelihood of assignment, even if the individual does not originate from any

of the reference populations. To address this issue, we developed a $z$-score metric for testing whether an individual could be from an unsampled population. The $z$-score is based on the individual's observed likelihood of assignment in relation to the expected likelihood from a hypothetical individual from the same population with the same allele count data as the individual being tested. The $z$-score metric functions as expected at higher genetic differentiation ($F_{ST}$ > 0.05) and with larger source populations by distinguishing the majority of individuals incorrectly assigned as having much lower $z$-scores (outside the 90% expected mass of the distribution of $z$-scores) than correctly assigned individuals. We recommend that any studies that may have incomplete sampling coverage of all genetically distinct populations test for correct assignment with the $z$-score metric. However, since this metric is limited by sample size and genetic differentiation, a robust approach toward using it would involve, first, observing the metric's behavior by testing it upon individuals of known origin, calculating $z$-scores both for the population they are from and the other populations.

For high assignment accuracy, source populations need to have sufficient effective sample sizes in relation to genetic differentiation among the populations. However, individual samples being assigned can have extremely low read depth for accurate assignment. Our results from downsampled Chinook salmon data showed that individuals were still correctly assigned when individual samples had average read depths as low as 0.001X. This has powerful implications for population assignment studies, especially those that are conducted at a large scale. For example, in the mid-2000's an arduous, international, multi-laboratory study was undertaken to standardize a DNA database of 13 microsatellite loci for genetic stock identification of Chinook salmon at a coast-wide scale (Seeb *et al.* 2007). With today's sequencing power, a low-coverage WGS approach could provide a cost-effective method for creating a reference baseline of known populations without the need for extensive standardization of genetic makers. Fish of unknown origin could be sequenced at very low read depth, and still be accurately assigned to populations from the reference baseline.

A potential benefit of low-coverage WGS over other sequence data for population assignment, is that low-coverage WGS provides more markers for assignment to weakly differentiated populations. Population assignment studies with RADseq data have commonly used SNP filtering methods for selecting the most informative loci for assignment to weakly differentiated populations (DeSaix *et al.* 2019; Ruegg *et al.* 2014; Benestan *et al.* 2015). Further identifying a subset of informative loci (e.g. < 200) can be cost-effective for genotyping large numbers of individuals for the purpose of assignment (Ruegg *et al.* 2014; Larison *et al.* 2021). However, our results highlight that high assignment accuracy is possible with low-coverage WGS data without the need for extensive analysis to determine the most informative loci. For example, high assignment accuracy was obtained with Yellow Warbler samples from weakly differentiated populations using 5,301,626 sites.

Furthermore, DNA quantity and quality requirements for RAD-seq methods—and even some chip-based genotyping methods—can be more stringent than they are for low-coverage whole genome sequencing. For example, reliable WGS data can be obtained from the tiny quantities of DNA adhering to the tip of a feather (Schweizer & DeSaix 2023), which is not possible with RAD-seq methods. Thus, being able to perform population assignment from low coverage whole genome sequencing data considerably expands the types of tissues available for sampling. And finally, using genotype data that is restricted to loci that are purposely biased toward detecting population structure (e.g. a SNP chip or hybridization-capture panel) limits the extent of analyses those data can be appropriately used for. Low-coverage WGS provides genome-wide data useful for population assignment in weakly differentiated populations, but it is also useful for demographic modeling, inference of population differentiation, detection of selection, and association studies (to name a few) because it has not been previously ascertained, and hence, biased.

*Accounting for population sample size and read depth with effective sample size*

Our development of the effective sample size metric provides a powerful tool for population genomics studies using low-coverage WGS data. Previous studies have provided recommendations for the number of individuals and sequencing depth required to accurately estimate allele frequencies with low-coverage WGS data (Buerkle & Gompert 2013; Lou *et al.* 2021; Fumagalli 2013). Effective sample size provides a metric to quantify these recommendations and determine the precision of allele frequency estimation needed for different applications. For example, the recommendation of (Lou *et al.* 2021) of at least 10 individuals with 1X average sequencing depth for allele frequency estimation can be quantified as an effective sample size of 2.3 individuals in the simulations from this study (Figure 7). For assignment to populations with moderate to strong differentiation ($F_{ST} > 0.01$), population effective sample sizes of at least 2.3 individuals are sufficient for achieving consistently high assignment accuracy (Figure 3). However, at weaker genetic differentiation among populations, effective sample size needs to be increased for accurate assignment. Furthermore, for similar levels of effective sample size, populations with 10 samples tend to perform worse than populations with more samples. These results suggest that sequencing more individuals at lower read depths can be a more effective study-design strategy than sequencing fewer individuals at higher read depths. One reason that using more individuals for source populations may improve assignment accuracy is that it increases the likelihood of detecting low-frequency alleles.

[Figure 7 about here.]

Effective sample size can facilitate population-assignment study design by determining target numbers of individuals and average read depth for source populations. Our results show how effective sample size quantifies different study design options. For example, in our simulations a population with 10 samples with mean read depths of 1X had a mean effective sample size of 2.3 individuals. Increasing the total read depth of the population from 10X to 50X could

be done by increasing the sequencing depth of the 10 individuals to 5X or increasing the sampled number of individuals to 50 and keeping the mean individual sequencing depth at 1X. The simulation results show that increasing the sequencing depth produces an effective sample size of 7.2 individuals, while increasing sample size results in an effective sample size of 17.1 individuals (Figure 7). Quantifying the amount of information gain for different study designs can inform researchers on how to more efficiently allocate resources for sequencing efforts.

Our simulation results show that disproportionate effective sample sizes among source populations can result in biased assignment of individuals to the populations with the highest effective sample sizes. We recommend that population assignment studies use the LOO assignment in WGSAssign to determine if biased assignment is occurring. If all individuals across populations have similar average read depths, then subsetting source populations to the same number of samples for allele frequency calculation should remove this bias. However, different populations may tend to have higher or lower read depths, especially if different DNA sources are used, which will result in different effective sample sizes despite equal numbers of individuals. In this case, the individual effective sample size (Equation 13) output from WGSAssign can be used to determine how many individuals to remove from the populations with the highest effective sample sizes. Alternatively, individuals could be further downsampled to reduce their effective sample size, which would decrease the overall population's effective sample size. Studies using low-coverage WGS data for population assignment can explore these different strategies with WGSAssign to determine what is most effective for their datasets.

*Further improvements for population assignment*

Currently in our implementation of WGSAssign, the issue of only a single allele being observed in a population, and thereby producing a likelihood of 0, is avoided by correcting a population with a minor allele frequency of 0 at a given locus to $\frac{1}{2n+2}$, where $n$ is the number of individuals in the population. Essentially, this treats the locus as having a rare allele that would be observed

in a single copy if another individual was to be sampled. Another approach that could potentially improve performance would be to specify a formal prior for the allele frequencies in each population (Rannala & Mountain 1997). Additionally, using a prior that accounts for the *a priori* expectation that allele frequencies at a locus are expected to be similar between weakly differentiated populations (Falush *et al.* 2003; Pella & Masuda 2006) may further improve performance of population assignment. We expect that the parameters of these more complex prior distributions could be estimated in an empirical Bayes approach (Maritz 2018) from the $n$-dimensional site frequency spectrum (Mas-Sandoval *et al.* 2022).

*Conclusion*

Low-coverage WGS is increasingly becoming more practical as sequencing costs decline and library preparation protocols are optimized for a wide-range of study systems (Schweizer & DeSaix 2023; Therkildsen & Palumbi 2017). In this paper, we present the WGSassign software which expands the types of analyses that can be done from genotype likelihoods. We demonstrate with simulated and empirical data that highly accurate and computationally efficient population assignment can be performed, even with weakly differentiated populations. We provide the software as open-source to facilitate further improvements on our developments in the field of molecular ecology.

## Acknowledgements

This study was funded by a Cooperative Agreement with the Alaska Department of Fish and Game (23-011) and an NSF CAREER award (008933-00002) to KCR. This work utilized the Alpine high performance computing resource at the University of Colorado Boulder. Alpine is jointly funded by the University of Colorado Boulder, the University of Colorado Anschutz, Colorado State University, and the National Science Foundation (award 2201538). We thank Isin Altinkaya for providing in-depth suggestions to modify their vcfgl software necessary for our simulations. For data input and allele-frequency estimation, WGSassign borrows from the well-organized and open-source code of PCAngsd. We thank members of the Fueggo Lab group at Colorado State University for providing intellectual support and suggestions throughout the development of the ideas in this manuscript. We are grateful to Ingrid Spies for providing extensive feedback on an early draft of the manuscript. A substantial portion of this manuscript was completed while MGD and ECA were scientists-in-residence at the mobile High Altitude Venue for Ecological Analysis, Genetics, and Statistics, on location in Moab, Utah for five days in March 2023 and again in April 2023. This is contribution number mHAVEAGAS–001. We gratefully acknowledge the services and the kind staff at the Grand County Public Library of Moab.

## References


Bay RA, Karp DS, Saracco JF, Anderegg WR, Frishkoff LO, Wiedenfeld D, Smith TB, Ruegg K (2021) Genetic variation reveals individual-level climate tracking across the annual cycle of a migratory bird. *Ecology Letters*, **24**, 819–828.

Benestan L, Gosselin T, Perrier C, Sainte-Marie B, Rochette R, Bernatchez L (2015) RAD genotyping reveals fine-scale genetic structuring and provides powerful population assignment in a widely distributed marine species, the American Lobster (Homarus americanus). *Molecular ecology*, **24**, 3299–3315.

Buerkle AC, Gompert Z (2013) Population genomics based on low coverage sequencing: how low should we go? *Molecular ecology*, **22**, 3028–3035.

Cornuet JM, Piry S, Luikart G, Estoup A, Solignac M (1999) New methods employing multilocus genotypes to select or exclude populations as origins of individuals. *Genetics*, **153**, 1989–2000.

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, *et al.* (2021) Twelve years of SAMtools and BCFtools. *Gigascience*, **10**, giab008.

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, **39**, 1–38.

DeSaix MG, Bulluck LP, Eckert AJ, Viverette CB, Boves TJ, Reese JA, Tonra CM, Dyer RJ (2019) Population assignment reveals low migratory connectivity in a weakly structured songbird. *Molecular Ecology*, **28**, 2122–2135.

Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, **164**, 1567–1587.

Fisher RA (1922) On the mathematical foundations of theoretical statistics. *Philosophical transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character*, **222**, 309–368.

Fumagalli M (2013) Assessing the effect of sequencing depth and sample size in population genetics inferences. *PloS one*, **8**, e79667.

Fumagalli M, Vieira FG, Korneliussen TS, Linderoth T, Huerta-Sánchez E, Albrechtsen A, Nielsen R (2013) Quantifying population genetic differentiation from next-generation sequencing data. *Genetics*, **195**, 979–992.

Fumagalli M, Vieira FG, Linderoth T, Nielsen R (2014) ngsTools: methods for population genetics analyses from next-generation sequencing data. *Bioinformatics*, **30**, 1486–1487.

Gibbs HL, Dawson RJ, Hobson KA (2000) Limited differentiation in microsatellite DNA variation among northern populations of the yellow warbler: evidence for male-biased gene flow? *Molecular Ecology*, **9**, 2137–2147.

Kelleher J, Etheridge AM, McVean G (2016) Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS computational biology*, **12**, e1004842.

Kim SY, Lohmueller KE, Albrechtsen A, Li Y, Korneliussen T, Tian G, Grarup N, Jiang T, Andersen G, Witte D, *et al.* (2011) Estimation of allele frequency and association mapping using next-generation sequencing data. *BMC bioinformatics*, **12**, 1–16.

Korneliussen TS, Albrechtsen A, Nielsen R (2014) ANGSD: analysis of next generation sequencing data. *BMC bioinformatics*, **15**, 1–13.

Korneliussen TS, Moltke I (2015) NgsRelate: a software tool for estimating pairwise relatedness from next-generation sequencing data. *Bioinformatics*, **31**, 4009–4011.

Larison B, Lindsay AR, Bossu C, Sorenson MD, Kaplan JD, Evers DC, Paruk J, DaCosta JM, Smith TB, Ruegg K (2021) Leveraging genomics to understand threats to migratory birds. *Evolutionary applications*, **14**, 1646–1658.

Larribe F, Fearnhead P (2011) On composite likelihoods in statistical genetics. *Statistica Sinica*, 43–69.

Li H (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, **27**, 2987–2993.

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. *bioinformatics*, **25**, 1754–1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R (2009) The sequence alignment/map format and SAMtools. *bioinformatics*, **25**, 2078–2079.

Lou RN, Jacobs A, Wilder AP, Therkildsen NO (2021) A beginner's guide to low-coverage whole genome sequencing for population genomics. *Molecular Ecology*, **30**, 5966–5993.

Manel S, Gaggiotti OE, Waples RS (2005) Assignment methods: matching biological questions with appropriate techniques. *Trends in Ecology & Evolution*, **20**, 136–142.

Maritz JS (2018) *Empirical Bayes methods with applications*, CRC Press.

Mas-Sandoval A, Pope NS, Nielsen KN, Altinkaya I, Fumagalli M, Korneliussen TS (2022) Fast and accurate estimation of multidimensional site frequency spectra from low-coverage high-throughput sequencing data. *GigaScience*, **11**.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, *et al.* (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research*, **20**, 1297–1303.

Meisner J, Albrechtsen A (2018) Inferring population structure and admixture proportions in low-depth NGS data. *Genetics*, **210**, 719–731.

Nielsen R, Korneliussen T, Albrechtsen A, Li Y, Wang J (2012) SNP calling, genotype calling, and sample allele frequency estimation from new-generation sequencing data. *PLoS ONE*.

Nielsen R, Paul JS, Albrechtsen A, Song YS (2011) Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics*, **12**, 443–451.

Paetkau D, Calvert W, Stirling I, Strobeck C (1995) Microsatellite analysis of population structure in Canadian polar bears. *Molecular Ecology*, **4**, 347–354.

Pella J, Masuda M (2006) The gibbs and split merge sampler for population mixture analysis from genetic data with incomplete baselines. *Canadian Journal of Fisheries and Aquatic Sciences*, **63**, 576–596.

R Core Team (2022) *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Rannala B, Mountain JL (1997) Detecting immigration by using multilocus genotypes. *Proceedings of the National Academy of Sciences*, **94**, 9197–9201.

Ruegg KC, Anderson EC, Paxton KL, Apkenas V, Lao S, Siegel RB, DeSante DF, Moore F, Smith TB (2014) Mapping migration in a songbird using high-resolution genetic markers. *Molecular Ecology*, **23**, 5726–5739.

Sayers EW, Cavanaugh M, Clark K, Pruitt KD, Schoch CL, Sherry ST, Karsch-Mizrachi I (2022) GenBank. *Nucleic acids research*, **50**, D161.

Schweizer TM, DeSaix MG (2023) Cost-effective library preparation for whole genome sequencing with feather DNA. *Conservation Genetics Resources*, 1–8.

Seeb L, Antonovich A, Banks MA, Beacham T, Bellinger M, Blankenship S, Campbell M, Decovich N, Garza J, Guthrie Iii C, *et al.* (2007) Development of a standardized DNA database for Chinook salmon. *Fisheries*, **32**, 540–552.

Skotte L, Korneliussen TS, Albrechtsen A (2013) Estimating individual admixture proportions from next generation sequencing data. *Genetics*, **195**, 693–702.

Smouse PE, Waples RS, Tworek JA (1990) A genetic mixture analysis for use with incomplete source population data. *Canadian Journal of Fisheries and Aquatic Sciences*, **47**, 620–634.

Therkildsen NO, Palumbi SR (2017) Practical low-coverage genomewide sequencing of hundreds of individually barcoded samples for population and evolutionary genomics in nonmodel species. *Molecular ecology resources*, **17**, 194–208.

Thompson NF, Anderson EC, Clemento AJ, Campbell MA, Pearse DE, Hearsey JW, Kinziger AP, Garza JC (2020) A complex phenotype in salmon controlled by a simple change in migratory timing. *Science*, **370**, 609–613.

## Data Accessibility

WGSAssign is available as a Python package with these associated links:

- Development version and entire revision history on GitHub: `https://github.com/mgdesaix/wgsassign`
- Zenodo archive of initial package release: `https://zenodo.org/record/7957898`
- Online version of data and scripts used in paper: `https://github.com/mgdesaix/WGSassign-manuscript-`
- Data repository with full datasets used in paper. UPDATED WHEN MER PROVIDES DOI. `https://dryad.something.or.other`

## Appendix A: Fisher Information

*Fisher Information from Genotype Likelihoods*

We focus on the information for the $\ell^{\text{th}}$ locus in the $k^{\text{th}}$ reference population. Accordingly we drop the $_{k,\ell}$ subscript from $\theta$ and the $\ell$ subscript from $g$. Furthermore, since $L(\theta)$ is a sum over the $n_k$ reference samples from $k$, we must simply find the derivative for the term in the sum corresponding to a single individual, knowing that the Fisher information will be the sum of that quantity over all $n_k$ individuals. To further ease notation, we will write $L_i(\theta)$ for the $i^{\text{th}}$ individual's term in the sum for $L(\theta)$, while we drop the superscript $^{(i)}$ from the $g$'s. Thus, we seek $-\frac{\partial^2 L_i(\theta)}{\partial \theta^2}$.

We start by finding the first derivative:

$$\frac{\partial L_i(\theta)}{\partial \theta} = \frac{\partial}{\partial \theta} \log\Big[g_0(1-\theta)^2 + g_1 2\theta(1-\theta) + g_2\theta^2\Big].$$

Let

$$\begin{aligned} u &= g_0(1-\theta)^2 + g_1 2\theta(1-\theta) + g_2\theta^2 \\ &= g_0(1-2\theta+\theta^2) + g_1(2\theta - 2\theta^2) + g_2\theta^2, \end{aligned}$$

and note that

$$\begin{aligned} \frac{\partial u}{\partial \theta} &= g_0(2\theta - 2) + g_1(2-4\theta) + g_2 2\theta \\ &= 2\theta(g_0 + g_2 - 2g_1) + 2(g_1 - g_0). \end{aligned}$$

Since $\partial \log(u)/\partial\theta = (\partial u/\partial\theta)u^{-1}$, we have that

$$\frac{\partial L_i(\theta)}{\partial \theta} = \Big(2\theta(g_0 + g_2 - 2g_1) + 2(g_1 - g_0)\Big)\Big(g_0(1-\theta)^2 + g_1 2\theta(1-\theta) + g_2\theta^2\Big)^{-1}.$$

Proceeding, define $v$ and $w$ as follows:

$$\begin{aligned} v &= 2\theta(g_{i,0} + g_{i,2} - 2g_{i,1}) + 2(g_{i,1} - g_{i,0}) && = \frac{\partial u}{\partial \theta} \\ w &= \Big(g_{i,0}(1-\theta)^2 + g_{i,1}2\theta(1-\theta) + g_{i,2}\theta^2\Big)^{-1} && = u^{-1}, \end{aligned}$$

and note that we can rewrite $\frac{\partial L_i(\theta)}{\partial \theta} = vw$, and take the derivative of that easily using the product rule: $(vw)' = v'w + w'v$. To do so, we first find the derivatives

$$\begin{aligned} v' &= \frac{\partial v}{\partial \theta} = 2(g_0 + g_2 - 2g_1) \\ w' &= \frac{\partial w}{\partial \theta} = -u^{-2}\frac{\partial u}{\partial \theta} = -u^{-2}v, \end{aligned}$$

then we put them together with the product rule

$$\frac{\partial^2 L_i(\theta)}{\partial \theta^2} = v'w + vw' = \frac{v'}{u} - \frac{v^2}{u^2}$$
$$= \frac{2(g_0 + g_2 - 2g_1)}{g_0(1-\theta)^2 + g_1 2\theta(1-\theta) + g_2\theta^2} - \left(\frac{2\theta(g_0 + g_2 - 2g_1) + 2(g_1 - g_0)}{g_0(1-\theta)^2 + g_1 2\theta(1-\theta) + g_2\theta^2}\right)^2.$$

Restoring the $_{k,\ell}$ subscript to $\theta$, and the $^{(i)}$ superscript and $\ell$ subscript to $g$, negating, taking the sum over the $n_k$ individuals and evaluating at the MLE yields $I_o^{(i)}(\theta_{k,\ell})$ in (10).

*Expected Fisher Information from Observed Genotypes*

Under Hardy-Weinberg equilibrium, the allelic type of the two gene copies within a locus are independent of one another, and thus a sample of $n$ diploids with fully observed genotypes is equivalent to a sample of $2n$ gene copies, each one an independent Bernoulli trial with success probability $\theta$. Finding the expected Fisher information in such a case is a standard exercise, but we repeat it here for completeness. For a single such variable $Y_i$, we have $P(Y_i = y|\theta) = \theta^y(1-\theta)^{1-y}$, so the log likelihood for that single observation is $L_i(\theta) = y\log\theta + (1-y)\log(1-\theta)$. It follows that

$$\frac{\partial}{\partial\theta}L_i(\theta) = \frac{y}{\theta} - \frac{1-y}{1-\theta} \qquad \text{and} \qquad \frac{\partial^2}{\partial\theta^2}L_i(\theta) = -\frac{y}{\theta^2} - \frac{1-y}{(1-\theta)^2}.$$

The expected Fisher information in a single gene copy is the expectation of the negative second derivative given the true value of $\theta$:

$$\mathbb{E}\left[-\frac{\partial^2}{\partial\theta^2}L_i(\theta)\right] = \mathbb{E}\left[\frac{y}{\theta^2} + \frac{1-y}{(1-\theta)^2}\right] = \frac{1}{\theta} + \frac{1}{1-\theta} = \frac{1}{\theta(1-\theta)}.$$

Since information from independent variables is additive, the information for $2n$ such Bernoulli variables is $2n[\theta(1-\theta)]^{-1}$. Evaluating the expectation under the assumption that the true value of $\theta$ is $\hat{\theta}_{k,\ell}$ gives $I_e(\theta_{k,\ell})$ in (11).

## Appendix B: $z$-Score Calculation

In order to assess whether an individual $A$'s genotype could not plausibly have come from one of the $K$ source populations, even though it was assigned to population $k$, we wish to compare $A$'s log read probability given that it originated from population $k$, $\log P(R^{(A)}|\theta_k)$, to the distribution of log read probability values expected from individuals that actually are from population $k$. Complicating matters, these log read probabilities are heavily influenced by the read depth, and to a lesser extent, by the relationship between allele depths (how many reads of each allele were seen) and the genotype likelihoods. So, in fact, we must compare $\log P(R^{(A)}|\theta_k)$ to the distribution of $\log P(R|\theta_k)$ expected from an individual that originates from source $k$, but also has read depths at each locus exactly the same as individual $A$, and also has genotype likelihoods that exhibit the same relationship to allele depths as those in individual $A$. (This relationship will be influenced by such factors as the base quality scores and the genotype likelihood model used).

In previous applications, with far fewer markers, determining such a distribution of the log probability of the observed data has been done through simulation, for example, in the "exclusion method" of Cornuet *et al.* (1999); however, with genomic-scale data it would be impractical to simulate thousands of new multilocus genotypes, each with potentially millions of loci, to assess whether each individual (with their own, specific read depth values) might be from a population not included among the source populations. Instead of simulation, we develop the expected distribution of log probabilities using a central limit theorem (CLT) approximation. Note that, since $P(R|\theta_k)$ is a product over many loci, $\log P(R|\theta_k)$ is a sum over loci. We will write the contribution of each locus to that sum as

$$W_\ell = \log[g_{\ell,0}(1-\theta_{k,\ell})^2 + g_{\ell,1}2(\theta_{k,\ell})(1-\theta_{k,\ell}) + g_{\ell,2}(\theta_{k,\ell})^2] = f(\boldsymbol{g}_\ell, \theta_{k,\ell})$$

where we include the notation $f(\boldsymbol{g}_\ell, \theta_{k,\ell})$ to emphasize the fact that $W_\ell$ is a deterministic function of $\theta_{k,\ell}$ and the vector of genotype likelihoods $\boldsymbol{g}_\ell = (g_{\ell,0}, g_{\ell,1}, g_{\ell,2})$. It is important to recognize in this context that $\theta_{k,\ell}$ is considered fixed while $\boldsymbol{g}_\ell$ is a random variable. By extension, then, so too is $W_\ell$ a random variable. By the CLT, the sum of very many independent $W_\ell$ random variables can be approximated by a normal

distribution with mean $\mu$ and variance $\sigma^2$ given by:

$$\mu = \sum_{\ell=1}^{L} \mathbb{E}(W_\ell)$$
$$\sigma^2 = \sum_{\ell=1}^{L} \mathrm{Var}(W_\ell).$$

Thus, we seek $\mathbb{E}(W_\ell)$ and $\mathrm{Var}(W_\ell)$.

The distribution of $W_\ell$ clearly depends on the distribution of $\boldsymbol{g}_\ell$. We develop such a distribution, hierarchically, based on the following assumptions:

1. $\boldsymbol{g}_\ell$ depends directly on the observed allele depths. Let $r_\ell$ be the number of reference alleles and $a_\ell$ the number of alternate alleles observed in the reads covering site $\ell$, and let $\gamma$ denote an individual-specific effect of base quality scores, etc., on the genotype likelihoods. Then we denote this conditional probability distribution as $P(\boldsymbol{g}_\ell | r_\ell, a_\ell, \gamma)$ and we will denote the set of values that $\boldsymbol{g}_\ell$ might take for a given pair $(r, a)$ as $\mathscr{G}_{r,a}$. Note that here we are asserting that given the allele depths, the genotype likelihood is independent of the genotype. This is a relatively unpalatable assumption, but we make it because we don't have access to the information we would need (knowledge of the true underlying genotypes) to easily relax this assumption, and it eases the computations considerably.

2. The read depths $r_\ell$ and $a_\ell$ depend on the genotype, $G^*_\ell$ at locus $\ell$ of the individual being sequenced and on a population-specific error rate, $\epsilon_k$. The model for this is simple binomial random sampling from a total read depth of $D_\ell$, with a probability $\epsilon_k$, independently for each read, that the base in question will be read incorrectly. Hence:

$$P(r_\ell, a_\ell | G^*_\ell, D_\ell) = \frac{D_\ell!}{r_\ell! a_\ell!} \times \begin{cases} (1-\epsilon_k)^r \epsilon_k^{a_\ell} & \text{if } G^*_\ell = 0 \\ (1/2)^{D_\ell} & \text{if } G^*_\ell = 1 \\ \epsilon_k^{r_\ell} (1-\epsilon_k)^{a_\ell} & \text{if } G^*_\ell = 2, \end{cases}$$

   where $a_\ell = D_\ell - r_\ell$, always. (We note that $r_\ell$ and $D_\ell$ completely determine $a_\ell$, but we leave both $r_\ell$ and $a_\ell$ in the preceding and following probability expressions for ease of explanation later.)

3. The frequency of $G^*_\ell$ in source population $k$ follows Hardy-Weinberg equilibrium with an allele frequency of $\theta_{k,\ell}$, so $P(G^*_\ell | \theta_{k,\ell})$ is given by (1).

With these assumptions, given the total read depth $D_\ell$, and $\gamma$ and $\epsilon_k$, the joint probability of the remaining variables is:

$$P(G^*_\ell, r_\ell, a_\ell, \boldsymbol{g}_\ell \mid \theta_{k,\ell}, D_\ell, \gamma, \epsilon_k) = P(G^*_\ell|\theta_{k,\ell})P(r_\ell, a_\ell|G^*_\ell, D_\ell)P(\boldsymbol{g}_\ell|r_\ell, a_\ell, \gamma)$$

The mean and the variance of $W_\ell$ can now be found from these by taking expectations:

$$\mathbb{E}\Big[W_\ell \mid \theta_{k,\ell}, D_\ell, \gamma, \epsilon_k\Big] = \bar{W}_\ell = \sum_{G=0}^{2} \sum_{\substack{(r,a):\\ r+a=D_\ell}} \sum_{\boldsymbol{g}\in\mathscr{G}_{r,a}} f(\boldsymbol{g}_\ell = \boldsymbol{g}, \theta_{k,\ell}) P(G^*_\ell = G, r_\ell = r, a_\ell = a, \boldsymbol{g}_\ell = \boldsymbol{g} \mid \theta_{k,\ell}, D_\ell, \gamma, \epsilon_k)$$

$$\mathrm{Var}\Big[W_\ell \mid \theta_{k,\ell}, D_\ell, \gamma, \epsilon_k\Big] = \sum_{G=0}^{2} \sum_{\substack{(r,a):\\ r+a=D_\ell}} \sum_{\boldsymbol{g}\in\mathscr{G}_{r,a}} [\bar{W}_\ell - f(\boldsymbol{g}_\ell = \boldsymbol{g}, \theta_{k,\ell})]^2 P(G^*_\ell = G, r_\ell = r, a_\ell = a, \boldsymbol{g}_\ell = \boldsymbol{g} \mid \theta_{k,\ell}, D_\ell, \gamma, \epsilon_k).$$

As there is no documented distribution for $P(\boldsymbol{g}_\ell|r_\ell, a_\ell, \gamma)$, we simply use the empirical distribution of $\boldsymbol{g}_\ell$ values across all loci within the individual having allele depths of $r$ and $a$. In practice, values of $\boldsymbol{g}$ for any particular pair $(r, a)$ are typically clustered around a single value, and we discretize that distribution into a histogram with a small number, $b$, of bins defined by the value of the largest of the three elements of $\boldsymbol{g}$, thus imagining $P(\boldsymbol{g}_\ell|r_\ell, a_\ell, \gamma)$ as a discrete distribution with weight on $b$ values of $\boldsymbol{g}$, each one the mean of the values of $\boldsymbol{g}$ within the bin. It is also possible to remove loci that have particularly odd values of $\boldsymbol{g}$. For example, GATK sometimes assigns a $\boldsymbol{g}_\ell$ of $(1/3, 1/3, 1/3)$ to loci with read depths $r = 1$, $a = 0$. Any such aberrant values can be removed, without penalty, since the $\mu$ and $\sigma^2$ that we seek are conditioned upon a set of loci. The parameter $\epsilon_k$ might be estimable, but for now we assume a value for it, like $\epsilon_k = 0.01$.

After all this, a sum over the loci included in the metric gives us the mean and variance of the normal distribution that the log genotype probabilities of a matched individual (same loci, same read depths, same relationship between allele depths and $\boldsymbol{g}$) from population $k$ would be expected to have:

$$\mu = \sum_{\ell=1}^{L} \delta_\ell \mathbb{E}\Big[W_\ell \mid \theta_{k,\ell}, D_\ell, \gamma, \epsilon_k\Big]$$

$$\sigma^2 = \sum_{\ell=1}^{L} \delta_\ell \mathrm{Var}\Big[W_\ell \mid \theta_{k,\ell}, D_\ell, \gamma, \epsilon_k\Big],$$

where $\delta_\ell = 1$ if the locus $\ell$ was included in the calculation, and 0 otherwise. Thus, the variable

$$z_k^{(A)} = \frac{\log P(R^{(A)}|\theta_k) - \mu}{\sigma}$$

should, by the CLT, have a normal distribution with mean 0 and variance 1.

Of course, there are several reasons why the actual distribution of $z_k^{(A)}$ might depart from a Normal$(0, 1)$: our calculations for the mean and variance of each locus are unlikely to be perfectly reliable, the rate of sequencing error might be higher or lower than we assume, or there might be genetic structure within population $k$, and hence also within the reference samples from population $k$. Thus, we correct the $z$-score so that it exhibits a mean of 0 and a variance of 1 for the reference samples, themselves, from population $k$. With $i = 1, \ldots, n_k$ denoting the reference samples from population $k$, we calculate

$$\bar{z}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} z_k^{(i)} \qquad \text{and} \qquad \bar{\sigma}_k^2 = \frac{1}{n_k - 1} \sum_{i=1}^{n_k} \left( z_k^{(i)} - \bar{z}_k \right)^2.$$

Then, we assess whether an unknown individual $A$ assigned to population $k$ may have come from an unsampled population using:

$$z_k^{*(A)} = \frac{z_k^{(A)} - \bar{z}_k}{\bar{\sigma}_k^2}.$$

781

# List of Figures

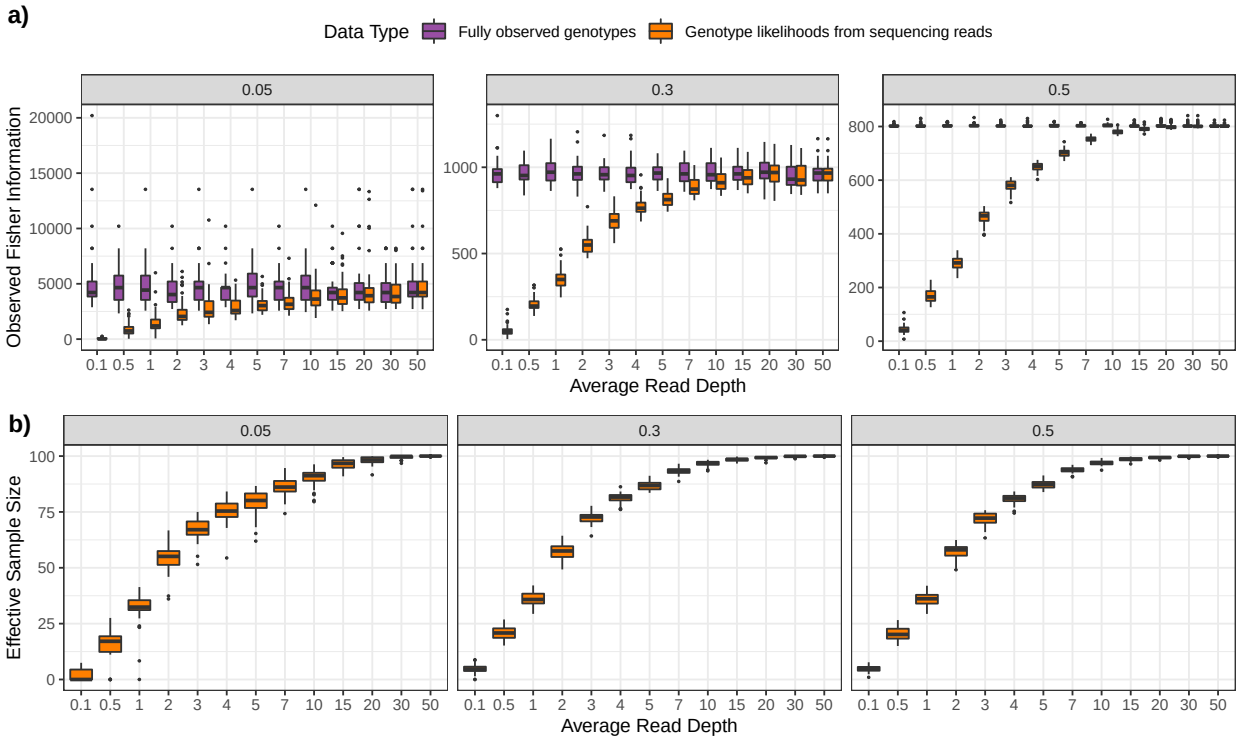


**Figure 1** **a)** Observed information calculated for simulated data summarized either as fully observed genotypes (purple) or as genotype likelihoods (orange) computed from sequencing read data of different depths simulated from the genotypes. Fully observed genotype data is not affected by read depth, but an independent set of fully observed genotypes was simulated for each different value of read depth, and these are all shown in the figure. **b)** Effective sample sizes calculated for simulated genotype likelihood data. In each figure the facet headers give the true population allele frequency, the $x$-axis gives the average read depth in the simulations, and the distribution of quantities in the $y$ direction are summarized as boxplots showing the median (dark line) the first and third quartiles (the edges of the boxes) the largest (or smallest) value no further than 1.5 $\times$ the interquartile range from the first (third) quartiles (the whiskers) and outliers beyond the whiskers (individual points).

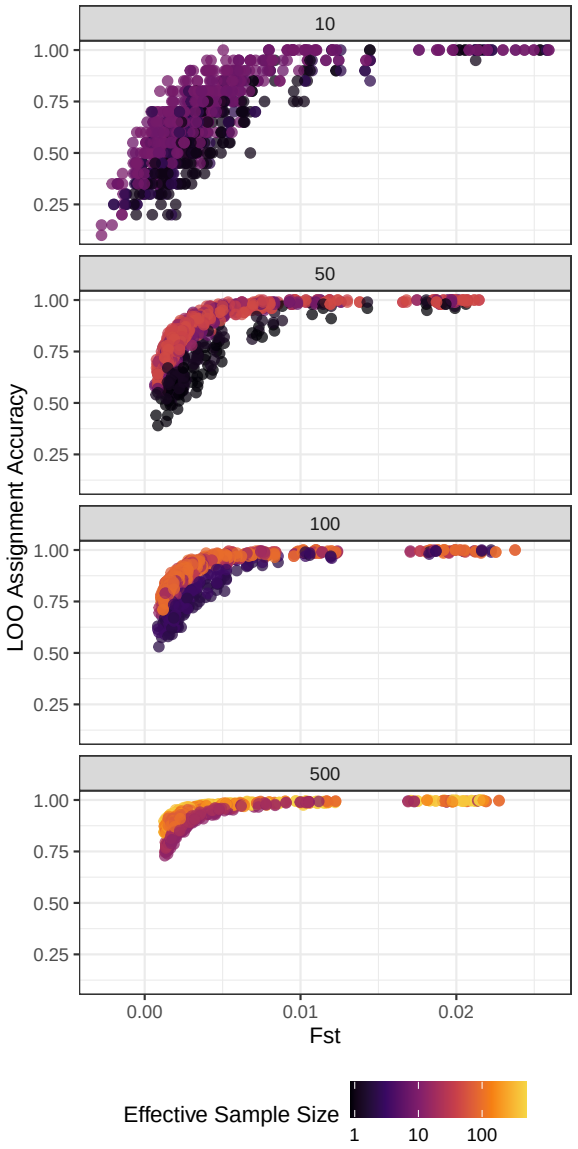


**Figure 2** Leave-one-out (LOO) assignment accuracy for known source individuals increases as genetic differentiation ($F_{ST}$) increases. Each point represents a single one of 4,633 simulation runs of the two-population island model when effective sample sizes were greater than 0.1 individuals. Panels are ordered by the number of individuals (10, 50, 100, 500) sampled from each of the two populations. The proportion of correctly assigned individuals, via LOO cross-validation for one population is given on the $y$-axis and genetic differentiation ($F_{ST}$) between the two populations is on the $x$-axis. The points are colored by effective sample size ($\log_{10}$ scale) of the population. Assignment accuracy in simulation runs with similar genetic differentiation tends to be greater for populations with greater effective sample size (lighter colors) than smaller effective sample sizes (darker colors). The variation in assignment accuracy decreases as more samples are used in the source population, with the highest amount of variation when 10 samples are used and the least amount of variation when 500 samples are used.

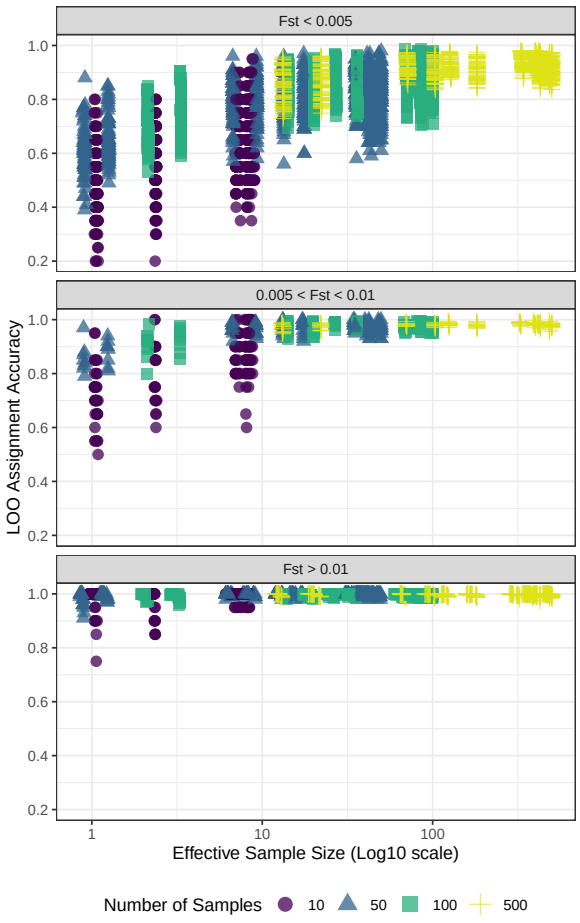


**Figure 3** Increasing effective sample size results in an increase in LOO assignment accuracy. The proportion of correctly assigned individuals, using LOO cross-validation, for one population, is given on the y-axis and effective sample size (log10 scale) of the population is on the x-axis. Similar values of effective sample size results in a similar range of assignment accuracy, however the number of samples also influences the accuracy at lower effective samples sizes and with weaker genetic differentiation. Some of the effect of sample size, separate from effective sample size, can be explained by LOO assignment removing an individual from the source population during assignment, which will disproportionately decrease the precision of allele frequency estimation for smaller sample sizes than larger sample sizes.

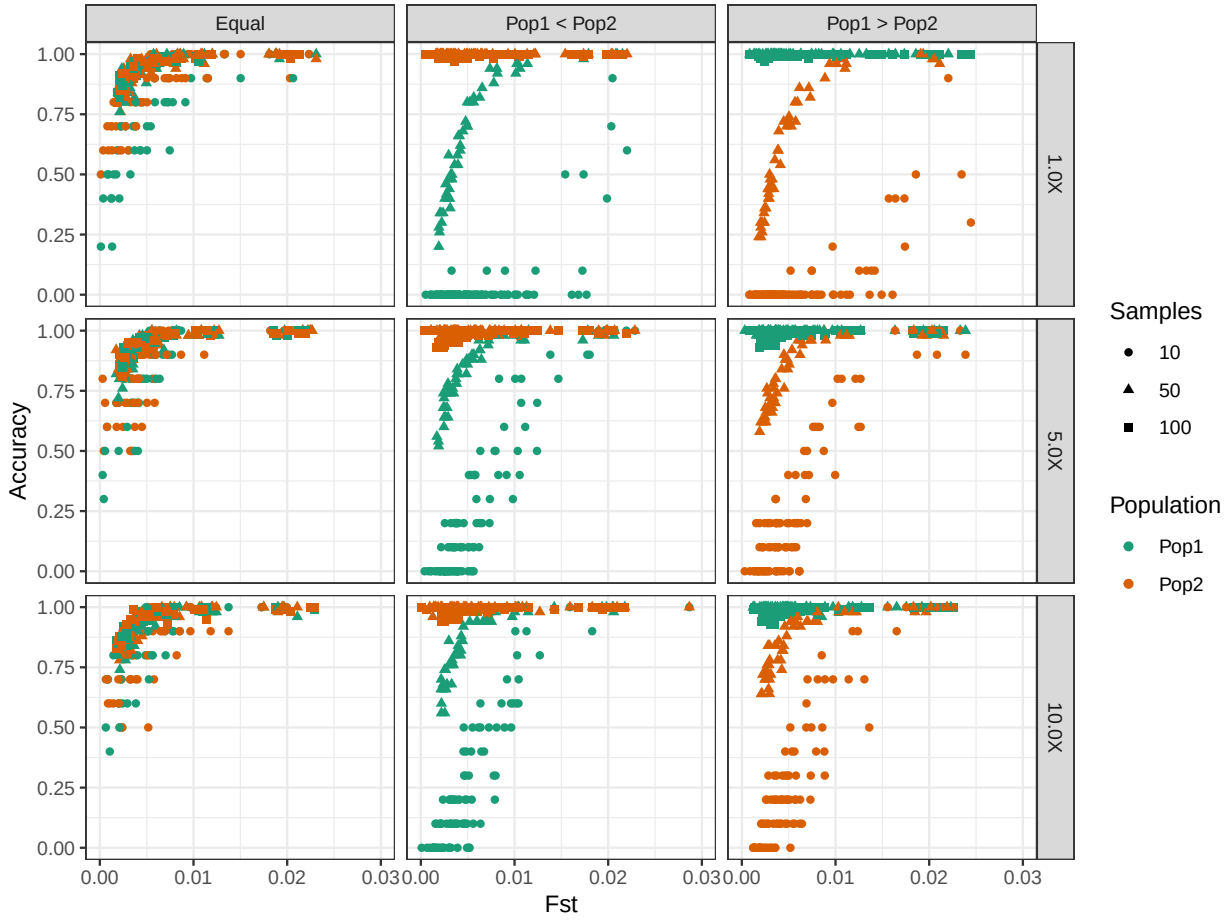


**Figure 4** Unequal sample sizes among source populations result in decreased assignment accuracy due to differences in the precision of allele frequency estimation among the populations. Here, the two populations had either 10, 50, or 100 samples used for estimating allele frequency and then assigned via leave-one-out. When both populations had the same number of samples ("Equal" column), assignment accuracy generally increased as Fst increased and was similar for either population. When Population 1 had fewer samples than Population 2 ("Pop1 < Pop2" column), the assignment accuracy of Population 1 was generally less than that of Population 2, and the reverse was demonstrated when Population 1 had more samples than Population 2 ("Pop1 > Pop2" column). The reduction in assignment accuracy from biased sample sizes was also more pronounced with lower read depth.

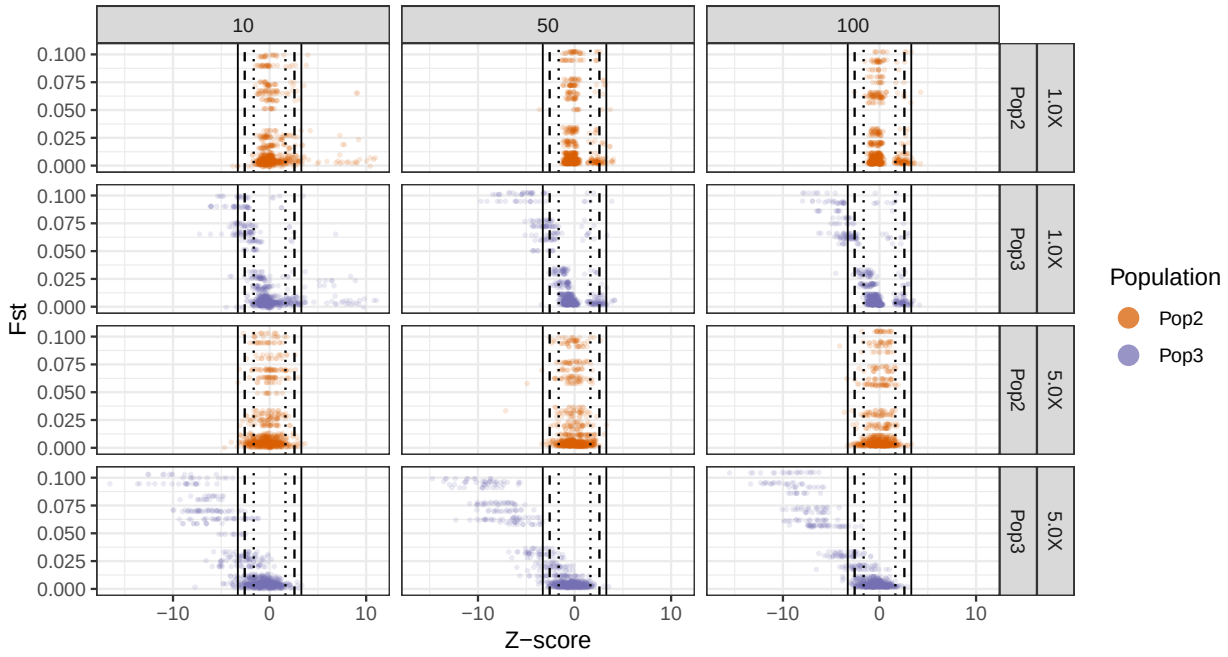


**Figure 5** Results from the three-population stepping-stone model demonstrate the behavior of the *z*-score metric in identifying individuals from an unsampled population (Pop3) assigned to a population in the reference compared to individuals correctly assigned to their source population of origin (Pop2). Symmetric lines subtending 90%, 99%, and 99.9% of the mass of a standard unit normal random variate are given by vertical lines (dotted, dashed, and solid, respectively).

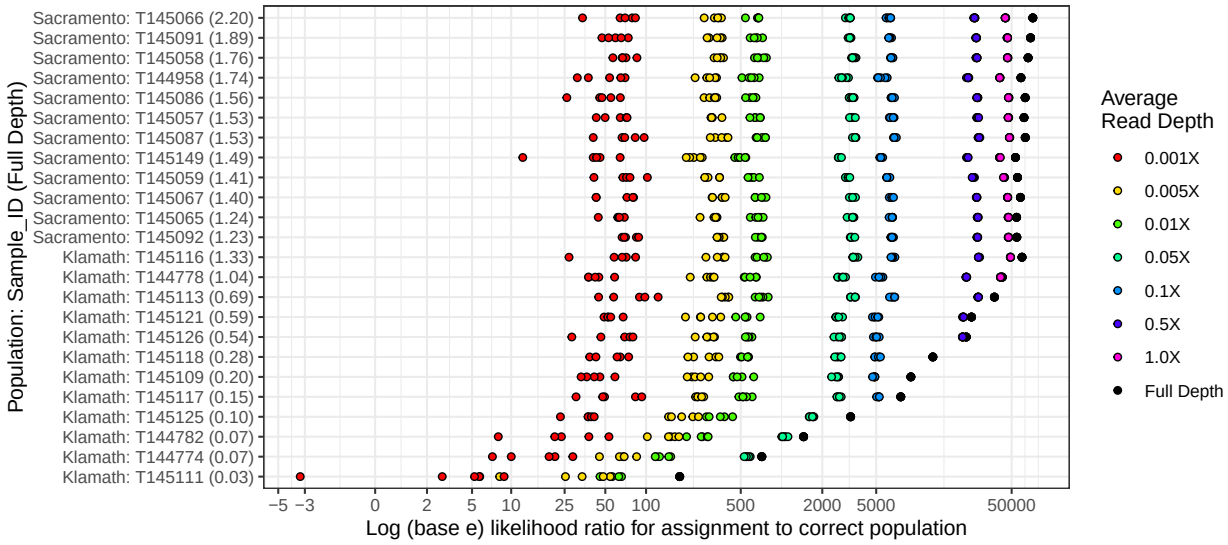


**Figure 6** Log likelihood ratios for assignment at different read depth levels for the Chinook salmon data. On the *y*-axis are different Chinook salmon samples, labeled by their population, a colon, their ID number, and then in parentheses the average read depth of their aligned data at full depth. On the *x*-axis is the log-likelihood ratio in favor of assignment to their own (correct) population on a "pseudo-log" scale that accommodates negative values. Positive numbers indicate correct assignment. Colors denote the read depths after downsampling. There are five points for each individual at each value of downsampling, reflecting the 5 different seeds used for downsampling.

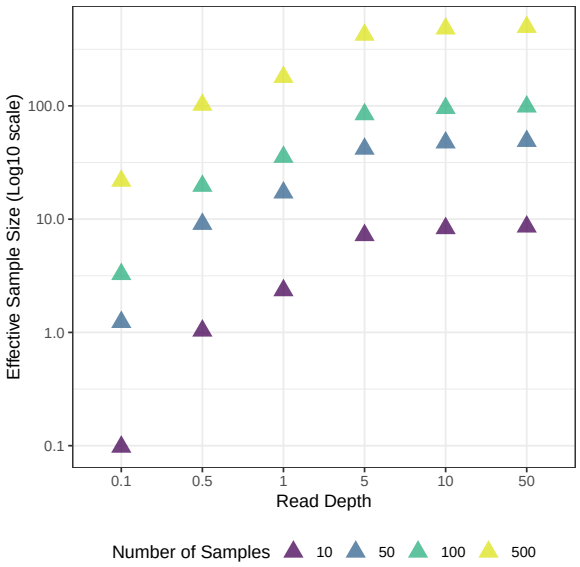


**Figure 7** The relation between read depth and number of samples in determining the effective sample size highlights the potential for different sampling design strategies for achieving similar effective sample size. For example, if the target effective sample size is 10, then sequencing 500 individuals at 0.1x would likely overshoot the target, 50 individuals at 0.5x would be close to the target, and 10 individuals at >10x coverage would be close to the target.