Received: 19 September 2023 | Accepted: 15 December 2023

DOI: 10.1111/2041-210X.14286

Methods in Ecology and Evolution

RESEARCH ARTICLE

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

Matthew G. DeSaix[1] | Marina D. Rodriguez[1] | Kristen C. Ruegg[1] | Eric C. Anderson[1,2,3]

[1]Department of Biology, Colorado State University, Fort Collins, Colorado, USA

[2]Fisheries Ecology Division, Southwest Fisheries Science Center, National Marine Fisheries Service NOAA, Santa Cruz, California, USA

[3]Department of Fisheries, Wildlife, and Conservation Biology, Colorado State University, Fort Collins, Colorado, USA

Correspondence
Matthew G. DeSaix
Email: mgdesaix@gmail.com

Funding information
National Science Foundation, Grant/Award Number: 008933-00002; Alaska Department of Fish and Game, Grant/Award Number: 23-011

Handling Editor: Oscar Gaggiotti

Abstract

1. 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 likelihoods.
2. Here, we present a probabilistic framework for using genotype likelihoods for standard population assignment applications. Additionally, we derive the Fisher information for allele frequency from genotype likelihoods 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.
3. Using simulated and empirical data sets, we demonstrate the behaviour 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.
4. 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.

KEYWORDS
Fisher information, genetic stock identification, genotype likelihoods, low-coverage whole-genome sequencing, next-generation sequencing, population assignment, population genomics, statistical genetics

1 | 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 (Buerkle & Gompert, 2013; Nielsen et al., 2011). 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 (Kim et al., 2011; Nielsen et al., 2011, 2012). 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 genetic 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. Benestan et al., 2015; DeSaix et al., 2019; Ruegg et al., 2014). 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. However, a challenge with using low-coverage WGS data for assignment tests is that the allele frequency estimates may be uncertain, which could lead to inaccurate assignment results. While this challenge is not unique to low-coverage WGS data, as low sample size also increases uncertainty regardless of sequencing coverage, the challenge of accurate allele frequency estimation is compounded for low-coverage WGS by low read depth. For accurate allele frequency estimation from low-coverage WGS data, 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 was 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. For population assignment tests, an information metric of this sort would allow researchers to more directly identify the necessary sample size and sequencing depth needed to perform accurate assignment given the genetic differentiation of their samples. Furthermore, given that unequal sample size among reference populations is a source of bias in assignment tests with called genotypes (Wang, 2017), an information metric would allow the identification and mitigation of biased assignment due to the combined influence of unequal sample sizes and sequencing depths among populations.

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) to provide common assignment methods that use genotype likelihoods, instead of called genotypes; (2) to evaluate the information available in low-read depth sequencing data for allele frequency estimation; and (3) to achieve computational efficiency for processing large numbers of samples with genome-wide data. WGSassign provides methods for individual assignment 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 (Casella & Berger, 2021) 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 data set. This calculation of effective sample size has broad utility for population genomics studies using low-coverage WGS.

We validate WGSassign and investigate its behaviour with an extensive set of simulations and demonstrate its use on two empirical data sets. In the first, we apply WGSassign to weakly differentiated groups of yellow warblers (*Setophaga 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.

2 | 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 (DeSaix, 2023).

2.1 | 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 ℓ ($1 \le \ell \le L$) 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 ℓ. Denote by $\theta_{k,\ell}$ the true—but typically unknown—frequency of the alternate allele at locus ℓ within source population k. Under the assumption of Hardy–Weinberg equilibrium, the probability of G_ℓ, 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_ℓ is not observed with certainty. Rather, evidence about the unknown genotype is obtained from sequencing reads covering the locus. Let R_ℓ denotes the sequencing read data from an individual at locus ℓ. The evidence for the state of G_ℓ 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:

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

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 programmes such as bcftools (Li, 2011; Li et al., 2009), 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).

Performing population assignment using read data from an individual (rather than from directly observed genotypes) requires, for each locus, ℓ, 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 Equations (1) and (2) using the law of total probability:

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

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.

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

where θ_k denotes the set of all L allele frequencies in population k. Of course, with WGS, some variants may be near one another and will then likely be in LD. In such a case, Equation (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). Given the unbiased nature of composite likelihood estimators, LD pruning of the WGS data is not necessary. For each individual of unknown origin, this likelihood can be computed for each source population, k, and the relative values of those likelihoods give the evidence that the individual came from each of the source populations. If the prior probability π_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:

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

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 ith reference sample, given that it came from source population k, is, following Equation (3),

$$P\left(R_\ell^{(i)} \mid \theta_{k,\ell}\right) = g_{\ell,0}^{(i)}\left(1-\theta_{k,\ell}\right)^2 + g_{\ell,1}^{(i)} 2\left(\theta_{k,\ell}\right)\left(1-\theta_{k,\ell}\right) + g_{\ell,2}^{(i)}\left(\theta_{k,\ell}\right)^2, \quad (6)$$

where the genotype likelihoods are now adorned with a superscript (i) to denote they are for the ith 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:

$$L\left(\theta_{k,\ell}\right) = \sum_{i=1}^{n_k} \log\ P\left(R_\ell^{(i)} \mid \theta_{k,\ell}\right). \quad (7)$$

In our implementation, we first use the expectation–maximization algorithm (Dempster et al., 1977) from ANGSD (Kim et al., 2011) and the code as implemented in PCangsd (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:

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

where, again, n_k is the number of reference samples from source population k. This provides a correction for cases in which 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. Without this correction, the $P\left(R_\ell^{(i)} \mid \theta_{k,\ell}\right) = 0$ in the absence of an allele and the $L\left(\theta_{k,\ell}\right)$ cannot be calculated. This approach is identical to the 'Frequency Criterion' used in GENECLASS 2.0 with the 'adjustable default value' set to $1/(2n+1)$. Another approach, due to Rannala and Mountain (1997), that places beta priors, independently for each population and locus, on the allele frequencies, has also been widely used in population assignment methods. Implementing that approach with genotype likelihoods is more computationally challenging than with observed genotypes, and since extensive simulations (not shown) revealed no substantial differences between the two methods, we adopted the 'Frequency Criterion' approach.

2.2 | Fisher information and effective sample size

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. In this section, we develop the 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. We define this metric as the *effective sample size*: The number of diploid individuals with called genotypes that provide the same amount of information for allele frequency as the observed information from the low-coverage WGS samples. 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:

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

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

$$I_o^{(i)}\left(\theta_{k,\ell}\right) = \left[\frac{2\left(g_{\ell,0}^{(i)} + g_{\ell,2}^{(i)} - 2g_{\ell,1}^{(i)}\right)}{g_{\ell,0}^{(i)}\left(1-\hat{\theta}_{k,\ell}\right)^2 + g_{\ell,1}^{(i)} 2\hat{\theta}_{k,\ell}\left(1-\theta_{k,\ell}\right) + g_{\ell,2}^{(i)}\hat{\theta}_{k,\ell}^2} + \left(\frac{2\hat{\theta}_{k,\ell}\left(g_{\ell,0}^{(i)} + g_{\ell,2}^{(i)} - 2g_{\ell,1}^{(i)}\right) + 2\left(g_{\ell,1}^{(i)} - g_{\ell,0}^{(i)}\right)}{g_{\ell,0}^{(i)}\left(1-\hat{\theta}_{k,\ell}\right)^2 + g_{\ell,1}^{(i)} 2\hat{\theta}_{k,\ell}\left(1-\hat{\theta}_{k,\ell}\right) + g_{i,2}^{(i)}\hat{\theta}_{k,\ell}^2}\right)^2\right]. \quad (10)$$

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

To derive $\tilde{n}_l$, our effective sample size metric for locus ℓ, 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 $\hat{\theta}_{k,\ell}$:

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

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

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

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 ℓ. 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 a population is then derived by taking the mean of $\tilde{n}_l$ across all loci, $\tilde{n} = \frac{1}{L}\sum_{l=1}^{L}\tilde{n}_l$. In practice, the estimates of information are highly variable for rare alleles; therefore, we recommend this calculation be done for loci with a minor allele frequency > 0.05.

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:

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

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

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

$\tilde{n}^{(i)}$ 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 likelihoods 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 (e.g. three) 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. The determination of the specific standard deviation cut-off for the *z*-score must be determined from the specific empirical data.

2.3 | 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_{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 θ 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 θ was the MLE from the genotype likelihoods and then used Equation (12).

We ran these simulations across values of $D_{ave} \in \{0.1,0.5,1,2,3,4,5,7,10,15,20,30,50\}$ 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.

2.4 | 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 data sets using the coalescent simulation program, msprime (Kelleher et al., 2016). The first simulation included two populations, each with an effective sizes of 1000, exchanging migrants. 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}, per site and per generation. To vary the genetic differentiation between populations, we varied the lineage migration rate parameter between 0.0005 and 0.05 in 10 equal increments. From both populations, we sampled 10, 50, 100 or 500 individuals. Pairwise F_{ST} was calculated between the two populations using the sampled individuals and the genetic variants were output in variant call format (VCF).

With the VCF file output from msprime, we used bcftools (Li, 2011; Li et al., 2009) to remove any SNPs with a minor allele frequency (MAF) less than 0.05 and randomly selected 100,000

of the remaining SNPs. 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 240 parameter combinations (10 migration rates, 4 sample sizes and 6 read depths), we simulated 10 replicates, for a total of 2400 simulated data sets. Genotype likelihood output was converted to Beagle file format with custom scripts, and we used these data as input into WGSassign.

To determine the influence 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 individual 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 include data being tested. The assigned population was determined by maximum likelihood.

In the second simulation, we conducted a deeper assessment of the behaviour of effective sample size and its influence on assignment accuracy. We implemented two-population island models as in the previous simulation, but included all sample combinations of 10, 30, 60 and 100 individuals for a population and read depths of 0.5X, 0.75X, 1X, 2X, 4X and 6X with 10 replicates for a total of 5760 simulations. We set migration rate at 0.005 for moderate genetic differentiation based on the previous simulation. In each run, we simulated an extra 20 individuals from each of the two populations and these individuals were held out from allele frequency calculations for the respective population and used for standard assignment accuracy. After performing initial assignment, if a population had a higher effective sample size than the other population, then individuals were removed to standardize the effective sample sizes, and assignment was performed again. In this simulation, all SNPs were used that had MAF > 0.05.

In the third simulation, we assessed the performance of the WGSassign *z*-score metric for determining whether an individual of unknown origin that is assigned to a population is actually from an unsampled population. We implemented a three-population stepping-stone model with 20, 60 or 110 individuals per population using msprime. We varied the migration rate parameter between 0.0001 and 0.01 in 20 equal increments. Individuals had simulated mean read depths of 1X or 5X. 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 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 behaviour 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).

Finally, to illustrate the relation of effective sample size to read depth and absolute sample size for the purpose of study design, we simulated from a two-population island-model coalescent to produce 10 replicates of all combinations of sample sizes in {10,12,15,20,30,60,80,120} and read depths in {0.5X,0.75X, 1X, 2X, 3X, 4X, 5X, 6X} for a total of 640 simulations. The two populations had the same number of samples and read depths, and the migration rate was set at 0.005. Effective sample size was calculated for all these replicate simulations. These values were chosen such that 'equal sequencing effort' could be compared, in this case for a total sequencing depth of 60X (e.g. 120 individuals at 0.5X to 10 individuals at 60X).

2.5 | 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 three 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 downsample 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) of mapping quality >30 and base quality >33. We retained SNPs with read data in at least 50% of individuals and an MAF > 0.05. The genetic data are stored at https://doi.org/10.5061/dryad.h9w0vt4pj (DeSaix et al., 2023).

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. 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 data set, 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 WGS 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 MAF > 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), were 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 total number sites in each data set was recorded, as was the number of informative sites (those with unequal likelihoods for the three different genotypes) within each individual. 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).

3 | RESULTS

3.1 | Effective sample size simulations

Fisher information and effective sample size are shown for three representative values of θ (0.05, 0.3 and 0.5) in Figure 1. 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 vary 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. 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; Fumagalli, 2013; Lou et al., 2021).

3.2 | Genetic simulations

In the first simulation, genetic differentiation between the sampled individuals from the two populations ranged from −0.003 to 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 $F_{ST} > 0.043$. 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).

Runtime for the simultaneous calculation of Fisher information, effective sample size and allele frequency for populations in WGSassign was fast. With two populations and 100,000 loci being analysed in parallel with 20 threads, runtime was less than 10 s for populations with 100 samples or less, and between 15 and 30 s 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 recalculation increases linearly with sample size. Accordingly, runtime for LOO

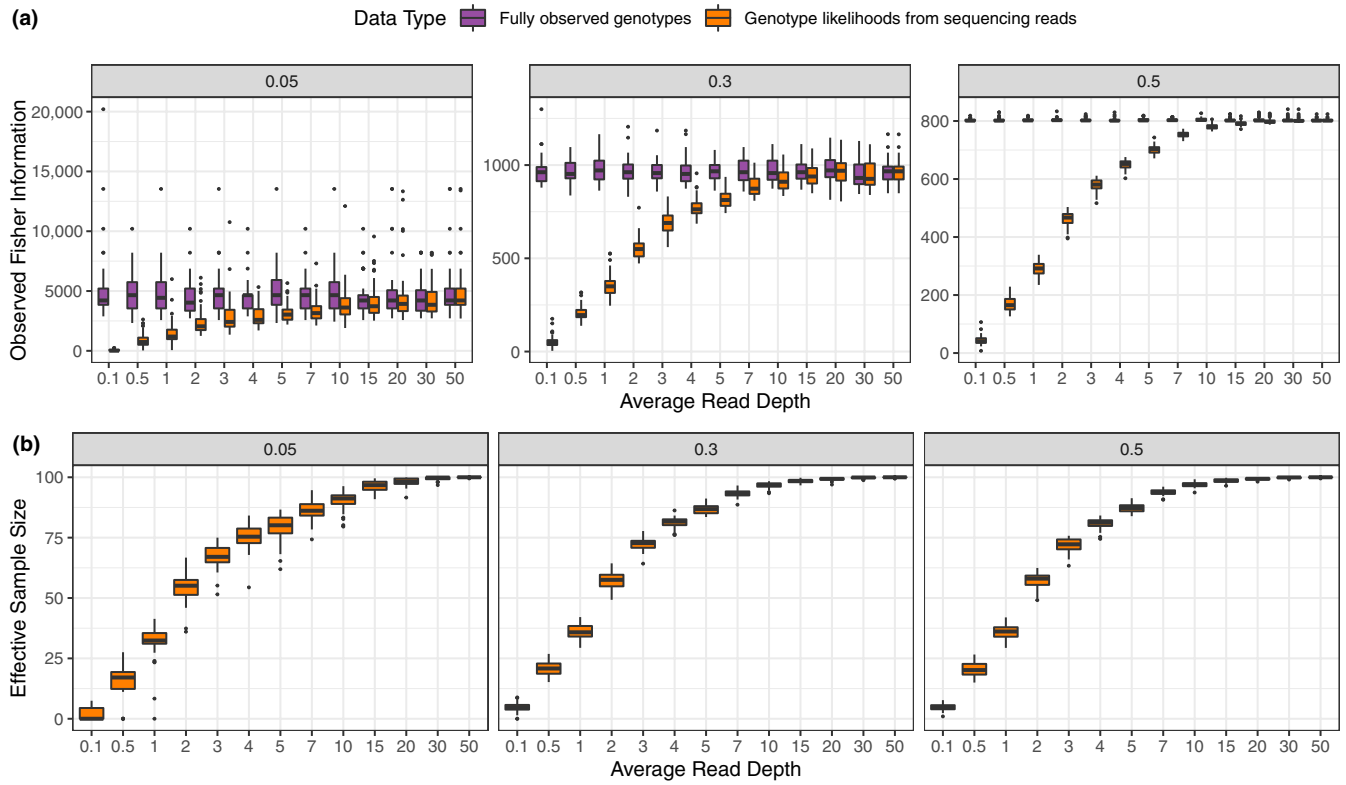

FIGURE 1 (a) Observed information calculated for simulated data summarized either as fully observed genotypes (purple) or as genotype likelihoods (orange) computed from sequencing read data of different depths simulated from the genotypes. Fully observed genotype data are not affected by read depth, 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 is 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 × the interquartile range from the first (third) quartiles (the whiskers) and outliers beyond the whiskers (individual points). All simulations had 100 individuals.

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 s and, for 500 samples, run time was 1743 s.

The second set of simulations showed that at weak to moderate genetic differentiation (mean $F_{ST} = 0.0055$), assignment accuracy was close to 100% when effective sample sizes of the two populations were equal and had at least eight effective individuals (Figure 3a). However, at higher measures of effective sample size, the two populations could have different effective samples sizes and still have high assignment accuracy (e.g. effective samples sizes of 20 vs. 100). Assignment bias occurred when there were sufficient differences between the effective sample sizes that individuals were only being incorrectly assigned from the lower effective sample size population (Figure 3b).

Importantly, when effective sample size is roughly equivalent between the two populations but the number of samples and read depth differ, assignment accuracy is still high and unbiased (Figure 3c,d). This pattern was apparent up through the maximum tested magnitude difference of 12 (Figure 3c,d).

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 4). 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 behaviour of the *z*-score metric (Figure 4). Generally, higher reference sample sizes and read depths allow individuals from unsampled populations to be distinctively identified from individuals that are truly from a sampled source population.

The simulations demonstrating the relationship of read depth and absolute sample size for producing effective sample size in a single population highlighted that prioritizing sample size over sequencing depth results in higher effective sample size. In the provided example of an equal sequencing effort of 60X (i.e. total sequencing depth of a single population), effective sample size increased as more samples at a lower read depth were used—with the lowest effective sample size of 7.8 for 10 individuals at 6X which increased threefold to 24.5 for 120 individuals at 0.5X (Figure 5). In other words, if a researcher

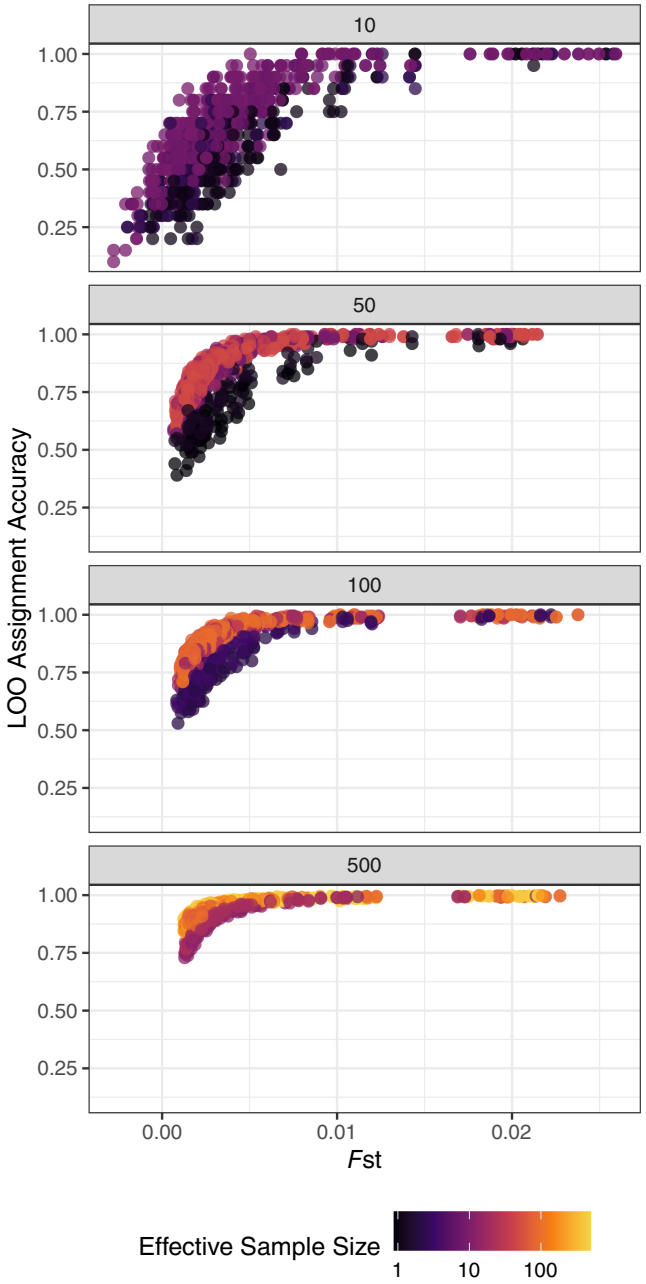

FIGURE 2 Each point represents a single simulation run of the two-population island model when effective sample sizes were greater than 0.1 individuals. Panels are ordered by the number of individuals (10, 50, 100, 500) sampled from each of the two populations. The proportion of correctly assigned individuals, 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 coloured by effective sample size ($\log_{10}$ scale) of the population. Assignment accuracy in simulation runs with similar genetic differentiation increases with greater effective sample sizes (lighter colours).

had the option to sequence 10 individuals at 6X from a population or 120 individuals at 0.5X, the latter strategy would provide over three times as much information regarding allele frequency estimation despite the same sequencing effort.

3.3 | Application to empirical data

Variant calling with the Yellow warbler samples identified 5,301,627 SNPs. Using all SNPs, 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 even when downsampled to similar read depth levels, suggesting that these samples suffer from factors other than low read depth, such as poor quality DNA or contamination. The number of informative sites per individual varied from 11,866 to 906,505 at full read depth, and from 370 to 3257 at 0.001X, while the total number of sites varied from 955,185 at full depth to 48,220 at 0.001X (Table 1). Evidently, at low read depths, each individual assignment relies on a set of informative SNPs that overlaps little with the informative SNPs in other individuals.

4 | 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.

4.1 | 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

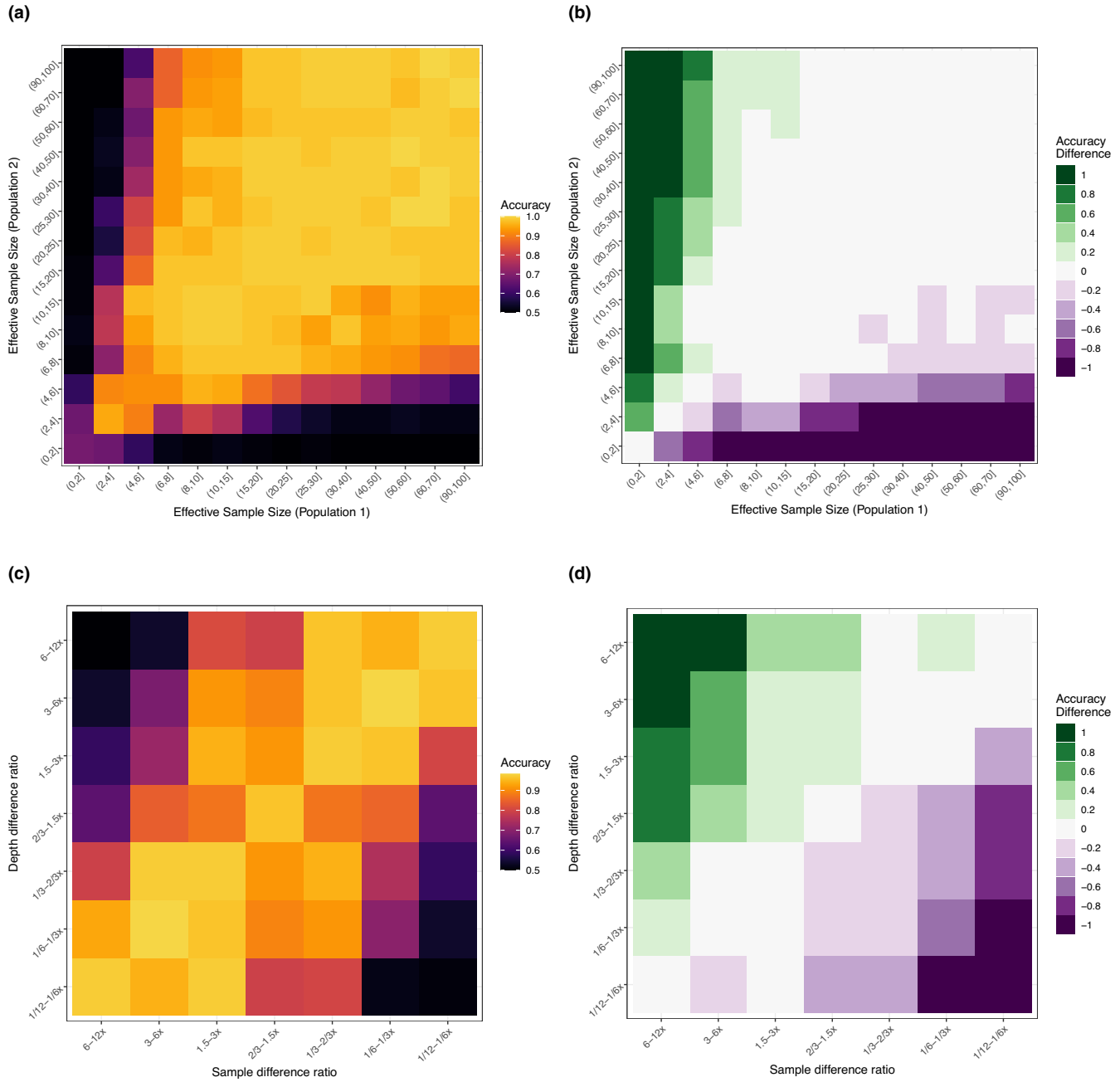

FIGURE 3 Mean assignment accuracy (a) and mean difference in assignment accuracy (b; assignment accuracy of population 2 – population 1) were compared for populations with an array of effective sample sizes, listed on the axes as ranges. Equal effective sample sizes are along the plots' diagonals. Assignment accuracy was high when effective sample sizes were sufficiently high, even when unbalanced. When effective sample sizes were approximately equal, assignment accuracy was high regardless of the combination of read depths and number of samples (c, d). Approximately equal effective sample sizes are found along the diagonal where the axes display a range of the magnitude of difference for depth (*y*-axis) and sample size (*x*-axis) for population 2 in relation to population 1 (e.g. 2 / 3 – 1.5*x* indicates the number of individuals in the sample from population 2 is between two-thirds and three-halves of the sample size from population 1). The centre tile of the plot, 2 / 3 – 1.5*x*, indicates when effective sample size is equal due to approximately similar sample numbers and read depth.

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.

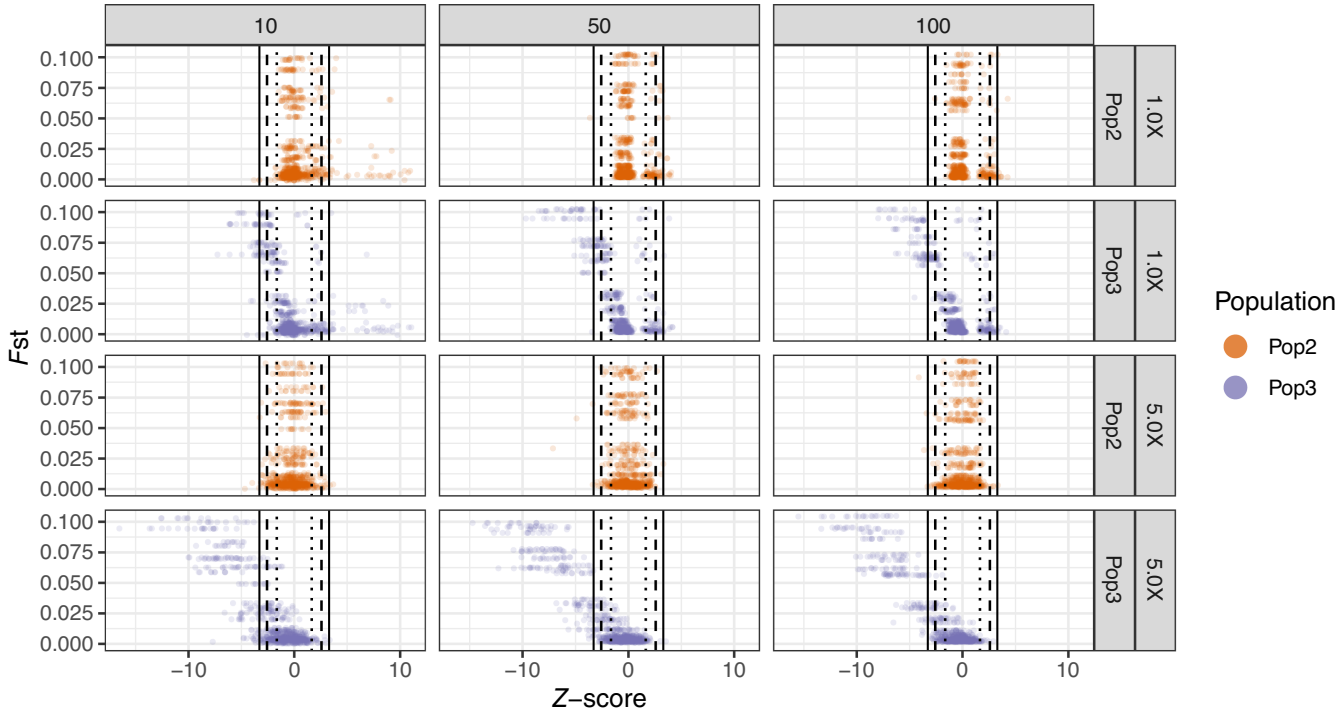

FIGURE 4 Results from the three-population stepping stone model demonstrate the behaviour of the *z*-score metric in identifying individuals from an unsampled population (Pop3) assigned to a population in the reference compared to individuals correctly assigned to their source population of origin (Pop2). The column facets list the number of samples used for the reference populations while the rows are the population of origin and sequenng depth. 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). In this simulation, Pop3 individuals are expected to be incorrectly assigned to Pop2 (since there are no Pop3 individuals in the reference set) and accordingly the z-score metric should depict this by falling outside the mass of the standard unit normal random variate.

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 reference samples 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 towards using it would involve, first, observing the metric's behaviour 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. For example, in our simulations for low to moderate levels of genetic differentiation (mean $F_{ST} = 0.0055$), an effective sample size of roughly eight individuals was sufficient when effective sample sizes were balanced (Figure 3). If the reference populations' effective sample sizes are sufficiently high for the given genetic differentiation, 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 to populations ($F_{ST} = 0.1$) when individual samples had average read depths as low as 0.001X. While the minimum sequencing coverage needed for highly accurate population assignment depends on genetic differentiation, this has powerful implications for population assignment studies, especially those that are conducted at a large scale. For example, in the mid-2000s, 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. Furthermore, using WGS data streamlines the process of adding new reference populations to compare to previous analyses because the loci used for assignment

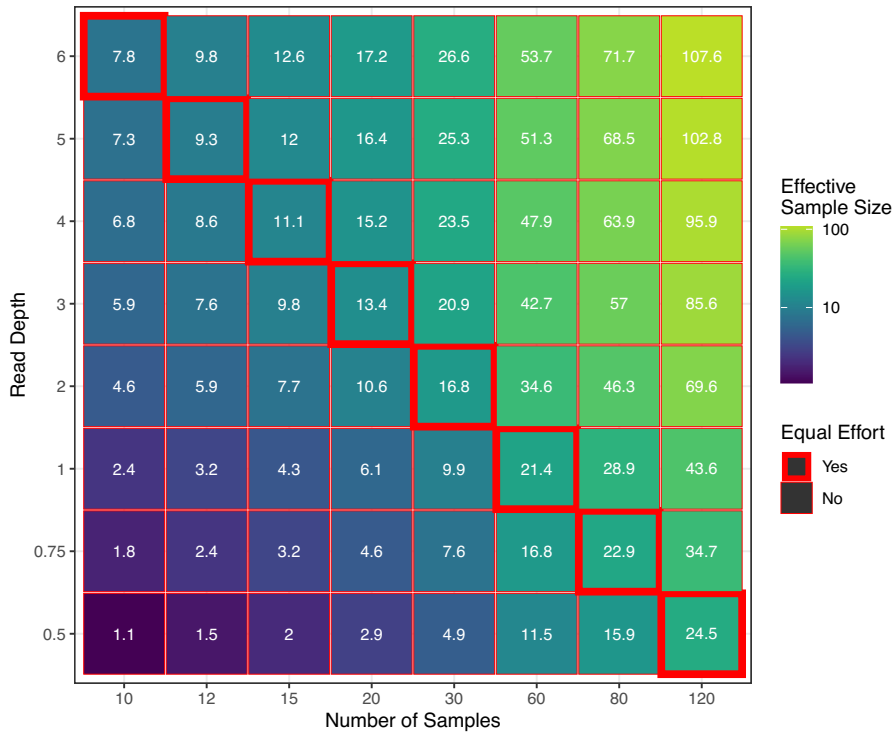

FIGURE 5 The relation between read depth and number of samples in determining the effective sample size for a single population highlights the potential for different sampling design strategies. Notably, effective sample increases more rapidly with changes in number of samples than read depth. The *x*-axis provides the number of samples from a single population, the *y*-axis is the mean read depth for the corresponding population, and the value listed is the mean effective sample size across 10 replicate simulations using all SNPs with minor allele frequency > 0.05 (126,019–158,871). Tiles outlined in red have equal sequencing effort based around 60 individuals at 1X. Given the same amount of sequencing effort, effective sample size increases when the number of samples is prioritized over sampling depth, with a low of 7.8 for 10 individuals at 6X and a high of 24.5 for 120 individuals at 0.5X. Off-diagonal values allow the comparison of sampling design strategies of different sequencing effort, for example, sequencing 20 individuals at 1X (effective sample size = 6.1) versus sequencing 10 individuals at 2X (effective sample size = 4.6).

are not pre-selected to maximize genetic differentiation and thereby potentially subject to ascertainment bias.

We note that WGSassign can be used in conjunction with other clustering approaches for low-coverage WGS data (e.g. PCangsd; Meisner & Albrechtsen, 2018). Notably, the formal population assignment implemented in WGSassign requires a priori delineation of populations. In species that live in discrete population groups, this can be done without genetic data. However, when species are distributed more continuously, then unsupervised clustering approaches in tandem with geography and other covariates (e.g. behaviour, morphology) can be used to delineate reference populations. Assignment accuracy from WGSassign on a set of hold-out individuals can be used to determine if the identified populations are informative for assignment. The use of complementary clustering methods is also informative for identifying if test samples are from populations not represented in the reference samples as well as identifying admixed individuals. Importantly, clustering methods for population structure can be biased by variation in sequencing depth among individuals (Lou et al., 2021), while WGSassign is less influenced by that variation in sequencing depth. Accordingly, WGSassign is expected to give more reliable assignment in the face of sequencing depth variation than unsupervised clustering approaches.

4.2 | 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 and informing study design. 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; Fumagalli, 2013; Lou et al., 2021). 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) at least 10 individuals with 1X average sequencing depth for allele frequency estimation can be quantified as an effective sample size of 2.4 individuals in the simulations from this study (Figure 5) and does correspond to sufficient precision to achieve accurate assignment at moderate genetic differentiation (Figure 3). However, at weaker genetic differentiation among populations, effective sample size needs to be increased for accurate assignment. Quantifying the amount of information gain for different study designs can inform researchers on how to more efficiently allocate resources for sequencing efforts.

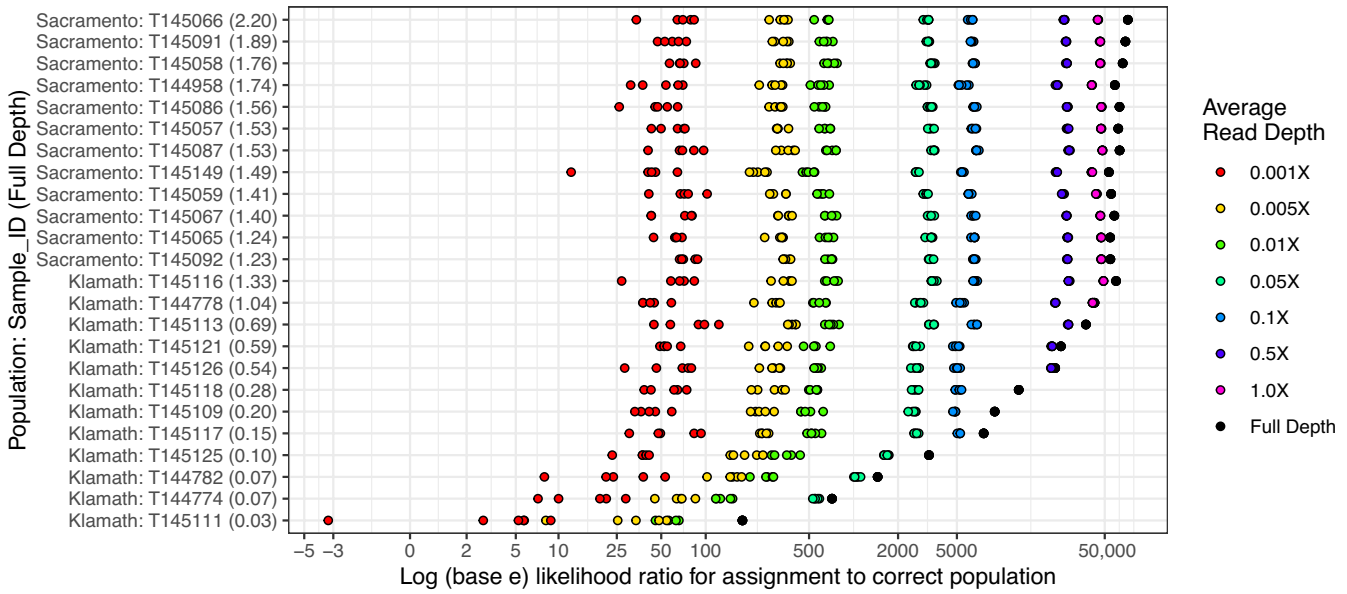

FIGURE 6 Log-likelihood ratios for assignment at different read depth levels for the Chinook salmon data. On the *y*-axis are different Chinook salmon samples, labelled by their population, a colon, their ID number and then in parentheses the average read depth of their aligned data at full depth. On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

TABLE 1 Numbers of informative SNPs (i.e. those covered by at least one read, such that the genotype likelihood is not equal for all three genotypes) at different downsampled coverage levels of the Chinook salmon data.

	Within individuals			
Coverage	Min	Mean	Max	Total
Full depth	11,866	577,982	906,505	955,185
1.0X	49,432	610,137	756,970	955,155
0.5X	31,032	426,405	554,470	955,018
0.1X	31,032	159,077	195,926	884,475
0.05X	11,866	88,712	114,431	734,813
0.01X	3769	21,337	28,044	307,815
0.005X	1882	11,126	14,807	186,384
0.001X	370	2326	3257	48,220

Note: 'Within Individuals' gives the minimum, mean and maximum number of informative SNPs within any single individual across the five downsampled replicates. 'Total' refers to the total number of variant sites in the downsampled data set. Individuals with a full read depth less than one of the downsampled levels, like 1.0X, were excluded from the downsampled data set.

Unbalanced 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 (and which) 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 data sets.

4.3 | 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 specifies a formal prior for the allele frequencies in each population (Rannala & Mountain, 1997). We note that the latter approach yields performance that is very similar to ours; however, implementing a prior for allele frequencies 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) could further improve performance of population assignment. In particular, we expect that it would ameliorate assignment bias with unequal sample sizes and also improve the distribution of posterior probabilities of assignment so that they more closely reflect the amount of uncertainty in each assignment. The parameters of these more complex prior distributions could likely be estimated very accurately using WGS data for use in an empirical Bayes approach (Maritz, 2018); however, we leave that for future research.

AUTHOR CONTRIBUTIONS

Matt DeSaix, Kristen Ruegg and Eric Anderson conceived the ideas; Matt DeSaix and Eric Anderson designed the methodology and developed the software; Matt DeSaix, Marina Rodriguez and Eric Anderson analysed the data; Matt DeSaix and Eric Anderson led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

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 Kristen Ruegg. 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 and to Daniel Wegmann for stimulating discussions. A substantial portion of this manuscript was completed while Matt DeSaix and Eric Anderson were scientists-in-residence at the mobile High Altitude Venue for Ecological Analysis, Genetics and Statistics, on location in Moab, Utah, for 5 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.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

PEER REVIEW

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1111/2041-210X.14286.

DATA AVAILABILITY STATEMENT

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/records/7957898.
- Online version of data and scripts used in paper: https://github.com/mgdesaix/WGSassign-manuscript-data.
- Data available via the Dryad Digital Repository https://doi.org/10.5061/dryad.h9w0vt4pj (DeSaix et al., 2023).

STATEMENT ON INCLUSION

Our study was predominantly based on simulated data.

ORCID

Matthew G. DeSaix https://orcid.org/0000-0002-5721-0311
Eric C. Anderson https://orcid.org/0000-0003-1326-0840

REFERENCES

Bay, R. A., Karp, D. S., Saracco, J. F., Anderegg, W. R., Frishkoff, L. O., Wiedenfeld, D., Smith, T. B., & 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, A. C., & Gompert, Z. (2013). Population genomics based on low coverage sequencing: How low should we go? *Molecular Ecology*, *22*, 3028–3035.

Casella, G., & Berger, R. L. (2021). *Statistical inference*. Cengage Learning.

Cornuet, J. M., 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, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*, giab008.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, *39*, 1–38.

DeSaix, M. (2023). *WGSassign*.

DeSaix, M., Rodriguez, M., Ruegg, K., & Anderson, E. (2023). Data from: Population assignment from genotype likelihoods for low-coverage whole-genome sequencing data. *Dryad Digital Repository*, 10.5061/dryad.h9w0vt4pj.

DeSaix, M. G., Bulluck, L. P., Eckert, A. J., Viverette, C. B., Boves, T. J., Reese, J. A., Tonra, C. M., & Dyer, R. J. (2019). Population assignment reveals low migratory connectivity in a weakly structured songbird. *Molecular Ecology*, *28*, 2122–2135.

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

Fisher, R. A. (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, F. G., Korneliussen, T. S., 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, F. G., Linderoth, T., & Nielsen, R. (2014). ngsTools: Methods for population genetics analyses from next-generation sequencing data. *Bioinformatics*, *30*, 1486–1487.
Gibbs, H. L., Dawson, R. J., & Hobson, K. A. (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, A. M., & McVean, G. (2016). Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS Computational Biology*, *12*, e1004842.
Kim, S. Y., Lohmueller, K. E., Albrechtsen, A., Li, Y., Korneliussen, T., Tian, G., Grarup, N., Jiang, T., Andersen, G., Witte, D., Jorgensen, T., Hansen, T., Pedersen, O., Wang, J., & Nielsen, R. (2011). Estimation of allele frequency and association mapping using next-generation sequencing data. *BMC Bioinformatics*, *12*, 1–16.
Korneliussen, T. S., Albrechtsen, A., & Nielsen, R. (2014). ANGSD: Analysis of next generation sequencing data. *BMC Bioinformatics*, *15*, 1–13.
Korneliussen, T. S., & Moltke, I. (2015). NgsRelate: A software tool for estimating pairwise relatedness from next-generation sequencing data. *Bioinformatics*, *31*, 4009–4011.
Larribe, F., & Fearnhead, P. (2011). On composite likelihoods in statistical genetics. *Statistica Sinica*, *21*, 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, R. N., Jacobs, A., Wilder, A. P., & Therkildsen, N. O. (2021). A beginner's guide to low-coverage whole genome sequencing for population genomics. *Molecular Ecology*, *30*, 5966–5993.
Manel, S., Gaggiotti, O. E., & Waples, R. S. (2005). Assignment methods: Matching biological questions with appropriate techniques. *Trends in Ecology & Evolution*, *20*, 136–142.
Maritz, J. S. (2018). *Empirical Bayes methods with applications*. CRC Press.
McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & DePristo, M. A. (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*, *7*, e37558.
Nielsen, R., Paul, J. S., Albrechtsen, A., & Song, Y. S. (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.
Rannala, B., & Mountain, J. L. (1997). Detecting immigration by using multilocus genotypes. *Proceedings of the National Academy of Sciences of the United States of America*, *94*, 9197–9201.
Ruegg, K. C., Anderson, E. C., Paxton, K. L., Apkenas, V., Lao, S., Siegel, R. B., DeSante, D. F., Moore, F., & Smith, T. B. (2014). Mapping migration in a songbird using high-resolution genetic markers. *Molecular Ecology*, *23*, 5726–5739.
Sayers, E. W., Cavanaugh, M., Clark, K., Pruitt, K. D., Schoch, C. L., Sherry, S. T., & Karsch-Mizrachi, I. (2022). GenBank. *Nucleic Acids Research*, *50*, D161–D164.
Schweizer, T. M., & DeSaix, M. G. (2023). Cost-effective library preparation for whole genome sequencing with feather DNA. *Conservation Genetics Resources*, 1–8. https://doi.org/10.21203/rs.3.rs-1871359/v1
Seeb, L., Antonovich, A., Banks, M. A., Beacham, T., Bellinger, M., Blankenship, S., Campbell, M., Decovich, N., Garza, J., Guthrie Iii, C., Lundrigan, T. A., Moran, P., Narum, S. R., Stephenson, J. J., Supernault, K. J., Teel, D. J., Templin, W. D., Wenburg, J. K., Young, S. F., & Smith, C. T. (2007). Development of a standardized DNA database for Chinook salmon. *Fisheries*, *32*, 540–552.
Skotte, L., Korneliussen, T. S., & Albrechtsen, A. (2013). Estimating individual admixture proportions from next generation sequencing data. *Genetics*, *195*, 693–702.
Smouse, P. E., Waples, R. S., & Tworek, J. A. (1990). A genetic mixture analysis for use with incomplete source population data. *Canadian Journal of Fisheries and Aquatic Sciences*, *47*, 620–634.
Thompson, N. F., Anderson, E. C., Clemento, A. J., Campbell, M. A., Pearse, D. E., Hearsey, J. W., Kinziger, A. P., & Garza, J. C. (2020). A complex phenotype in salmon controlled by a simple change in migratory timing. *Science*, *370*, 609–613.
Wang, J. (2017). The computer program structure for assigning individuals to populations: Easy to use but easier to misuse. *Molecular Ecology Resources*, *17*, 981–990.

How to cite this article: DeSaix, M. G., Rodriguez, M. D., Ruegg, K. C., & Anderson, E. C. (2024). Population assignment from genotype likelihoods for low-coverage whole-genome sequencing data. *Methods in Ecology and Evolution*, *00*, 1–18. https://doi.org/10.1111/2041-210X.14286

APPENDIX A: FISHER INFORMATION

Fisher information from genotype likelihoods

We focus on the information for the ℓth locus in the kth reference population. Accordingly, we drop the k,ℓ subscript from θ and the ℓ 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 ith 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:

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

Let

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

and note that

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

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

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

Proceeding, define v and w as follows:

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

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

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

Then, we put them together with the product rule

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

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

Expected fisher information from observed genotypes

Under Hardy–Weinberg equilibrium, the allelic type of the two gene copies within a locus is 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 θ. 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

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

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

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

Since information from independent variables is additive, the information for $2n$ such Bernoulli variables is $2n\left[\theta(1-\theta)\right]^{-1}$. Evaluating the expectation under the assumption that the true value of θ is $\widehat{\theta}_{k,\ell}$ gives $I_e\left(\theta_{k,\ell}\right)$ in Equation (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, $\log P\left(R^{(A)}|\,\theta_k\right)$, 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 $\log P\left(R^{(A)}|\,\theta_k\right)$ to the distribution of $\log P\left(R|\,\theta_k\right)$ 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 $P\left(R|\,\theta_k\right)$ is a product over many loci, $\log P\left(R|\,\theta_k\right)$ is a sum over loci. We will write the contribution of each locus to that sum as:

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

where we include the notation $f\left(g_\ell, \theta_{k,\ell}\right)$ to emphasize the fact that W_ℓ is a deterministic function of $\theta_{k,\ell}$ and the vector of genotype likelihoods $g_\ell = \left(g_{\ell,0}, g_{\ell,1}, g_{\ell,2}\right)$. It is important to recognize in this context that $\theta_{k,\ell}$ is considered fixed while g_ℓ is a random variable. By extension, then, so too is W_ℓ a random variable. By the CLT, the sum of very many independent W_ℓ random variables can be approximated by a normal distribution with mean μ and variance σ^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 $\mathrm{Var}(W_\ell)$.

With these assumptions, given the total read depth D_ℓ, and γ and ϵ_k, the joint probability of the remaining variables is:

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

The mean and the variance of W_ℓ can now be found from these by taking expectations:

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

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

The distribution of W_ℓ clearly depends on the distribution of g_ℓ. We develop such a distribution, hierarchically, based on the following assumptions:

1. g_ℓ depends directly on the observed allele depths. Let r_ℓ be the number of reference alleles and a_ℓ the number of alternate alleles observed in the reads covering site ℓ, and let γ denotes an individual-specific effect of base quality scores, etc., on the genotype likelihoods. Then, we denote this conditional probability distribution as $P(g_\ell \mid r_\ell, a_\ell, \gamma)$ and we will denote the set of values that g_ℓ might take for a given pair (r, a) as $\mathcal{G}_{r,a}$. 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 do not 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_ℓ and a_ℓ depend on the genotype, G^*_ℓ at locus ℓ of the individual being sequenced and on a population-specific error rate, ϵ_k. The model for this is simple binomial random sampling from a total read depth of D_ℓ, with a probability ϵ_k, independently for each read, that the base in question will be read incorrectly. Hence:

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

   where $a_\ell = D_\ell - r_\ell$, always. (We note that r_ℓ and D_ℓ completely determine a_ℓ, but we leave both r_ℓ and a_ℓ in the preceding and following probability expressions for ease of explanation later.)
3. The frequency of G^*_ℓ in source population k follows Hardy–Weinberg equilibrium with an allele frequency of $\theta_{k,\ell}$, so $P(G^*_\ell \mid \theta_{k,\ell})$ is given by Equation (1).

As there is no documented distribution for $P(g_\ell \mid r_\ell, a_\ell, \gamma)$, we simply use the empirical distribution of g_ℓ values across all loci within the individual having allele depths of r and a. In practice, values of 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 g, thus imagining $P(g_\ell \mid r_\ell, a_\ell, \gamma)$ as a discrete distribution with weight on b values of g, each one the mean of the values of g within the bin. It is also possible to remove loci that have particularly odd values of g. For example, GATK sometimes assigns a g_ℓ 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 μ and σ^2 that we seek are conditioned upon a set of loci. The parameter ϵ_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 g) from population k would be expected to have:

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

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

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

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

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

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

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

As in the likelihoods calculated by WGSassign, values of $\tilde{\theta}_{k,\ell}$ are used in place of values of $\theta_{k,\ell}$ in all of the above calculations. Furthermore, when calculating the z scores for each individual from the reference samples, the value of $\theta_{k,\ell}$ used must be one estimated while leaving the individual out of the sample (analogous to the LOO procedure described in the paper).