Molecular Biology and Evolution, 2025, **42**, 1–11
https://doi.org/10.1093/molbev/msaf002
Advance access publication 23 January 2025
Methods

MBE OXFORD

Modeling Biases from Low-Pass Genome Sequencing to Enable Accurate Population Genetic Inferences

Emanuel M. Fonseca ,* Linh N. Tran , Hannah Mendoza, Ryan N. Gutenkunst *

Department of Molecular and Cellular Biology, University of Arizona, Tucson, AZ 85721, USA

***Corresponding authors**: E-mails: emanuelmfonseca@arizona.edu; rgutenk@arizona.edu.

Associate editor: Russell Corbett-Detig

Abstract

Low-pass genome sequencing is cost-effective and enables analysis of large cohorts. However, it introduces biases by reducing heterozygous genotypes and low-frequency alleles, impacting subsequent analyses such as model-based demographic history inference. Several approaches exist for inferring an unbiased allele frequency spectrum (AFS) from low-pass data, but they can introduce spurious noise into the AFS. Rather than correcting the AFS, here, we developed an approach that incorporates low-pass biases into the demographic modeling and directly analyzes the AFS from low-pass data. Our probabilistic model captures biases from the Genome Analysis Toolkit multisample calling pipeline, and we implemented it in the population genomic inference software dadi. We evaluated the model using simulated low-pass datasets and found that it alleviated low-pass biases in inferred demographic parameters. We further validated the model by downsampling 1000 Genomes Project data, demonstrating its effectiveness on real data. Our model is widely applicable and substantially improves model-based inferences from low-pass population genomic data.

Keywords: demography inference, inbreeding, low-pass sequencing, allele frequency spectrum, GATK multisample calling

Introduction

Enabled by reduced sequencing costs, population genetics has experienced a revolution, from focusing on a limited number of loci to now encompassing entire genomes (Maddison et al. 1992; Reid et al. 2016; Marchi et al. 2022). Yet researchers must still trade off (i) the extent of the genome to be sequenced, (ii) the depth of coverage for each sample, and (iii) the number of sequenced samples (Lou et al. 2021; Martin et al. 2021; Duckett et al. 2023). One way to address this trade off is to sequence one reference sample at high coverage depth while sequencing others at lower depth (Lou et al. 2021). Low-pass sequencing, in which the genome is sequenced at a lower depth of coverage, avoids many of the financial, methodological, and computational challenges of high-pass sequencing (Li et al. 2011). Furthermore, limited availability of DNA can also make high depth impractical, especially for ancient samples and museum or herbarium specimens (Mota et al. 2023).

Despite its advantages, low-pass sequencing may lead to an incomplete and biased representation of genetic diversity within a population (e.g. Vieira et al. 2013; Han et al. 2014; Fox et al. 2019). Low-frequency genomic variants may not be detected (Fumagalli 2013), and genotypes may be less accurate (Nielsen et al. 2011). Low-pass sequencing increases the likelihood of miscalling heterozygous loci as homozygous (Duitama et al. 2011; Gorjanc et al. 2015), due to a lack of sufficient reads on homologous chromosomes to distinguish between different alleles at a given locus. These issues can then bias downstream analyses. It is thus important for analysis methods to accommodate low-pass sequencing (see Carstens et al. 2022 for a discussion of related issues).

The allele frequency spectrum (AFS) is a powerful summary of population genomic data (Sawyer and Hartl 1992; Wakeley 2009). Briefly, the AFS is matrix which records the number alleles observed at given frequencies in a sample of individuals from one or more populations. The AFS is often the basis for inferring demographic history (Gutenkunst et al. 2009) or distributions of fitness effects (Kim et al. 2017). In low-pass sequencing, the loss of alleles and the excess of homozygosity can bias the estimation of the AFS (Fumagalli 2013) and thus those inferences.

To address the challenges of low-pass data, several tools have emerged (Bryc et al. 2013; Blischak et al. 2018; Meisner and Albrechtsen 2018) to estimate the AFS from low-pass data. One of the most widely adopted is ANGSD (Korneliussen et al. 2014), which offers a diverse range of analyses tailored for low-pass sequencing data. To infer an AFS, ANGSD uses sample allele frequency likelihoods, which can be computed either directly from raw data or, more frequently, from genotype likelihoods (Nielsen et al. 2012). These likelihoods quantify the probability of observing the complete set of read data for multiple individuals at specific genomic sites, given particular sample allele frequencies (Nielsen et al. 2012; Korneliussen et al. 2014), enabling ANGSD to estimate allele frequencies. However, as the number of samples increases, ANGSD becomes computationally inefficient and numerically unstable (Han et al. 2015). To address this, a score-limited dynamic programming algorithm was introduced, offering significantly greater efficiency by scaling linearly with the number of genomes, unlike ANGSD's quadratic complexity (Han et al. 2015). Another method, winsfs, uses a stochastic expectation–maximization (EM) algorithm developed to address common

Received: July 19, 2024. **Revised:** January 6, 2025. **Accepted:** January 7, 2025

problems like overfitting, high-memory usage, and long runtime associated with standard methods for estimating AFS from low-pass sequencing data (Rasmussen et al. 2022). A related tool, the Bayesian genotype caller, calls genotypes from high-throughput sequencing, including low-pass sequencing, using population-level information and sequencing error rates to improve accuracy (Maruki and Lynch 2017).

While tools for inferring an unbiased AFS have demonstrated substantial utility, they also present notable limitations. For instance, many analyses depend on differentiating between variant types, such as synonymous versus nonsynonymous sites. But such differentiation requires calling genotypes, which the developers of ANGSD advise against (Korneliussen et al. 2014). In cases of low or moderate sequencing depth, AFS inference process can introduce significant uncertainty due to limited data, potentially leading to errors or biases in downstream analyses (Lou et al. 2021). Furthermore, we observe here that these tools sometimes fall short of fully correcting for biases inherent in low-pass sequencing.

Rather than attempting to estimate an unbiased AFS from low-pass data, we developed a probabilistic model of low-pass AFS biases. We incorporated it into the population genomic inference software dadi (Gutenkunst et al. 2009), so the biased low-pass AFS can be directly analyzed. Our model is based on the multisample genotype calling pipeline of the Genome Analysis Toolkit (GATK), the most widely used tool for calling variants from read data (McKenna et al. 2010; Van der Auwera and O'Connor 2020). We assessed the accuracy of our model using simulated low-depth data as well as subsampled data from the 1000 Genomes Project (Fairley et al. 2020, https://www.internationalgenome.org/). We found that our model accurately captures low-depth biases in the AFS and enables accurate inference of demographic history from low-pass data.

Model for Low-Pass Biases

Our approach is to incorporate low-pass sequencing biases into the model inference process. From a user perspective, our approach begins when the AFS is computed from an input Variant Call Format (VCF) file containing population genomic data. While parsing the VCF, population-specific distributions for depth of coverage per site per individual are also extracted. (Note that our model does not incorporate quality scores, beyond the filtering done by GATK. The defaults for those filters ensure only high-quality reads are included in our analysis.) These distributions are passed to a function which wraps the demographic model function implemented in dadi. The demographic model function takes in an assumed set of demographic model parameters and returns the corresponding model AFS. The wrapper function takes as fixed inputs n_{seq}, the number of sequenced individuals; n_{sub}, the number of individuals the AFS is subsampled down to; and $\mathbb{D}(d)$, the distribution of sequencing depths. (See Table 1 for a summary of key mathematical notation.) The wrapper function acts on the AFS returned by the demographic model function and applies the transformations described below to generate an AFS that is biased by low-pass sequencing. The user uses this wrapped function in place of the original demographic model function during parameter optimization, to maximize the composite likelihood of their (biased) AFS data.

In the dadi command-line interface, dadi-cli (Huang et al. 2023), low-pass bias correction is seamlessly incorporated into the workflow. When using GenerateFs, users can specify—calc-coverage to create a .coverage.pickle file containing per-sample depth of coverage data. During demographic inference, the InferDM command accepts this file through the—coverage-model flag, allowing the LowPass model to adjust the AFS based on observed coverage, thereby addressing low-pass sequencing biases directly within the optimization process. Integration with dadi and dadi-cli require no additional free parameters, enabling an efficient yet thorough adjustment for low-pass bias correction developed in this study.

Table 1. Key mathematical notation.

n_{seq}	Number of sequenced individuals
n_{sub}	Number of individuals AFS is subsampled to
$\mathbb{D}(d)$	Distribution of sequencing depths d
f	True alternate allele count
$\mathbb{P}_{n\text{seq}}(f)$	Genotype partition function
n_0, n_1, n_2	Numbers of true reference homozygotes, heterozygotes, and alternate homozygotes
$P_a(n)$	Probability of observing n alternative reads
$P^{\text{het}}_{\text{mis}}$	Probability a heterozygous individual is miscalled

When biases arises from low-pass sequencing, the AFS may be affected by both the loss of low-frequency variants and the misidentification of heterozygous individuals as homozygous. These two effects result in a deficit of variant sites and misleading shifts in allele frequencies, respectively. Moreover, the data must often be subsampled to generate an AFS for analysis, because not all individuals will be called at all sites. We account for these distortions by sequentially modeling the probabilities of a variable site being called, of that site having enough called individuals for subsampling, and of having its allele frequency misestimated.

The specific choices in our model are motivated by the default GATK multisample calling algorithm, in which information from all samples is used to identify whether a site is variant. In particular, we assume that a site will only be called as variant if at least two alternate allele reads are observed. Once a site is identified as variant, an individual will be called as missing if zero reads are observed, homozygous if all reads correspond to a single allele, and heterozygous if at least one reference and one alternate read are observed. For simplicity, we first describe the case of sequencing n_{seq} individuals from a single population.

First, we calculate the probability that a true variant site is called as variant. Consider a site in which the true alternate allele count within our sample of n_{seq} individuals is f. Those f alternate alleles can be distributed among the $2n_{\text{seq}}$ sampled alleles in many ways. To quantify those ways, we define the partition function $\mathbb{P}_{n\text{seq}}(f)$, which is an array of integer partitions with n entries that sum to the allele frequency f such that all entries in the partition are 0, 1, or 2 (corresponding to the possible genotype values). For example, the partitions defined by $\mathbb{P}_4(3)$ are [2, 1, 0, 0] and [1, 1, 1, 0]. Each possible partition within $\mathbb{P}_{n\text{seq}}(f)$ can occur in $\frac{n!}{n_0!n_1!n_2!}2^{n_1}$ ways, where n_0, n_1, and n_2 denote the number of partition entries equal to 0, 1, or 2. (The factor of 2^{n_1} accounts for the two possible haplotypes the alternate allele could lie on in each heterozygote.) The corresponding probability of each partition within $\mathbb{P}_{n\text{seq}}(f)$ is then the number of ways it can occur divided by the total over all partitions within $\mathbb{P}_{n\text{seq}}(f)$. The relation between the sample allele frequency and potential genotypes has previously been explored to develop exact tests of Hardy–Weinberg equilibrium (Wigginton et al. 2005).

Let $\mathbb{D}$ denote the distribution of read depth d within the population sample, which we assume to be shared among all individuals. For an individual homozygous for the alternate allele, the probability of observing a alternate reads is simply $P_a^{\text{hom}}(a) = \mathbb{D}(a)$. For a heterozygous individual, the probability of zero alternate reads is

$$P_a^{\text{het}}(0) = \sum_d \mathbb{D}(d)\left(\frac{1}{2}\right)^d. \tag{1}$$

Here, we sum over the distribution of depths, and at each depth each read has a 1/2 chance of containing the reference allele, so the probability of all reads being reference is $(1/2)^d$. Similarly, the probability of exactly one alternate read is

$$P_a^{het}(1) = \sum_d \binom{d}{1}\mathbb{D}(d)\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)^{d-1}. \tag{2}$$

Note that for depth d, there are d possible configurations with one alternate read and $d-1$ reference reads.

For a given partition within $\mathbb{P}_{n\text{seq}}(f)$ that has true genotype counts n_0, n_1, and n_2, there are multiple ways of failing to identify the variant site. The probability of zero reads supporting the alternate allele is

$$P_a^{\text{part}}(0) = P_a^{\text{het}}(0)^{n_1}\ P_a^{\text{hom}}(0)^{n_2}. \tag{3}$$

The probability of exactly one read supporting the alternate allele is

$$\begin{aligned} P_a^{\text{part}}(1) = {} & n_1 P_a^{\text{het}}(1) P_a^{\text{het}}(0)^{n_1-1}\ P_a^{\text{hom}}(0)^{n_2} \\ & + P_a^{\text{het}}(0)^{n_1}\ n_2 P_a^{\text{hom}}(1) P_a^{\text{hom}}(0)^{n_2-1}. \end{aligned} \tag{4}$$

Here, the two terms account for the probability that the alternate read occurs in one of the heterozygotes or homozygotes, respectively. The overall probability of not calling a variant site for a given partition is thus $P_a^{\text{part}}(0) + P_a^{\text{part}}(1)$. And the overall probability of not calling a variant site with a given true allele frequency f is the sum of these probabilities over partitions $\mathbb{P}_{n\text{seq}}(f)$, weighted by the partition probabilities. For any given coverage distribution, the probability of calling a variant site increases rapidly with allele frequency f (supplementary fig. S1, Supplementary Material online).

Second, we calculate the probability that a site has enough called individuals to enter the AFS. When analyzing low-pass data, generating an AFS for the full sample size n_{seq} may result in the loss of many sites where not all individuals were called. Consequently, it is common to subsample the data to some lower sample size n_{sub}; only sites with calls for at least n_{sub} individuals can then be analyzed. To calculate the probability that a site can be subsampled, we want to calculate the probability that at least n_{sub} individuals have at least one read, conditional on the site having been called as variant. This conditional probability is complex, so we make a simple approximation. If a site has been called as variant, then at least one individual was called, so we sum the probability that an additional c individuals are called out of the remaining $n_{\text{seq}} - 1$, where c is at least $n_{\text{sub}} - 1$:

$$\sum_{c=n_{\text{sub}}-1}^{n_{\text{seq}}-1} \frac{(n_{\text{seq}}-1)!}{c!(n_{\text{seq}}-1-c)!}\mathbb{D}(0)^{n_{\text{seq}}-1-c}\ (1-\mathbb{D}(0))^c. \tag{5}$$

The projection of the model AFS from sample size n_{seq} down to sample size n_{sub} is already implemented in dadi, based on sampling without replacement as initially described by Marth et al. (2004). Equation (5) is an additional overall factor by which the AFS is reduced, due to sites that cannot be projected downward because they do not have at least n_{sub} called individuals. From this point onward, we consider partitions $\mathbb{P}_{\text{nsub}}(f')$ over the subsampled individuals, where f' is the true alternate allele count in the subsample.

Lastly, we calculate the distortion in estimated allele frequencies. Once a site as called as variant, low-pass sequencing can bias the estimation of the allele frequency at that site, if one or more heterozygotes are miscalled because all their reads are reference or alternate. For each heterozygous individual, this occurs with total probability

$$P_{\text{mis}}^{\text{het}} = 2\sum_{d\geq 1}\mathbb{D}'(d)\left(\frac{1}{2}\right)^d, \tag{6}$$

where $\mathbb{D}'(d) = \mathbb{D}(d)/\sum_{d\geq 1}\mathbb{D}(d)$, is the distribution of depths conditioned on having at least one read. For a partition with n_1 true heterozygotes, the number of miscalled heterozygotes $N_{\text{mis}}^{\text{het}}$ is binomially distributed with mean $n_1 P_{\text{mis}}^{\text{het}}$. Each miscalled heterozygote has equal chance of being called as homozygous reference or alternate, so the number of miscalls to homozygous reference $N_{\to\text{ref}}^{\text{het}}$ is binomially distributed with mean $N_{\text{mis}}^{\text{het}}/2$, and the number of miscalls to homozygous alternate is $N_{\to\text{alt}}^{\text{het}} = N_{\text{mis}}^{\text{het}} - N_{\to\text{ref}}^{\text{het}}$. The net change in estimated alternative allele frequency is then $N_{\to\text{alt}}^{\text{het}} - N_{\to\text{ref}}^{\text{het}}$.

The biases caused by low-pass sequencing do not depend on the underlying AFS; for each true allele frequency a given fraction will always, on average, be miscalled as any given other allele frequency. The correction above can be thus be calculated once for a given data set then applied to all model AFS generated, for example, during demographic parameter optimization. For efficiency, we calculate and cache an n_{seq} by n_{nub} transition matrix that can be multiplied by any given model AFS for n_{seq} individuals to apply the low coverage correction. When analyzing multiple populations, we calculate and apply transitions matrices for each population, because variant calling is independent among populations once a variant has been identified. Variant identification is, however, not independent among populations, which we address using simulated calling described next.

When calculating the probability of miscalling a heterozygote (Equation (6)), the correct distribution of depth is not simply $\mathbb{D}(d)$; rather it is the distribution conditional on the site being identified as variant. The lower the true sample alternate allele count, the more these distributions will differ. The conditional distribution is complex to calculate, particularly when multiple populations are involved. Instead, for true allele frequencies for which the probability of not identifying is above a user-defined threshold (by default 10^{-2}), we simulate the calling process rather than using our analytic results. For multiple populations, we calculate this threshold assuming that a variant must be identified independently in all populations, which gives a lower bound on the true probability of not identifying. To simulate calling, for a given true allele frequency (or combination in the multipopulation case) we simulate reads (default 1,000) using the coverage distribution $\mathbb{D}(d)$ and simulate variant identification and genotype calling for each potential partition of genotypes across the populations, proportional to its probability. For each combination of input true allele frequencies simulated, we estimate and store probability of each potential output allele frequency. These distortions are then applied in place of the transition matrices from the analytic model.

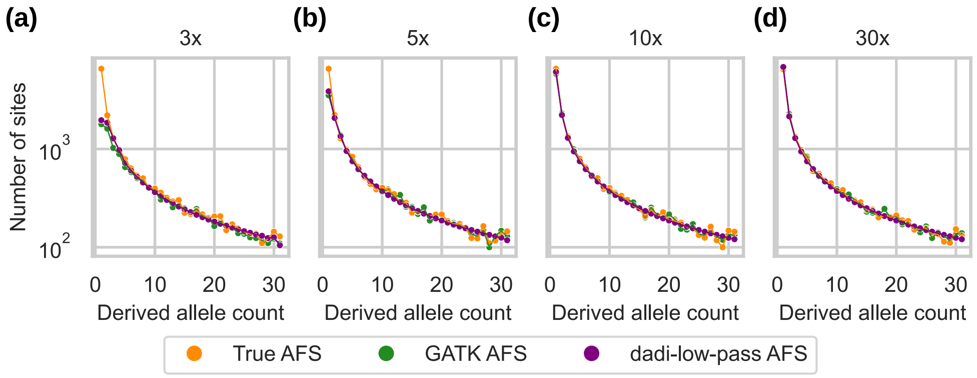

Fig. 1. The low-pass AFS is biased, which our model captures. Simulated sequence data from an exponential growth demographic model for 20 individuals were called by GATK and subsampled to 16 individuals (to accommodate missing data at low depth). The GATK-called AFS (green) is biased compared to the true AFS (orange), and our dadi model for low-pass sequencing (purple) fits those biases well. Coverage was a) 3×, b) 5×, c) 10×, and d) 30×.

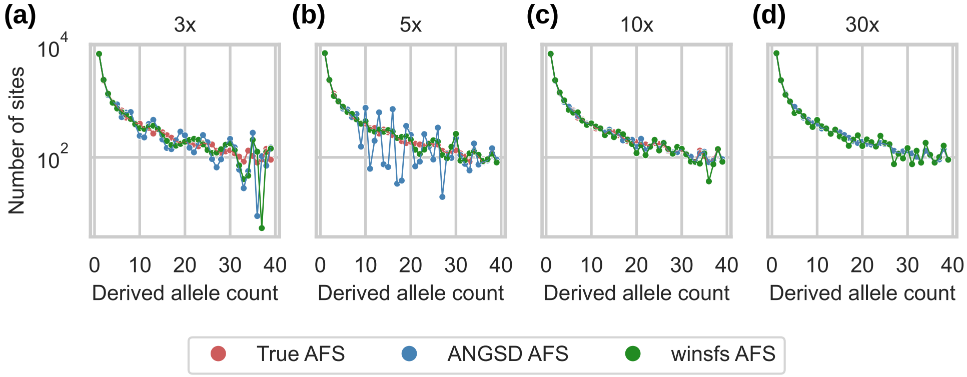

Fig. 2. ANGSD and winsfs correct for low-pass bias of the AFS, but ANGSD introduces fluctuations. For the same simulations as Fig. 1, ANGSD (blue) and winsfs (green) were used to reconstruct the true AFS (red). Coverage was a) 3×, b) 5×, c) 10×, and d) 30×.

For inbred populations, there is an excess of homozygotes compared to the Hardy–Weinberg expectation, which reduces biases associated with low-pass sequencing. In this case, we follow Blischak et al. (2020) and within each genotype partition calculate the probability of reference homozygotes, heterozygotes, and alternate homozygotes using results from Balding and Nichols (1995, 1997), given the inbreeding coefficient F. The partition probability is then multinomial given these probabilities. In these calculations, we approximate the population allele frequency by the true sample allele frequency. Because calculation of the low-pass correction is expensive compared to typical normal model AFS calculation, we precalculate and cache transition matrices and calling simulations. But inbreeding is often an inferred model parameter, to be optimized during analysis. In this case, users can specify an assumed inbreeding parameter for the low-pass model, optimize the inbreeding parameter in their demographic model, update the inbreeding coefficient assumed in the low-pass model, and iterate until convergence.

Results

Low-Pass Sequencing Biases the AFS

We used simulated data to assess the biases introduced by low-pass sequencing with GATK multisample calling, along with our model of those biases. For a simulated population undergoing growth (supplementary fig. S2a, Supplementary Material online), low-pass sequencing reduces the number of observed low-frequency alleles (Fig. 1). Our model accurately captures these biases (Fig. 1). In contrast with our model, ANGSD attempts to reconstruct the true AFS from low-pass data. In our simulations, ANGSD reconstructed the mean shape of the AFS well, but it introduced dramatic fluctuations into the reconstructed AFS at low depth (Fig. 2). winsfs performed better than ANGSD, producing smoother AFS estimates and reducing errors at low coverage (Fig. 2).

When a pair of populations undergoing a split and isolation (supplementary fig. S2b, Supplementary Material online) is analyzed through a joint AFS, similar low coverage biases occur (supplementary fig. S3, Supplementary Material online). Again, our model corrects those biases well (supplementary fig. S3, Supplementary Material online). Similar to the single-population case, ANGSD also introduces large fluctuations in the joint AFS (supplementary fig. S4, Supplementary Material online). While winsfs enhances AFS estimation at low depth of coverage, it tends to introduce more noise as the depth of coverage increases (supplementary fig. S4, Supplementary Material online).

Low-pass biases are expected to be smaller in inbred populations, due to the reduction of heterozygosity. In a simulated population recovering from a bottleneck with inbreeding (supplementary fig. S2c, Supplementary Material online), biases are still observed, which our model corrects

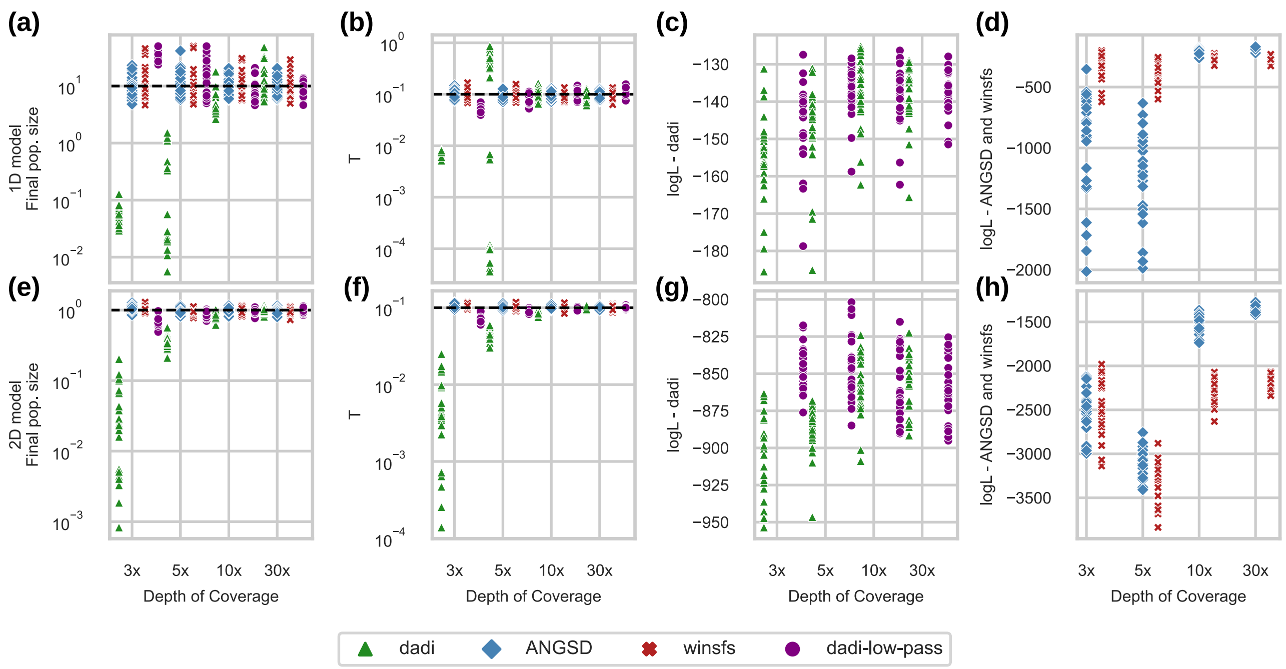

Fig. 3. Our low-pass model, ANGSD and winsfs enable accurate demographic parameter inference. a, b) From data simulated under a single-population growth model, the final population size and time of growth onset (*T*) were accurately inferred using our low-pass model and a GATK-called AFS or using normal dadi and an ANGSD and winsfs-called AFS. But they were biased if low depth was not accounted for when fitting a GATK-called AFS. (Dashed horizontal lines are simulated true values.) c) The likelihoods using the GATK AFS were similar whether or not low-pass biases were modeled. d) The fluctuations introduced into the AFS by ANGSD and winsfs caused low likelihoods at low depth of coverage for ANGSD. e–h) For a two-population split with isolation model, similar results were found, although inferences from our low-pass model were slightly biased at 3× coverage.

(supplementary fig. S5, Supplementary Material online). Again, ANGSD introduced large fluctuations in low-pass AFS, beyond those expected from inbreeding (supplementary fig. S6, Supplementary Material online).

Demographic History Inference from Low-Pass AFS

To assess effects on inference, we first fit demographic models to single-population data simulated under the same growth model as our prior simulations (supplementary fig. S2a, Supplementary Material online). When not modeling low-pass biases, the final population size was underestimated (Fig. 3a), consistent with a deficit of low-frequency alleles. The timing of growth onset was also inaccurately inferred, underestimated at 3× depth and overestimated at 5× depth (Fig. 3b). When the same data were fit with our low-pass model, both model parameters were accurately recovered (Fig. 3a, b) even at the lowest depth. Fits to the AFS reconstructed by ANGSD and winsfs also yielded accurate model parameters (Fig. 3a, b).

The logarithm of the likelihood is commonly used to assess the quality of model fit. ANGSD and winsfs reconstruct the AFS for the full sequenced sample size, while we subsample in our approach to deal with missing genotypes, so the likelihoods are not directly comparable. The likelihoods of models fit to the subsampled GATK data were similar whether or not low-pass biases were modeled (Fig. 3c), suggesting that the likelihood itself cannot be used to detect unmodeled low-pass bias. When fitting AFS estimated by ANGSD, likelihoods were much lower at low coverage than high coverage (Fig. 3d), likely driven by the fluctuations ANGSD introduced into the estimated AFS (Fig. 2). Conversely, likelihood estimates were more stable with winsfs, as the variance in likelihoods decreased with increasing depth of coverage (Fig. 3d).

For two-population data simulated under an isolation model (supplementary fig. S2b, Supplementary Material online), similar results were found. Fitting the observed low-pass AFS with our model enabled accurate parameter inference (although there was some bias at 3× coverage) as did fitting the AFS estimated by ANGSD and winsfs (Fig. 3e, f). As in the single-population case, likelihoods were substantially lower when fitting the ANGSD and winsfs-estimated AFS, consistent with introduced fluctuations in the AFS (Fig. 3g, h). For both models, at 3× coverage, ANGSD and winsfs showed slightly better results than dadi-low-pass in parameter inference (Fig. 3a, e).

For one-population data simulated under a growth model with inbreeding (supplementary fig. S2c, Supplementary Material online), failing to correct for low-pass biases at low inbreeding ($F = 0.1$ or $F = 0.5$) led to similar biases as with no inbreeding, which our low-pass model corrected (supplementary fig. S7, Supplementary Material online). For high inbreeding ($F = 0.9$), the impact of low-pass sequencing on accuracy was smaller, because inbreeding reduces heterozygosity (supplementary fig. S7, Supplementary Material online).

When applying our low-pass bias correction, the user must specify a value for inbreeding, while they may separately estimate it during demographic parameter optimization. We tested the impact of misspecifying inbreeding in the low-bias correction using data simulated with moderate inbreeding of $F = 0.5$. Large inbreeding values were inferred if inbreeding was initially underestimated in the low-coverage model, and small values were inferred if inbreeding was initially overestimated (supplementary fig. S8c, Supplementary Material online). A substantial difference between the inbreeding coefficient used for correction and the inferred value thus suggests that the assumed inbreeding coefficient was

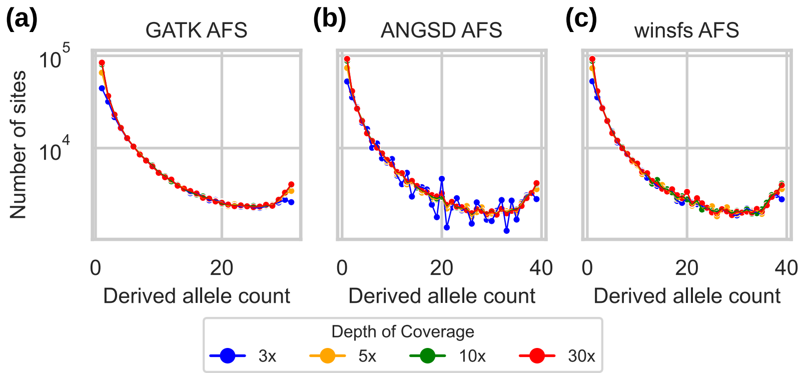

Fig. 4. Allele frequency spectra from 20 YRI samples versus subsampled sequencing depth. a) Spectra generated using the GATK pipeline and subsampled to 32 haplotypes to accommodate missing genotypes. b) Spectra generated using ANGSD genotype likelihood optimization with BAM files input. c) Spectra generated using winsfs genotype likelihood optimization.

Table 2. One-population YRI model analysis results.

			Depth			
Parameter	AFS	Model	30×	10×	5×	3×
ν_{YRI}	GATK	dadi	1.82	1.76	1.55	0.05
	GATK	Low-pass	1.83	1.79	1.74	1.57
	ANGSD	dadi	1.82	1.83	1.71	0.08
	winsfs	dadi	1.80	1.84	1.72	0.08
T	GATK	dadi	0.43	0.51	0.91	0.001
	GATK	Low-pass	0.43	0.47	0.47	0.44
	ANGSD	dadi	0.60	0.73	1.02	0.001
	winsfs	dadi	0.55	0.74	1.03	0.001
$\theta\ (\times 10^4)$	GATK	dadi	5.14	5.07	4.77	5.99
	GATK	Low-pass	5.15	5.11	5.12	5.16
	ANGSD	dadi	5.48	5.24	4.93	6.80
	winsfs	dadi	5.52	5.21	4.90	6.81
Log-likelihood	GATK	dadi	−289	−258	−557	−1,301
	GATK	Low-pass	−294	−289	−327	−347
	ANGSD	dadi	−467	−495	−1,188	−5,674
	winsfs	dadi	−497	−584	−1,207	−3,443

Inferred demographic parameters in dadi using empirical GATK and ANGSD AFS. We analyzed GATK empirical spectra without (dadi) and with low-pass correction (low-pass).

not optimal. Users can thus iterate and update the value assumed in the low-pass correction to converge to a best inference of inbreeding.

Analysis of Human Data

To empirically validate our approach and compare with ANGSD and winsfs, we used chromosome 20 sequencing data from the 1000 Genomes Project, focusing on two sets of samples: Yoruba from Ibadan, Nigeria (YRI) and Utah residents of Northern and Western European ancestry (CEU). We inferred a single-population two-epoch demographic model (supplementary figure S9a, Supplementary Material online) from the YRI samples, and a two-population isolation-with-migration model (supplementary fig. S9b, Supplementary Material online) from the combined YRI and CEU samples. To mimic low-pass sequencing, we subsampled the original high-depth data (which averaged 30× per site per individual) to create data with low to medium depth.

As with simulated data, the observed AFS from low-pass subsampled data was biased compared to high-pass data (Fig. 4a). Using the GATK pipeline, low-pass data yielded few low- and high-frequency derived alleles. In contrast to the simulated data, on these real data ANGSD and winsfs failed to recover the correct number of low-frequency alleles at 3× and 5× depth, while still introducing large fluctuations at intermediate frequencies (Fig. 4b, c).

If low-pass biases were corrected for, we expected the inferred demographic parameters from subsampled low-pass data to match those from the original high-pass data. For the two-epoch model fit to YRI data, we found that with a GATK-called AFS and no low-pass model (Table 2), the inferred population sizes were biased downward and the times were inaccurate, similar to the growth model fit to simulated data. With the low-pass model, inferred values for low depth were similar to those for high depth, with some deviation at 3× (Table 2). Results from fitting ANGSD and winsfs-estimated spectra were similar to not modeling low depth, suggesting that ANGSD and winsfs are ineffective for these data (Table 2). In the simulated datasets, only ANGSD showed notably low likelihoods at low depth of coverage, whereas in the real data, both ANGSD and winsfs exhibited poor performance.

For the isolation-with-migration model fit to YRI and CEU data, the results were broadly similar (supplementary table S1, Supplementary Material online). For population sizes and the divergence time, inferences were more stable from GATK genotyping and our low-pass model than from ANGSD and winsfs-estimated AFS. By contrast, the inferred migration rate was similar across analyses. For both the two-epoch and isolation-with-migration models, the selection of GATK parameters had negligible influence on the results, as evidenced by the consistent similarity in the outcomes when using `minMapQ = 30` (Table 2 and supplementary table S1, Supplementary Material online) and `minMapQ = 1` (supplementary tables S2 and S3, Supplementary Material online).

Discussion

We assessed the biases introduced by low-pass sequencing using GATK multisample genotype calling and developed a model to mitigate these biases. In contrast to existing approaches (e.g. Korneliussen et al. 2014; Han et al. 2015; Mas-Sandoval et al. 2022), which attempt to estimate an unbiased AFS to which standard model-based analyses can be applied, we directly analyzed the low-pass AFS by accounting for low-pass biases in our modeling framework. In a simulated population undergoing growth, we found that low-pass sequencing reduced the presence of low-frequency alleles (Fig. 1), consistent with findings from previous studies (Han et al. 2014). Our model accounted for these biases, contrasting with ANGSD, which created fluctuations in the AFS at low depth (Fig. 2). In scenarios involving two populations, we observed similar biases, which our model effectively corrected, whereas ANGSD introduced additional noise (supplementary figs S3 and S4, Supplementary Material online). For demographic inference, using our model enabled accurate parameter estimates even at low-pass depths, while neglecting low-depth biases resulted in substantial inaccuracies (Fig. 3). ANGSD also yielded accurate estimates, but worse likelihoods. Empirical testing using human data from the 1000 Genomes Project showcased the accuracy of our correction method in improving demographic inference from low-pass data, outperforming both uncorrected analysis and ANGSD results (Fig. 4 and Table 2 and supplementary table S1, Supplementary Material online).

While ANGSD is recognized for its effectiveness in managing low-pass sequencing, our results showed its difficulties in modeling medium-frequency alleles. This is reflected in lower likelihood scores, particularly when comparing low-pass datasets to high-pass ones (Fig. 3). Despite their utility in incorporating uncertainty related to low-pass sequencing (Nielsen et al. 2011; Fumagalli 2013; Korneliussen et al. 2014), genotype likelihoods might not always accurately capture the entire range of allele frequencies. Despite the AFS fluctuations, ANGSD yielded reliable parameter estimates for simulated data. But ANGSD was unable to accurately estimate the demographic parameters of real datasets, as demonstrated in the analysis of the 1000 Genomes Project data (Table 2 and supplementary table S1, Supplementary Material online). This underscores the need for rigorous and critical assessments of results by evaluating the likelihood of the model and conducting uncertainty analysis. Similarly, winsfs showed the same pattern as ANGSD, performing well in simulations and with less noise in AFS estimation (Figs. 2 and 3), but producing inaccurate results in the analysis of 1000 Genomes data (Table 2 and supplementary table S1, Supplementary Material online).

The observed discrepancy between simulated and empirical data performance in ANGSD and winsfs likely stems from multiple underlying factors. While simulated data offers well-controlled and known parameters, which facilitate accurate model fitting, real-world datasets such as the 1kGP introduce complexities like sequencing biases or population structure. These elements may not be fully captured by ANGSD's model assumptions, contributing to its underperformance with empirical data. Specifically, ANGSD's reliance on genotype likelihoods, which are sensitive to the accurate modeling of sequencing errors, can pose challenges when real data exhibit issues such as base quality miscalibration or variable sequencing depth. A more detailed exploration of ANGSD's assumptions, particularly regarding error profiles and filtering strategies, is required to fully understand these discrepancies and improve performance with empirical datasets.

Variant discovery using GATK involves two main approaches: multisample (classic joint-calling) and single-sample calling (Nielsen et al. 2011). We modeled multisample calling, which has higher statistical power compared to single-sample calling (Nielsen et al. 2011; Poplin et al. 2018). But multisample calling can become computationally burdensome with larger sample sizes, leading to the development of incremental single-calling as a scalable alternative (McKenna et al. 2010; Van der Auwera and O'Connor 2020). When our model was applied to incremental single-calling AFS from subsampled 1000 Genomes Project data, parameter inference was poor (supplementary table S4, Supplementary Material online). Therefore, our model should only be used with multisample calling, and a slightly different model may need to be developed for incremental single-calling.

We present a GATK multisample calling model designed to compensate for AFS biases introduced by low-pass sequencing. Although tailored for GATK, our model's design allows for its extension to different pipelines with modifications to address the unique aspects of each calling algorithm. For example, our model currently assumes that a site is called when at least two reads supporting the alternative allele are found (Equations (3) and (4)), but this could be modified for other pipelines with different calling criteria. Our approach can thus be generalized to other calling pipelines, including those using short reads, long reads, and hybrid approaches (e.g. Bankevich et al. 2012; Poplin et al. 2018). Note that our mathematical model assumes a shared read depth distribution among all individuals, and some studies may vary depth among individuals. Simulations suggest, however, that our model remain accurate with uneven depths (supplementary fig. S10, Supplementary Material online).

Our approach can also be integrated into other AFS-based inference tools such as moments (Portik et al. 2017; Leaché et al. 2019), fastsimcoal2 (Excoffier et al. 2013, 2021), GADMA (Noskova et al. 2020), and delimitR (Smith and Carstens 2020), because our approach modifies the model AFS, independent of how it is computed. Our approach may also be useful in approximate Bayesian computation (Beaumont 2010; Csilléry et al. 2012) and machine learning workflows (Pudlo et al. 2016; Smith and Carstens 2020), facilitating simulation of low-pass datasets. Note, however, that we model bias in the mean shape of the AFS under low-pass sequencing, not its full variance (supplementary fig. S11, Supplementary Material online). Furthermore, AFS-based analyses are used not only for demographic studies but also to examine natural selection, including inferring the distribution of fitness effects of new mutations (Eyre-Walker and Keightley 2007; Huang et al. 2021). Our approach can thus facilitate population genomics research across tools, approaches, and problem domains.

In conclusion, we have developed a robust correction for low-pass sequencing biases, significantly enhancing the accuracy of demographic parameter estimation at various coverage depths. As the genetic research community continues to address challenges associated with low-pass data (Bryc et al. 2013; Korneliussen et al. 2014; Blischak et al. 2018; Meisner and Albrechtsen 2018), especially when constrained by economics or sample availability, our methodology provides enables more reliable genetic analysis.

Material and Methods

Simulating AFS Under Low-Pass Sequencing

We used msprime (Kelleher et al. 2016; Baumdicker et al. 2022) to generate SNP datasets via coalescent simulations. We simulated two demographic models. The demographic models were visualized using demesdraw (Gower et al. 2022). The first model, single-population exponential growth (supplementary fig. S2a, Supplementary Material online), involved two parameters: the relative population size $\nu_1 = 10$ and time of past growth $T = 0.1$ (in units of two times the effect population size generations). The second model, two-population isolation (supplementary fig. S2b, Supplementary Material online), involved three parameters: equal relative sizes of populations 1 and 2, $\nu_1 = \nu_2 = 1$, and divergence time in the past T = 0.1. For each model, we conducted 25 independent simulations. For the exponential growth model, we sampled 20 diploid individuals, whereas for the isolation model, we sampled 10 individuals per population. Both demographic scenarios used an ancestral effective population size N_e of 10,000, a sequence length of 10^7 bp, a mutation rate of $\mu = 10^{-7}$ per site per generation, and recombination rate of $r = 10^{-7}$ per site per generation.

For simulations incorporating inbreeding, we used SLiM 4 (Messer 2013; Haller and Messer 2023). Datasets were generated under a bottleneck and growth model (supplementary figure S2c, Supplementary Material online), with a population bottleneck of $\nu_B = 0.25$, followed by a population expansion to $\nu_F = 1.0$. The time of the past bottleneck was set at $T = 0.2$, and the level of inbreeding was varied with $F \in \{0.1, 0.5, 0.9\}$. Inbreeding was introduced using the selfing rate, set to $s = \frac{2F}{1+F}$. Twenty-five independent simulations were conducted, with 20 individuals sampled for each replicate. Simulation parameters were $N_e = 1{,}000$, $L = 2 \times 10^6$ bp, $\mu = 5 \times 10^{-6}$, and $r = 2.5 \times 10^{-6}$, with a burn-in of 10,000 generations.

To create low-pass datasets, we used synthetic diploid genomes. For each simulation replicate, we generated a random reference genome spanning 10 Mb with a GC content of 40%, resembling the human genome. Mutations were incorporated by altering single nucleotides at the positions observed in the SNP matrix generated during each simulation, assuming that all sites were biallelic. Diploid individual genomes were generated by randomly selecting two chromosomes from the population pool.

Using the synthetic individual genomes as templates, we simulated 126 bp paired-end short reads for each individual with InSilicoSeq v2.0.1. (Gourlé et al. 2019). We calculated the number of reads per scenario as LC/R, where L is the genome length, C the coverage depth, and R the read length. Reads for each diploid chromosome were simulated with equal probability. Depth of coverage per individual was sampled from a normal distribution with means of 3, 5, 10, and 30 and corresponding standard deviations of 0.3, 0.5, 1, and 3 to explore coverage variability, which increased with coverage levels. These standard deviations were selected based on preliminary simulations that suggested they offer a realistic variance for each coverage level.

For each individual, we aligned simulated reads to the reference genome using BWA v0.7.17 (Li et al. 2009). We then processed the aligned reads with SAMTools v1.10 (Li 2013) to perform sorting, indexing, and pileup generation. To generate GATK spectra, we used the GATK multisample approach via HaplotypeCaller v4.2 (McKenna et al. 2010; Van der Auwera and O'Connor 2020). To minimize false positives, the identified variants underwent filtering based on GATK's Best Practices guidelines, with thresholds tailored to expected error rates and variant quality. These thresholds included depth-normalized variant confidence (QD < 2.0), mapping quality (MQ < 40), strand bias estimate (FS > 60.0), and strand bias (SOR > 10.0). The filtered SNP VCF files were subsequently used in demographic inference analyses to estimate population parameters based on the AFS of these variants. To generate ANGSD spectra, we used the BAM files containing information about each individual with reads aligned to the reference genome. Subsequently, realAFS was used to estimate a maximum-likelihood AFS through the EM algorithm. The ANGSD v0.94 analysis was conducted with the following parameters: `doSaf = 1` enabled the calculation of site allele frequency likelihoods, `minMapQ = 30` set a minimum mapping quality score of 30 to filter low-quality alignments, `minQ = 20` applied a minimum base quality score of 20 to exclude low-quality bases, and `GL = 2` specified the genotype likelihood model (using model 2, which is the GATK-like model).

Empirical Subsampling of Human Data

We used high-quality whole-genome sequencing data (30×) from the 1000 Genomes Project (1kGP), sourced from The International Genome Sample Resource data portal (https://www.internationalgenome.org/ Fairley et al. 2020). The data comprised CRAM files aligned to the GRCh38 human reference genome. We focused on two sets of samples for our analysis: 10 randomly selected individuals from the Yoruba from Ibadan, Nigeria (YRI) samples and 10 from the Utah residents with Northern and Western European ancestry (CEU) samples. The specific individuals included for the YRI were NA18486, NA18499, NA18510, NA18853, NA18858, NA18867, NA18878, NA18909, NA18917, NA18924, and for the CEU NA07037, NA11829, NA11892, NA11918, NA11932, NA11994, NA12004, NA12144, NA12249, and NA12273. Additionally, for a single-population demographic model, 20 YRI individuals were analyzed, which includes the initial 10 plus an additional 10 samples: NA19092, NA19116, NA19117, NA19121, NA19138, NA19159, NA19171, NA19184, NA19204, and NA19223.

Initially, we converted the CRAM files to BAM format and indexed them using Picard tools (https://broadinstitute.github.io/picard/). We then isolated reads from chromosome 20 at the original 30× coverage, which we subsequently subsampled to 10×, 5×, and 3× coverage using samtools v.1.10 (Li 2013) to emulate varying sequencing depths. Next, using GATK version 4.2.5 HaplotypeCaller (McKenna et al. 2010; Van der Auwera and O'Connor 2020), we called SNPs and indels from these varying coverage depths for each population. We employed multisample SNP calling, merging BAM files with identical coverage prior to processing with HaplotypeCaller. This approach yielded a raw output VCF file.

We also carried out a single-sample calling procedure. For this, individual BAM files were used directly as inputs for the GATK HaplotypeCaller with the `-ERC GVCF` flag to enable GVCF mode. Following this, we used GATK GenomicsDBImport to compile the individual variant calls into a cohesive data structure. This setup allowed us to conduct joint genotyping using GATK GenotypeGVCFs, ultimately producing a multisample VCF.

Following SNP calling, we employed GATK SelectVariants to filter out indels for both approaches, retaining only SNPs. Quality filtering of SNPs was conducted using GATK VariantFiltration, applying criteria such as depth-normalized variant confidence ($QD < 2.0$), mapping quality ($MQ < 40$), strand bias estimate ($FS > 60.0$), and overall strand bias ($SOR > 10.0$). After quality filtering, the VCF files were annotated with ancestral allele information using the `vcftools fill-aa` module, based on data from the Ensembl Release 110 Database (Danecek et al. 2011).

Finally, we used ANGSD to generate an AFS by using BAM files as input. The sample allele frequencies were first estimated using ANGSD's `-doSaf` flag, using GATK genotype likelihoods. These likelihoods were then used to calculate the AFS via the EM algorithm using ANGSD's realAFS program. In this way, we maintained the original sample sizes from the BAM files, resulting in AFS for 40 chromosomes in the single-population analysis and 20 chromosomes per population in the two-population analysis. ANGSD v0.94 analysis was executed with the following settings: `doSaf = 1`, `minMapQ = 30`, `minQ = 20`, and `GL = 2`. We used winsfs v0.7 to generate AFS based on site allele frequency likelihoods calculated with ANGSD, using its default parameters. We also generate AFS using winsfs, based on site allele frequency likelihoods computed with ANGSD.

Demographic Inference Using dadi

We used dadi (Gutenkunst et al. 2009) to fit demographic models to simulated and empirical datasets. For the GATK spectra, we used the VCF files as input and subsampled individuals to accommodate missing data. For the ANGSD spectra, we used them as input directly. Within dadi, we used three demographic models for the simulated datasets: (i) an exponential growth model: `dadi.Demographics1D.growth`; (ii) a divergence model with migration fixed to zero: `dadi.Demographics2D.split_mig`; (iii) an bottleneck then exponential growth model modified to incorporate inbreeding: `dadi.Demographics1D.bottlegrowth`. For the human datasets, we used two models: (i) a divergence with migration model: `dadi.Demographics2D.split_mig` and (ii) an instantaneous growth model: `dadi.Demographics1D.two_epoch`. The extrapolation grid points were set using the formula $[\max(ns) + 120, \max(ns) + 130, \max(ns) + 140]$, where ns is the sample size of the AFS. Our low-pass correction is also implemented in dadi-cli (Huang et al. 2023).

Supplementary Material

Supplementary material is available at *Molecular Biology and Evolution* online.

Funding

This work was supported by the National Institute of General Medical Sciences of the National Institutes of Health (R01GM127348 and R35GM149235 to R.N.G.).

Conflict of interests

None declared.

Data Availability

The correction for low-pass sequencing is implemented in the dadi Python package (https://bitbucket.org/gutenkunstlab/dadi) and in the corresponding dadi-cli command-line interface (https://github.com/xin-huang/dadi-cli). The codebase for creating and analyzing both simulated and empirical datasets is on GitHub at https://github.com/emanuelmfonseca/low-coverage-sfs and https://github.com/lntran26/low-coverage-sfs/tree/main/empirical_analysis. Furthermore, we provide a user guide to help users implement our methodology on the dadi webpage (https://dadi.readthedocs.io/en/latest/user-guide/low-pass). The empirical data were downloaded from The International Genome Sample Resource data portal (https://www.internationalgenome.org Fairley et al. 2020).

References

Van der Auwera GA, O'Connor BD. *Genomics in the cloud: using Docker, GATK, and WDL in Terra*. 1st ed. Beijing Boston Farnham Sebastopol Tokyo: O'Reilly; 2020.

Balding DJ, Nichols RA. A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica*. 1995:96(1-2):3–12. https://doi.org/10.1007/BF01441146.

Balding DJ, Nichols RA. Significant genetic correlations among Caucasians at forensic DNA loci. *Heredity (Edinb)*. 1997:78(6): 583–589. https://doi.org/10.1038/hdy.1997.97.

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol*. 2012:19(5):455–477. https://doi.org/10.1089/cmb.2012.0021.

Baumdicker F, Bisschop G, Goldstein D, Gower G, Ragsdale AP, Tsambos G, Zhu S, Eldon B, Ellerman EC, Galloway JG, et al. Efficient ancestry and mutation simulation with msprime 1.0. *Genetics*. 2022:220(3): iyab229. https://doi.org/10.1093/genetics/iyab229.

Beaumont MA. Approximate Bayesian computation in evolution and ecology. *Annu Rev Ecol Evol Syst*. 2010:41(1):379–406. https://doi.org/10.1146/ecolsys.2010.41.issue-1.

Blischak PD, Barker MS, Gutenkunst RN. Inferring the demographic history of inbred species from genome-wide SNP frequency data. *Mol Biol Evol*. 2020:37(7):2124–2136. https://doi.org/10.1093/molbev/msaa042.

Blischak PD, Kubatko LS, Wolfe AD. SNP genotyping and parameter estimation in polyploids using low-coverage sequencing data. *Bioinformatics*. 2018:34(3):407–415. https://doi.org/10.1093/bioinformatics/btx587.

Bryc K, Patterson N, Reich D. A novel approach to estimating heterozygosity from low-coverage genome sequence. *Genetics*. 2013:195(2): 553–561. https://doi.org/10.1534/genetics.113.154500.

Carstens BC, Smith ML, Duckett DJ, Fonseca EM, Thomé MTC. Assessing model adequacy leads to more robust phylogeographic inference. *Trends Ecol Evol*. 2022:37(5):402–410. https://doi.org/10.1016/j.tree.2021.12.007.

Csilléry K, François O, Blum MGB. abc: an R package for approximate Bayesian computation (ABC). *Methods Ecol Evol*. 2012:3(3): 475–479. https://doi.org/10.1111/j.2041-210X.2011.00179.x.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al. The variant call format and VCFtools. *Bioinformatics*. 2011:27(15): 2156–2158. https://doi.org/10.1093/bioinformatics/btr330.

Duckett DJ, Calder K, Sullivan J, Tank DC, Carstens BC. Reduced representation approaches produce similar results to whole genome sequencing for some common phylogeographic analyses. *PLoS One*. 2023:18(11):e0291941. https://doi.org/10.1371/journal.pone.0291941.

Duitama J, Kennedy J, Dinakar S, Hernández Y, Wu Y, M⍰ndoiu II. Linkage disequilibrium based genotype calling from low-coverage shotgun sequencing reads. *BMC Bioinformatics*. 2011:12(S1):S53. https://doi.org/10.1186/1471-2105-12-S1-S53.

Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M. Robust demographic inference from genomic and SNP data. *PLoS Genet*. 2013:9(10):e1003905. https://doi.org/10.1371/journal.pgen.1003905.

Excoffier L, Marchi N, Marques DA, Matthey-Doret R, Gouy A, Sousa VC. fastsimcoal2: demographic inference under complex evolutionary scenarios. *Bioinformatics*. 2021:37(24):4882–4885. https://doi.org/10.1093/bioinformatics/btab468.

Eyre-Walker A, Keightley PD. The distribution of fitness effects of new mutations. *Nat Rev Genet*. 2007:8(8):610–618. https://doi.org/10.1038/nrg2146.

Fairley S, Lowy-Gallego E, Perry E, Flicek P. The international genome sample resource (IGSR) collection of open human genomic variation resources. *Nucleic Acids Res*. 2020:48(D1):D941–D947. https://doi.org/10.1093/nar/gkz836.

Fox EA, Wright AE, Fumagalli M, Vieira FG. ngsLD: evaluating linkage disequilibrium using genotype likelihoods. *Bioinformatics*. 2019:35(19):3855–3856. https://doi.org/10.1093/bioinformatics/btz200.

Fumagalli M. Assessing the effect of sequencing depth and sample size in population genetics inferences. *PLoS One*. 2013:8(11):e79667. https://doi.org/10.1371/journal.pone.0079667.

Gorjanc G, Cleveland MA, Houston RD, Hickey JM. Potential of genotyping-by-sequencing for genomic selection in livestock populations. *Genet Sel Evol*. 2015:47(1):12. https://doi.org/10.1186/s12711-015-0102-z.

Gourlé H, Karlsson-Lindsjö O, Hayer J, Bongcam-Rudloff E. Simulating Illumina metagenomic data with InSilicoSeq. *Bioinformatics*. 2019:35(3):521–522. https://doi.org/10.1093/bioinformatics/bty630.

Gower G, Ragsdale AP, Bisschop G, Gutenkunst RN, Hartfield M, Noskova E, Schiffels S, Struck TJ, Kelleher J, Thornton KR. Demes: a standard format for demographic models. *Genetics*. 2022:222:iyac131. https://doi.org/10.1093/genetics/iyac131.

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet*. 2009:5: e1000695. https://doi.org/10.1371/journal.pgen.1000695.

Haller BC, Messer PW. SLiM 4: multispecies eco-evolutionary modeling. *Am Nat*. 2023:201:E127. https://doi.org/10.1086/723601.

Han E, Sinsheimer JS, Novembre J. Characterizing bias in population genetic inferences from low-coverage sequencing data. *Mol Biol Evol*. 2014:31:723. https://doi.org/10.1093/molbev/mst229.

Han E, Sinsheimer JS, Novembre J. Fast and accurate site frequency spectrum estimation from low coverage sequence data. *Bioinformatics*. 2015:31:720. https://doi.org/10.1093/bioinformatics/btu725.

Huang X, Fortier AL, Coffman AJ, Struck TJ, Irby MN, James JE, León-Burguete JE, Ragsdale AP, Gutenkunst RN. Inferring genome-wide correlations of mutation fitness effects between populations. *Mol Biol Evol*. 2021:38:4588. https://doi.org/10.1093/molbev/msab162.

Huang X, Struck TJ, Davey SW, Gutenkunst RN. dadi-cli: automated and distributed population genetic model inference from allele frequency spectra. 2023.

Kelleher J, Etheridge AM, McVean G. Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS Comput Biol*. 2016:12:e1004842. https://doi.org/10.1371/journal.pcbi.1004842.

Kim BY, Huber CD, Lohmueller KE. Inference of the distribution of selection coefficients for new nonsynonymous mutations using large samples. *Genetics*. 2017:206(1):345–361. https://doi.org/10.1534/genetics.116.197145.

Korneliussen TS, Albrechtsen A, Nielsen R. ANGSD: analysis of next generation sequencing data. *BMC Bioinformatics*. 2014:15:356. https://doi.org/10.1186/s12859-014-0356-4.

Leaché AD, Portik DM, Rivera D, Rödel M, Penner J, Gvoždík V, Greenbaum E, Jongsma GFM, Ofori-Boateng C, Burger M, et al. Exploring rain forest diversification using demographic model testing in the African foam-nest treefrog *Chiromantis rufescens*. *J Biogeogr*. 2019:46(12):2706–2721. https://doi.org/10.1111/jbi.13716.

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM, arXiv, arXiv:1303.3997 [q-bio], preprint: not peer reviewed. https://doi.org/10.48550/arXiv.1303.3997

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The sequence alignment/map format and SAMtools. *Bioinformatics (Oxford, England)*. 2009:25:2078. https://doi.org/10.1093/bioinformatics/btp352.

Li Y, Sidore C, Kang HM, Boehnke M, Abecasis GR. Low-coverage sequencing: implications for design of complex trait association studies. *Genome Res*. 2011:21(6):940–951. https://doi.org/10.1101/gr.117259.110.

Lou RN, Jacobs A, Wilder AP, Therkildsen NO. A beginner's guide to low-coverage whole genome sequencing for population genomics. *Mol Ecol*. 2021:30(23):5966–5993. https://doi.org/10.1111/mec.v30.23.

Maddison DR, Ruvolo M, Swofford DL. Geographic origins of human mitochondrial DNA: phylogenetic evidence from control region sequences. *Syst Biol*. 1992:41(1):111–124. https://doi.org/10.1093/sysbio/41.1.111.

Marchi N, Winkelbach L, Schulz I, Brami M, Hofmanová Z, Blöcher J, Reyna-Blanco CS, Diekmann Y, Thiéry A, Kapopoulou A, et al. The genomic origins of the world's first farmers. *Cell*. 2022:185:1842. https://doi.org/10.1016/j.cell.2022.04.008.

Marth GT, Czabarka E, Murvai J, Sherry ST. The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. *Genetics*. 2004:166:351. https://doi.org/10.1534/genetics.166.1.351.

Martin AR, Atkinson EG, Chapman SB, Stevenson A, Stroud RE, Abebe T, Akena D, Alemayehu M, Ashaba FK, Atwoli L, et al. Low-coverage sequencing cost-effectively detects known and novel variation in underrepresented populations. *Am J Hum Genet*. 2021:108:656. https://doi.org/10.1016/j.ajhg.2021.03.012.

Maruki T, Lynch M. Genotype calling from population-genomic sequencing data. *G3 (Bethesda)*. 2017:7:1393. https://doi.org/10.1534/g3.117.039008.

Mas-Sandoval A, Pope NS, Nielsen KN, Altinkaya I, Fumagalli M, Korneliussen TS. Fast and accurate estimation of multidimensional site frequency spectra from low-coverage high-throughput sequencing data. *Gigascience*. 2022:11:giac032. https://doi.org/10.1093/gigascience/giac032.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 2010:20:1297. https://doi.org/10.1101/gr.107524.110.

Meisner J, Albrechtsen A. Inferring population structure and admixture proportions in low-depth NGS data. *Genetics*. 2018:210:719. https://doi.org/10.1534/genetics.118.301336.

Messer PW. SLiM: simulating evolution with selection and linkage. *Genetics*. 2013:194:1037. https://doi.org/10.1534/genetics.113.152181.

Mota BS, Rubinacci S, Cruz Dávalos DI, G Amorim CE, Sikora M, Johannsen NN, Szmyt MH, Włodarczak P, Szczepanek A, Przybyła MM, et al. Imputation of ancient human genomes. *Nat Commun*. 2023:14:3660. https://doi.org/10.1038/s41467-023-39202-0.

Nielsen R, Korneliussen T, Albrechtsen A, Li Y, Wang J. SNP calling, genotype calling, and sample allele frequency estimation from new-generation sequencing data. *PLoS One*. 2012:7:e37558. https://doi.org/10.1371/journal.pone.0037558.

Nielsen R, Paul JS, Albrechtsen A, Song YS. Genotype and SNP calling from next-generation sequencing data. *Nat Rev Genet*. 2011:12: 443. https://doi.org/10.1038/nrg2986.

Noskova E, Ulyantsev V, Koepfli KP, O'Brien SJ, Dobrynin P. GADMA: genetic algorithm for inferring demographic history of multiple populations from allele frequency spectrum data. *Gigascience*. 2020:9: giaa005. https://doi.org/10.1093/gigascience/giaa005.

Poplin R, Chang PC, Alexander D, Schwartz S, Colthurst T, Ku A, Newburger D, Dijamco J, Nguyen N, Afshar PT, et al. A universal SNP and small-indel variant caller using deep neural networks. *Nat Biotechnol*. 2018:36(10):983–987. https://doi.org/10.1038/nbt.4235.

Portik DM, Leaché AD, Rivera D, Barej MF, Burger M, Hirschfeld M, Rödel M, Blackburn DC, Fujita MK. Evaluating mechanisms of diversification in a Guineo-Congolian tropical forest frog using demographic model selection. *Mol Ecol*. 2017:26(19):5245–5263. https://doi.org/10.1111/mec.2017.26.issue-19.

Pudlo P, Marin JM, Estoup A, Cornuet JM, Gautier M, Robert CP. Reliable ABC model choice via random forests. *Bioinformatics*. 2016:32(6):859–866. https://doi.org/10.1093/bioinformatics/btv684.

Rasmussen MS, Garcia-Erill G, Korneliussen TS, Wiuf C, Albrechtsen A. Estimation of site frequency spectra from low-coverage sequencing data using stochastic EM reduces overfitting, runtime, and memory usage. *Genetics*. 2022:222(4):iyac148. https://doi.org/10.1093/genetics/iyac148.

Reid NM, Proestou DA, Clark BW, Warren WC, Colbourne JK, Shaw JR, Karchner SI, Hahn ME, Nacci D, Oleksiak MF, et al. The genomic landscape of rapid repeated evolutionary adaptation to toxic pollution in wild fish. *Science*. 2016:354(6317):1305–1308. https://doi.org/10.1126/science.aah4993.

Sawyer SA, Hartl DL. Population genetics of polymorphism and divergence. *Genetics*. 1992:132(4):1161–1176. https://doi.org/10.1093/genetics/132.4.1161.

Smith ML, Carstens BC. Process-based species delimitation leads to identification of more biologically relevant species. *Evolution*. 2020:74(2):216–229. https://doi.org/10.1111/evo.v74.2.

Vieira FG, Fumagalli M, Albrechtsen A, Nielsen R. Estimating inbreeding coefficients from NGS data: impact on genotype calling and allele frequency estimation. *Genome Res*. 2013:23(11):1852–1861. https://doi.org/10.1101/gr.157388.113.

Wakeley J. *Coalescent theory: an introduction*. Greenwood Village (CO): Roberts & Co. Publishers; 2009.

Wigginton JE, Cutler DJ, Abecasis GR. A note on exact tests of Hardy-Weinberg equilibrium. *Am J Hum Genet*. 2005:76(5): 887–893. https://doi.org/10.1086/429864.