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 author:** E-mail: emanuelmfonseca@arizona.edu; rgutenk@arizona.edu

The correction for low-pass sequencing is performed using the publicly available dadi Python package, which can be accessed at https://bitbucket.org/gutenkunstlab/dadi. Additionally, the codebase for creating and analyzing both simulated and empirical datasets, ensuring reproducibility, is readily accessible 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 illustrative examples to assist users in implementing our methodology.

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 demographic history inference. We developed a probabilistic model of low-pass biases from the Genome Analysis Toolkit (GATK) multi-sample 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.

Key words: demography inference, inbreeding, low-pass sequencing, allele frequency spectrum, GATK multi-sample 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 a) the extent of the genome to be sequenced, b) the depth of coverage for each sample, and c) 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; 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 & 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 & Albrechtsen 2018), with one of the most widely adopted being ANGSD (Korneliussen et al. 2014). ANGSD offers a diverse range of analyes 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. While ANGSD has proven its utility, limitations exist. For example, many analyses rely on distinguishing different types of variant sites (such a synonymous versus nonsymonyous) which the developers of ANGSD recommend against. Moreover, in some cases unbiased estimation of the AFS may be difficult or impossible.

Rather than attempting to estimate an unbiased AFS from low-pass data, we developed a probabilistic model of low-pass AFS biases and incorporated it into the population genomic inference software dadi (Gutenkunst et al. 2009). Our model is based on the multi-sample genotype calling pipeline of the Genome Analysis Toolkit (GATK), the most widely used tool for calling variants from read data (McKenna et al. 2010; Auwera & 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

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 multi-sample 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.

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 $2n_{seq}$ sampled alleles in many ways. To quantify those ways, we define the partition function $\mathbb{P}_{nseq}(f)$, 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 $\mathbb{P}_{nseq}(f)$ 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 $\mathbb{P}_{nseq}(f)$ is then the number of ways it can occur divided by the total over all partitions within $\mathbb{P}_{nseq}(f)$.

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 $P_a^{hom}(a) = \mathbb{D}(a)$. For a heterozygous individual, the probability of zero alternate

reads is

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

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

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

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

For a given partition within $\mathbb{P}_{nseq}(f)$ 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

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

The probability of exactly one read supporting the alternate allele is

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

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 $P_a^{part}(0) + P_a^{part}(1)$. 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 $\mathbb{P}_{nseq}(f)$, weighted by the partition probabilities. For any given coverage distribution, the probability of calling a variant site increases rapidly with allele frequency f (Fig. S1).

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. The probability a site can be analyzed is independent of the allele frequency and is

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

Here we sum the probability that exactly c individuals have at least one read at this site over all potential values of the number of covered individuals $n_{sub} \le c \le n_{seq}$. From this point onward, we consider partitions $\mathbb{P}_{nsub}(f)$ over the subsampled individuals.

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

$$P_{mis}^{het} = 2 \sum_d \mathbb{D}(d) \left(\frac{1}{2}\right)^d = 2\, P_a^{het}(0). \tag{6}$$

For a partition with n_1 true heterozygotes, the number of miscalled heterozygotes N_{mis}^{het} is binomially distributed with mean $n_1 P_{mis}^{het}$. Each miscalled heterozygote has equal chance of being called as homozygous reference or alternate, so the number of miscalls to homozygous reference $N_{\to ref}^{het}$ is binomially distributed with mean $N_{mis}^{het}/2$, and the number of miscalls to homozygous alternate is $N_{\to alt}^{het} = N_{mis}^{het} - N_{\to ref}^{het}$. The net change in estimated alternative allele frequency is then $N_{\to alt}^{het} - N_{\to ref}^{het}$.

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 (Eq. 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 allele frequency, 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 multi-population case) we simulate reads (default 1000) 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.

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 & Nichols (1995, 1997), given the inbreeding coeffficient 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 pre-calculate 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 multi-sample calling, along with our model of those biases. For a simulated population undergoing growth (Fig. S2A), 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).

When a pair of populations undergoing a split and isolation (Fig. S2B) is analyzed through a joint AFS, similar low coverage biases occur (Fig. S3). Again, our model corrects those biases well (Fig. S3). Similar to the single-population case, ANGSD also introduces large fluctuations in the joint AFS S4).

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 (Fig. S2C), biases are still observed, which our model corrects (Fig. S5). Again, ANGSD introduced large fluctuations in low-pass AFS, beyond those expected from inbreeding (Fig. S6).

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 (Fig. S2A). When not modeling low-pass biases, the final

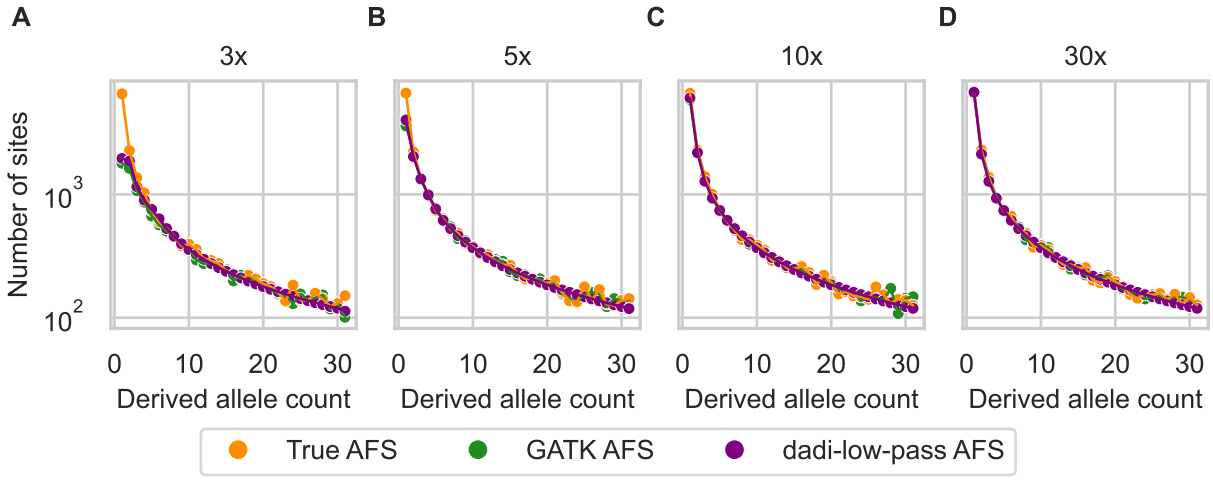

Figure 1: The low-pass AFS is biased, which our model captures. Simulated sequence data from an exponential growth demographic model for 20 individuals were called by GATK and subsampled to 16 individuals (to accommodate missing data at low depth). The GATK-called AFS (green) is biased compared to the 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×.

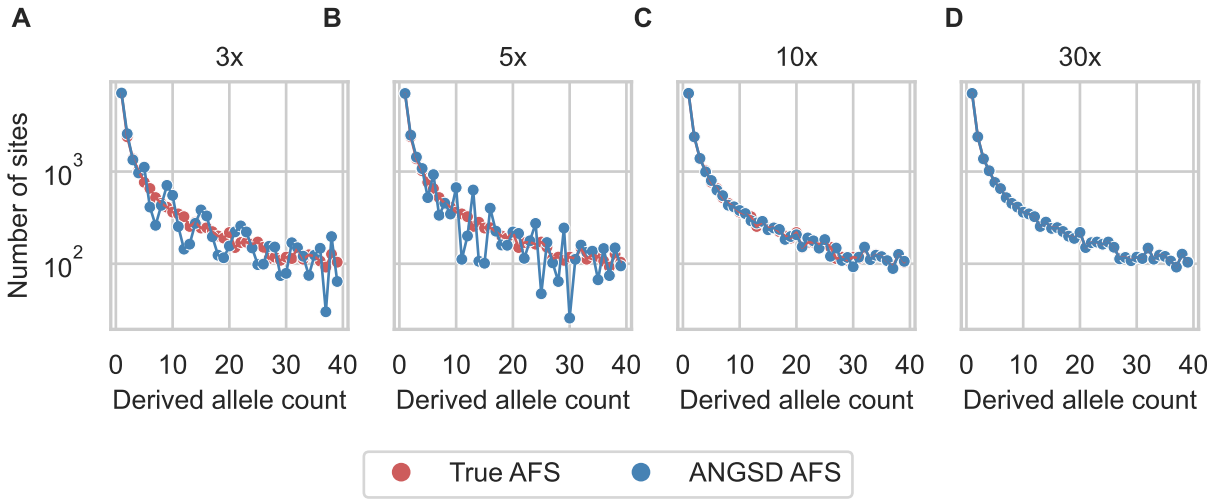

Figure 2: ANGSD corrects for low-pass bias of the AFS, but introduces fluctuations. For the same simulations as Fig. 1, ANGSD (blue) was used to reconstruct the true AFS (red). Coverage was (A) 3×, (B) 5×, (C) 10×, and (D) 30×.

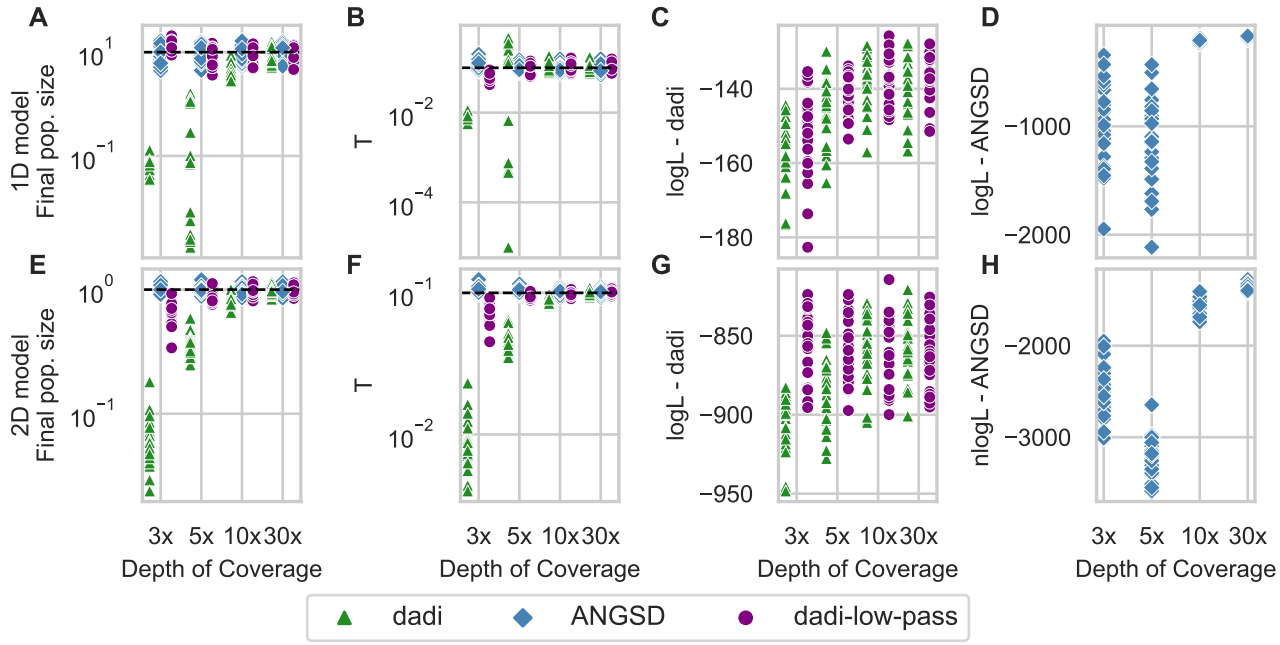

Figure 3: Our low-pass model and ANGSD enable accurate demographic parameter inference. A&B: From data simulated under a single-population growth model, the final population size and time of growth onset (T) were accurately inferred using our low-pass model and a GATK-called AFS or using normal dadi and an ANGSD-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 caused low likelihoods at low depth. 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.

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 also yielded accurate model parameters (Fig. 3A&B).

The logarithm of the likelihood is commonly used to assess the quality of model fit. ANGSD reconstructs 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).

For two-population data simulated under an isolation model (Fig. S2B), 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 ANGS (Fig. 3D). As in the single-population case, likelihoods were substantially lower when fitting the ANGSD-estimated AFS, consistent with introduced fluctuations in the AFS (Fig. S4).

For one-population data simulated under a growth model with inbreeding (Fig. S2C), 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 (Fig. S7). For high inbreeding ($F = 0.9$), the impact of low-pass sequencing

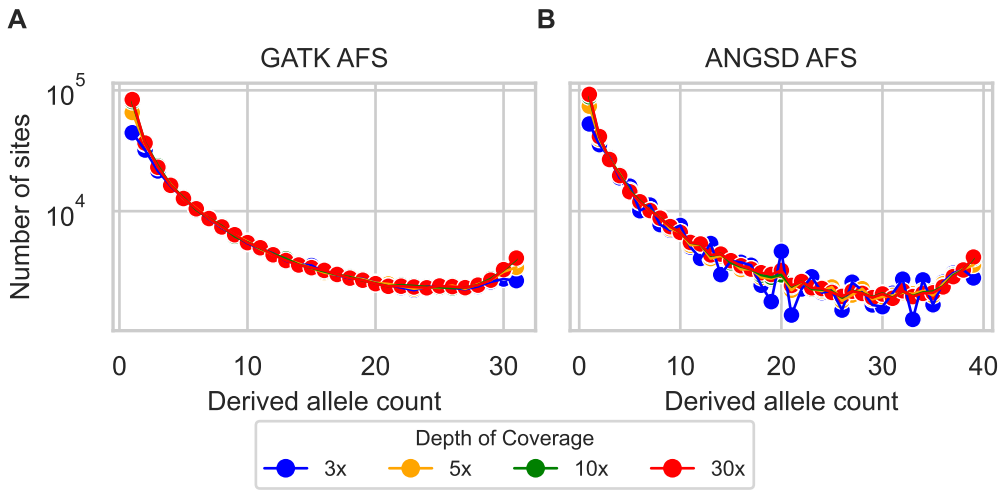

Figure 4: Allele frequency spectra from 20 YRI samples versus subsampled sequencing depth. A: Spectra generated using the GATK pipeline and subsampled to 32 haplotypes to accomodate missing genotypes. B: Spectra generated using ANGSD genotype likelihood optimization with BAM files input.

on accuracy was smaller, because inbreeding reduces heterozygosity (Fig. S7).

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 (Fig. S8C). A substantial difference between the inbreeding coefficient used for correction and the inferred value thus suggests that the assumed inbreeding coefficient was 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, 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 (Fig. S9A) from the YRI samples, and a two-population isolation-with-migration model (Fig. S9B) 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 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).

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 1), 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 1). Results from fitting ANGSD-estimated spectra were similar to not modeling low depth,

suggesting that ANGSD is ineffective for these data (Table 1). As with simulated data, the likelihoods for ANGSD at low depth were also low.

For the isolation-with-migration model fit to YRI and CEU data, the results were broadly similar (Table S1.) For population sizes and the divergence time, inferences were more stable from GATK genotyping and our low-pass model than from ANGSD-estimated AFS. By contrast, the inferred migration rate was similar across analyses.

Table 1: One-population YRI model analysis results. 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).

			depth			
parameter	AFS	model	30×	10×	5×	3 ×
ν_{YRI}	GATK	dadi	1.82	1.76	1.54	0.05
	GATK	low-pass	1.83	1.77	1.70	1.54
	ANGSD	dadi	1.81	1.80	1.67	0.08
T	GATK	dadi	0.43	0.51	0.88	0.001
	GATK	low-pass	0.42	0.49	0.51	0.43
	ANGSD	dadi	0.55	0.68	0.96	0.001
θ ($\times 10^4$)	GATK	dadi	5.15	5.08	4.81	5.99
	GATK	low-pass	5.16	5.10	5.10	5.20
	ANGSD	dadi	5.52	5.32	5.03	6.78
log-likelihood	GATK	dadi	-297	-253	-533	-1312
	GATK	low-pass	-302	-259	-283	-339
	ANGSD	dadi	-475	-486	-1120	-5905

Discussion

We assessed the biases introduced by low-pass sequencing using GATK multi-sample genotype calling and developed a model to mitigate these biases. In a simulated population undergoing growth, we found that low-pass sequencing reduced the presence of low-frequency alleles (Fig. 1). 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 (Fig. S3 and S4). 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 Tables 1 and S1).

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 (Tables 1 and S1). This underscores the need for rigorous and critical assessments of results by evaluating the likelihood of the model and conducting uncertainty analysis.

Variant discovery using GATK involves two main approaches: multi-sample (classic joint-calling) and single-sample calling (Nielsen et al. 2011). We modeled multi-sample calling, which has higher statistical

power compared to single-sample calling (Nielsen et al. 2011; Poplin et al. 2018). But multi-sample 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; Auwera & O'Connor 2020). When our model was applied to incremental single-calling AFS from subsampled 1000 Genomes Project data, parameter inference was poor (Table S2). Therefore, our model should only be used with multi-sample calling, and a slightly different model may need to be developed for incremental single-calling.

We present a GATK multi-sample 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 (Eq. 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 (Fig. S10).

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 & 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 & 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 (Fig. S11). 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 & 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 & 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, singe-population exponential growth (Fig. S2A), 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 (Fig. S2B), 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 & Messer 2023). Datasets were generated under a bottleneck and growth model (Fig. S2C), 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 $N_e = 1000$, $L = 2 \times 10^6$ 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 multi-sample approach via HaplotypeCaller v4.2 (McKenna et al. 2010; Auwera & 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 Expectation-Maximization algorithm. ANGSD v0.94 analysis was executed with the following settings: `doSaf = 1`, `minMapQ = 1`, `minQ = 20`, and `GL = 2`.

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, 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; Auwera & O'Connor 2020), we called SNPs and indels from these varying coverage depths for each population. We employed multi-sample 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

multi-sample 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 Expectation-Maximization 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.

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-coverage correction is also implemented in dadi-cli (Huang et al. 2023).

Acknowledgements

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.).

References

Auwera GAVd, O'Connor BD (2020) *Genomics in the cloud: using Docker, GATK, and WDL in Terra.* O'Reilly, Beijing Boston Farnham Sebastopol Tokyo, first edition edition.

Balding DJ, Nichols RA (1995) A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. Genetica 96:3.

Balding DJ, Nichols RA (1997) Significant genetic correlations among Caucasians at forensic DNA loci. Heredity 78:583.

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA (2012) SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. Journal of Computational Biology 19:455.

Baumdicker F, Bisschop G, Goldstein D, Gower G, Ragsdale AP, Tsambos G, Zhu S, Eldon B, Ellerman EC, Galloway JG, Gladstein AL, Gorjanc G, Guo B, Jeffery B, Kretzschumar WW, Lohse K, Matschiner M, Nelson D, Pope NS, Quinto-Cortés CD, Rodrigues MF, Saunack K, Sellinger T, Thornton K, Van Kemenade H, Wohns AW, Wong Y, Gravel S, Kern AD, Koskela J, Ralph PL, Kelleher J (2022) Efficient ancestry and mutation simulation with msprime 1.0. Genetics 220:iyab229.

Beaumont MA (2010) Approximate Bayesian computation in evolution and ecology. Annual Review of Ecology, Evolution, and Systematics 41:379.

Blischak PD, Barker MS, Gutenkunst RN (2020) Inferring the demographic history of inbred species from genome-wide SNP frequency data. Molecular Biology and Evolution 37:2124.

Blischak PD, Kubatko LS, Wolfe AD (2018) SNP genotyping and parameter estimation in polyploids using low-coverage sequencing data. Bioinformatics 34:407.

Bryc K, Patterson N, Reich D (2013) A novel approach to estimating heterozygosity from low-coverage genome sequence. Genetics 195:553.

Carstens BC, Smith ML, Duckett DJ, Fonseca EM, Thomé MTC (2022) Assessing model adequacy leads to more robust phylogeographic inference. Trends in Ecology & Evolution 37:402.

Csilléry K, François O, Blum MGB (2012) abc: an R package for approximate Bayesian computation (ABC). Methods in Ecology and Evolution 3:475.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R, 1000 Genomes Project Analysis Group (2011) The variant call format and VCFtools. Bioinformatics 27:2156.

Duckett DJ, Calder K, Sullivan J, Tank DC, Carstens BC (2023) Reduced representation approaches produce similar results to whole genome sequencing for some common phylogeographic analyses. PLOS ONE 18:e0291941.

Duitama J, Kennedy J, Dinakar S, Hernández Y, Wu Y, Măndoiu II (2011) Linkage disequilibrium based genotype calling from low-coverage shotgun sequencing reads. BMC Bioinformatics 12:S53.

Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M (2013) Robust demographic inference from genomic and SNP data. PLoS Genetics 9:e1003905.

Excoffier L, Marchi N, Marques DA, Matthey-Doret R, Gouy A, Sousa VC (2021) fastsimcoal2: demographic inference under complex evolutionary scenarios. Bioinformatics 37:4882.

Eyre-Walker A, Keightley PD (2007) The distribution of fitness effects of new mutations. Nature Reviews Genetics 8:610.

Fairley S, Lowy-Gallego E, Perry E, Flicek P (2020) The International Genome Sample Resource (IGSR) collection of open human genomic variation resources. Nucleic Acids Research 48:D941.

Fox EA, Wright AE, Fumagalli M, Vieira FG (2019) ngsLD: evaluating linkage disequilibrium using genotype likelihoods. Bioinformatics 35:3855.

Fumagalli M (2013) Assessing the effect of sequencing depth and sample size in population genetics inferences. PLoS ONE 8:e79667.

Gorjanc G, Cleveland MA, Houston RD, Hickey JM (2015) Potential of genotyping-by-sequencing for genomic selection in livestock populations. Genetics Selection Evolution 47:12.

Gourlé H, Karlsson-Lindsjö O, Hayer J, Bongcam-Rudloff E (2019) Simulating Illumina metagenomic data with InSilicoSeq. Bioinformatics 35:521.

Gower G, Ragsdale AP, Bisschop G, Gutenkunst RN, Hartfield M, Noskova E, Schiffels S, Struck TJ, Kelleher J, Thornton KR (2022) Demes: a standard format for demographic models. Genetics 222:iyac131.

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD (2009) Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genetics 5:e1000695.

Haller BC, Messer PW (2023) SLiM 4: multispecies eco-evolutionary modeling. The American Naturalist 201:E127.

Huang X, Fortier AL, Coffman AJ, Struck TJ, Irby MN, James JE, León-Burguete JE, Ragsdale AP, Gutenkunst RN (2021) Inferring genome-wide correlations of mutation fitness effects between populations. Molecular Biology and Evolution 38:4588.

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

Kelleher J, Etheridge AM, McVean G (2016) Efficient coalescent simulation and genealogical analysis for large sample sizes. PLOS Computational Biology 12:e1004842.

Kim BY, Huber CD, Lohmueller KE (2017) Inference of the distribution of selection coefficients for new nonsynonymous mutations using large samples. Genetics 206:345.

Korneliussen TS, Albrechtsen A, Nielsen R (2014) ANGSD: Analysis of Next Generation Sequencing Data. BMC Bioinformatics 15:356.

Leaché AD, Portik DM, Rivera D, Rödel M, Penner J, Gvoždík V, Greenbaum E, Jongsma GFM, Ofori-Boateng C, Burger M, Eniang EA, Bell RC, Fujita MK (2019) Exploring rain forest diversification using demographic model testing in the African foam-nest treefrog Chiromantis rufescens. Journal of Biogeography 46:2706.

Li H (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ArXiv:1303.3997 [q-bio].

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics (Oxford, England) 25:2078.

Li Y, Sidore C, Kang HM, Boehnke M, Abecasis GR (2011) Low-coverage sequencing: implications for design of complex trait association studies. Genome Research 21:940.

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

Maddison DR, Ruvolo M, Swofford DL (1992) Geographic origins of Human mitochondrial DNA: phylogenetic evidence from control region sequences. Systematic Biology 41:111.

Marchi N, Winkelbach L, Schulz I, Brami M, Hofmanová Z, Blöcher J, Reyna-Blanco CS, Diekmann Y, Thiéry A, Kapopoulou A, Link V, Piuz V, Kreutzer S, Figarska SM, Ganiatsou E, Pukaj A, Struck TJ, Gutenkunst RN, Karul N, Gerritsen F, Pechtl J, Peters J, Zeeb-Lanz A, Lenneis E, Teschler-Nicola M, Triantaphyllou S, Stefanović S, Papageorgopoulou C, Wegmann D, Burger J, Excoffier L (2022) The genomic origins of the world's first farmers. Cell 185:1842.

Martin AR, Atkinson EG, Chapman SB, Stevenson A, Stroud RE, Abebe T, Akena D, Alemayehu M, Ashaba FK, Atwoli L, Bowers T, Chibnik LB, Daly MJ, DeSmet T, Dodge S, Fekadu A, Ferriera S, Gelaye B, Gichuru S, Injera WE, James R, Kariuki SM, Kigen G, Koenen KC, Kwobah E, Kyebuzibwa J, Majara L, Musinguzi H, Mwema RM, Neale BM, Newman CP, Newton CR, Pickrell JK, Ramesar R, Shiferaw W, Stein DJ, Teferra S, Van Der Merwe C, Zingela Z (2021) Low-coverage sequencing cost-effectively detects known and novel variation in underrepresented populations. The American Journal of Human Genetics 108:656.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA (2010) The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research 20:1297.

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

Messer PW (2013) SLiM: simulating evolution with selection and linkage. Genetics 194:1037.

Mota BS, Rubinacci S, Cruz Dávalos DI, G Amorim CE, Sikora M, Johannsen NN, Szmyt MH, Włodarczak P, Szczepanek A, Przybyła MM, Schroeder H, Allentoft ME, Willerslev E, Malaspinas AS, Delaneau O (2023) Imputation of ancient human genomes. Nature Communications 14:3660.

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 JS, Albrechtsen A, Song YS (2011) Genotype and SNP calling from next-generation sequencing data. Nature Reviews Genetics 12:443.

Noskova E, Ulyantsev V, Koepfli KP, O'Brien SJ, Dobrynin P (2020) GADMA: Genetic algorithm for inferring demographic history of multiple populations from allele frequency spectrum data. GigaScience 9:giaa005.

Poplin R, Chang PC, Alexander D, Schwartz S, Colthurst T, Ku A, Newburger D, Dijamco J, Nguyen N, Afshar PT, Gross SS, Dorfman L, McLean CY, DePristo MA (2018) A universal SNP and small-indel variant caller using deep neural networks. Nature Biotechnology 36:983.

Portik DM, Leaché AD, Rivera D, Barej MF, Burger M, Hirschfeld M, Rödel M, Blackburn DC, Fujita MK (2017) Evaluating mechanisms of diversification in a Guineo-Congolian tropical forest frog using demographic model selection. Molecular Ecology 26:5245.

Pudlo P, Marin JM, Estoup A, Cornuet JM, Gautier M, Robert CP (2016) Reliable ABC model choice via random forests. Bioinformatics 32:859.

Reid NM, Proestou DA, Clark BW, Warren WC, Colbourne JK, Shaw JR, Karchner SI, Hahn ME, Nacci D, Oleksiak MF, Crawford DL, Whitehead A (2016) The genomic landscape of rapid repeated evolutionary adaptation to toxic pollution in wild fish. Science 354:1305.

Sawyer SA, Hartl DL (1992) Population genetics of polymorphism and divergence. Genetics 132:1161.

Smith ML, Carstens BC (2020) Process-based species delimitation leads to identification of more biologically relevant species. Evolution 74:216.

Vieira FG, Fumagalli M, Albrechtsen A, Nielsen R (2013) Estimating inbreeding coefficients from NGS data: Impact on genotype calling and allele frequency estimation. Genome Research 23:1852.

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

Supplementary Material

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 author:** E-mail: emanuelmfonseca@arizona.edu; rgutenk@arizona.edu

The correction for low-pass sequencing is performed using the publicly available dadi Python package, which can be accessed at https://bitbucket.org/gutenkunstlab/dadi. Additionally, the codebase for creating and analyzing both simulated and empirical datasets, ensuring reproducibility, is readily accessible 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 illustrative examples to assist users in implementing our methodology.

Table S1: Two-population model analysis results. 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).

			depth			
parameter	AFS	model	30×	10×	5×	3 ×
ν_{YRI}	GATK	dadi	1.79	1.63	1.18	0.61
	GATK	low-pass	1.82	1.62	1.67	1.69
	ANGSD	dadi	1.69	1.58	1.26	0.87
ν_{CEU}	GATK	dadi	0.38	0.37	0.31	0.17
	GATK	low-pass	0.38	0.38	0.36	0.34
	ANGSD	dadi	0.39	0.38	0.33	0.22
T	GATK	dadi	0.21	0.22	0.18	0.06
	GATK	low-pass	0.21	0.23	0.20	0.16
	ANGSD	dadi	0.21	0.22	0.20	0.07
m	GATK	dadi	1.80	2.00	2.24	1.68
	GATK	low-pass	1.80	2.00	1.89	1.66
	ANGSD	dadi	1.99	2.12	2.44	1.91
θ ($\times 10^4$)	GATK	dadi	5.42	5.44	5.56	5.65
	GATK	low-pass	5.43	5.42	5.45	5.40
	ANGSD	dadi	6.04	6.01	6.03	6.25
log-likelihood	GATK	dadi	-2588	-2378	-2329	-2663
	GATK	low-pass	-2590	-2479	-2224	-1850
	ANGSD	dadi	-5518	-5595	-7074	-11029

Table S2: One-population model analysis results with single-sample calling using empirical GATK AFS. We analyzed GATK empirical single-sample call spectra without (dadi) and with low-pass correction (low-pass).

		depth			
parameter	model	30×	10×	5×	3 ×
ν_{YRI}	dadi	1.85	1.87	1.82	1.56
	low-pass	1.86	1.93	2.73	3.60
T	dadi	0.43	0.45	0.51	0.48
	low-pass	0.42	0.40	0.24	0.24
θ ($\times 10^3$)	dadi	5.13	5.05	4.62	4.31
	low-pass	5.14	5.10	4.96	4.49
log-likelihood	dadi	-284	-280	-457	-1755
	low-pass	-291	-317	-597	-1005

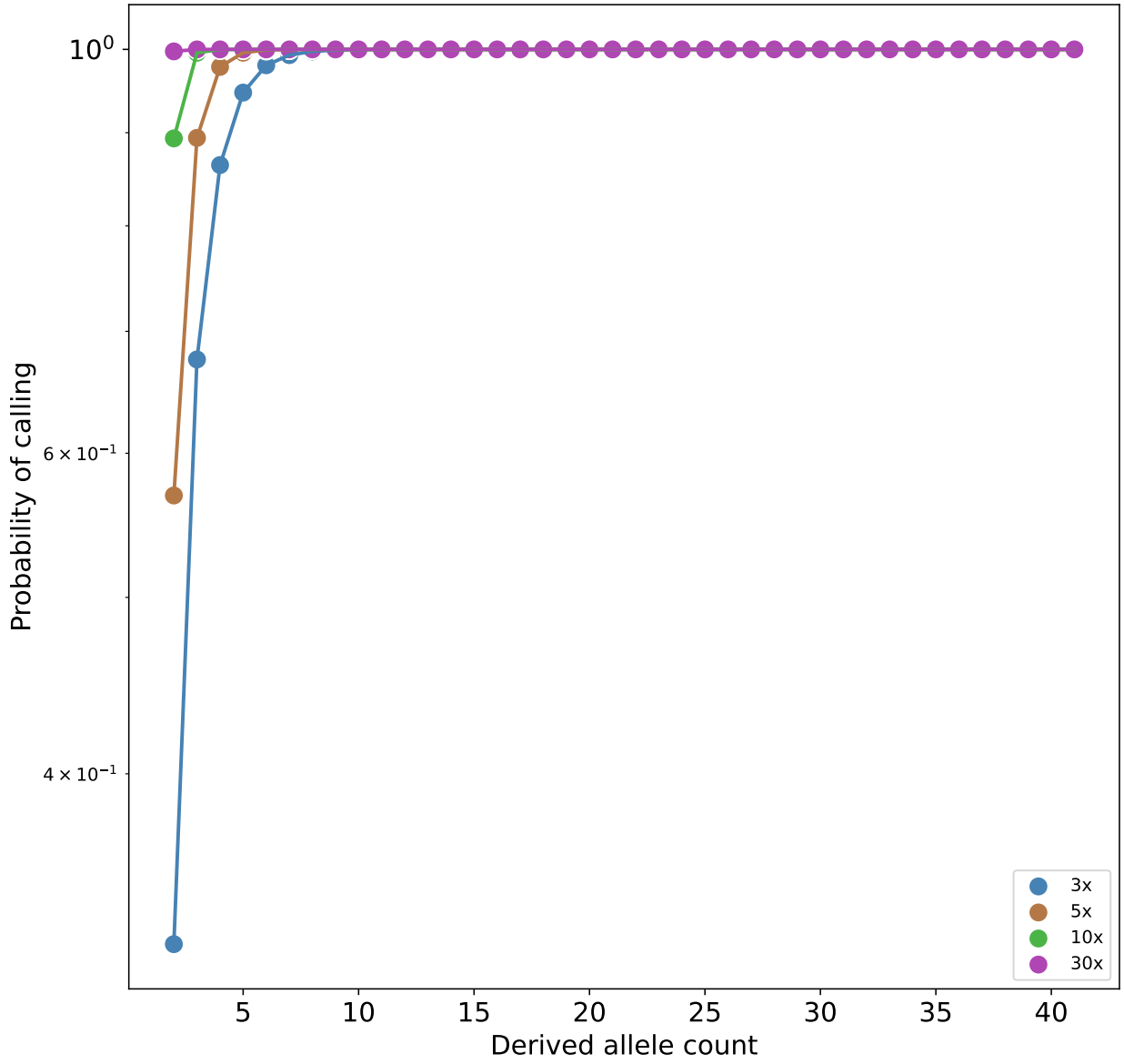

Figure S1: Probability of calling a variant site versus true allele frequency and coverage depth.

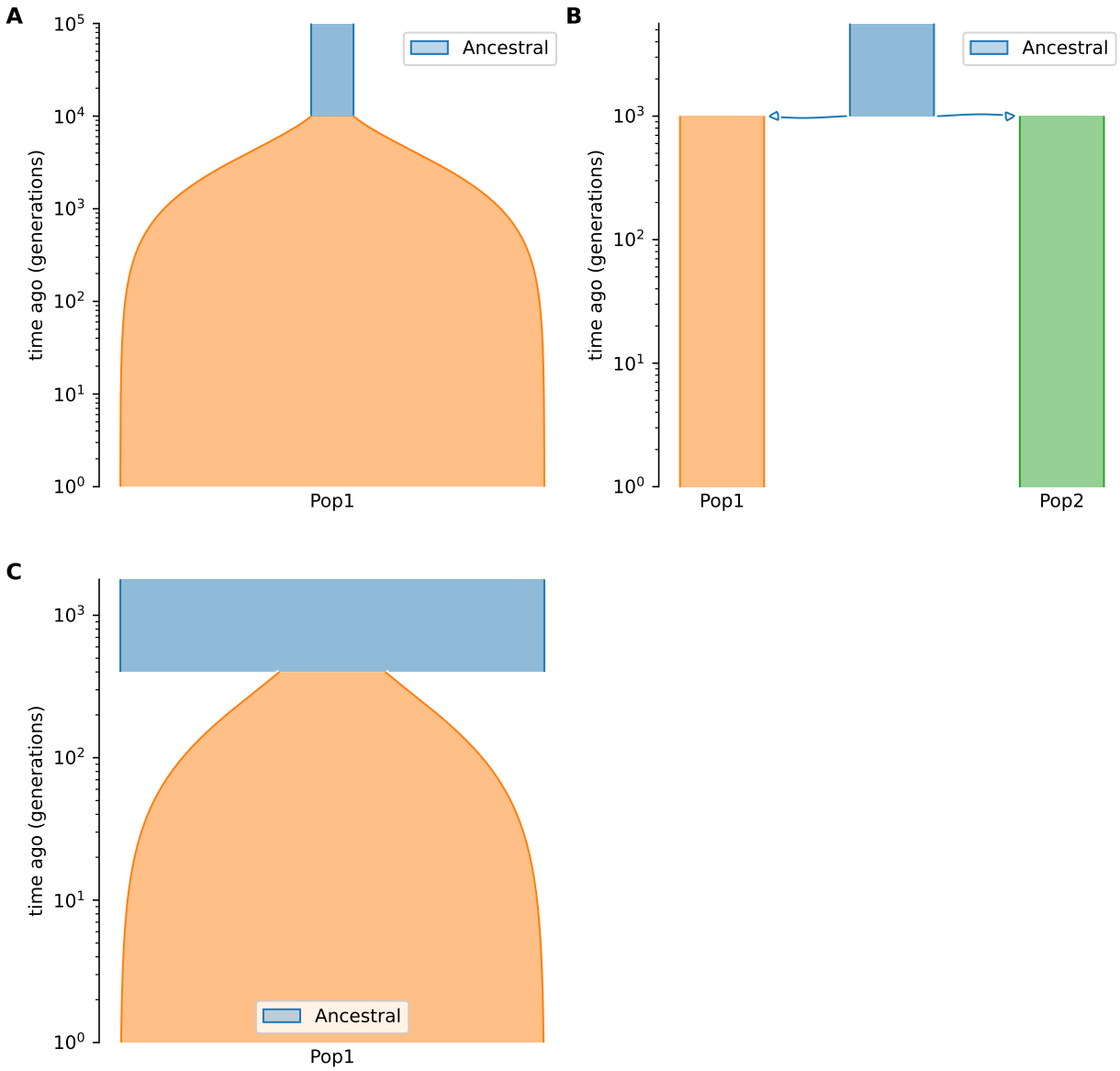

Figure S2: Representation of the demographic models used in the simulations: (A) single-population exponential growth model with parameters $\nu_1 = 10$ and $T = 0.1$, (B) two-population isolation model with $\nu_1 = \nu_2 = 1$ and $T = 0.1$, (C) single-population exponential growth model with inbreeding with parameters $\nu_1 = 4$, $T = 0.4$, and $F \in \{0.1, 0.5, 0.9\}$. ν, T, F represent relative population size, time in the past, and inbreeding coefficient, respectively. This plot was created with Demes (Gower et al. 2022)

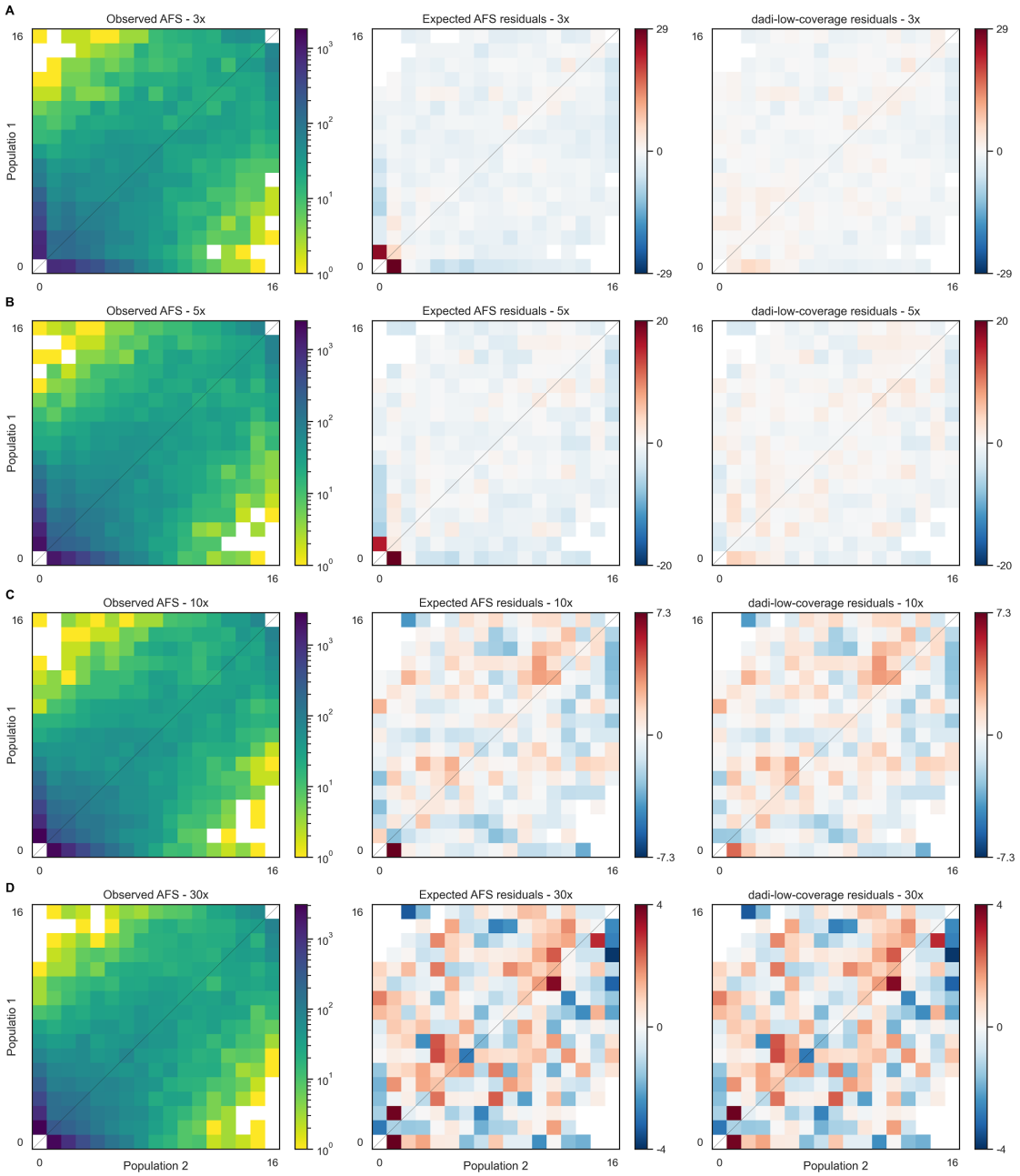

Figure S3: The observed 2D AFS is biased by low coverage. Deviation between the observed low-coverage AFS (first column) and the expected AFS (calculated by dadi) for the isolation demographic scenario is visualized through the residual plot (second column). Dark red residuals indicate that the observed low-coverage AFS is deficient in low-frequency alleles compared to the expectation. By contrast, the residuals between the observed AFS and the low-coverage model are much smaller. At 30× coverage (D) the residuals become small and random, indicating agreement between all three spectra. Coverage depths compared are (A) 3×, (B) 5×, (C) 10×, and (D) 30×.

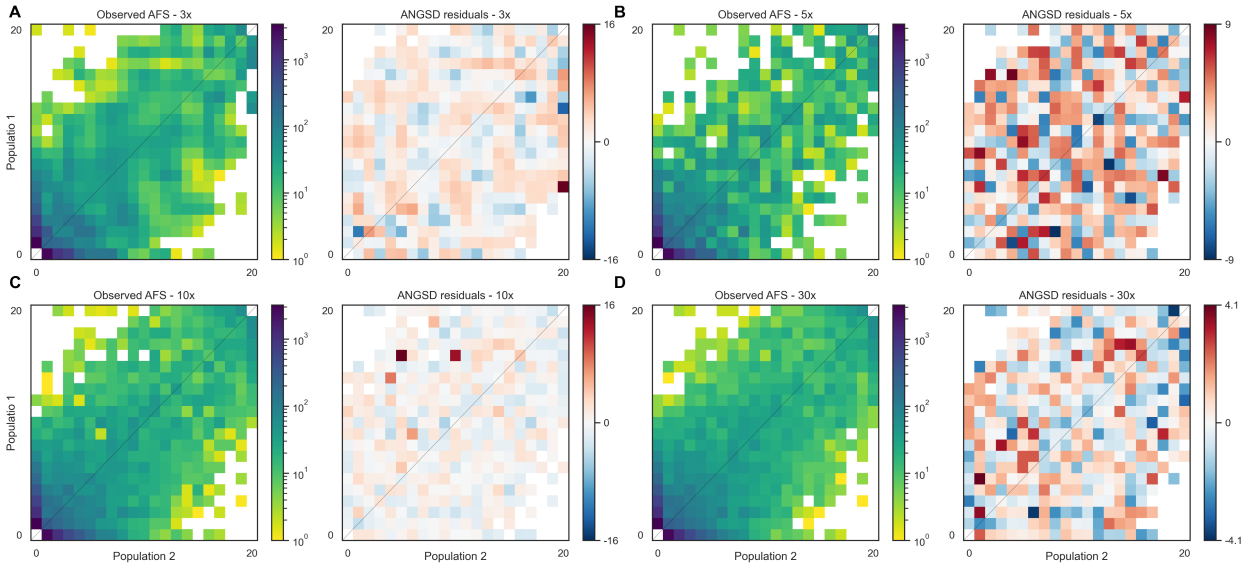

Figure S4: ANGSD creates fluctuations in the joint AFS. The joint AFS output by ANGSD exhibits sporadic very large residuals when compared with the true simulated AFS, similar to the oscillations seen in the single population AFS (Fig. 2). Coverage depths compared are (A) 3×, (B) 5×, (C) 10×, and (D) 30×.

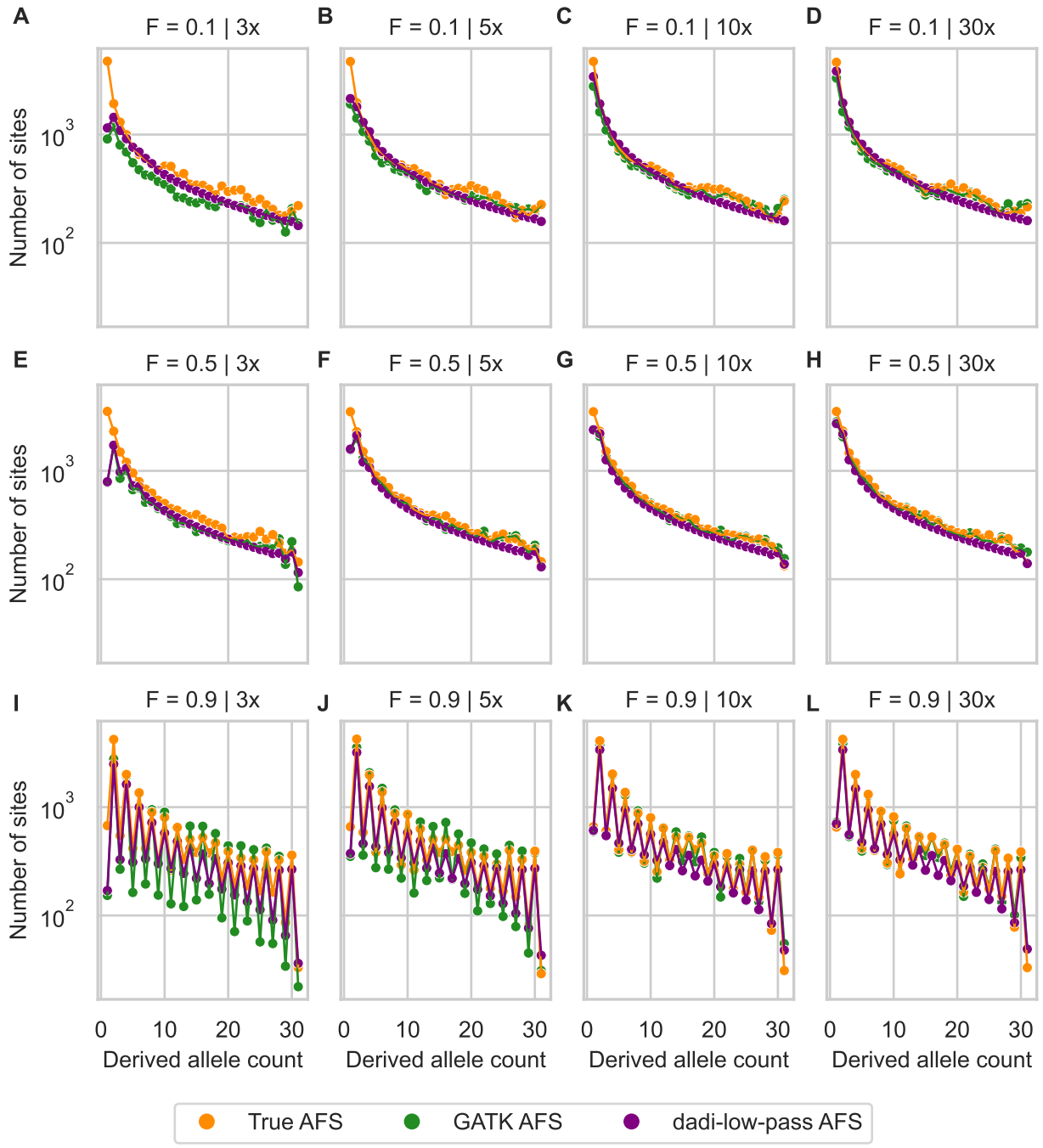

Figure S5: The observed AFS is impacted by low-pass sequencing (3×, 5×, 10×, and 30×) and inbreeding ($F \in \{0.1, 0.5, 0.9\}$). This figure presents a comparison of the observed AFS from low-pass variant calling with simulations in both the standard dadi and dadi-low-pass frameworks, using the true parameter values for a single-population model.

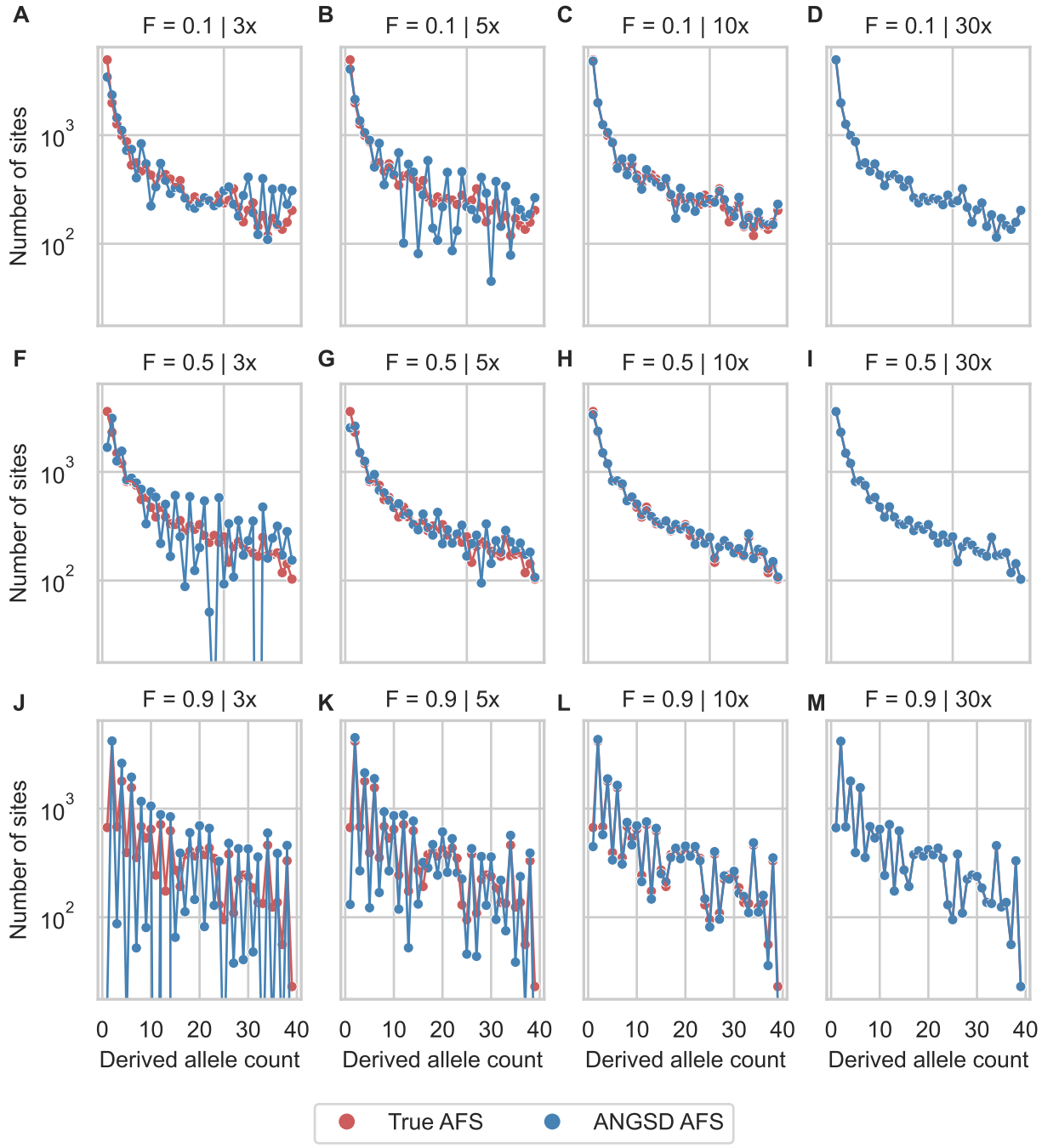

Figure S6: ANGSD corrects for the low-pass bias of the AFS, but it introduces fluctuations in inbreeding models. For the same simulations as Fig. S5, ANGSD (blue) was used to reconstruct the simulated AFS (red). Coverages were 3×, 5×, 10×, and 30×) and inbreeding 0.1, 0.5, and 0.9.

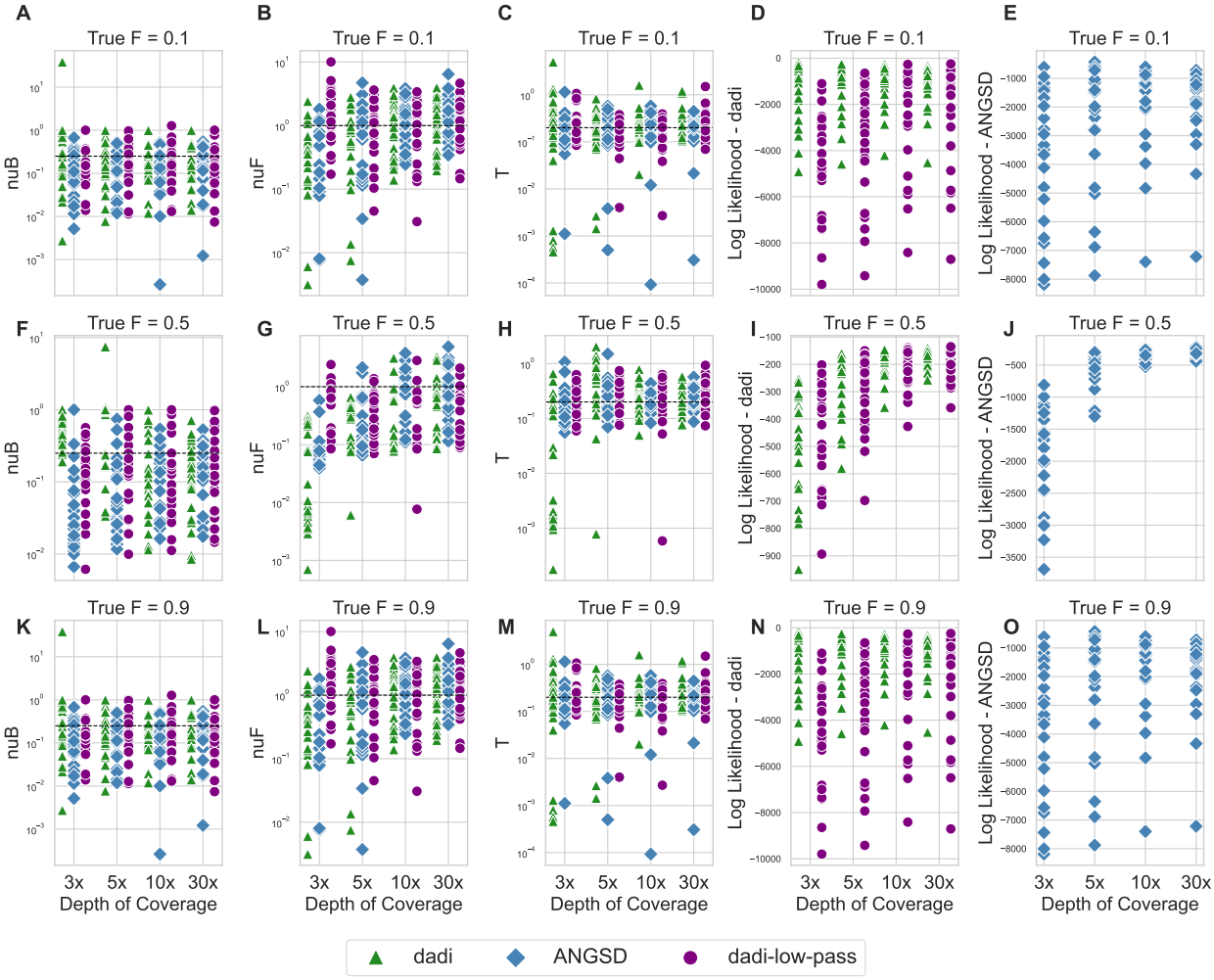

Figure S7: Graph showcasing the accuracy of parameter and likelihood estimations across various sequencing depths (3 ×, 5 ×, 10 ×, and 30 ×) and inbreeding ($F \in \{0.1, 0.5, 0.9\}$) for a population bottleneck and growth model. The inbreeding parameters were kept fixed for both the low-pass calculation and the optimization process. Parameters were obtained through different methods, including dadi, both with and without corrections for low coverage, as well as ANGSD. Details of the graph include: (A), (F), (K) the estimated size after population bottleneck; (B), (G), (L) the estimated size after population expansion; (C), (H), (M) the time of population expansion; (D), (I), (N) log-likelihood calculations from dadi, highlighting the distinction between corrected and uncorrected model for low coverage; and (E), (J), (O) log-likelihood calculations from ANGSD. The black line present in the plots for (A), (B), (E), (F), (I), (J) and indicates the true value of the parameter, providing a standard for evaluating the accuracy of different approaches.

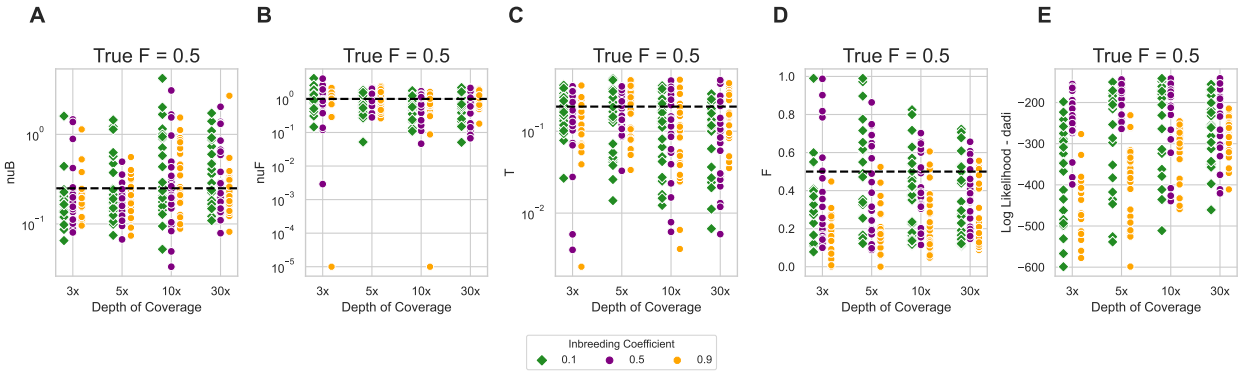

Figure S8: Graph showcasing the accuracy of parameter and likelihood estimations across various sequencing depths (3 ×, 5 ×, 10 ×, and 30 ×) and inbreeding ($F \in \{0.1, 0.5, 0.9\}$) for a population expansion model under a true inbreeding value of 0.5. The inbreeding parameters used for the low-pass calculation were 0.1, 0.5, and 0.9. Parameters were obtained using dadi-low-pass. Details of the graph include: (A) the estimated size after population bottleneck; (B) the estimated size after population expansion; (C) the time of population expansion; (D) inferred inbreeding coefficient; (E) log-likelihood calculations from dadi-low-pass. The black line present in the plots for (A), (B), (C), and (D) indicates the true value of the parameter, providing a standard for evaluating the accuracy of different approaches.

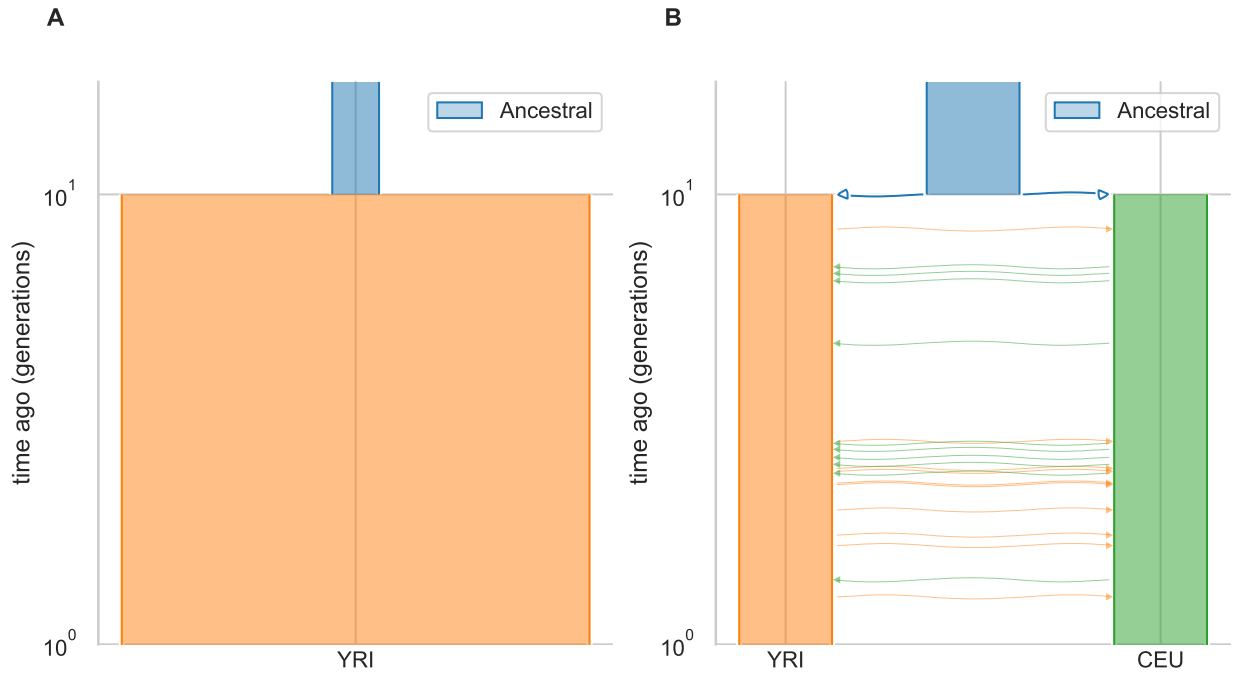

Figure S9: Representation of the demographic models used to analyse 1000 genomes datasets: (A) single-population two-epoch growth model with parameters, (B) two-population isolation with migration model. This plot was created with Demes (Gower et al. 2022)

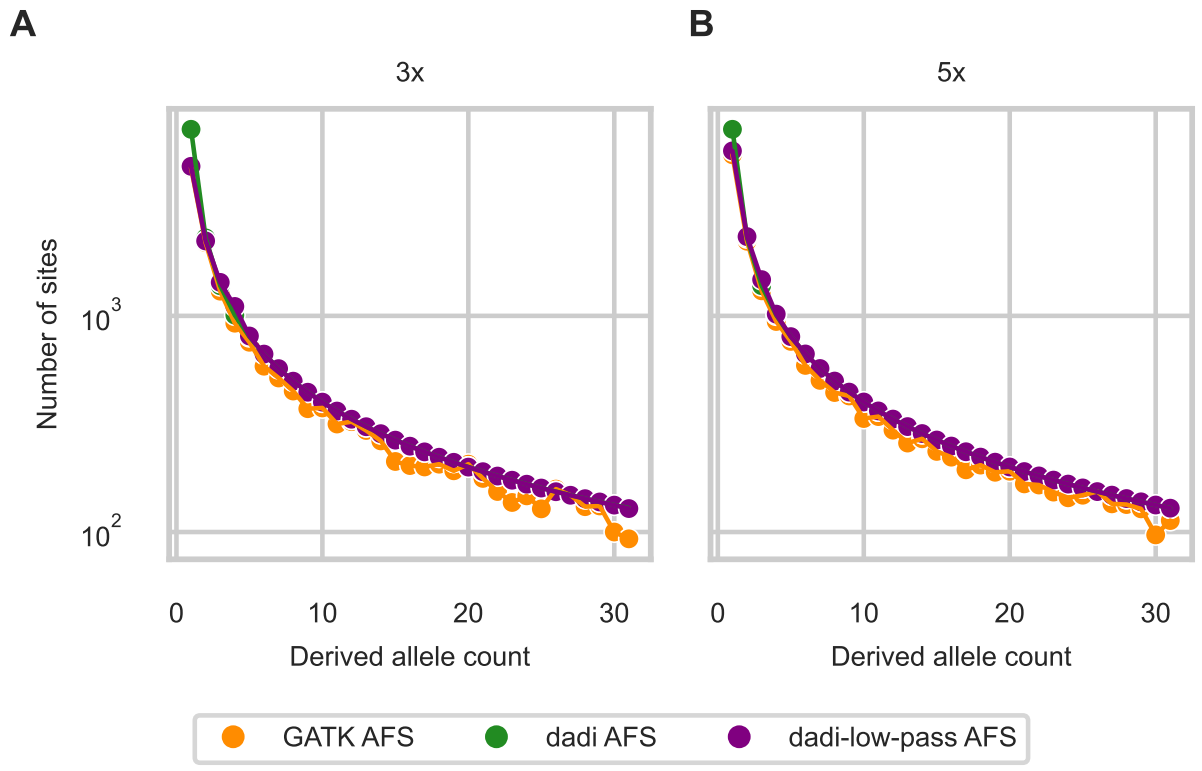

Figure S10: Unbalanced depth of coverage does not bias the dadi-low-pass model. Simulations were performed using 20 individuals, with half simulated under low-coverage conditions (A: 3× or B: 5×) and the other half under high-depth coverage (30 ×).

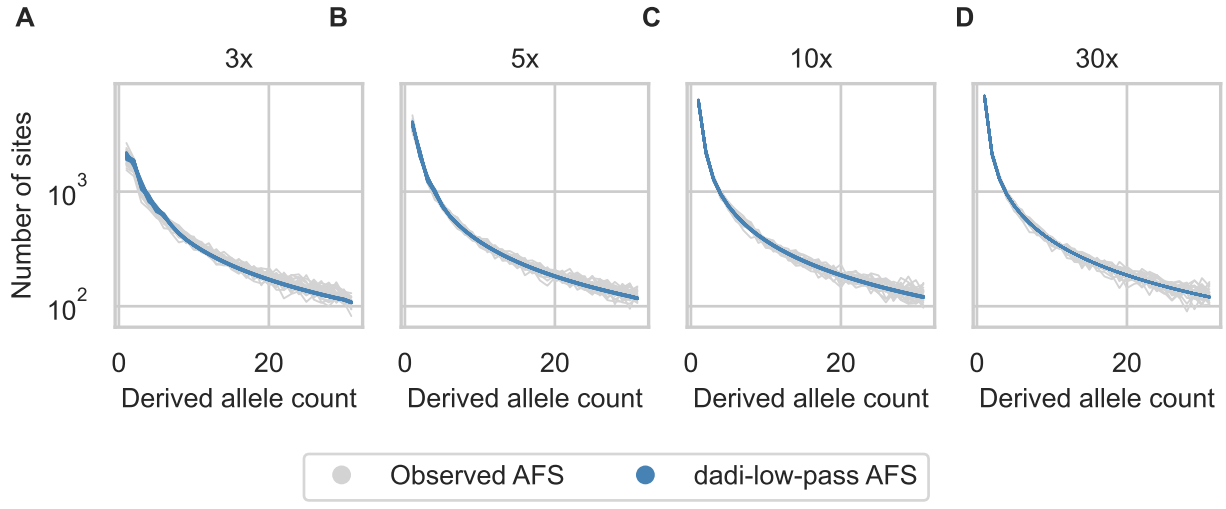

Figure S11: The simulated AFS under the low-pass model shows less variance compared to that observed in the simulated datasets. We generated 25 AFS for each condition.