bioRxiv preprint doi: https://doi.org/10.1101/2024.07.19.604366; this version posted July 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

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

bioRxiv preprint doi: https://doi.org/10.1101/2024.07.19.604366; this version posted July 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

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

bioRxiv preprint doi: https://doi.org/10.1101/2024.07.19.604366; this version posted July 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

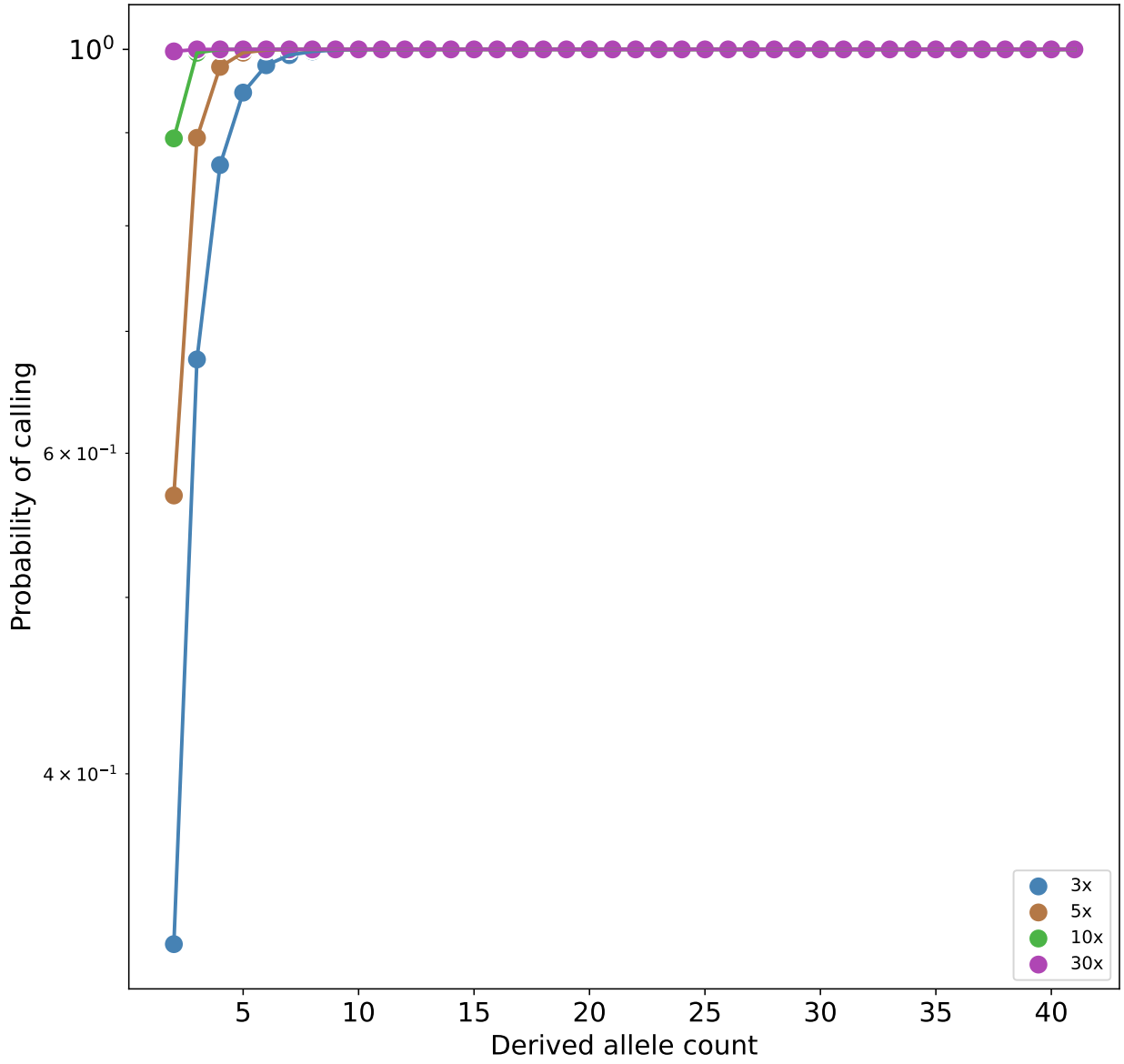

Figure S1: Probability of calling a variant site versus true allele frequency and coverage depth.

bioRxiv preprint doi: https://doi.org/10.1101/2024.07.19.604366; this version posted July 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

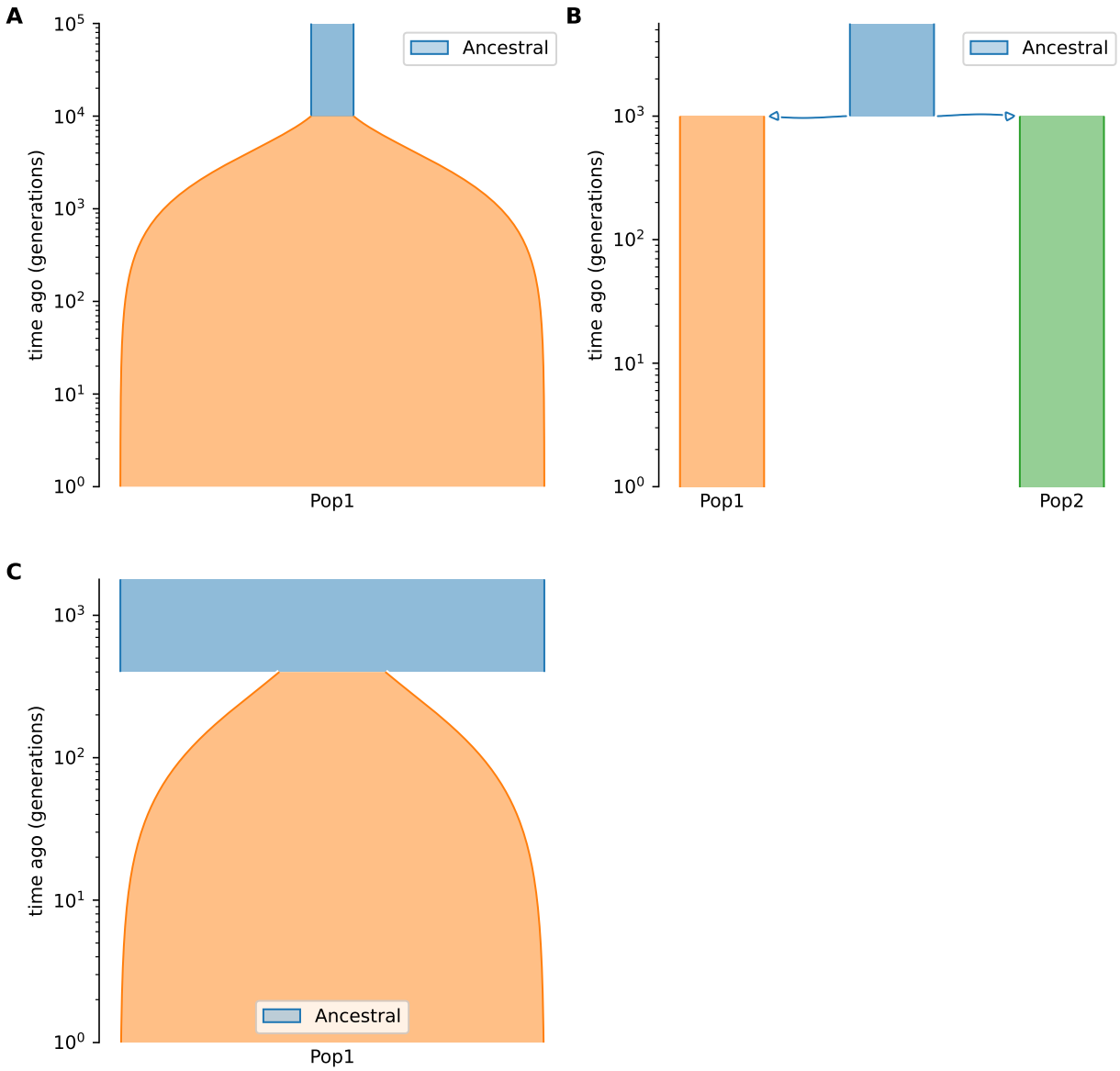

Figure S2: Representation of the demographic models used in the simulations: (A) single-population exponential growth model with parameters $\nu_1 = 10$ and $T = 0.1$, (B) two-population isolation model with $\nu_1 = \nu_2 = 1$ and $T = 0.1$, (C) single-population exponential growth model with inbreeding 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)

bioRxiv preprint doi: https://doi.org/10.1101/2024.07.19.604366; this version posted July 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

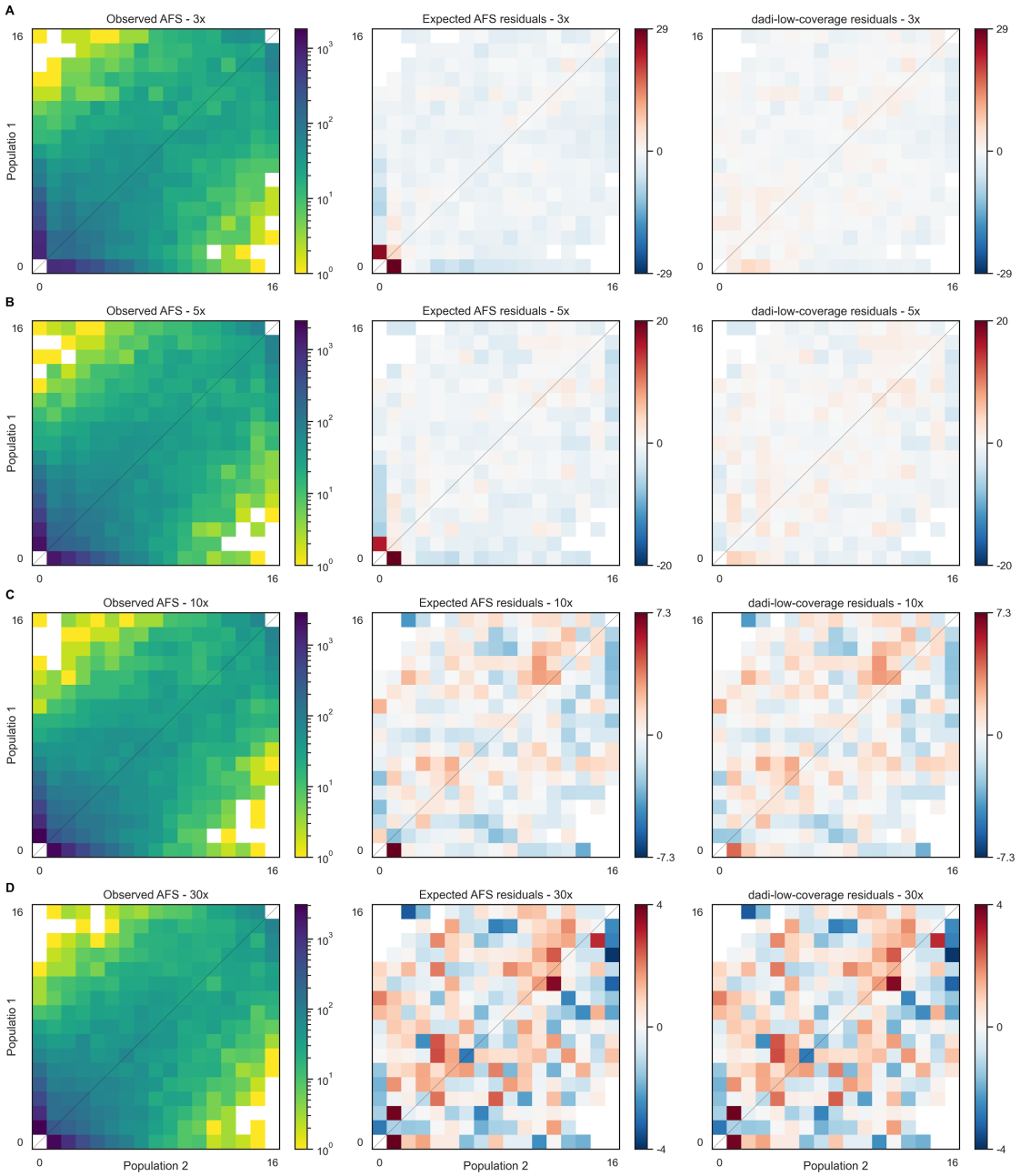

Figure S3: The observed 2D AFS is biased by low coverage. Deviation between the observed low-coverage AFS (first column) and the expected AFS (calculated by dadi) for the isolation demographic scenario is visualized through the residual plot (second column). Dark red residuals indicate that the observed 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×.

bioRxiv preprint doi: https://doi.org/10.1101/2024.07.19.604366; this version posted July 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

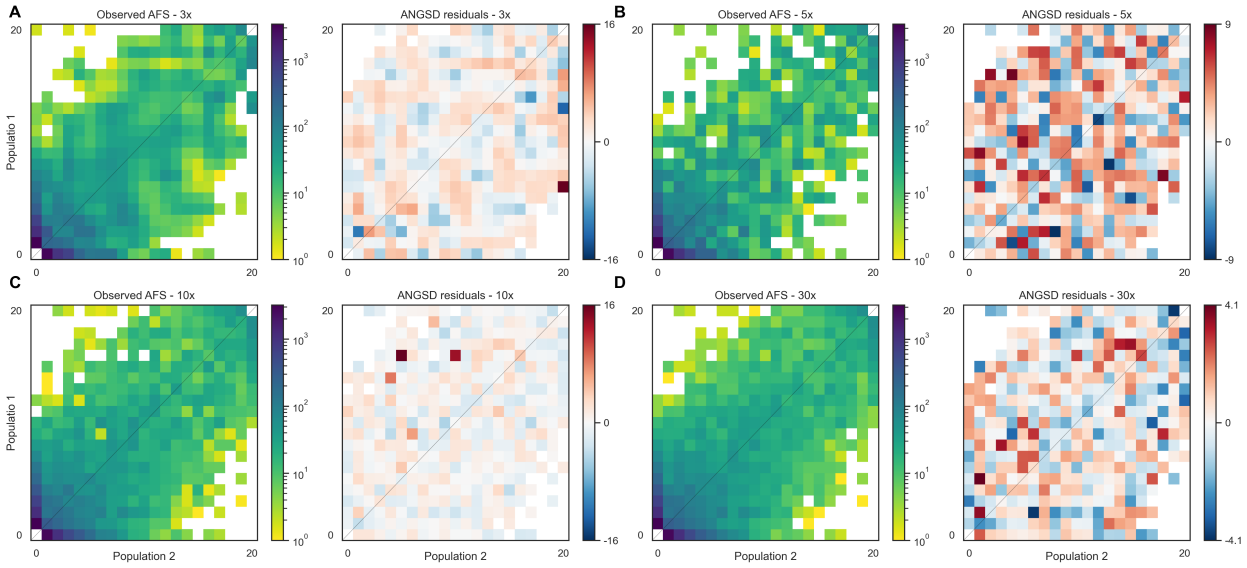

Figure S4: ANGSD creates fluctuations in the joint AFS. The joint AFS output by ANGSD exhibits sporadic very large residuals when compared with the true simulated AFS, similar to the oscillations seen in the single population AFS (Fig. 2). Coverage depths compared are (A) 3×, (B) 5×, (C) 10×, and (D)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

bioRxiv preprint doi: https://doi.org/10.1101/2024.07.19.604366; this version posted July 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

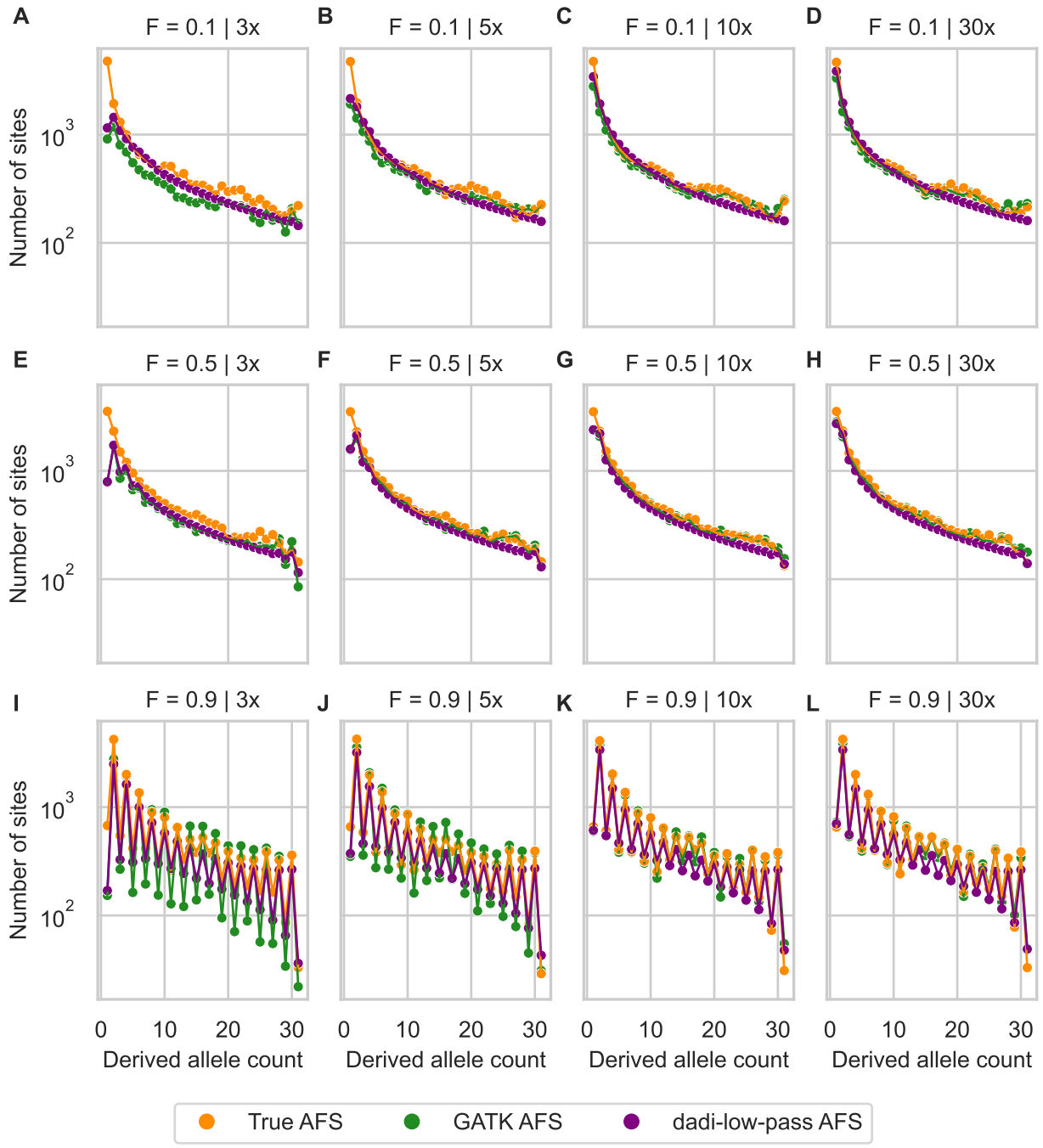

Figure S5: The observed AFS is impacted by low-pass sequencing (3×, 5×, 10×, and 30×) and inbreeding ($F \in \{0.1, 0.5, 0.9\}$). This figure presents a comparison of the observed AFS from low-pass variant calling with simulations in both the standard dadi and dadi-low-pass frameworks, using the true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

bioRxiv preprint doi: https://doi.org/10.1101/2024.07.19.604366; this version posted July 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

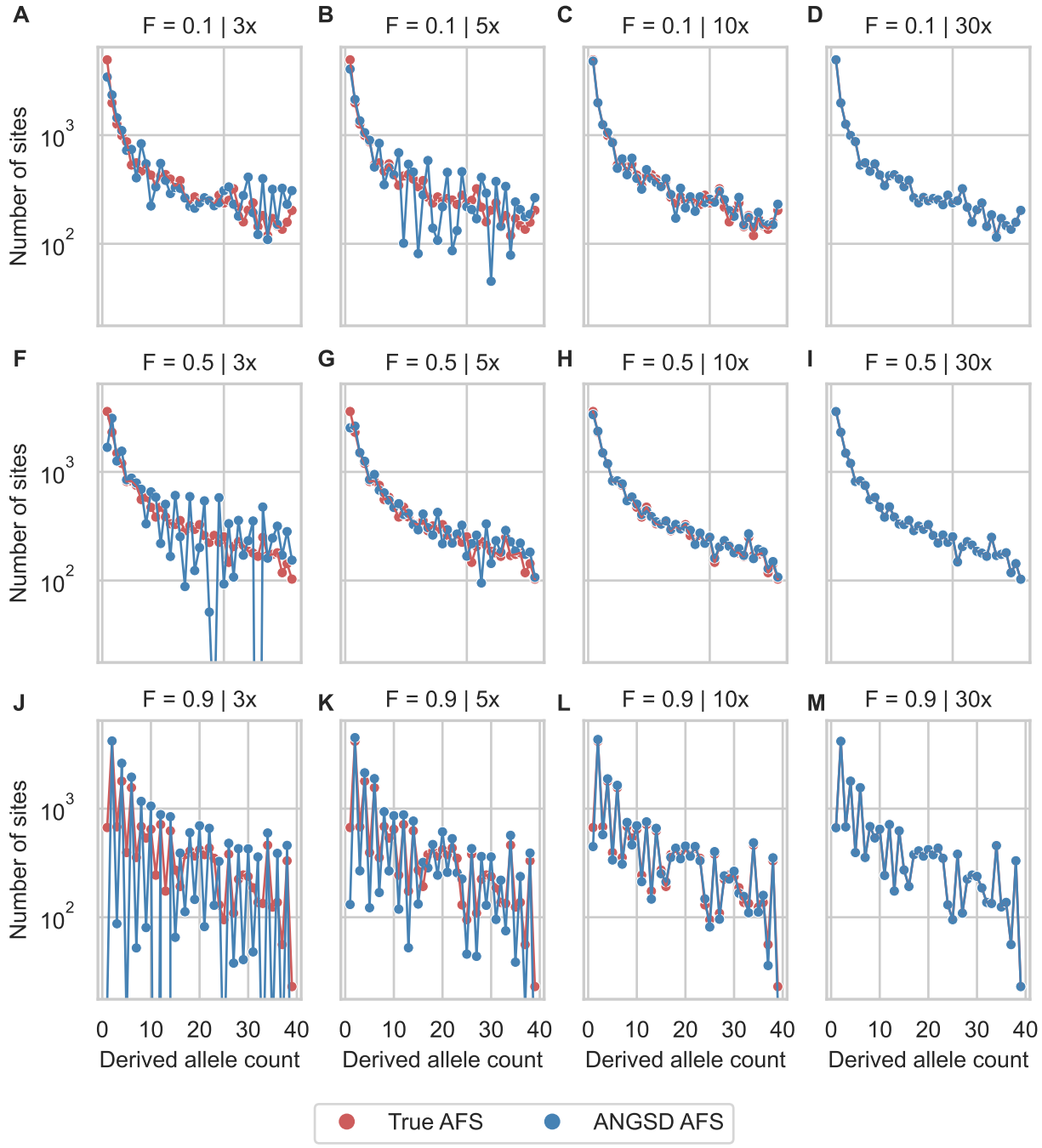

Figure S6: ANGSD corrects for the low-pass bias of the AFS, but it introduces fluctuations in inbreeding models. For the same simulations as Fig. S5, ANGSD (blue) was used to reconstruct the simulated AFS (red). Coverages were 3×, 5×, 10×, and 30×) and inbreeding 0.1, 0.5, and 0.9.

bioRxiv preprint doi: https://doi.org/10.1101/2024.07.19.604366; this version posted July 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

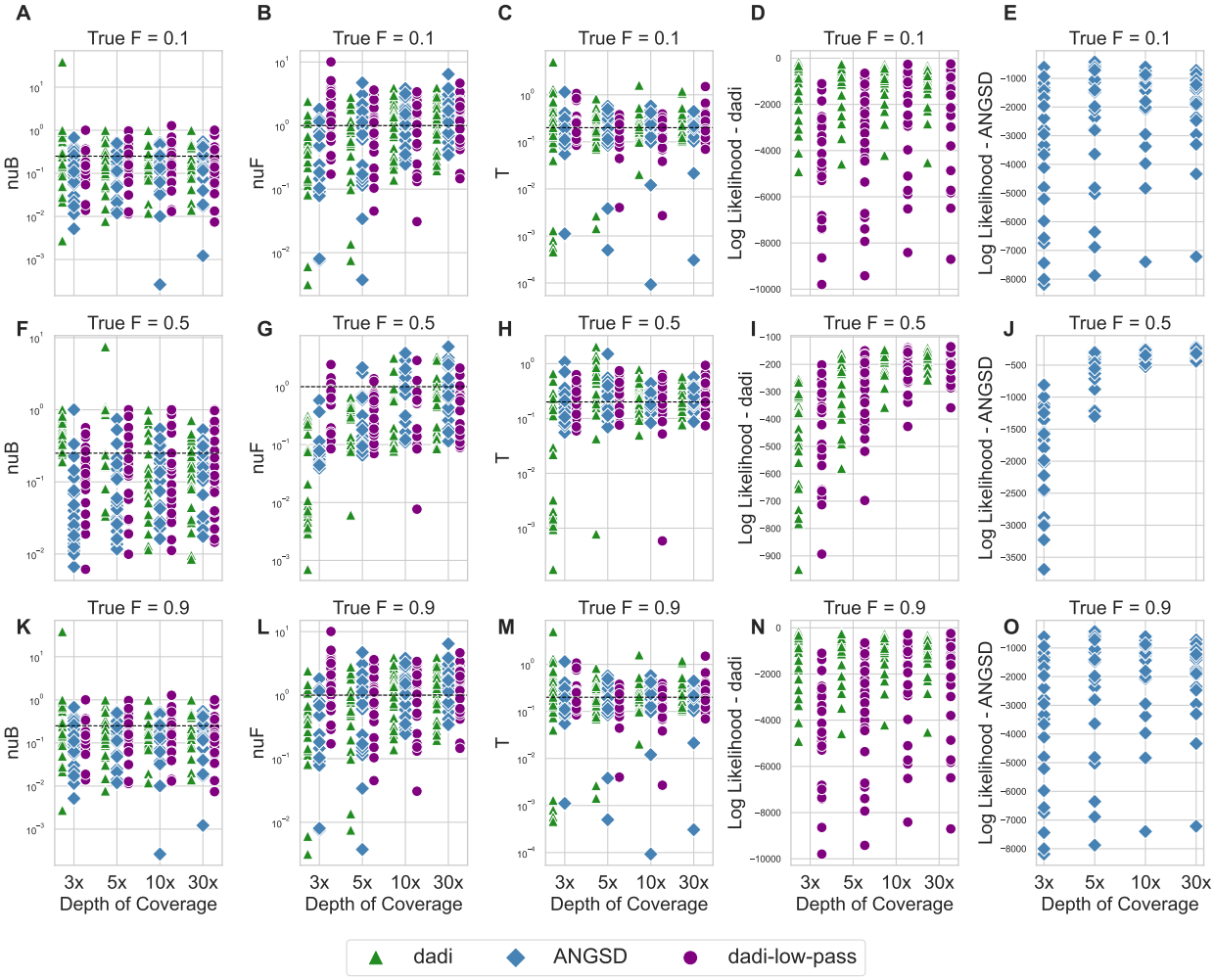

Figure S7: Graph showcasing the accuracy of parameter and likelihood estimations across various sequencing depths (3 ×, 5 ×, 10 ×, and 30 ×) and inbreeding ($F \in \{0.1, 0.5, 0.9\}$) for a population bottleneck and growth model. The inbreeding parameters were kept fixed for both the low-pass calculation 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.

bioRxiv preprint doi: https://doi.org/10.1101/2024.07.19.604366; this version posted July 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

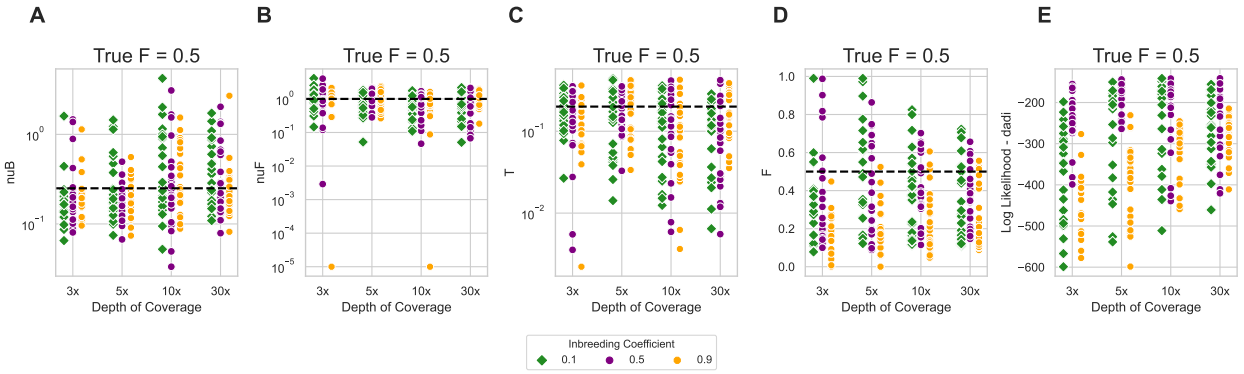

Figure S8: Graph showcasing the accuracy of parameter and likelihood estimations across various sequencing depths (3 ×, 5 ×, 10 ×, and 30 ×) and inbreeding ($F \in \{0.1, 0.5, 0.9\}$) for a population expansion model under a true inbreeding value of 0.5. The inbreeding parameters used for the low-pass 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.

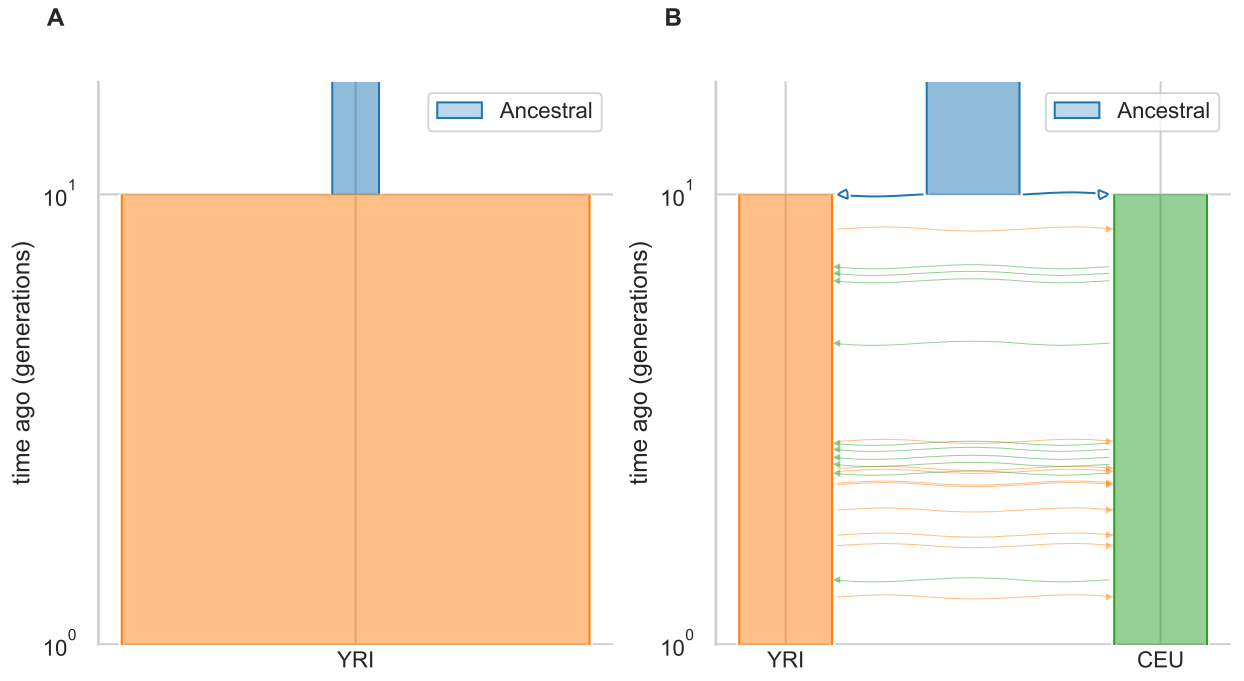

Figure S9: Representation of the demographic models used to analyse 1000 genomes datasets: (A) single-population two-epoch growth model with parameters, (B) two-population isolation with migration model. This plot was created with Demes (Gower et al. 2022)

bioRxiv preprint doi: https://doi.org/10.1101/2024.07.19.604366; this version posted July 23, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

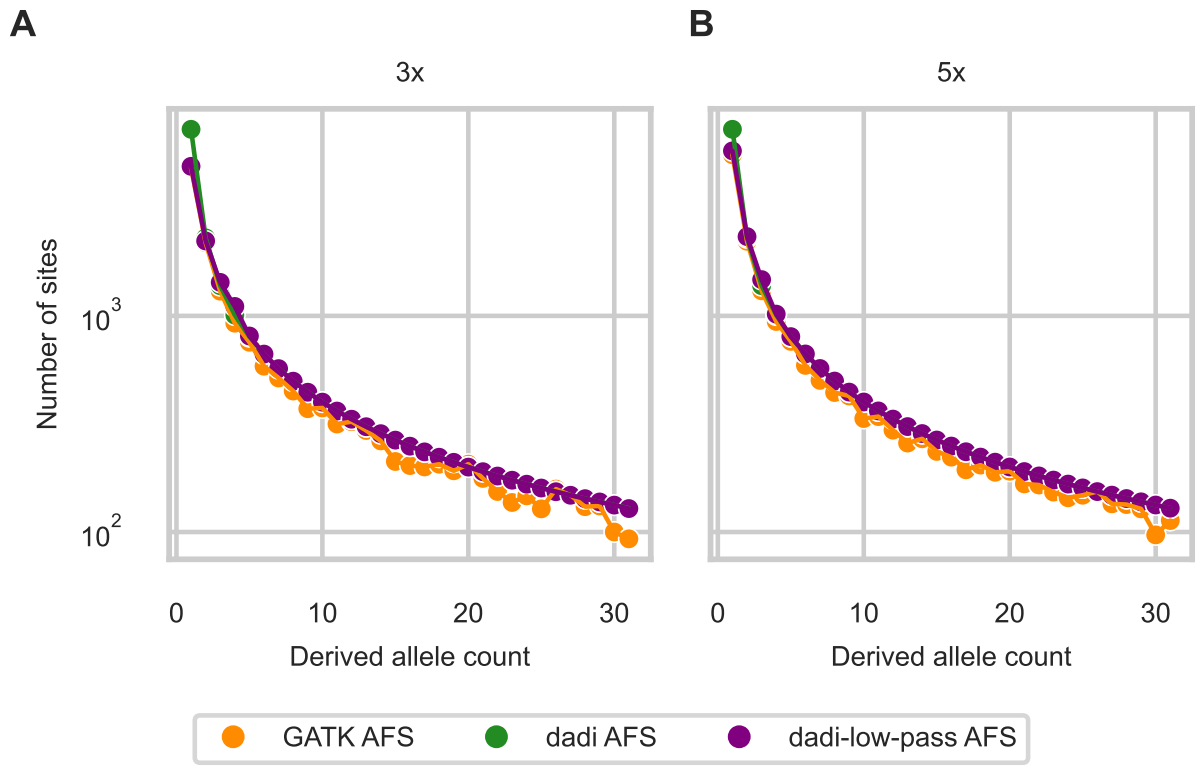

Figure S10: Unbalanced depth of coverage does not bias the dadi-low-pass model. Simulations were performed using 20 individuals, with half simulated under low-coverage conditions (A: 3× or B: 5×) and the other half under high-depth coverage (30 ×).

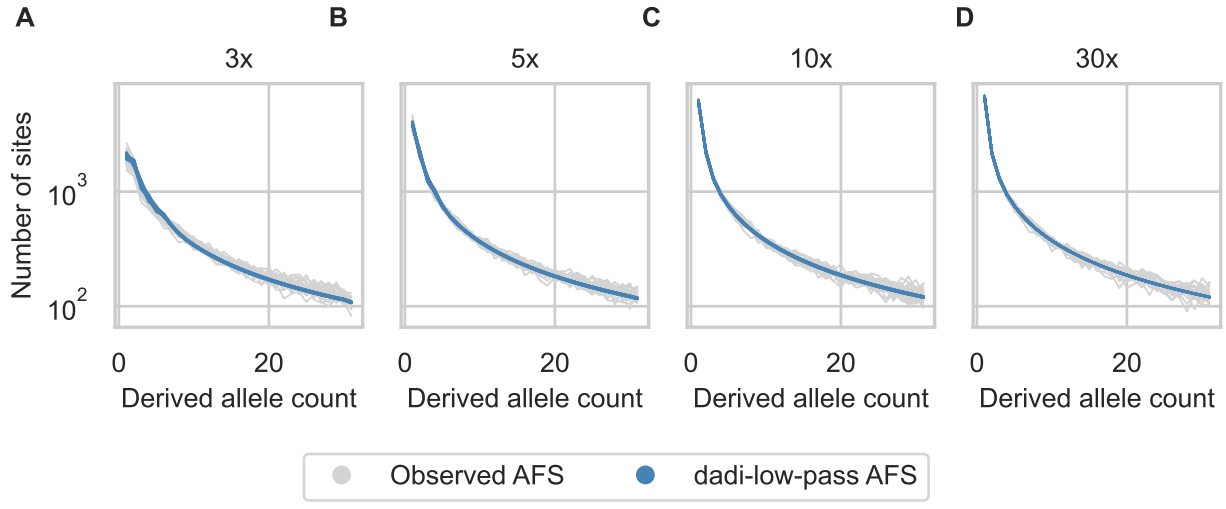

Figure S11: The simulated AFS under the low-pass model shows less variance compared to that observed in the simulated datasets. We generated 25 AFS for each condition.