*Bioinformatics Advances*, 2026, **6**(1), vbag125
https://doi.org/10.1093/bioadv/vbag125
Published: 4 May 2026
**Application Note | Genetics and population analysis**

OXFORD

# EGGS: Empirical Genotype Generalizer for Samples

**T. Quinn Smith**[1,*], **Amatur Rahman**[1], **Zachary A. Szpiech**[1,*]

[1]Department of Biology, The Pennsylvania State University, University Park, PA 16802, United States
*Corresponding authors. T. Quinn Smith, Department of Biology, The Pennsylvania State University, University Park, PA 16802, United States.
E-mail: tqs5778@psu.edu; Zachary A. Szpiech, Department of Biology, The Pennsylvania State University, University Park, PA 16802, United States.
E-mail: zps5164@psu.edu
Associate Editor: Scott Hazelhurst

## Abstract

***Motivation:*** **We introduce Empirical Genotype Generalizer for Samples (EGGS) which accepts empirical genotypes with missing data and replicates the distribution of missing genotypes along the empirical segment in other replicates. The empirical segment must have a number of sites less than the replicate. In addition, EGGS can remove phase, remove polarization, simulate deamination, simulate sequencing error, create pseudohaploids, and convert between Variant Call Format (VCF), *ms*-style replicates, and EIGENSTRAT/ANCESTRYMAP. When producing VCF files, EGGS is not limited to biallelic sites and assumes all samples are diploid.**

***Availability:*** **EGGS is written in the C programming language. Precompiled executables, source code, the manual, and the analysis conducted in the paper are available at https://github.com/TQ-Smith/EGGS**

## 1 Introduction

Recent advantages in both forward and backward in-time simulations have made it possible to generate synthetic genotypes under varying evolutionary assumptions and complicated demographic scenarios (Kelleher *et al*. 2016, Kern and Schrider 2016, Lauterbur *et al*. 2023, Haller *et al*. 2026). Such simulations are essential for rigorously testing computational methods, and more recently, training machine learning models (Schrider and Kern 2018, Flagel *et al*. 2019, Amin *et al*. 2024, Rahman *et al*. 2025). In addition, simulations are often used to test hypotheses concerning the evolutionary histories of human and nonhuman species (Amin *et al*. 2024, Rayner *et al*. 2024).

Simulated samples are reported under idyllic conditions where each genotype is phased, and in the case of biallelic sites, each site has a known ancestral allele. Simulated genotypes are devoid of technical artifacts that introduce variant uncertainty, such as missing alleles and deamination. Empirical data are often error-prone and violate the assumptions of simulated data. Ignoring this discrepancy in data quality can potentially lead to erroneous inference and unreliable results. Robust inference in the presence of low-quality genotypes is of particular interest when studying ancient DNA (aDNA) (Hofreiter *et al*. 2001, Williams and Huber 2025). This has motivated the development of methods to account for genotype uncertainty (Achaz 2008, Hellmann *et al*. 2008, Johnson and Slatkin 2008, Korneliussen *et al*. 2014, Korunes and Samuk 2021, Bailey *et al*. 2025).

Missing alleles are caused by compounding sources, including sample quality, sequencing technology, and data processing. The inherent stochasticity and complexity of such error are randomly introduced into simulated data (Pandey *et al*. 2024). This approach assumes that the proportion of missing samples at each site viewed in aggregate reflects the underlying structure of missing genotypes. However, a cumulative summary of missing sites for each sample ignores stretches of missing genotypes across the empirical segment, especially in regions of low complexity. In addition, the random modelling of missing data using a predefined distribution may inaccurately capture the empirical distribution. These difficulties are often exacerbated in non-model organisms lacking a high-quality reference genome (Mora-Márquez *et al*. 2025). Several recent studies introduce missingness blocks that correspond to regions of low mappability and variant callability (Skov *et al*. 2018, Arnab *et al*. 2025). Other current methods that introduce missing genotypes require a FASTA file (Kern and Schrider 2018), ignore the position of the missing sites within the simulated segment, are restricted to specific simulation frameworks, simulate error prone reads directly (Renaud *et al*. 2017), or are not generalizable to any supplied Variant Call Format (VCF) file (Goulart and Samuk 2026).

We introduce Empirical Genotype Generalizer for Samples (EGGS) which extracts the underlying distribution and tracks of missing sites across a genomic region and introduces similar patterns of missing genotypes in replicates. To avoid confusion, we define a missing genotype at a locus where a sample is missing

**Received:** 29 October 2025. **Revised:** 13 April 2026. **Accepted:** 25 April 2026

both alleles. In a VCF, this is denoted with ‘./.’. EGGS works on replicates with a number of sites less than the empirical segment and any number of samples. We demonstrate the use of EGGS and illustrate its ability to capture realistic patterns of missing genotypes for a simulated set of samples.

# 2 Methods

## 2.1 Replicating missing sites

Consider $S$ samples genotyped at $N$ sites. Our goal is to introduce the structure of missing genotypes in a synthetic replicate. Let $T$ be the number of samples in the synthetic replicate, where the $T$ samples are genotyped at $M < N$ sites. $T$ is not restricted by $S$. A missing genotype is a site where the sample is missing both alleles. We partition the $N$ sites into $M$ blocks. Each block contains $\lfloor \frac{N}{M} \rfloor$ sites, except the first $N(\mathrm{mod}\, M)$ blocks which contain $\lfloor \frac{N}{M} \rfloor + 1$ sites.

For each sample $1 \leq i \leq S$, we calculate the average number of missing genotypes within a block. This quantity is denoted by $p_{ij}$, where $1 \leq j \leq M$. We introduce missing genotypes into the synthetic replicate by repeating the following process for each of the $T$ samples: For each site $1 \leq j \leq M$, randomly choose one of the $S$ samples $s \sim Uniform(1, \ldots, S)$. We label the site as missing with probability $p_{sj}$.

Representing missing sites in $M < N$ sites forces some amount of information to be lost. By partitioning the $N$ sites into $M$ blocks and summarizing the average number of missing genotypes for each sample within each block, we are capturing the larger trends of missing genotypes, and essentially, compressing the pattern of missing genotypes. The resampling procedure allows the missing genotypes to be recreated among many synthetic samples. Our method allows empirical segments containing regions with large amounts of missingness to be recreated in smaller sized synthetic replicates. This is analogous to how diploS/HIC introduces missing genotypes in VCF files (Kern and Schrider 2018). Instead of using two VCF files, diploS/HIC partitions a FASTA given a VCF file and labels the site as missing for all samples depending on the current FASTA partition’s ‘N’ content (Kern and Schrider 2018).

## 2.2 Implementation

EGGS accepts variants in VCF (Danecek *et al*. 2011) or *ms*-style formatting (Hudson 2002). In addition, we provide an option for the conversion from EIGENSTRAT/ANCESTRYMAP format, the format used by The Allen Ancient DNA Resource (AADR), to VCF (Mallick and Reich 2023, Mallick *et al*. 2024). In the case of multiple *ms*-style replicates as input, replicates are split to separate VCF files. We provide an option that splits each lineage to a separate sample set. This allows each lineage in an *ms*-style replicate or in a VCF file to be treated as its own sample in the resulting VCF. When creating VCF files, EGGS produces diploid samples.

To remove phase, EGGS provides two options. The first option switches the left and right allele at each genotype with equal probability. The second option places the smaller-index allele first in the genotype. In the case of VCF output, each unphased genotype is marked with ‘/’ as opposed to ‘|’. Computational methods that rely on phase typically ignore the VCF specification (Danecek *et al*. 2011) and rely on the position of the left and right allele at a genotype. Therefore, changing ‘|’ to ‘/’ is not sufficient to erase phase information. The reference allele at each site is typically treated as the ancestral allele, which in turn, is used as the outgroup’s state in many analyses. The choice of the ancestral allele is known to affect inference (Patterson *et al*. 2012). Treating an allele as the ancestral is known as polarization. To remove polarization from biallelic sites, EGGS swaps the ancestral and derived alleles for all samples’ alleles by a user supplied probability. EGGS can introduce missing genotypes randomly according to the approach of Pandey *et al.* (2024). The proportion of samples with both alleles missing is calculated per site. The mean and standard deviation of the proportion of missing genotypes is calculated over all sites and is used to parameterize a $\beta$-distribution. The $\beta$-distribution is used to randomly choose the proportion of missing samples at each site in the supplied genotypes. The user can choose to directly give the mean and standard deviation to parameterize the $\beta$-distribution or a VCF from which the values are calculated.

Another complication commonly arising in working with aDNA is deamination. Deamination is a type of damage, when usually, a cytosine is incorrectly read as a thymine during sequencing (Hofreiter *et al*. 2001). Deamination is simulated according to Harney *et al.* (2021). The user supplies two values: the probability that a site is a transition and the probability that a called reference allele will deaminate to the alternative allele. Deamination is synthetically introduced into the dataset by applying the above to every sample’s alleles at all site. An additional complication arising with aDNA is the presence of pseudohaploids. The low quality of aDNA makes confidently calling heterozygotes difficult. Instead, at sites with mapped reads showing multiple alleles, one allele is randomly chosen (Pandey *et al*. 2024). Pseudohaploids are created by randomly choosing each sample to be homozygous for either the reference or alternative allele with equal probability at sites where the sample contains both the reference and alternative allele. For sites where one of the alleles are missing, the non-missing allele is made homozygous.

Modern-day genomes are less degraded compared to aDNA and contain much lower error rates on next-generation sequencing machines (Stoler and Nekrutenko 2021). We introduce sequencing error by switching the state of every sample’s called allele at a biallelic site with a user supplied probability. Finally, summary statistics regarding missing genotypes can be calculated similar to those implemented in PLINK (Purcell *et al*. 2007). EGGS can produce *ms*-style output when missing data are not introduced. Some options cannot be used together to avoid ambiguity in precedence. For example, we do not allow sequencing error to be introduced when deamination is being introduced. This is described in the manual. EGGS is written in the C programming language to enable fast processing of thousands of replicates. Precompiled executables, source code, and manual are available at https://github.com/TQ-Smith/EGGS.

We used EGGS to convert the AADR (Mallick and Reich 2023) to VCF, and we extracted chromosome 1 of the 217 samples exclusively from Mathieson *et al.* (2018) with bcftools (Danecek *et al*. 2021). This resulted in 93 166 sites. The sites had a mean

**Table 1** Results of DTW.[a]

| Length of segment (Mb) | Number of segregating sites | Block size | $\beta$-Distribution | EGGS |
|---|---|---|---|---|
| 1 | 3327 | 28 | 57.49 | 57.38 |
| 2 | 6612 | 14 | 55.79 | 55.02 |
| 3 | 9657 | 9 | 54.57 | 53.09 |
| 4 | 12 659 | 7 | 53.46 | 51.39 |
| 5 | 16 468 | 5 | 52.51 | 49.2 |
| 6 | 19 930 | 4 | 51.53 | 47.33 |
| 7 | 23 012 | 4 | 51.08 | 45.67 |
| 8 | 25 682 | 3 | 50.45 | 44.29 |
| 9 | 29 391 | 3 | 49.84 | 42.39 |
| 10 | 33 023 | 2 | 49.27 | 40.92 |

[a] The first column is the length of the simulated segment in megabases. The second column is the number of segregating sites in the simulated segment. The third column is the minimal number of segregating sites in the block. The fourth column is the distance as calculated by the DTW between the missingness in the empirical segment and the missingness introduced in the simulated replicate using the $\beta$-distribution. The fifth column is the distance as calculated by the DTW between the missingness in the empirical segment and the missingness introduced in the simulated replicate using EGGS's method. For the $\beta$-distribution and EGGS's method, the DTW was calculated with the average proportion of missing samples at each site in the simulated segment over 10 runs.

missingness of 0.55 and a standard deviation of 0.21. We refer to this set as the ancient samples.

Using *msprime* (Kelleher *et al*. 2016), we simulated 200 diploid samples on segments ranging from 1 Mb to 10 Mb in 1 Mb increments. The segments were generated with the effective population size $Ne = 10,000$, mutation rate $\mu = 1.25 \times 10^{-8}$ per base pair, and the recombination rate $\rho = 1 \times 10^{-8}$ per base pair. We generated a range of segment lengths because the number of segregating sites in each replicate is the important variable when evaluating EGGS. We ran EGGS 10 times for each replicate introducing missing sites with EGGS's method and using the $\beta$-distribution method. At each site, we average the proportion of samples missing for each method across the 10 runs. The average proportion of samples missing at each site for each segment length is used for comparison with the ancient samples.

Dynamic time warping (DTW) is a signal processing technique used to compare two signals of different lengths assuming that one signal is a slowed down or sped up version of the other. DTW is often used to evaluate compression or interpolation methods and is analogous to sequence alignment in bioinformatics (Senin 2008). We use DTW to evaluate our method. We compute the proportion of missing samples at each of the 93 166 sites in our ancient samples. We treat this as a signal and use DTW to compare it to the average signal produced by our simulated replicates. DTW computes the Euclidean distance between two aligned time points. A lower value indicates that the two signals are more similar. We purposefully omit the units of this value as the interpretation is difficult in this context. The results of the DTW for the replicates is shown in Table 1.

For all segment lengths, we see that EGGS's method captures the proportion of missing samples along the empirical chromosome better than the approach of the $\beta$-distribution. The discrepancy between the two methods is smaller for shorter segments and becomes larger for longer segments. This suggests that introducing missing data into shorter segments according to a $\beta$-distribution may be sufficient when missing data are not a concern for downstream analysis; however, as the number of segregating sites in the simulated segment approaches the number of segregating sites in the empirical segment, EGGS's method better realizes fluctuations in the proportion of missing samples.

# 3 Conclusion

We introduced EGGS that artificially removes genotypes in patterns similar to empirical data as opposed to treating the distribution of missing data across a segment as purely random. Combining EGGS's ability to remove genotypes and other evolutionary assumptions, we can quickly make simulations more realistic, enabling more robust downstream analyses.

EGGS's method to replicate missing genotype patterns becomes limited when there is a large difference in the number of loci between the set of variation containing the patterns to replicate and the set of variation in which missing genotypes will be introduced. Such a discrepancy between the number of loci causes the resolution of EGGS's stratification method to decline. In addition, we expect that in situations where the rate of missing genotypes is high and the segment is small (less than 1 Mb), EGGS's approach will not yield significantly different results than modelling missingness with a beta-distribution. We also assume that adjacent missing sites for samples are not correlated. While this assumption is justified for aDNA (Pandey *et al*. 2024), this is not justified for less error prone data. In such a case, our approach could be replaced with more sophisticated signal processing techniques and offers a path for future work (Sayood 2017). Finally, we note that our approach could be adapted to replicating other characteristics of genotype uncertainty, such as homozygosity and ascertainment (Clark *et al*. 2005, Ramirez-Soriano and Nielsen 2009, Lachance and Tishkoff 2013).

# Acknowledgements

The authors would like to thank Abigail N. Sequeira, Ana V Leon-Apodaca, Anna Maria Calderon, and three reviewers for helpful suggestions. Computations for this research were performed

using the Pennsylvania State University's Institute for Computational Data Sciences' Roar supercomputer.

## Author contributions

T. Quinn Smith and Zachary A. Szpiech conceived the experiment(s), T. Quinn Smith conducted the experiment(s) and analyzed the results. T. Quinn Smith, Zachary A. Szpiech, and Amatur Rahman wrote and reviewed the manuscript.

## Conflicts of interest

The authors declare that they have no competing interests.

## Funding

This work was supported by the National Institute of General Medical Sciences of the National Institutes of Health award number R35GM146926 to Z.A.S.

## Data availability

The data underlying this article are available in the Allen Ancient DNA Resource (AADR) at https://dataverse.harvard.edu/dataverse/reich_lab.

## References

Achaz G. Testing for neutrality in samples with sequencing errors. *Genetics* 2008;**179**:1409–24. https://doi.org/10.1534/genetics.107.082198

Amin MR, Hasan M, DeGiorgio M. Digital image processing to detect adaptive evolution. *Mol Biol Evol* 2024;**41**:msae242. https://doi.org/10.1093/molbev/msae242

Arnab SP, Campelo dos Santos AL, Fumagalli M *et al.* Efficient detection and characterization of targets of natural selection using transfer learning. *Mol Biol Evol* 2025;**42**:msaf094. https://doi.org/10.1093/molbev/msaf094

Bailey N, Stevison L, Samuk K. Correcting for bias in estimates of $\theta w$ and Tajima's $D$ from missing data in next-generation sequencing. *Mol Ecol Resour* 2025;**25**:e14104. https://doi.org/10.1111/1755-0998.14104

Clark AG, Hubisz MJ, Bustamante CD *et al.* Ascertainment bias in studies of human genome-wide polymorphism. *Genome Res* 2005;**15**:1496–502. https://doi.org/10.1101/gr.4107905

Danecek P, Auton A, Abecasis G *et al.*; 1000 Genomes Project Analysis Group. The variant call format and vcftools. *Bioinformatics* 2011;**27**:2156–8. https://doi.org/10.1093/bioinformatics/btr330

Danecek P, Bonfield JK, Liddle J *et al.* Twelve years of samtools and bcftools. *Gigascience* 2021;**10**:giab008. https://doi.org/10.1093/gigascience/giab008

Flagel L, Brandvain Y, Schrider DR. The unreasonable effectiveness of convolutional neural networks in population genetic inference. *Mol Biol Evol* 2019;**36**:220–38. https://doi.org/10.1093/molbev/msy224

Goulart P, Samuk K. vcfsim: flexible simulation of all-sites vcfs with missing data. *BMC Bioinformatics* 2026. https://doi.org/10.1186/s12859-026-06453-9

Haller BC, Ralph PL, Messer PW. Slim 5: eco-evolutionary simulations across multiple chromosomes and full genomes. *Mol Biol Evol* 2026;**43**:msaf313. https://doi.org/10.1093/molbev/msaf313

Harney É, Patterson N, Reich D *et al.* Assessing the performance of qpadm: a statistical tool for studying population admixture. *Genetics* 2021;**217**:iyaa045. https://doi.org/10.1093/genetics/iyaa045

Hellmann I, Mang Y, Gu Z *et al.* Population genetic analysis of shotgun assemblies of genomic sequences from multiple individuals. *Genome Res* 2008;**18**:1020–9. https://doi.org/10.1101/gr.074187.107

Hofreiter M, Jaenicke V, Serre D *et al.* Dna sequences from multiple amplifications reveal artifacts induced by cytosine deamination in ancient dna. *Nucleic Acids Res* 2001;**29**:4793–9. https://doi.org/10.1093/nar/29.23.4793

Hudson RR. Generating samples under a wright–fisher neutral model of genetic variation. *Bioinformatics* 2002;**18**:337–8. https://doi.org/10.1093/bioinformatics/18.2.337

Johnson PL, Slatkin M. Accounting for bias from sequencing error in population genetic estimates. *Mol Biol Evol* 2008;**25**:199–206. https://doi.org/10.1093/molbev/msm239

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

Kern AD, Schrider DR. Discoal: flexible coalescent simulations with selection. *Bioinformatics* 2016;**32**:3839–41. https://doi.org/10.1093/bioinformatics/btw556

Kern AD, Schrider DR. diplos/hic: an updated approach to classifying selective sweeps. *G3 (Bethesda)* 2018;**8**:1959–70. https://doi.org/10.1534/g3.118.200262

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

Korunes KL, Samuk K. pixy: unbiased estimation of nucleotide diversity and divergence in the presence of missing data. *Mol Ecol Resour* 2021;**21**:1359–68. https://doi.org/10.1111/1755-0998.13326

Lachance J, Tishkoff SA. Snp ascertainment bias in population genetic analyses: why it is important, and how to correct it. *Bioessays* 2013;**35**:780–6. https://doi.org/10.1002/bies.201300014

Lauterbur ME, Cavassim MIA, Gladstein AL *et al.* Expanding the stdpopsim species catalog, and lessons learned for realistic genome simulations. *Elife* 2023;**12**:RP84874. https://doi.org/10.7554/eLife.84874

Mallick S, Micco A, Mah M *et al.* The allen ancient dna resource (aadr) a curated compendium of ancient human genomes. *Sci Data* 2024;**11**:182. https://doi.org/10.1038/s41597-024-03031-7

Mallick S, Reich D. *The Allen Ancient DNA Resource (AADR): a curated compendium of ancient human genomes*, 2023. https://doi.org/10.7910/DVN/FFIDCW (17 February 2026, date last accessed).

Mathieson I, Alpaslan-Roodenberg S, Posth C *et al.* The genomic history of southeastern Europe. *Nature* 2018;**555**:197–203. https://doi.org/10.1038/nature25778

Mora-Márquez F, Nuño JC, Soto Á *et al.* Missing genotype imputation in non-model species using self-organizing maps. *Mol Ecol Resour* 2025;**25**:e13992. https://doi.org/10.1111/1755-0998.13992

Pandey D, Harris M, Garud NR *et al.* Leveraging ancient dna to uncover signals of natural selection in Europe lost due to admixture or drift. *Nat Commun* 2024;**15**:9772. https://doi.org/10.1038/s41467-024-53852-8

Patterson N, Moorjani P, Luo Y *et al.* Ancient admixture in human history. *Genetics* 2012;**192**:1065–93. https://doi.org/10.1534/genetics.112.145037

Purcell S, Neale B, Todd-Brown K *et al.* Plink: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007;**81**:559–75. https://doi.org/10.1086/519795

Rahman A, Smith TQ, Szpiech ZA. Fast and memory-efficient dynamic programming approach for large-scale ehh-based selection scans. *Mol Biol Evol* 2025;**42**:msaf275. https://doi.org/10.1093/molbev/msaf275

Ramirez-Soriano A, Nielsen R. Correcting estimators of $\theta$ and tajima's d for ascertainment biases caused by the single-nucleotide polymorphism discovery process. *Genetics* 2009;**181**:701–10. https://doi.org/10.1534/genetics.108.094060

Rayner JG, Eichenberger F, Bainbridge JV *et al.* Competing adaptations maintain nonadaptive variation in a wild cricket population. *Proc Natl Acad Sci USA* 2024;**121**:e2317879121. https://doi.org/10.1073/pnas.2317879121

Renaud G, Hanghøj K, Willerslev E *et al.* Gargammel: a sequence simulator for ancient dna. *Bioinformatics* 2017;**33**:577–9. https://doi.org/10.1093/bioinformatics/btw670

Sayood K. *Introduction to Data Compression*. Cambridge, MA, USA: Morgan Kaufmann, 2017.

Schrider DR, Kern AD. Supervised machine learning for population genetics: a new paradigm. *Trends Genet* 2018;**34**:301–12. https://doi.org/10.1016/j.tig.2017.12.005

Senin P. *Dynamic Time Warping Algorithm Review*, Vol. **855**. Information and Computer Science Department University of Hawaii at Manoa Honolulu, USA, 2008, p. 40.

Skov L, Hui R, Shchur V *et al.* Detecting archaic introgression using an unadmixed outgroup. *PLoS Genet* 2018;**14**:e1007641. https://doi.org/10.1371/journal.pgen.1007641

Stoler N, Nekrutenko A. Sequencing error profiles of illumina sequencing instruments. *NAR Genom Bioinform* 2021;**3**:lqab019. https://doi.org/10.1093/nargab/lqab019

Williams MP, Huber CD. The genomic footprints of migration: how ancient dna reveals our history of mobility. *Genome Biol* 2025;**26**:357–29. https://doi.org/10.1186/s13059-025-03706-3


Bioinformatics Advances, 2026, 6, 1–5
https://doi.org/10.1093/bioadv/vbag125
Application Note