EGGS: Empirical Genotype Generalizer for Samples

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

[1]Department of Biology, The Pennsylvania State University, University Park, PA 16802
[*]Correspondence: tqs5778@psu.edu, zps5164@psu.edu

Abstract

Summary We introduce Empirical Genotype Generalizer for Samples (EGGS) which accepts empirical genotypes with missing data and replicates the underlying dispersal and distribution of missing genotypes in other replicates. In addition, EGGS can remove phase, remove polarization, simulate deamination, create pseudohaploids, and convert between VCF, *ms*-style replicates, and EIGENSTRAT.

Availability and Implementation EGGS is written in the C programming language. Precompiled executables, source code, and the manual 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 [1, 2, 3, 4]. Such simulations are essential for rigorously testing computational methods, and more recently, training machine learning models [5, 6]. In addition, simulations are often used to test hypotheses concerning the evolutionary histories of human and nonhuman species [6, 7].

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 [8]. This has motivated the development of methods to account for genotype uncertainty [9].

Missing genotypes 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 genotypes [10]. This approach assumes that the proportion of missing samples at each site viewed in aggregate reflects the underlying structure of missingness. However, a cumulative summary of missingness ignores the dispersal of missingness across the empirical segment, which is known to be uneven in reality, especially in regions of low complexity. In addition, the random modeling 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 [11]. Several recent studies introduce missingness blocks that correspond to regions of low mappability and variant callability [12, 13]. Other current methods that introduce missingness ignore the position of the missingness, are restricted to specific simulation frameworks, and are not generalizable to any supplied VCF [14].

We introduce Empirical Genotype Generalizer for Samples (EGGS) which extracts the underlying distribution and dispersal of missing sites across a genomic region and introduces similar patterns of missingness in replicates. EGGS works for any number of sites and number of samples. We demonstrate the use of EGGS and illustrate its ability to capture realistic patterns of missing genotypes.

2 Materials and Methods

2.1 Replicating Dispersal of Missingness

Suppose we are given N sites such that the proportion of samples with both genotypes missing at site $1 \leq i \leq N$ is x_i. We wish to replicate the dispersal of missingness summarized by x across M loci. Let y_j for $1 \leq j \leq M$ be the result. We consider two cases: $M \leq N$ and $M > N$. In the first case, we partition x into blocks of $\lceil \frac{N}{M} \rceil$ loci. The last block could contain fewer loci. We assign y_j, for $1 \leq j \leq M$, by randomly selecting a proportion from the jth block of x. In the second case, when $M > N$, we partition y into blocks of $\lceil \frac{M}{N} \rceil$ loci. Let $U(a, b)$ be the uniform distribution between $[a, b)$. For the first $1 \leq j \leq N-1$ blocks of y, we fill the block of y from $U(min(x_j, x_{j+1}), max(x_j, x_{j+1}))$. For the last block, we choose from $U(min(x_{N-1}, x_N), max(x_{N-1}, x_N))$. This implicitly assumes that the missingness pattern at the end of x can be extended.

2.2 Implementation

EGGS accepts variants in Variant Call Format (VCF) [15] or *ms*-style formatting [16]. In addition, we provide an option for the conversion from EIGENSTRAT format, the format used by The Allen Ancient DNA Resource (AADR), to VCF [17, 18]. In the case of multiple *ms*-style replicates as input, replicates are split to separate files VCF files. EGGS assumes all samples are diploid.

To remove phase, EGGS swaps the left and right genotype with equal probability. To remove polarization from biallelic sites, EGGS swaps the ancestral and derived alleles with equal probability. EGGS can introduce missing genotypes randomly according to the approach of Pandey et al. [10]. 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.

Deamination can be introduced according to Harney et al. [19]. The user supplies the probability that a site is a transition, and the probability that the transition will deaminate to the alternative allele if the sample originally holds the reference allele. Pseudohaploids are created by randomly choosing each sample to be homozygous for either the reference or alternative allele at biallelic sites with equal probability. Finally, summary statistics regarding missing genotypes can be calculated similar to those implemented in PLINK [20]. EGGS can produce *ms*-style output when missing data is not introduced. 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`.

3 Results

We used EGGS to convert the AARD [17] to VCF, and we extracted the autosomes of the 217 samples exclusively from Mathieson et al. [21] with bcftools [22]. This resulted in 1150639 VCF records. The sites had a mean missingness of 0.56 and a standard deviation of 0.21. We refer to this set as the ancient samples. Using *msprime* [2], we simulated a neutrally evolving 10 Mb region for 200 diploid samples with $Ne = 10,000$, $\mu = 1.29 \times 10^{-8}$, and $\rho = 1 \times 10^{-8}$. The simulated segment contained 33766 segregating sites. We introduced missing genotypes into the simulated segment randomly according to Pandey et al. [10] using a beta-distribution and with EGGS's method.

We calculated the proportion of sites with the fraction of samples with missing genotypes for the set of ancient samples (Figure 1a), for the simulated missingness using the beta-distribution (Figure 1b), and for the simulated missingness using EGGS's method (Figure 1c). Comparing Figure 1a and Figure 1b, we notice that Figure 1b does not possess a similar shape to that seen in Figure 1a. In addition, Figure 1b introduces sites with missingness proportions not observed in Figure 1a. Comparing Figure 1a to Figure 1c, we see that Figure 1c preserves the domain of missingness proportions and overall shape.

The Kullback–Leibler (KL) divergence measures the difference between two distributions [23]. We used the KL divergence to quantify the observed differences in Figure 1. The KL divergence between Figure 1a and Figure 1b is 13.81, and the KL divergence between Figure 1a and Figure 1c is 9.11. This supports our interpretation that the distribution of missingness created by EGGS is closer to the missingness in the ancient samples than the missingness created by the beta distribution.

4 Discussion

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

EGGS's method to replicate missingness 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 lose resolution. In addition, we expect that in situations where the rate of missing genotypes is high and region agnostic, EGGS's approach will not yield significantly different than modeling missingness with a beta-distribution. Our approach could be replaced with more sophisticated compression and interpolation methods [23]. These methods can be tuned to preserve specific patterns of missing genotypes that trend along a segment. Finally, we note that our approach could be adapted to replicating other characteristics, such as homozygosity and ascertainment.

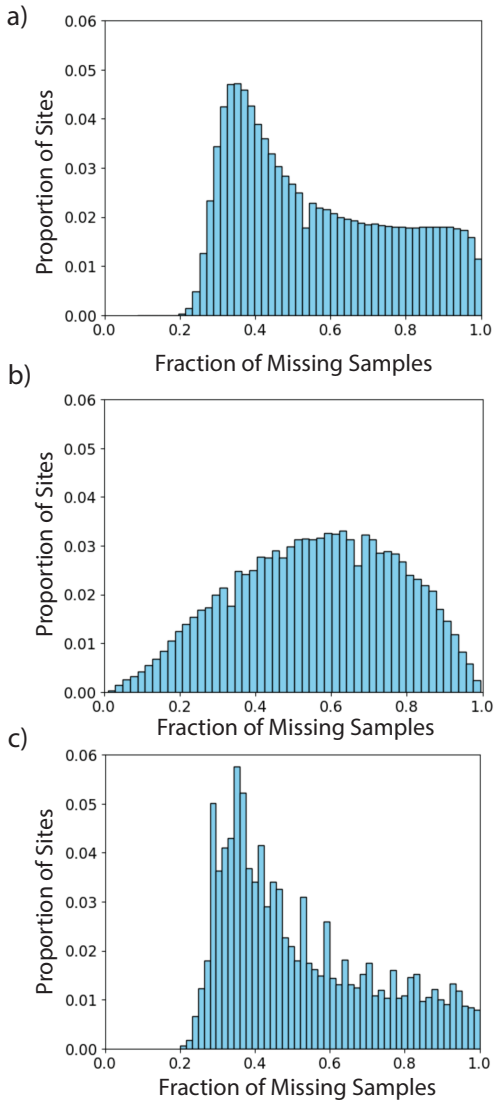


Figure 1: **Distribution of Missingness.** Each histogram was created using 50 bins. **Panel a)** The ancient samples. **Panel b)** Simulated missingness with the beta-distribution. **Panel c)** Simulated missingness with EGGS's method.

5 Acknowledgments

The authors would like to thank Abigail N. Sequeira and Ana V Leon-Apodaca for helpful suggestions. Computations for this research were performed using the Pennsylvania State University's Institute for Computational Data Sciences' Roar supercomputer. This work was supported by the National Institute of General Medical Sciences of the National Institutes of Health award number R35GM146926 (ZAS).

References

[1] Benjamin C Haller, Peter L Ralph, and Philipp W Messer. Slim 5: Eco-evolutionary simulations across multiple chromosomes and full genomes. *bioRxiv*, pages 2025–08, 2025.

[2] Jerome Kelleher, Alison M Etheridge, and Gilean McVean. Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS computational biology*, 12(5):e1004842, 2016.

[3] M Elise Lauterbur, Maria Izabel A Cavassim, Ariella L Gladstein, Graham Gower, Nathaniel S Pope, Georgia Tsambos, Jeffrey Adrion, Saurabh Belsare, Arjun Biddanda, Victoria Caudill, et al. Expanding the stdpopsim species catalog, and lessons learned for realistic genome simulations. *Elife*, 12:RP84874, 2023.

[4] Andrew D Kern and Daniel R Schrider. Discoal: flexible coalescent simulations with selection. *Bioinformatics*, 32(24):3839–3841, 2016.

[5] Amatur Rahman, Thomas Quinn Smith, and Zachary A Szpiech. Fast and memory-efficient dynamic programming approach for large-scale ehh-based selection scans. *bioRxiv*, pages 2025–04, 2025.

[6] Md Ruhul Amin, Mahmudul Hasan, and Michael DeGiorgio. Digital image processing to detect adaptive evolution. *Molecular Biology and Evolution*, 41(12):msae242, 2024.

[7] Jack G Rayner, Franca Eichenberger, Jessica VA Bainbridge, Shangzhe Zhang, Xiao Zhang, Leeban H Yusuf, Susan Balenger, Oscar E Gaggiotti, and Nathan W Bailey. Competing adaptations maintain nonadaptive variation in a wild cricket population. *Proceedings of the National Academy of Sciences*, 121(32):e2317879121, 2024.

[8] Matthew P Williams and Christian D Huber. The genomic footprints of migration: how ancient dna reveals our history of mobility. *Genome Biology*, 26(1):1–29, 2025.

[9] Nick Bailey, Laurie Stevison, and Kieran Samuk. Correcting for bias in estimates of θ w θ _w and tajima's d d from missing data in next-generation sequencing. *Molecular Ecology Resources*, page e14104, 2025.

[10] Devansh Pandey, Mariana Harris, Nandita R Garud, and Vagheesh M Narasimhan. Leveraging ancient dna to uncover signals of natural selection in europe lost due to admixture or drift. *Nature Communications*, 15(1):9772, 2024.

[11] Fernando Mora-Márquez, Juan Carlos Nuño, Álvaro Soto, and Unai López de Heredia. Missing genotype imputation in non-model species using self-organizing maps. *Molecular Ecology Resources*, 25(3):e13992, 2025.

[12] Sandipan Paul Arnab, Andre Luiz Campelo dos Santos, Matteo Fumagalli, and Michael DeGiorgio. Efficient detection and characterization of targets of natural selection using transfer learning. *Molecular Biology and Evolution*, 42(5):msaf094, 2025.

[13] Laurits Skov, Ruoyun Hui, Vladimir Shchur, Asger Hobolth, Aylwyn Scally, Mikkel Heide Schierup, and Richard Durbin. Detecting archaic introgression using an unadmixed outgroup. *PLoS genetics*, 14(9):e1007641, 2018.

[14] Paimon Goulart and Kieran Samuk. vcfsim: flexible simulation of all-sites vcfs with missing data. *bioRxiv*, pages 2025–01, 2025.

[15] Petr Danecek, Adam Auton, Goncalo Abecasis, Cornelis A Albers, Eric Banks, Mark A DePristo, Robert E Handsaker, Gerton Lunter, Gabor T Marth, Stephen T Sherry, et al. The variant call format and vcftools. *Bioinformatics*, 27(15):2156–2158, 2011.

[16] Richard R Hudson. Generating samples under a wright–fisher neutral model of genetic variation. *Bioinformatics*, 18(2):337–338, 2002.

[17] Swapan Mallick and David Reich. The Allen Ancient DNA Resource (AADR): A curated compendium of ancient human genomes, 2023. URL `https://doi.org/10.7910/DVN/FFIDCW`.

[18] Swapan Mallick, Adam Micco, Matthew Mah, Harald Ringbauer, Iosif Lazaridis, Iñigo Olalde, Nick Patterson, and David Reich. The allen ancient dna resource (aadr) a curated compendium of ancient human genomes. *Scientific Data*, 11(1):182, 2024.

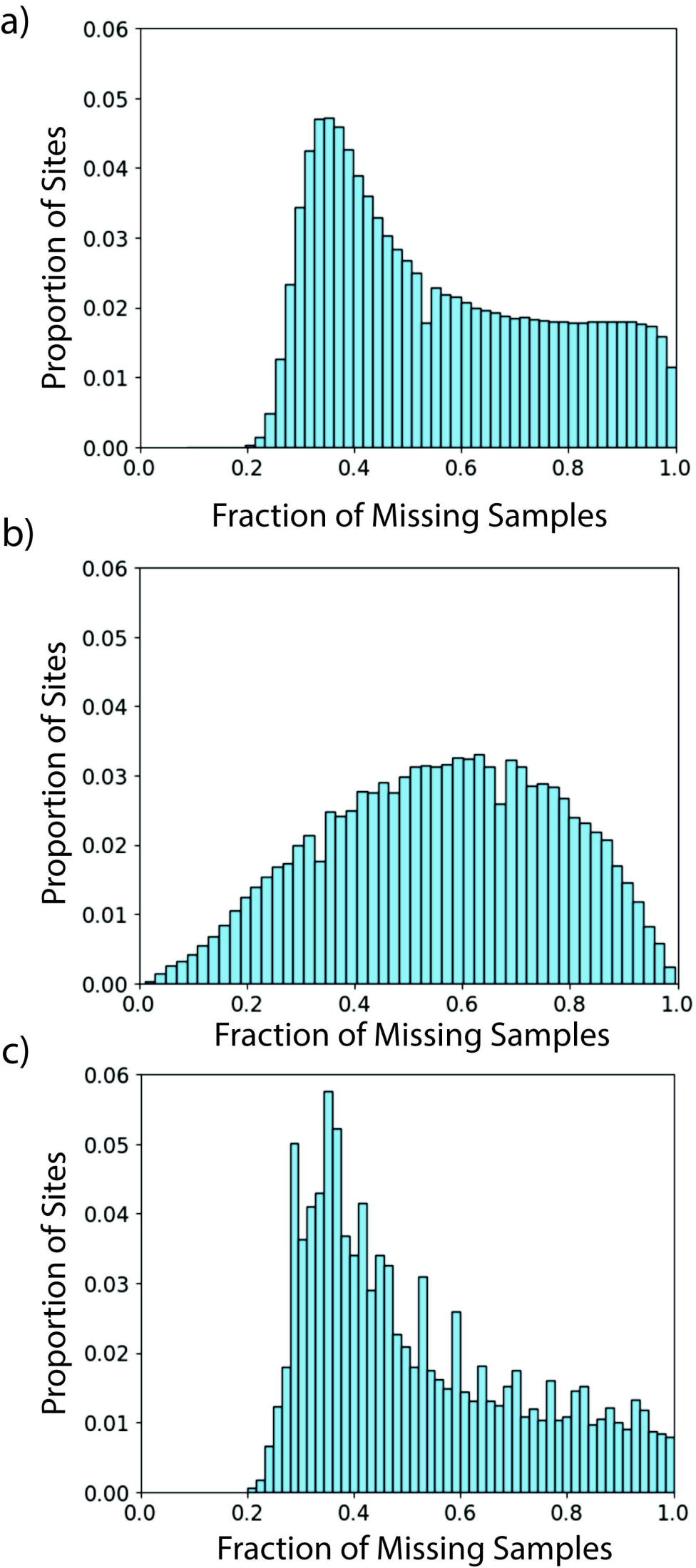
[19] Éadaoin Harney, Nick Patterson, David Reich, and John Wakeley. Assessing the performance of qpadm: a statistical tool for studying population admixture. *Genetics*, 217(4):iyaa045, 2021.

[20] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel AR Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul IW De Bakker, Mark J Daly, et al. Plink: a tool set for whole-genome association and population-based linkage analyses. *The American journal of human genetics*, 81(3):559–575, 2007.

[21] Iain Mathieson, Songül Alpaslan-Roodenberg, Cosimo Posth, Anna Szécsényi-Nagy, Nadin Rohland, Swapan Mallick, Iñigo Olalde, Nasreen Broomandkhoshbacht, Francesca Candilio, Olivia Cheronet, et al. The genomic history of southeastern europe. *Nature*, 555(7695):197–203, 2018.

[22] Petr Danecek, James K Bonfield, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin O Pollard, Andrew Whitwham, Thomas Keane, Shane A McCarthy, Robert M Davies, et al. Twelve years of samtools and bcftools. *Gigascience*, 10(2):giab008, 2021.

[23] Khalid Sayood. *Introduction to data compression*. Morgan Kaufmann, 2017.


a)
Proportion of Sites
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.0
0.2
0.4
0.6
0.8
1.0
Fraction of Missing Samples
b)
Proportion of Sites
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.0
0.2
0.4
0.6
0.8
1.0
Fraction of Missing Samples
c)
Proportion of Sites
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.0
0.2
0.4
0.6
0.8
1.0
Fraction of Missing Samples