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 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. When producing VCF files, EGGS assumes all samples are diploid.

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, 7, 8]. In addition, simulations are often used to test hypotheses concerning the evolutionary histories of human and nonhuman species [6, 9].

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) [10, 11]. This has motivated the development of methods to account for genotype uncertainty [12, 13, 14, 15, 16, 17].

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 [18]. 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, 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 [19]. Several recent studies introduce missingness blocks that correspond to regions of low mappability and variant callability [20, 21]. Other current methods that introduce missing genotypes require a FASTA file [22], ignore the position of the missing sites within the simulated segment, are restricted to specific simulation frameworks, or are not generalizable to any supplied Variant Call Format (VCF) file [23].

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 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 Materials and 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 $s \sim Uniform(1, ..., S)$. 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 gneotypes in VCF files [22]. 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 [22].

2.2 Implementation

EGGS accepts variants in VCF [24] or *ms*-style formatting [25]. 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 [26, 27]. 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 samples. This allows each lineage in an *ms*-style replicate or in an VCF file to be

treated as its own sample in the resulting VCF. When creating VCF files, EGGS produces diploid samples. This avoids confusion and possible parsing errors in downstream analyses. In the case when a genotype only contains one allele, the other allele is marked as absent ('.').

To remove phase, EGGS swaps the left and right allele with equal probability, and in the case of VCF output, marks each genotype with '/' as opposed to '|'. Computational methods that rely on phase, typically rely on the position of the left and right allele at a genotype. Therefore, changing '|' to '/' is not sufficient to erase phase information. To remove polarization from biallelic sites, EGGS swaps the ancestral and derived alleles with equal probability for all samples' alleles. 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 [28]. EGGS can introduce missing genotypes randomly according to the approach of Pandey et al. [18]. 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 β-distribution. The β-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 β-distribution or a VCF from which the values are calculated.

Another complication commonly arising in working with aDNA is demaination. Deamination is a type of damage, when usually, a cytosine is incorrectly read as a thymine during sequencing [11]. Deamination is simulated according to Harney et al. [29]. 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 [18]. 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 degredated compared to aDNA and contain much lower error rates

on next-generation sequencing machines [30]. 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 [31]. EGGS can produce *ms*-style output when missing data is 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`.

3 Results

We used EGGS to convert the AARD [26] to VCF, and we extracted chromosome 1 of the 217 samples exclusively from Mathieson et al. [32] with bcftools [33]. This resulted in 93166 sites. The sites had a mean missingness of 0.55 and a standard deviation of 0.21. We refer to this set as the ancient samples.

Using *msprime* [2], we simulated 200 diploid samples on segments ranging from 1Mb to 10Mb in 1Mb 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 β-distribution method. At each site, we average the proportion of samples missing for each method across the 10 runs.

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 [34]. We use DTW to evaluate our method. We compute the proportion of missing samples at each of the 93166 sites in our ancient samples. We treat this as a signal and use DTW to compare it to the average of the signals produced by our simulated replicates. DTW computes a similarity measure, and 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 β-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 β-distribution may be sufficient when missing data is 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.

Table 1: **Results of DTW** 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 DWT between the missingness in the empirical segment and the missingness introduced in the simulated replicate using the β-distribution. The fifth column is the distance as calculated by the DWT between the missingness in the empirical segment and the missingness introduced in the simulated replicate using EGGS's method. For the β-distribution and EGGS's method, the DWT was calculated with the average proportion of missing samples at each site in the simulated segment over 10 runs.

Length of Segment (Mb)	Number of Segregating Sites	Block Size	β-Distribution	EGGS
1	3327	28	57.49	57.38
2	6612	14	55.79	55.02
3	9657	9	54.57	53.09
4	12659	7	53.46	51.39
5	16468	5	52.51	49.2
6	19930	4	51.53	47.33
7	23012	4	51.08	45.67
8	25682	3	50.45	44.29
9	29391	3	49.84	42.39
10	33023	2	49.27	40.92

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 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 lose resolution. In addition, we expect that in situations where the rate of missing genotypes is high and the segment is small (less than 1Mb), EGGS's approach will not yield significantly different results than modeling missingness with a beta-distribution. We also assume that adjacent missing sites for samples are not correlated. While this assumption is justified for aDNA [18], 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 [35]. Finally, we note that our approach could be adapted to replicating other characteristics of genotype uncertainty, such as homozygosity and ascertainment [36, 37, 38].

5 Acknowledgments

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. 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] Lex Flagel, Yaniv Brandvain, and Daniel R Schrider. The unreasonable effectiveness of convolutional neural networks in population genetic inference. *Molecular biology and evolution*, 36 (2):220–238, 2019.

[8] Daniel R Schrider and Andrew D Kern. Supervised machine learning for population genetics: a new paradigm. *Trends in Genetics*, 34(4):301–312, 2018.

[9] 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.

[10] 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.

[11] Michael Hofreiter, Viviane Jaenicke, David Serre, Arndt von Haeseler, and Svante Pääbo. Dna sequences from multiple amplifications reveal artifacts induced by cytosine deamination in ancient dna. *Nucleic acids research*, 29(23):4793–4799, 2001.

[12] Katharine L Korunes and Kieran Samuk. pixy: Unbiased estimation of nucleotide diversity and divergence in the presence of missing data. *Molecular ecology resources*, 21(4):1359–1368, 2021.

[13] 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.

[14] Thorfinn Sand Korneliussen, Anders Albrechtsen, and Rasmus Nielsen. Angsd: analysis of next generation sequencing data. *BMC bioinformatics*, 15(1):356, 2014.

[15] Guillaume Achaz. Testing for neutrality in samples with sequencing errors. *Genetics*, 179(3): 1409–1424, 2008.

[16] Philip LF Johnson and Montgomery Slatkin. Accounting for bias from sequencing error in population genetic estimates. *Molecular biology and evolution*, 25(1):199–206, 2008.

[17] Ines Hellmann, Yuan Mang, Zhiping Gu, Peter Li, Francisco M De La Vega, Andrew G Clark, and Rasmus Nielsen. Population genetic analysis of shotgun assemblies of genomic sequences from multiple individuals. *Genome research*, 18(7):1020–1029, 2008.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] Andrew D Kern and Daniel R Schrider. diplos/hic: an updated approach to classifying selective sweeps. *G3: Genes, Genomes, Genetics*, 8(6):1959–1970, 2018.

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

[24] 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.

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

[26] 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`.

[27] 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.

[28] Nick Patterson, Priya Moorjani, Yontao Luo, Swapan Mallick, Nadin Rohland, Yiping Zhan, Teri Genschoreck, Teresa Webster, and David Reich. Ancient admixture in human history. *Genetics*, 192(3):1065–1093, 2012.

[29] 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.

[30] Nicholas Stoler and Anton Nekrutenko. Sequencing error profiles of illumina sequencing instruments. *NAR genomics and bioinformatics*, 3(1):lqab019, 2021.

[31] 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.

[32] 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.

[33] 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.

[34] Pavel Senin. Dynamic time warping algorithm review. *Information and Computer Science Department University of Hawaii at Manoa Honolulu, USA*, 855(1-23):40, 2008.

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

[36] Andrew G Clark, Melissa J Hubisz, Carlos D Bustamante, Scott H Williamson, and Rasmus Nielsen. Ascertainment bias in studies of human genome-wide polymorphism. *Genome research*, 15(11):1496–1502, 2005.

[37] Joseph Lachance and Sarah A Tishkoff. Snp ascertainment bias in population genetic analyses: why it is important, and how to correct it. *Bioessays*, 35(9):780–786, 2013.

[38] Anna Ramirez-Soriano and Rasmus Nielsen. Correcting estimators of θ and tajima's d for ascertainment biases caused by the single-nucleotide polymorphism discovery process. *Genetics*, 181(2):701–710, 2009.