

The missing data problem in population genomics and statistical methods to address them

Arun Sethuraman  *

Department of Biology, San Diego State University, San Diego, CA 92182, United States

*Corresponding author: Email: asethuraman@sdsu.edu

Keywords: genomics; population genetics; missing data

The “Missing Data” problem is prevalent across all statistical inference, owing to the “absence of some part of a familiar data structure” (Efron 1994). Population genomic datasets are riddled with missing data (Fig. 1)—broadly classified as data missing at random (e.g. due to degradation, sequencing errors), data missing “on purpose” (e.g. due to sequencing strategies like genotyping by sequencing), and data missing due to unknown evolutionary history (e.g. introgression from ancestral ghost populations). Editors and scientific contributors to both the GSA’s journals, *Genetics* and *G3* have continually highlighted statistical issues and pitfalls with inference in the presence of missing data (McIntyre 2025), particularly in an age of Biobank scale population genomic datasets. Here I highlight studies, including those that have been recently published in *Genetics* and *G3* towards systematically assessing the effects of missing data problems and addressing them towards inference in a variety of population genomics questions.

The missing data problems

Genotyping and sequencing errors might lead to incorrectly called nucleotide bases (Nielsen and Signorovitch 2003). These erroneous base calls are often either discarded or imputed with respect to a reference genome but could lead to ascertainment biases in calling variant sites and potentially incorrect inferences. Ascertainment biases during variant discovery, mapping biases during read alignment, and imputation biases during missing data inference are also exacerbated in ancient genomic datasets, which suffer from systematic degradation from DNA damage, and cytosine deamination (described by Günther et al. 2025).

A second type of missing population genomic data arises due to molecular methodological issues with genotyping or sequencing. This is particularly exacerbated in methods that utilize reduced representation genomic sequencing, often called Genotyping By Sequencing (GBS) methods such those using SNP-chips, which suffer from ascertainment bias. For instance, Gómez-Palacio et al. 2024 in a recent issue of *G3* discuss the “robustness” of demographic inference from SNP chip vs whole genome sequencing (WGS) data in *Aedes aegypti*, highlighting the utility of SNP-chips as a lower cost alternative to WGS in non-model systems. Other lower coverage approaches such as Ultra-Conserved Element-based sequencing (Portik and Wiens 2021), PoolSeq (Kofler et al. 2011; Carvalho

et al. 2023), and Restriction Site Associated DNA sequencing (RADseq—Baird et al. 2008) remain powerful low-cost strategies for phylogenetic inference as described in several recent studies in *G3* (e.g. Forest et al. 2024 in woodpeckers, Parys et al. 2025 in bees), variant calling (Chen et al. 2025), identifying meiotic drive loci (Barbash et al. 2023), and in inference of evolutionary history, despite missing data issues. These issues arise because of allele dropout due to un/under-sequenced genomic regions, along with polymorphisms present across restriction sites, being absent in subsequent sequencing output (Andrews et al. 2016). Allele dropout introduces systematic biases and limitations (Fu 2014) while using RADseq data to perform inference of evolutionary history (Nunziata and Weisrock 2018).

Thirdly, genomic data could also be missing due to the presence of unsampled genomic variation from populations that are either extinct, severely depleted, or impossible to sample. Termed “ghost” populations, ancestral or continuing gene flow from these “ghosts” leaves signatures on extant and/or sampled genomes that often go undetected. “Ghost” variation in sampled genomes also introduces biases in estimates of evolutionary history, if unaccounted for (Beerli 2004; Slatkin 2005; Tricou et al. 2022; Sethuraman et al. 2025).

Solving missing data problems

There are three common approaches to handling missing data in genomics that are widely implemented in popularly used tools in the field of population genomics, several of which are described in *Genetics* and *G3*—(1) missing data is considered as data missing at random (e.g. popular tools like STRUCTURE—Pritchard et al. 2000, ADMIXTURE—Alexander et al. 2009), and therefore not accounting for in the inference model, (2) imputing missing data—i.e. filling in missing data with respect to a reference genome or panel (e.g. see a detailed treatment of the issue of “switch-errors” during local ancestry inference in Avadhanam and Williams 2025, or while imputing ancient DNA data in Ausmees et al. 2022), and (3) mean imputation—i.e. filling in missing data with a mean population statistic (e.g. DAPC/ADEGENET—Jombart 2008). Nonetheless, all three approaches offer little improvement on the performance of model-based inference methods (Le Morvan and Varoquax 2024) and can bias non-model-based inferences (e.g. PCA—see Yi and Latch 2022).

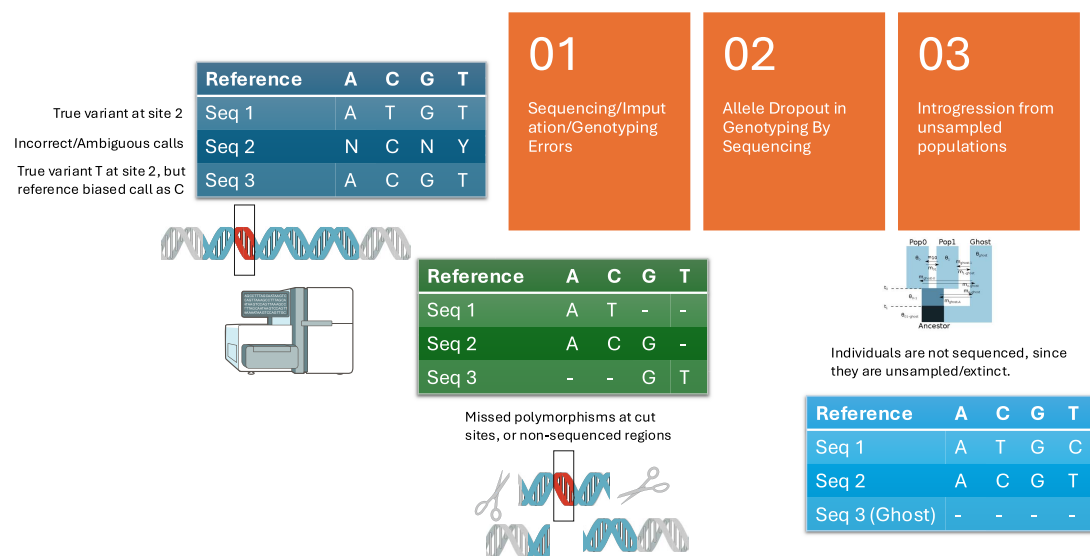


Fig. 1. Infographic showing three types of missing data problems in population genomics, concerning (1) sequencing/imputation/genotyping errors, (2) allele dropout in genotyping by sequencing, and (3) missing data due to introgression from unsampled populations.

A few researchers have developed corrected method-of-moments estimators of population genetic summary statistics, such as allele frequencies and ancestry proportions (Günther et al. 2025), effective population size (N_e —Peel et al. 2013), Watterson's θ , Tajima's D (Bailey et al. 2025), Fay and Wu's H (Ferretti et al. 2012), nucleotide diversity (π —Korunes and Samuk 2021) that account for genotyping errors. Others have developed methods that include missing genotypic data in the inference model, specifically while inferring haplotypic phase (e.g. Montero-Tena et al. 2024), effective population size (e.g. Hong et al. 2024) in genome-wide association studies (e.g. Jørsboe and Albrechtsen 2022), and ancestry estimation (e.g. Vitale et al. 2025).

Recently, we (Sethuraman et al. 2025) present an account of systematic biases in the estimation of evolutionary history while not accounting for the presence of gene flow from unsampled ghost populations; we developed theoretical expectations for estimates of commonly utilized population genomic summary statistics such as nucleotide diversity, genetic differentiation, and Tajima's D , while applying extensive simulations to discuss how increased degree of ancestral gene flow from unsampled ghosts will bias inference from sampled extant populations. We are continuing to develop methods to address these biases in estimation of evolutionary history while making evolutionary inferences.

Going forward, I see two important venues for the population genomics community to undertake.

- 1) Developing simulators of missingness: Simulators of population-scale genomic data come in two flavors—backward-time, or coalescent simulators, such as *msprime* (Baumdicker et al. 2022) and *fsc28* (Excoffier et al. 2021), and forward-time simulators, such as SLiM (Haller and Messer 2019). Both flavors have been used extensively in fundamental studies of evolutionary population genomics, especially to create “null” and “alternate” hypotheses for testing (a) validity of new evolutionary inference methods, and (b) establishing expectations under neutral/selective evolutionary processes. However, in reality, data generated using genomic sequencing platforms are replete with missing data of varying types, that are not addressed in simulators. Therefore, to address this important gap in simulations, we need to develop missingness simulators in population genomics that can account for sequencing/genotyping errors and in

restriction enzyme associated sequencing methods (RADseq). For instance, the RADinitio method of Rivera-Colón et al. 2021 describes a new simulator for systematic processing of forward-time simulations via SLiM to include missingness. Thereon, we should be characterizing the effects of missingness and imputation in estimation of evolutionary history via systematic assessments of simulated and empirical datasets to understand the effects of missingness in large population genomic datasets under a variety of evolutionary demographic models. Another method, developed by Goulart and Samuk (2025) utilizes the *msprime* pipeline to accurately introduce missing data across various ploidy levels.

- 2) Develop statistical methods to include missing data in the inference: Current methods for estimating evolutionary history from population-scale genomic data are yet simplistic, and for the most part, fail to account for the presence of missing data. As elegantly discussed by Nielsen and Signorovitch 2003, we require new methods that will incorporate different kinds of missingness in the generative models—I hope that this editorial will be an invitation for researchers to consider the effects of missingness while developing such methods.

Funding

This work was funded by NSF CAREER 2147182 and NIH 1R15GM143700-01 to Arun Sethuraman.

Conflicts of interest

None declared.

Literature cited

- Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* 19: 1655–1664. <https://doi.org/10.1101/gr.094052.109>.
- Andrews KR et al. 2016. Harnessing the power of RADseq for ecological and evolutionary genomics. *Nat Rev Genet.* 17:81–92. <https://doi.org/10.1038/nrg.2015.28>.

- Ausmees K et al. 2022. An empirical evaluation of genotype imputation of ancient DNA. *G3* (Bethesda). 12:jkac089. <https://doi.org/10.1093/g3journal/jkac089>.
- Avadhanam S, Williams AL. 2025. Phase-free local ancestry inference mitigates the impact of switch errors on phase-based methods. *G3* (Bethesda). 15:jkaf122. <https://doi.org/10.1093/g3journal/jkaf122>.
- Bailey N, Stevison L, Samuk K. 2025. Correcting for bias in estimates of θ_w and Tajima's D from missing data in next-generation sequencing. *Mol Ecol Resour*. 25:e14104. <https://doi.org/10.1111/1755-0998.14104>.
- Baird NA et al. 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One*. 3:e3376. <https://doi.org/10.1371/journal.pone.0003376>.
- Barbash DA et al. 2023. Testing a candidate meiotic drive locus identified by pool sequencing. *G3* (Bethesda). 13:jkad225. <https://doi.org/10.1093/g3journal/jkad225>.
- Baumdicker F et al. 2022. Efficient ancestry and mutation simulation with msprime 1.0. *Genetics*. 220:iyab229. <https://doi.org/10.1093/genetics/iyab229>.
- Beerli P. 2004. Effect of unsampled populations on the estimation of population sizes and migration rates between sampled populations. *Mol Ecol*. 13:827–836. <https://doi.org/10.1111/j.1365-294X.2004.02101.x>.
- Carvalho J, Morales HE, Faria R, Butlin RK, Sousa VC. 2023. Integrating pool-seq uncertainties into demographic inference. *Mol Ecol Resour*. 23:1737–1755. <https://doi.org/10.1111/1755-0998.13834>.
- Chen D et al. 2025. A cost-effective, high-throughput, highly accurate genotyping method for outbred populations. *G3* (Bethesda). 15:jkaf291. <https://doi.org/10.1093/g3journal/jkaf291>.
- Efron B. 1994. Missing data, imputation, and the bootstrap. *J Am Stat Assoc*. 89:463–475. <https://doi.org/10.1080/01621459.1994.10476768>.
- Excoffier L et al. 2021. Fastsimcoal2: demographic inference under complex evolutionary scenarios. *Bioinformatics*. 37:4882–4885. <https://doi.org/10.1093/bioinformatics/btab468>.
- Ferretti L, Raineri E, Ramos-Onsins S. 2012. Neutrality tests for sequences with missing data. *Genetics*. 191:1397–1401. <https://doi.org/10.1534/genetics.112.139949>.
- Forest T et al. 2024. Chromosome-level genome assembly of the European green woodpecker *Picus viridis*. *G3* (Bethesda). 14:jkaf042. <https://doi.org/10.1093/g3journal/jkaf042>.
- Fu YB. 2014. Genetic diversity analysis of highly incomplete SNP genotype data with imputations: an empirical assessment. *G3* (Bethesda). 4:891–900. <https://doi.org/10.1534/g3.114.010942>.
- Gómez-Palacio A et al. 2024. Robustness in population-structure and demographic-inference results derived from the *Aedes aegypti* genotyping chip and whole-genome sequencing data. *G3* (Bethesda). 14:jkaf082. <https://doi.org/10.1093/g3journal/jkaf082>.
- Goulart P, Samuk K. 2025. vcfSim: flexible simulation of all-sites VCFs with missing data [preprint]. *bioRxiv* 2025-01. <https://doi.org/10.1101/2025.01.29.635540>.
- Günther T, Goldberg A, Schraiber JG. 2025. Estimating allele frequencies, ancestry proportions and genotype likelihoods in the presence of mapping bias. *G3* (Bethesda). 15:jkaf172. <https://doi.org/10.1093/g3journal/jkaf172>.
- Haller BC, Messer PW. 2019. SLiM 3: forward genetic simulations beyond the Wright–Fisher model. *Mol Biol Evol*. 36:632–637. <https://doi.org/10.1093/molbev/msy228>.
- Hong A et al. 2024. ONEsAMP 3.0: estimation of effective population size via single nucleotide polymorphism data from one population. *G3* (Bethesda). 14:jkaf153. <https://doi.org/10.1093/g3journal/jkaf153>.
- Jombart T. 2008. ADEGENET: a R package for the multivariate analysis of genetic markers. *Bioinformatics*. 24:1403–1405. <https://doi.org/10.1093/bioinformatics/btn129>.
- Jørnsboe E, Albrechtsen A. 2022. Efficient approaches for large-scale GWAS with genotype uncertainty. *G3* (Bethesda). 12:jkab385. <https://doi.org/10.1093/g3journal/jkab385>.
- Kofler R et al. 2011. Popoolation: a toolbox for population genetic analysis of next generation sequencing data from pooled individuals. *PLoS One*. 6:e15925. <https://doi.org/10.1371/journal.pone.0015925>.
- Korunes KL, Samuk K. 2021. Pixy: unbiased estimation of nucleotide diversity and divergence in the presence of missing data. *Mol Ecol Resour*. 21:1359–1368. <https://doi.org/10.1111/1755-0998.13326>.
- Le Morvan M, Varoquaux G. 2024. Imputation for prediction: beware of diminishing returns [preprint]. *arXiv preprint arXiv:2407.19804*. <https://doi.org/10.48550/arXiv.2407.19804>.
- McIntyre LM. 2025. Biobanks in GENETICS and G3: tackling the statistical challenges. *Genetics*. 229:iyaf046. <https://doi.org/10.1093/genetics/iyaf046>.
- Montero-Tena JA et al. 2024. haploMAGIC: accurate phasing and detection of recombination in multiparental populations despite genotyping errors. *G3* (Bethesda). 14:jkaf109. <https://doi.org/10.1093/g3journal/jkaf109>.
- Nielsen R, Signorovitch J. 2003. Correcting for ascertainment biases when analyzing SNP data: applications to the estimation of linkage disequilibrium. *Theor Popul Biol*. 63:245–255. [https://doi.org/10.1016/S0040-5809\(03\)00005-4](https://doi.org/10.1016/S0040-5809(03)00005-4).
- Nunziata SO, Weisrock DW. 2018. Estimation of contemporary effective population size and population declines using RAD sequence data. *Heredity* (Edinb). 120:196–207. <https://doi.org/10.1038/s41437-017-0037-y>.
- Parys KA, et al.. 2025. Chromosome-level genome assembly of *Protandrena* (Anthemurgus) *passiflorae* (Hymenoptera: Andrenidae), a host-plant specialist bee. *G3* (Bethesda). 15:jkaf096. <https://doi.org/10.1093/g3journal/jkaf096>.
- Peel D et al. 2013. Accounting for missing data in the estimation of contemporary genetic effective population size (N_e). *Mol Ecol Resour*. 13:243–253. <https://doi.org/10.1111/1755-0998.12049>.
- Portik DM, Wiens JJ. 2021. Do alignment and trimming methods matter for phylogenomic (UCE) analyses? *Syst Biol*. 70:440–462. <https://doi.org/10.1093/sysbio/syaa064>.
- Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. *Genetics*. 155:945–959. <https://doi.org/10.1093/genetics/155.2.945>.
- Rivera-Colón AG, Rochette NC, Catchen JM. 2021. Simulation with RADinitio improves RADseq experimental design and sheds light on sources of missing data. *Mol Ecol Resour*. 21:363–378. <https://doi.org/10.1111/1755-0998.13163>.
- Sethuraman A, Lynch M, Wanjiku M, Kuzminskiy M. 2025. Accounting for gene flow from unsampled ghost populations while estimating evolutionary history. *G3* (Bethesda). 15:jkaf180. <https://doi.org/10.1093/g3journal/jkaf180>.
- Slatkin M. 2005. Seeing ghosts: the effect of unsampled populations on migration rates estimated for sampled populations. *Mol Ecol*. 14:67–73. <https://doi.org/10.1111/j.1365-294X.2004.02393.x>.
- Tricou T, Tannier E, de Vienne DM. 2022. Ghost lineages highly influence the interpretation of introgression tests. *Syst Biol*. 71:1147–1158. <https://doi.org/10.1093/sysbio/syaa011>.
- Vitale D et al. 2025. GenoTools: an open-source Python package for efficient genotype data quality control and analysis. *G3* (Bethesda). 15:jkaf268. <https://doi.org/10.1093/g3journal/jkaf268>.
- Yi X, Latch EK. 2022. Nonrandom missing data can bias principal component analysis inference of population genetic structure. *Mol Ecol Resour*. 22:602–611. <https://doi.org/10.1111/1755-0998.13498>.