Article:
Carvalho, J. orcid.org/0000-0002-1728-0075, Faria, R. orcid.org/0000-0001-6635-685X, Butlin, R.K. orcid.org/0000-0003-4736-0954 et al. (1 more author) (2023) poolHelper: an R package to help in designing Pool-seq studies. Methods in Ecology and Evolution, 14 (9). pp. 2300-2307. ISSN 2041-210X

https://doi.org/10.1111/2041-210x.14185

Received: 31 January 2023 | Accepted: 22 June 2023

DOI: 10.1111/2041-210X.14185

Methods in Ecology and Evolution

APPLICATION

poolHelper: An R package to help in designing Pool-Seq studies

João Carvalho[1] | Rui Faria[2,3] | Roger K. Butlin[4,5] | Vitor C. Sousa[1]

[1]cE3c-Centre for Ecology, Evolution and Environmental Changes & CHANGE-Global Change and Sustainability Institute, Departamento de Biologia Animal, Faculdade de Ciências, Universidade de Lisboa, Lisboa, Portugal

[2]CIBIO-Centro de Investigação em Biodiversidade e Recursos Genéticos, InBIO, Laboratório Associado, Universidade do Porto, Vairão, Portugal

[3]BIOPOLIS Program in Genomics, Biodiversity and Land Planning, CIBIO, Campus de Vairão, Vairão, Portugal

[4]Ecology and Evolutionary Biology, School of Biosciences, University of Sheffield, Sheffield, UK

[5]Department of Marine Sciences, University of Gothenburg, Gothenburg, Sweden

Correspondence
João Carvalho
Email: jgcarvalho@fc.ul.pt

Funding information
Fundação para a Ciência e a Tecnologia, Grant/Award Number: 2020.00275.CEECIND, CEECINST/00032/2018/CP1523/CT0008, PD/BD/128350/2017, PTDC/BIA-EVL/1614/2021, UIDB/00329/2020 and 2021.09795.CPCA; H2020 European Research Council, Grant/Award Number: ERC-2015-AdG-693030-BARRIERS; Human Frontier Science Program, Grant/Award Number: RGY0081/2020

Handling Editor: Oscar Gaggiotti

Abstract

1. Next-generation sequencing of pooled samples (Pool-seq) is an important tool in population genomics and molecular ecology. In Pool-seq, the relative number of reads with an allele reflects the allele frequencies in the sample. However, unequal individual contributions to the pool and sequencing errors can lead to inaccurate allele frequency estimates, influencing downstream analysis. When designing Pool-seq studies, researchers need to decide the pool size (number of individuals) and average depth of coverage (sequencing effort). An efficient sampling design should maximise the accuracy of allele frequency estimates while minimising the sequencing effort. We describe a novel tool to simulate single nucleotide polymorphism (SNP) data using coalescent theory and account for sources of uncertainty in Pool-seq.
2. We introduce an R package, poolHelper, enabling users to simulate Pool-seq data under different combinations of average depth of coverage and pool size, accounting for unequal individual contributions and sequencing errors, modelled by adjustable parameters. The mean absolute error is computed by comparing the sample allele frequencies obtained based on individual genotypes with the frequency estimates obtained with Pool-seq.
3. poolHelper enables users to simulate multiple combinations of pooling errors, average depth of coverage, pool sizes and number of pools to assess how they influence the error of sample allele frequencies and expected heterozygosity. Using simulations under a single population model, we illustrate that increasing the depth of coverage does not necessarily lead to more accurate estimates, reinforcing that finding the best Pool-seq study design is not straightforward. Moreover, we show that simulations can be used to identify different combinations of parameters with similarly low mean absolute errors. This can help users to define an effective sampling design by using those combinations of parameters that minimise the sequencing effort.
4. The poolHelper package provides tools for performing simulations with different combinations of parameters (e.g. pool size, depth of coverage, unequal individual contribution) before sampling and generating data, allowing users to define sampling schemes based on simulations. This allows researchers to focus on the best

sampling scheme to answer their research questions. poolHelper is comprehensively documented with examples to guide effective use.

KEYWORDS
experimental design, open source, Pool-seq, R package, simulations

1 | INTRODUCTION

Next generation sequencing (NGS) is an important tool for many biologists, providing access to polymorphism data across a wide range of model and non-model species (Ellegren, 2014). Although the cost of sequencing is continuously decreasing, high coverage sequencing of multiple individuals is still expensive. Furthermore, it is challenging to obtain individual genomic data for certain species (e.g. small organisms) or in evolve-and-resequence experiments involving a large number of populations or many points along a time series. In those instances, next-generation sequencing of pooled samples (Pool-seq) might be the only viable alternative, as it requires less DNA per individual. Pool-seq is a sequencing technique that provides a cost-effective approach to quantify genetic variation within a population. It involves pooling multiple individual DNA samples together and sequencing them collectively. A typical Pool-seq analysis requires several steps. First, researchers should determine the pool size (i.e. the number of individuals included in the pool) and the desired sequencing depth of coverage during the experimental design step. Next, DNA extracted from individual samples is combined into pools. In situations where obtaining DNA from each individual sample is impractical, an alternative approach is to group several individuals together prior to DNA extraction. For instance, muscle tissue from multiple individuals can be combined, extracting DNA from the entire group of individuals (Morales et al., 2019; Ross et al., 2019). Then, DNA extracted from multiple groups of individuals can be merged into a single, final pool. Non-equimolar quantities of DNA between these groups of multiple individuals, or between individuals within a group, can lead to unequal contributions. This disparity in contribution may result in certain groups of individuals having a disproportionate impact on the overall allele frequencies, leading to inaccurate estimation of sample allele frequencies, potentially affecting downstream analysis (Anderson et al., 2014; Ellegren, 2014). Subsequently, for each pool, a single library is generated prior to sequencing with NGS technologies. Note that Pool-seq does not require individual tagging of sequences, reducing the laboratory work required for library preparation, while still generating population-level genomic data (Schlotterer et al., 2014). During this library preparation step, stochastic variation in amplification efficiency can also result in unequal contributions of individuals, and lead to inaccurate sample allele frequencies. Finally, the pooled libraries are sequenced. This step also introduces uncertainties in the analysis due to variation in sequencing depth along the genome, and sequencing errors. The next steps, such as quality control, read alignment and variant calling, are similar to individual-based sequencing.

Despite these potential sources of uncertainty (e.g. unequal individual contribution), Pool-seq has been extensively used in a variety of settings (Begun et al., 2007; Ferretti et al., 2013; Prescott et al., 2015; Zhou et al., 2011). This has lead to the development of tools such as the R package poolSeq (Taus et al., 2017) and the DIYABC—RF software (Collin et al., 2021) that simulate Pool-seq data, as well as data analysis tools (e.g. Kofler et al., 2011). Nonetheless, to the best of our knowledge, no tool currently exists that can simultaneously and explicitly account for variation in depth of coverage, unequal contribution and sequencing errors, which are known sources of Pool-seq uncertainty (see Table S1 for more details). It is worth noting that unequal contribution occurs due to variations in DNA concentration or amplification efficiency among the pooled samples, resulting in an uneven representation of genetic material from each sample. Here, we use the term pooling error to quantify the error caused by unevenly combining multiple DNA samples into a single pool, which we explicitly model as the dispersion around the expected proportion of reads from each sample. This pooling error can introduce biases in estimates of sample allele frequencies. As mentioned, two key parameters in the experimental design step of a Pool-seq study are the number of individuals in each pool, and the average depth of coverage. These two parameters determine how much the sample allele frequencies are affected by Pool-seq associated errors. On one hand, increasing the number of individuals allows estimating more accurate allele frequencies, but more individuals in the pool might not avoid errors associated with unequal individual contribution when the pooling error is high. On the other hand, increasing the depth of coverage should lead to more reliable estimates but it can amplify pooling errors and increase the frequency of sequencing errors, which can make it challenging to differentiate true low-frequency variants from sequencing errors. Moreover, due to its costs, the depth of coverage is typically the limiting resource. Simulations of single nucleotide polymorphism (SNP) data accounting for sources of uncertainty with Pool-seq data under different sampling schemes can thus provide a tool to help researchers design Pool-seq experiments and to minimise the error associated with the sample allele frequencies.

Here, we introduce an R package (Team, 2020), poolHelper, to simulate Pool-seq data according to different sampling designs. Our approach relies on coalescent simulations under neutrality using *scrm* (Staab et al., 2015). The poolHelper package provides tools and functions to simulate Pool-seq datasets, accounting for potential sources of error in the Pool-seq analysis process. Importantly, these errors are modelled by parameters that users can adjust. poolHelper models the unequal contribution resulting from differences in DNA

concentration and amplification efficiency during DNA extraction and library preparation. Additionally, it accounts for sequencing depth variation across SNPs, sequencing errors, and mapping errors during read alignment. This allows comparing the allele frequencies obtained directly from the simulated individual genotypes with the frequencies obtained from Pool-seq data. Since R is a free and collaborative project, users can use available tools to handle, analyse and visualise genomic datasets. Our goal is to provide a flexible method of simulating Pool-seq data, allowing researchers to design their experiments with a better a priori knowledge of possible errors associated with Pool-seq, thus contributing to the recognition of Pool-seq as a valuable source of data to reconstruct the evolutionary history of populations.

2 | IMPLEMENTATION

The main steps of our pipeline follow a relatively simple scheme: coalescent simulations of individual genotypes under a single population model with a constant size, computation of alternative allele frequencies directly from the genotypes, simulation of Pool-seq given the genotypes, and computation of alternative allele frequencies from the Pool-seq data, assuming that it corresponds to the proportion of reads with that allele. To measure the error associated with Pool-seq we computed the average absolute difference between the actual allele frequencies based on individual genotypes in the sample and the allele frequencies obtained with Pool-seq. Thus, note that we measure the difference between two estimates of the allele frequencies in the sample, one based on the sampled individual genotypes and the other obtained with Pool-seq of the same sample. The poolHelper package provides functions to simulate Pool-seq data, under a variety of user-defined conditions. More specifically, users can vary the average and variance of the depth of coverage, the pool size, sequencing error and the pooling error (see below). Additionally, they can also vary the number of groups of individuals contributing to the final sequenced pool. By varying all of these conditions, it is possible to assess how they influence the accuracy of allele frequency estimations. No external R objects are needed to use the package. Users can use the implemented coalescent simulations to obtain genotypes, or provide genotypes directly. The resulting Pool-seq data can be outputed as R objects with counts of reads, or converted to commonly used file formats (.vcf and .sync), allowing users to analyse simulated Pool-seq data with existing downstream methods.

2.1 | Coalescent simulations of individual genotypes

To obtain individual genotypes, we used *scrm* to simulate coalescent gene trees under a model of a single population with constant effective size N_e. To model different effective population sizes and mutation rates, users can vary $\theta = 4N_e\mu$, where μ is the neutral mutation rate per locus per generation. This allows to investigate Pool-seq associated uncertainties in populations with varying levels of expected genetic diversity, which is proportional to θ. We assumed that the sample size was the same for each locus, corresponding to the total number of individuals sampled in the Pool-seq experiment. The effective size of the population from which the sample is taken is defined by θ, which users can modify. Additionally, we assumed that the actual haplotypes of all individuals in the pool were known. The effect of pooling is simulated in posterior steps (see next section). To obtain individual genotypes, we assumed random mating in the population and paired haplotypes at each locus at random for each biallelic single nucleotide polymorphic (SNP) site.

2.2 | Simulation of Pool-seq data

We follow a series of steps (Figure 1) to model and simulate allele frequencies obtained with Pool-seq for biallelic SNPs, as described in Carvalho et al. (2023). The variation in depth of coverage across SNPs is assumed to follow a negative binomial distribution (*nBin*,

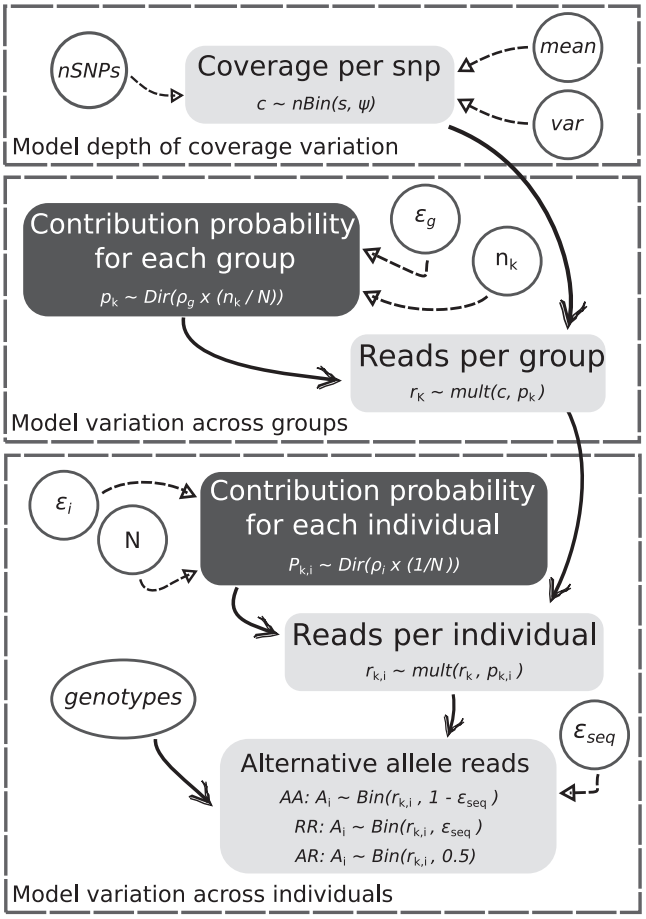

FIGURE 1 Diagram of the required steps to simulate Pool-seq data. The steps related to contribution probabilities are depicted by dark coloured boxes, while circles represent the required inputs for each corresponding step. Each box contains the relevant formulas for its corresponding step.

following e.g. Hardcastle & Kelly, 2010). Thus, the number of reads c at each site is

$$c \sim nBin(s, \psi), \tag{1}$$

where $s = \text{mean}(c) / \text{var}(c)$ and $\psi = \text{mean}(c)^2 / (\text{var}(c) - \text{mean}(c))$. The mean($c$) and var($c$) represent, respectively, the mean and variance of the depth of coverage across all SNPs. We assumed that the sequenced pool can have resulted from merging DNA extracted from K different groups of individuals, where each group could have a different number of individuals. To account for variability in the contribution of each individual to the pool, we assumed that the number of reads follows a multinomial-Dirichlet distribution. That is, at each site, reads from the ith individual in the kth group ($r_{k,i}$) follow a multinomial distribution

$$r_{k,i} \sim \text{mult}(c, p_{k,i}), \tag{2}$$

where $p_{k,i}$ denotes the proportion of reads from individual i in group k, which is assumed to follow a Dirichlet distribution,

$$p_{k,i} \sim \text{Dir}\left(\rho_i \frac{1}{N}\right), \tag{3}$$

where N denotes the total number of sequenced individuals in group k, and ρ_i models the variance of contribution, reflecting the unequal contribution of individuals. Note that the contribution is expected to be equal for all individuals. If DNA extraction is performed for K groups of individuals that are then combined into a larger pool, uneven contributions between these groups of individuals may also occur. To account for this, we modelled the unequal contribution of each group of individuals by assuming that the number of reads from the kth group (r_k) follows a multinomial-Dirichlet distribution, such that $r_k \sim \text{mult}(c, p_k)$, where p_k is the proportion of reads from a given group, assumed to follow a Dirichlet distribution,

$$p_k \sim \text{Dir}\left(\rho_g \frac{n_k}{N}\right), \tag{4}$$

where n_k is the number of individuals in group k, and ρ_g models the variance of contribution due to unequal contribution of groups of individuals. Following Gautier et al. (2013), we model explicitly the pooling error due to unequal contribution with the parameters ρ_i and ρ_g, which reflect the variance of contribution of individuals and groups of individuals, respectively, as

$$\rho_i = \frac{N - 1 - \varepsilon_i^2}{\varepsilon_i^2}, \tag{5}$$

$$\rho_g = \frac{(N / n_k) - 1 - \varepsilon_g^2}{\varepsilon_g^2}, \tag{6}$$

where ρ_i and ρ_g are the unequal contribution parameters for individuals within a group, and among groups of individuals, respectively. All groups of individuals are assumed to have the same ρ_g, and all individuals are assumed to have the same ρ_i. These depend on pooling error parameters ε_i and ε_g for individuals and groups of individuals, respectively (Gautier et al., 2013). Larger values of ε_i and ε_g lead to a larger dispersion, resulting in more unequal contributions. The variance of contribution depends on the experimental error as $\text{var}(p_{k,i}) = (\varepsilon_i E[p_{k,i}])^2$ and $\text{var}(p_k) = (\varepsilon_g E[p_k])^2$. Although the selection of an appropriate pooling error might be potentially hard, given its unknown nature, we previously estimated values ranging from 24 to 236 (Carvalho et al., 2023). Furthermore, previous studies have also considered values ranging from 0 to 250 (Gautier et al., 2013). Thus, the pooling errors used here are within the reasonable ranges for this parameter (see Figure S1 for an example of how different pooling errors impact individual contribution). Note that the ε_i and ε_g that reflect the maximum dispersion, that is maximum unequal contribution when just one individual or one group of individuals contribute to the pool, correspond to ρ_i and ρ_g of zero. This implies that the upper limit for ε_i^2 is $N - 1$ (Equation 5), and for ε_g^2 is $(N / n_k) - 1$ (Equation 6). Users can use these values as a reference to determine the maximum error values based on their sample sizes.

We also accounted for sequencing and mapping errors by assuming that the reference allele R may be incorrectly called as an alternative allele A or vice versa with an error rate ε_{seq}. We modelled the number of reads A_i with the alternative allele for the ith individual at a particular site following a binomial distribution: we assumed $A_i \sim Bin(r_{k,i}, \varepsilon_{seq})$ if the individual is homozygous for the reference allele and $A_i \sim Bin(r_{k,i}, 1 - \varepsilon_{seq})$ if the individual is homozygous for the alternative allele. We also assumed that there are only two alleles at each site and that each base has an equal chance of being miscalled. Therefore, for heterozygous individuals, each read originates from either the reference or alternative allele with equal probability (Li et al., 2012) and $A_i \sim Bin(r_{k,i}, 0.5)$, where $r_{k,i}$ represents the total number of reads contributed by an individual. A commonly used filter can also be applied, discarding SNPs with less than the required number of minor-allele reads. The allele frequencies estimated for the Pool-seq data correspond to the proportion of reads with the alternative allele.

2.3 | Measuring error of estimates

To measure the error of Pool-seq estimates of allele frequencies or expected heterozygosity, we compared the estimates obtained from the individual genotypes in the sample with the estimates obtained from Pool-seq. We calculate the mean absolute error as

$$\varepsilon = \frac{1}{n} \times \sum |\, y_i - x_i \,|, \tag{7}$$

where n indicates the total number of SNPs. When calculating the error of Pool-seq estimates of allele frequencies, x_i and y_i correspond to the frequencies of the alternative allele at the i^{th} SNP in the sample, obtained with individual genotypes (x_i) or with Pool-seq (y_i). When measuring the error of expected heterozygosity, x_i and y_i represent the expected heterozygosity obtained based on the sample of either individual genotypes (x_i) or Pool-seq (y_i).

2.4 | Main functionality

The poolHelper package allows users to compute the mean absolute error of allele frequencies and expected heterozygosity under a variety of conditions. Users can vary the mean depth of coverage and the associated variance, the value of the pooling error and the number of sampled individuals. Additionally, it is possible to evaluate the effect of combinations of parameters, for instance, various mean depths of coverage combined with several pooling error values. Thus, the poolHelper package provides users with a tool to aid in the design of pooled sequencing experiments, by allowing researchers to evaluate the best strategy, in terms of pool sizes or depth of coverage, to obtain accurate estimates of allelic frequencies, while minimising the sampling effort and costs.

2.5 | Effect of combining multiple groups of individuals

An important consideration is whether DNA extraction should involve multiple groups of individuals, which are then combined into a final pool for library preparation and sequencing, or if DNA should be extracted individually from each sample and subsequently combined into a final pool. Users can test the effect of this choice by using the "maePool" function. This function computes the mean absolute error for a given sample size sequenced using a pool with a single group of individuals or a pool combining multiple groups of individuals (Figure S2). By varying the mean coverage and the pooling error, it is possible to evaluate the effect of using a single or multiple groups under different conditions.

2.6 | Impact of mean depth of coverage

Another critical decision is defining the mean depth of coverage used to sequence a pool of individuals. The "maeFreqs" function implements the calculation of the mean absolute error between allele frequencies computed from genotypes and Pool-seq allele frequencies simulated under different mean depth of coverage. By varying the mean depth of coverage and the associated variance, users can determine which coverage produces more accurate allele frequency estimates for a given sample size and pooling error (Figure 2).

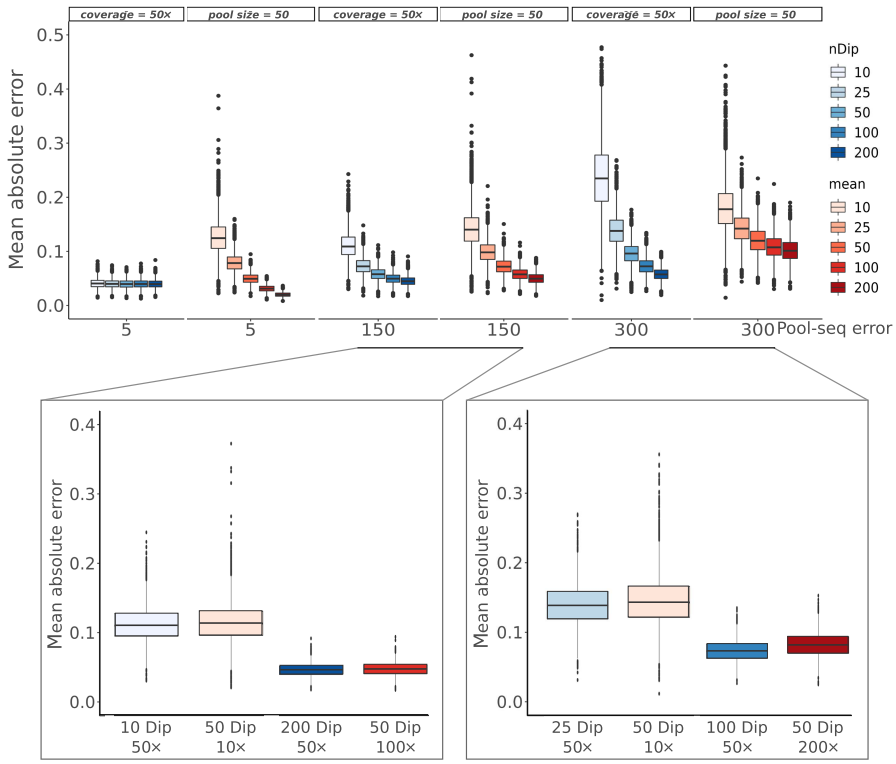

FIGURE 2 Mean absolute error between the allele frequencies obtained from the individual genotypes in the sample and those obtained from Pool-seq data under a variety of conditions. For all conditions, sites with less than two minor-allele reads were removed. In all plots, the *y*-axis represents the mean absolute error between the allele frequencies estimates. The top panel shows the mean absolute error for three different pooling error values (*x*-axis). For each plot, either the pool size or the coverage were fixed (the fixed value is indicated on the top of each plot). Thus, when pool size was fixed, the average coverage varied and vice-versa. In the bottom panel, we highlight comparisons that lead to similar mean absolute errors for intermediate values of pooling error (150 in the bottom left panel) and high pooling error (300 in the bottom right panel). In all plots, the pool size, defined by the *nDip* parameter, is represented in shades of blue, with darker shades indicating a larger pool and the average coverage, defined by the *mean* parameter, is represented in shades of red, with darker shades indicating higher coverage.

2.7 | Impact of pool sizes

When designing a Pool-seq experiment, it is essential to define the number of individuals to include in the pool, that is the pool size. The calculation of the mean absolute error between allele frequencies for different pools sizes can be carried out using the "maeFreqs" function. This allows users to evaluate what is the optimal pool size for a fixed coverage and/or pooling error (Figure 2). Thus, the "maeFreqs" function allows users to decide how many individuals to pool to obtain the most accurate allele frequencies estimates for a given mean depth of coverage.

2.8 | Example of an effective Pool-seq design using simulations

By performing simulations in a single panmictic population, assuming that pooling error is intermediate to high (150 or 300) and after applying a commonly used filter (removing sites with less than two minor-allele reads), it is not obvious that one should always increase the average depth of coverage per individual in the pool (Figure 2). For instance, when pooling error is 150, we observe the same mean absolute error with a pool of 50 individuals sequenced at 10× than with a pool of 10 individuals sequenced at 50×. This suggests that it may be more cost-effective to use a pool of 50 individuals at 10× (expected individual contribution of 10/50) than using fewer individuals with a higher expected coverage per individual. This holds true for larger pool sizes and depths of coverage, given that we also get the same mean absolute error when comparing a pool of 200 individuals sequenced at 50× with a pool of 50 individuals sequenced at 100× (Figure 2). If pooling error is even higher (i.e. 300) a pool of 100 individuals sequenced at 100× leads to a slightly lower mean absolute error than a pool of 50 individuals sequenced at double the coverage (200×; Figure 2). Thus, similar errors of allele frequencies in the sample can be obtained with different combinations of pool sizes and average depth of coverage. Therefore, the design of an effective Pool-seq study is not straightforward and an a priori simulation study can help assess an efficient sampling scheme to obtain accurate allele frequencies while minimising the sequencing effort (mean depth of coverage).

3 | CONCLUSIONS

We present an R package, poolHelper, to simulate pooled sequencing data under a model of a single panmictic population and compute the error in sample allele frequencies and expected heterozygosity obtained with Pool-seq for different study designs and commonly used filters (e.g. filters on minimum and maximum depth of coverage and on minimum number of minor-allele reads). The package relies on coalescent simulations performed with *scrm* (Staab et al., 2015). Currently, data are simulated under a single population with a constant effective population size. However, our package allows users to simulate genotypes under different models and use those genotypes as input to compute the mean absolute error or simulate Pool-seq data. This enables users to focus on their specific scenarios of interest and then simulate Pool-seq data under a wide range of user-defined parameters. This package is implemented in the R environment, providing tools for data visualisation, allowing users to produce graphics and quickly visualise the effect of multiple combinations of Pool-seq parameters. The poolHelper package's vignette contains a comprehensive explanation of the functions in the package, as well as examples detailing its usage.

AUTHOR CONTRIBUTIONS

João Carvalho and Vítor C. Sousa conceptualised and designed poolHelper; João Carvalho implemented the package in R with supervision and input of Vítor C. Sousa; Roger K. Butlin and Rui Faria discussed preliminary results and provided suggestions that improved the functionality of the package; João Carvalho wrote the first draft, with comments from all authors. All authors contributed critically to the drafts and gave final approval for publication.

ACKNOWLEDGEMENTS

We thank the editor and two anonymous reviewers for their comments and suggestions and Beatriz Portinha for suggesting the package name. This work was funded by the strategic project UIDB/00329/2020 granted to cE3c from the Portuguese National Science Foundation—Fundação para a Ciência e a Tecnologia (FCT). João Carvalho was supported by an FCT Ph.D. scholarship (PD/BD/128350/2017). Rui Faria is funded by a FCT CEEC (Fundação para a Ciência e a Tecnologia, Concurso Estímulo ao Emprego Científico) contract (2020.00275.CEECIND) and by a FCT research project (PTDC/BIA-EVL/1614/2021). Roger K. Butlin was funded by the European Research Council (ERC-2015-AdG-693030-BARRIERS). Vítor C. Sousa was supported by FCT (CEECINST/00032/2018/CP1523/CT0008) and by the Human Frontier Science Program (RGY0081/2020). We thank the National Network for Advanced Computing (RNCA) and INCD (https://incd.pt/) for use of the computing infrastructure, funded by FCT to Vítor C. Sousa (2021.09795.CPCA).

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

PEER REVIEW

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1111/2041-210X.14185.

DATA AVAILABILITY STATEMENT

The package poolHelper source code is hosted at https://github.com/joao-mcarvalho/poolHelper, along with the package tutorials and vignettes. The source code is archived with Zenodo at https://doi.org/10.5281/zenodo.7520303 (Carvalho et al., 2023). poolHelper is

available on the Comprehensive R Archive Network (CRAN; https://cran.r-project.org/package=poolHelper). There is no other data associated with this paper.

ORCID

João Carvalho https://orcid.org/0000-0002-1728-0075
Rui Faria https://orcid.org/0000-0001-6635-685X
Roger K. Butlin https://orcid.org/0000-0003-4736-0954
Vitor C. Sousa https://orcid.org/0000-0003-3575-0875

REFERENCES

Anderson, E. C., Skaug, H. J., & Barshis, D. J. (2014). Next-generation sequencing for molecular ecology: A caveat regarding pooled samples. *Molecular Ecology*, *23*(3), 502–512. https://doi.org/10.1111/mec.12609

Begun, D. J., Holloway, A. K., Stevens, K., Hillier, L. W., Poh, Y.-P., Hahn, M. W., Nista, P. M., Jones, C. D., Kern, A. D., Dewey, C. N., Pachter, L., Myers, E., & Langley, C. H. (2007). Population genomics: Whole-genome analysis of polymorphism and divergence in drosophila simulans. *PLoS Biology*, *5*(11), e310. https://doi.org/10.1371/journal.pbio.0050310

Carvalho, J., Faria, R., Butlin, R. K., & Sousa, V. C. (2023). *poolHelper*. https://doi.org/10.5281/zenodo.7520303

Carvalho, J., Morales, H. E., Faria, R., Butlin, R. K., & Sousa, V. C. (2023). Integrating pool-seq uncertainties into demographic inference. *Molecular Ecology Resources*. https://doi.org/10.1111/1755-0998.13834

Collin, F.-D., Durif, G., Raynal, L., Lombaert, E., Gautier, M., Vitalis, R., Marin, J.-M., & Estoup, A. (2021). Extending approximate bayesian computation with supervised machine learning to infer demographic history from genetic polymorphisms using diyabc random forest. *Molecular Ecology Resources*, *21*(8), 2598–2613. https://doi.org/10.1111/1755-0998.13413

Ellegren, H. (2014). Genome sequencing and population genomics in nonmodel organisms. *Trends in Ecology & Evolution*, *29*(1), 51–63. https://doi.org/10.1016/j.tree.2013.09.008

Ferretti, L., Ramos-Onsins, S. E., & Pérez-Enciso, M. (2013). Population genomics from pool sequencing. *Molecular Ecology*, *22*(22), 5561–5576. https://doi.org/10.1111/mec.12522

Gautier, M., Foucaud, J., Gharbi, K., Cézard, T., Galan, M., Loiseau, A., Thomson, M., Pudlo, P., Kerdelhué, C., & Estoup, A. (2013). Estimation of population allele frequencies from next-generation sequencing data: Pool-versus individual-based genotyping. *Molecular Ecology*, *22*(14), 3766–3779. https://doi.org/10.1111/mec.12360

Hardcastle, T. J., & Kelly, K. A. (2010). bayseq: Empirical bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics*, *11*(422), 1–14. https://doi.org/10.1186/1471-2105-11-422

Kofler, R., Pandey, R. V., & Schlötterer, C. (2011). Popoolation2: Identifying differentiation between populations using sequencing of pooled dna samples (pool-seq). *Bioinformatics*, *27*(24), 3435–3436. https://doi.org/10.1093/bioinformatics/btr589

Li, B., Chen, W., Zhan, X., Busonero, F., Sanna, S., Sidore, C., Cucca, F., Kang, H. M., & Abecasis, G. R. (2012). A likelihood-based framework for variant calling and de novo mutation detection in families. *PLoS Genetics*, *8*(10), 1–12. https://doi.org/10.1371/journal.pgen.1002944

Morales, H. E., Faria, R., Johannesson, K., Larsson, T., Panova, M., Westram, A. M., & Butlin, R. K. (2019). Genomic architecture of parallel ecological divergence: Beyond a single environmental contrast. *Science Advances*, *5*(12), eaav9963. https://doi.org/10.1126/sciadv.aav9963

Prescott, N. J., Lehne, B., Stone, K., Lee, J. C., Taylor, K., Knight, J., Papouli, E., Mirza, M. M., Simpson, M. A., Spain, S. L., Lu, G., Fraternali, F., Bumpstead, S. J., Gray, E., Amar, A., Bye, H., Green, P., Chung-Faye, G., Hayee, B., … UK IBD Genetics Consortium. (2015). Pooled sequencing of 531 genes in inflammatory bowel disease identifies an associated rare variant in btnl2 and implicates other immune related genes. *PLoS Genetics*, *11*(2), 1–19. https://doi.org/10.1371/journal.pgen.1004955

R Core Team. (2020). *R: A language and environment for statistical computing [computer software manual]*. R Foundation for Statistical Computing. https://www.R-project.org/

Ross, P. A., Endersby-Harshman, N. M., & Hoffmann, A. A. (2019). A comprehensive assessment of inbreeding and laboratory adaptation in *Aedes aegypti* mosquitoes. *Evolutionary Applications*, *12*(3), 572–586. https://doi.org/10.1111/eva.12740

Schlotterer, C., Tobler, R., Kofler, R., & Nolte, V. (2014). Sequencing pools of individuals—Mining genome-wide polymorphism data without big funding. *Nature Reviews Genetics*, *15*(11), 749–763. https://doi.org/10.1038/nrg3803

Staab, P. R., Zhu, S., Metzler, D., & Lunter, G. (2015). scrm: Efficiently simulating long sequences using the approximated coalescent with recombination. *Bioinformatics*, *31*(10), 1680–1682. https://doi.org/10.1093/bioinformatics/btu861

Taus, T., Futschik, A., & Schlötterer, C. (2017). Quantifying selection with pool-seq time series data. *Molecular Biology and Evolution*, *34*(11), 3023–3034. https://doi.org/10.1093/molbev/msx225

Zhou, D., Udpa, N., Gersten, M., Visk, D. W., Bashir, A., Xue, J., Frazer, K. A., Posakony, J. W., Subramaniam, S., Bafna, V., & Haddad, G. G. (2011). Experimental selection of hypoxia-tolerant drosophila melanogaster. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(6), 2349–2354. https://doi.org/10.1073/pnas.1010643108

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

Table S1. Comparison of different tools to simulate Pool-seq data. We compared the functionalities of three different packages/tools that can generate simulated Pool-seq data. We assessed if they can be used to model known sources of Pool-seq uncertainty, such as variation in depth of coverage, unequal individual contribution and sequencing errors.

Figure S1. Impact of pooling errors in individual contribution. We simulated Pool-seq data obtained with pools of either 10 or 50 individuals, sequenced at a coverage of 200×. Three different levels of pooling errors were considered: 5, 150 and 300. In A and B, the expected contribution of each individual is indicated by the dashed line. The distribution of the number of reads contributed by each individual for a pool of 10 (A) or 50 (B) individuals shows that higher pooling errors result in deviations from the expected value and an increased number of individuals with zero (or near-zero) reads. Note that, with a pooling error of 300, some individuals contributed ~200 reads. The impact of higher pooling errors is also clear when we analyse the proportion of individuals contributing zero reads or twice the expected number of reads for pools of both 10 (C) and 50 (D) individuals.

Figure S2. Mean absolute error between the allele frequencies obtained from the individual genotypes in the sample and those obtained from Pool-seq data using a single or multiple groups of individuals. Pool-seq data were simulated for 100 individuals. Sequencing was performed with an average coverage of 100× using either a single group of individuals or a pool combining 10 groups of

individuals, with each group containing 10 individuals. Two scenarios were considered: one assuming a low pooling error rate of 5, and the other assuming a high pooling error rate of 300. The same pooling error value was used to model the dispersion among pools and individuals. The *y*-axis represents the mean absolute error between the allele frequencies estimates and the *x*-axis indicates the number of groups used to sequence the sample.

How to cite this article: Carvalho, J., Faria, R., Butlin, R. K., & Sousa, V. C. (2023). POOLHELPER: An R package to help in designing Pool-Seq studies. *Methods in Ecology and Evolution*, *00*, 1–8. https://doi.org/10.1111/2041-210X.14185