PhaseTypeR: an R package for phase-type distributions in population genetics

Iker Rivas-González [1*¶], **Lars Nørvang Andersen** [2*], **and Asger Hobolth** [2*]

1 Bioinformatics Research Centre, Aarhus University, Denmark **2** Department of Mathematics, Aarhus University, Denmark ¶ Corresponding author * These authors contributed equally.

Summary

Phase-type distributions describe the time until absorption of a continuous or discrete-time Markov chain (Bladt & Nielsen, 2017). The probabilistic properties of phase-type distributions (i.e., the probability density function, cumulative distribution function, quantile function, moments and generating functions) are well-described and analytically tractable using matrix manipulations.

Phase-type distributions have traditionally been used in actuarial sciences and queuing theory, and, more recently, in population genetics. In order to facilitate the use of phase-type theory in population genetics, we present PhaseTypeR, a general-purpose and user-friendly R (R Core Team, 2021) package which contains all key functions —mean, (co)variance, probability density function, cumulative distribution function, quantile function and random sampling— for both continuous and discrete phase-type distributions. Importantly, univariate and multivariate reward transformations are implemented for continuous and discrete phase-type distributions. Multivariate reward transformations have great potential for applications in population genetics, and we have included two examples. The first is concerned with the easy calculation of the variance-covariance matrix for the site frequency spectrum (SFS) of the n-coalescent, and the second is concerned with the correlation between tree heights in the two-locus ancestral recombination graph.

Statement of need

In recent years, the usefulness of phase-type theory in population genetics has become evident. Important quantities in population genetics such as the time until the most recent ancestor, the total tree length, the total number of segregating sites, and the site frequency spectrum follow phase-type distributions (Hobolth et al., 2019). There are already several other R packages that model phase-type distributions, such as actuar (Dutang et al., 2008), mapfit (Hiroyuki Okamura, 2015; Hiroyuki Okamura & Dohi, 2015; H. Okamura & Dohi, 2016) or matrixdist (Albrecher et al., 2022; Albrecher & Bladt, 2019). However, these packages only model univariate continuous phase-type distributions, do not include reward transformations, and are tailored to actuarial sciences and queuing theory.

To overcome these limitations, we present PhaseTypeR, an R package with easy-to-use, general-purpose phase-type functions, which is particularly well suited for population genetics. The package has already been used in Hobolth et al. (2021) to model the site frequency spectrum using multivariate phase-type theory, and we believe that its intuitive implementation will encourage more population geneticists to use phase-type theory.

Overview

PhaseTypeR contains general implementations of core functions for continuous and discrete phase-type distributions, for both the univariate and multivariate cases. These include the mean and (co)variance of phase-type distributions; their density, probability and quantile functions; functions for random draws; and functions for reward transformations.

Table 1 provides an overview of the PhaseTypeR functions for a univariate continuous phase-type distribution $\tau \sim \mathrm{PH}(\alpha, T)$, where α is the initial distribution and T the sub-intensity matrix. Let $\{X(t) : t \geq 0\}$ denote the corresponding continuous-time Markov chain (CTMC). The reward transformation is then given by $Y = \int_0^\tau r(X(t))dt$, where τ is the time to absorption, and Y is also phase-type distributed (Bladt & Nielsen, 2017). If the CTMC has p transient states, then the reward function $r(i), i = 1, \ldots, p$, is a vector of length p.

Table 1: formulas and corresponding PhaseTypeR functions for univariate continuous phase-type distributions. The vector a determines the initial probabilities, T is the sub-intensity matrix, e is a vector with 1 in every entry, and r is the reward vector.

Quantity	Formula	Function
PH object	$\tau \sim \mathrm{PH}(a, T)$	`PH(T, a)`
Mean	$\mathsf{E}[\tau] = a(-T)^{-1}e$	`mean(PH)`
Variance	$\mathsf{V}[\tau] = \mathsf{E}[\tau^2] - \mathsf{E}[\tau]^2$	`var(PH)`
Density	$f(x) = a \exp(Tx)(-Te),$ $x \geq 0$	`dPH(x, PH)`
Cumulative distribution	$F(x) = 1 - a \exp(Tx)e,$ $x \geq 0$	`pPH(x, PH)`
Quantile function		`qPH(p, PH)`
Random sampling		`rPH(n, PH)`
Random sampling of full path		`rFullPH(n, PH)`
Reward transformation	$Y = \int_0^\tau r(X(t))dt$	`reward_phase_type(PH, r)`

Table 2 provides an overview of the PhaseTypeR functions for the univariate discrete phase-type distribution. Here, the reward transformation is given by Campillo Navarro (2018). Table 3 gives an overview of the multivariate phase-type distribution. A multivariate phase-type distribution is the joint distribution of $(Y_1, \ldots, Y_k)$ where $Y_j = \int_0^\tau r_j(X(t))dt$ for $j = 1, \ldots, k$. We summarize the rewards $r_j(i)$ in a matrix R with p rows (corresponding to the transient states) and k columns (corresponding to the k reward functions) with entries $R_{ij} = r_j(i)$.

Table 2: Formulas and corresponding PhaseTypeR functions for univariate discrete phase-type distributions. The vector a determines the initial probabilities, T is the sub-transition matrix, e is a vector with one in every entry, I is the identity matrix, and r is the reward vector.

Quantity	Formula	Function
DPH object	$\tau \sim \mathrm{DPH}(a, T)$	`DPH(T, a)`
Mean	$\mathsf{E}[\tau] = \pi(I - T)^{-1}e$	`mean(DPH)`
Variance	$\mathsf{V}[\tau] = \mathsf{E}[\tau^2] - \mathsf{E}[\tau]^2$	`var(DPH)`
Density	$f(x) = \pi T^{x-1}t, \; x \geq 1$	`dDPH(x, DPH)`
Cumulative distribution	$F(x) = 1 - \pi T^x e, \; x \geq 1$	`pDPH(x, DPH)`
Quantile function		`qDPH(p, DPH)`
Random sampling		`rDPH(n, DPH)`
Random sampling of full path		`rFullDPH(n, DPH)`
Reward transformation	$Y = \sum_{m=0}^{\tau-1} r(X_m)$	`reward_phase_type(DPH, r)`

Table 3: `PhaseTypeR` functions for multivariate continuous and multivariate discrete phase-type distributions. `a` is the vector of initial probabilities, `T` is the sub-intensity matrix and `R` is the reward matrix. For information about the formulas for calculating the covariances, please see Bladt & Nielsen (2017).

Quantity	Continuous	Discrete
Multivariate PH object	`MPH(T, a, R)`	`MDPH(T, a, R)`
Mean	`mean(MPH)`	`mean(MDPH)`
(Co)variance	`var(MPH)`	`var(MDPH)`
Density	`dMPH(x, MPH)`	`dMDPH(x, MDPH)`
Cumulative distribution	`pMPH(x, MPH)`	`pMDPH(x, MDPH)`
Quantile function	`qMPH(p, MPH)`	`qMDPH(p, MDPH)`
Random sampling	`rMPH(n, MPH)`	`rMDPH(n, MDPH)`
Random sampling of full path	`rFullMPH(n, MPH)`	`rFullMDPH(n, MDPH)`

Example 1: variance-covariance matrix of the SFS

This section concerns reproducing the table associated with Theorem 2.2 in Durrett (2008), which can be used to derive the variance of the elements of the site frequency spectrum (SFS) and the covariance between pairs of elements of the SFS.

Let ξ_i be the i'th element of the site frequency spectrum, i.e., ξ_1 is the number of singletons, ξ_2 is the number of doubletons, etc. Let's also define L_i, which is the total branch length leading to ξ_i. Theorem 3.1 in Hobolth et al. (2019) states that the vector $L = (L_1, ..., L_{n-1})$ has a multivariate phase-type distribution

$$L \sim \text{MPH}(e_1, T, R),$$

where R and T are respectively the state-space and the sub-transition matrix of the so-called "block-counting process", and $e_1 = (1, 0, ..., 0)$. Conditionally on L, the ξ_i's are independent and Poisson distributed, $\xi_i \mid L_i \sim \text{Poisson}\left(L_i \frac{\theta}{2}\right)$, where θ is the underlying mutation rate (Wakeley, 2009). By the law of total variance, we have

$$\text{Var}[\xi] = \frac{\theta^2}{4}\Sigma + \frac{\theta}{2}\text{diag}(\text{E}[L]).$$

where $\text{diag}(\cdot)$ is the diagonal matrix whose entries are given by $\text{E}[L]$. It is well-known that $\text{E}[L_i] = 1/i$, but the expressions for the entries of Σ are fairly complicated (Durrett, 2008; Fu, 1995). However as seen below, numeric calculation of Σ is straightforward using phase-type theory, since we just need to specify the subintensity matrix T and the reward matrix R for the block-counting process.

Accompanying our package are a number of vignettes, which illustrate the use of phase-type distribution in population genetics. As part of one of these vignettes, we include a function that calculates R and T for the block-counting process with sample size n, which is shown below:

```r
RateMAndStateSpace <- function(n){
  ## --------- State space ---------
  # Size of the state space (number of states)
  nSt <- partitions::P(n)
  # Definition of the state space
  StSpM <- matrix(ncol=n,nrow=nSt)
  # Set of partitions of [n]
  x <- partitions::parts(n)
  # Rewriting the partitions as (a1,...,an)
  for (i in 1:nSt) {
```

```r
    st <- x[,i]
    StSpM[i,] <- tabulate(x[,i],nbins=n)
  }
  # Reordering
  StSpM <- StSpM[order(rowSums(StSpM),decreasing=TRUE),]
  # Below the diagonal the entries are always zero
  ## --------- Intensity matrix ---------
  RateM <- matrix(0,ncol=nSt,nrow=nSt)
  for (i in 1:(nSt-1)){
    for (j in (i+1):nSt){
      cvec <- StSpM[i,]-StSpM[j,]
      # Two branches are merged, i.e. removed from state i
      check1 <- sum(cvec[cvec>0])==2
      # One new branch is created, i.e. added in state from j
      check2 <- sum(cvec[cvec<0])==-1
      if (check1 & check2){
        # Size(s) of the block(s) and the corresponding rates
        tmp <- StSpM[i,which(cvec>0)]
        RateM[i,j] <- ifelse(length(tmp)==1,tmp*(tmp-1)/2,prod(tmp))
      }
    }
  }
  # Diagonal part of the rate matrix
  for (i in 1:nSt){
    RateM[i,i] <- -sum(RateM[i,])
  }
  list(RateM=RateM,StSpM=StSpM)
}
```

We can now define a multivariate phase-type distribution such that $L \sim \mathrm{MPH}(\alpha, T, R)$. This is straightforward to build in `PhaseTypeR` with the `MPH()` function. For $n = 8$:

```r
n <- 8
RMASS <- RateMAndStateSpace(n)
m <- dim(RMASS$RateM)[1]
# Obtain subintensity matrix
subintensity_matrix <- RMASS$RateM[1:(m-1),1:(m-1)]
# The reward matrix is the state space matrix of the block counting process
rew_mat <- RMASS$StSpM[1:(m-1),1:(n-1)]
# The initial probability vector
init_probs <- c(1, rep(0, n-2))
# Define MPH object
ph_rew_obj <- MPH(subintensity_matrix, init_probs, rew_mat)
```

$\Sigma/4$ can now be directly calculated using `var()`:

```r
var_covar_mat <- var(ph_rew_obj)
round(0.25*var_covar_mat, 4)

        [,1]    [,2]    [,3]    [,4]    [,5]    [,6]    [,7]
[1,]  0.3211 -0.0358 -0.0210 -0.0141 -0.0103 -0.0079  0.1384
[2,] -0.0358  0.2495 -0.0210 -0.0141 -0.0103  0.1328 -0.0356
[3,] -0.0210 -0.0210  0.2076 -0.0141  0.1283 -0.0346 -0.0267
[4,] -0.0141 -0.0141 -0.0141  0.3173 -0.0359 -0.0275 -0.0216
[5,] -0.0103 -0.0103  0.1283 -0.0359  0.1394 -0.0230 -0.0183
[6,] -0.0079  0.1328 -0.0346 -0.0275 -0.0230  0.1310 -0.0159
[7,]  0.1384 -0.0356 -0.0267 -0.0216 -0.0183 -0.0159  0.1224
```

Rivas-González et al. (2023). PhaseTypeR: an R package for phase-type distributions in population genetics. *Journal of Open Source Software*, 4
8(82), 5054. https://doi.org/10.21105/joss.05054.

This yields the same variance-covariance matrix as in Theorem 2.2 in Durrett ([2008]) without the need for analytical derivations.

Example 2: the coalescent with recombination

The traditional procedure for deriving the correlation between the branch lengths in two loci for a sample of size two is by a first-step analysis (e.g., section 7 in Wakeley, 2009). In this section, we exemplify the easy use of PhaseTypeR to obtain the same result.

The state space and transition rates for the two-locus ancestral recombination graph is shown in Figure 1. The filled circles represent material ancestral to the sample, and the crosses indicate that the most recent common ancestor has been found. The lines between the circles or crosses indicate if the ancestral material is present on the same chromosome. The starting state is state 1 at present day with two samples from the same chromosome.

The time τ when both loci have found their common ancestor is $\text{PH}(e_1, S)$ distributed with $e_1 = (1, 0, \ldots, 0)$ and

$$
S = \begin{pmatrix}
-(1 + 2\rho/2) & 2\rho/2 & 0 & 0 & 0 \\
1 & -(3 + \rho/2) & \rho/2 & 1 & 1 \\
0 & 4 & -6 & 1 & 1 \\
0 & 0 & 0 & -1 & 0 \\
0 & 0 & 0 & 0 & -1
\end{pmatrix},
$$

where ρ is the recombination rate.

The tree height T_{left} in the left locus is the first time the ancestral process $\{X(t) : t \geq 0\}$ enters state 4 or state 6 or, equivalently, the time spent in state 1, 2, 3 and 5 before absorption in state 6. We therefore have

$$
T_{\text{left}} = \min\{t \geq 0 : X(t) \in \{4, 6\}\} = \int_0^\tau r_{\text{left}}(X_t) dt
$$

with the reward vector $r_{\text{left}} = (1, 1, 1, 0, 1)$. Similarly, the tree height T_{right} in the right locus is the first time the ancestral process enters state 5 or state 6 or, equivalently, the time spent in state 1, 2, 3 and 4 before absorption in state 6. We therefore have

$$
T_{\text{right}} = \min\{t \geq 0 : X(t) \in \{5, 6\}\} = \int_0^\tau r_{\text{right}}(X_t) dt
$$

with the reward vector $r_{\text{right}} = (1, 1, 1, 1, 0)$. A classical result in population genetics gives the covariance between the two tree heights

$$
\text{Cov}(T_{\text{left}}, T_{\text{right}}) = \frac{\rho + 18}{\rho^2 + 13\rho + 18},
$$

and we note that for large recombination rates $\text{Cov}(T_{\text{left}}, T_{\text{right}})$ is close to zero, and for small recombination rates it is close to one. Also note that T_{left} and T_{right} are both exponentially distributed with a rate of 1, so $\text{Var}(T_{\text{left}}) = \text{Var}(T_{\text{right}}) = 1$, and, consequently, $\text{Cor}(T_{\text{left}}, T_{\text{right}}) = \text{Cov}(T_{\text{left}}, T_{\text{right}})$ (see also equation 3.10 in Wakeley, 2009). Moreover, as shown by a simple proof in Wilton et al. ([2015]), we have that $P(T_{\text{left}} = T_{\text{right}}) = \text{Cov}(T_{\text{left}}, T_{\text{right}})$.
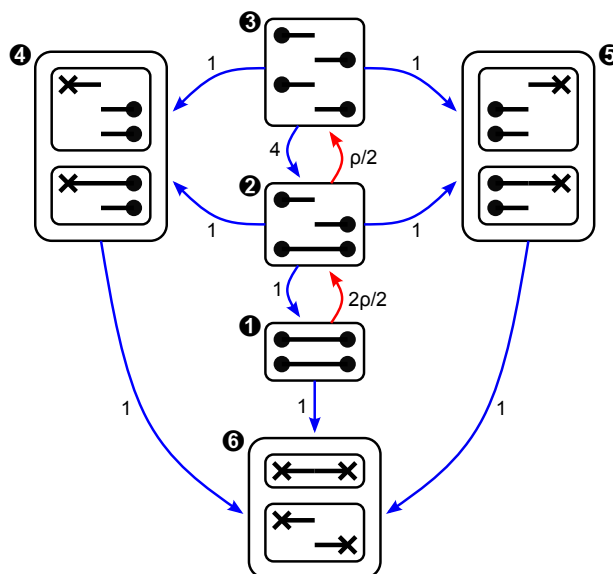
Figure 1: Two-locus ancestral recombination graph. Filled circles represent uncoalesced sites, while crosses represent coalesced sites. ρ is the recombination rate.

An implementation using PhaseTypeR simply consists of specifying the initial distribution, rate matrix for the ancestral process, rewards for the two tree heights, and calling the variance function `var()` for the multivariate phase-type distribution.

```r
recomb_rate <- 0.3
ARG_subint_mat <- function(recomb_rate) {
  matrix(
    c(-(1+2*recomb_rate/2),  2*recomb_rate/2,   0,            0, 0,
       1,                    -(3+recomb_rate/2), recomb_rate/2, 1, 1,
       0,                     4,                 -6,            1, 1,
       0,                     0,                  0,           -1, 0,
       0,                     0,                  0,            0, -1),
    nrow=5, byrow=TRUE)
}
subintensity_matrix <- ARG_subint_mat(recomb_rate)
initial_probabilities <- c(1, 0, 0, 0, 0)
# T_left: T_MRCA in left locus
reward_left <- c(1, 1, 1, 0, 1)
# T_right: T_MRCA in right locus
reward_right <- c(1, 1, 1, 1, 0)
# Joint distribution T_joint of T_left and T_right
T_joint <- MPH(subintensity_matrix,
               initial_probabilities,
               matrix(c(reward_left, reward_right), nrow = 5))
var(T_joint)[1, 2]
```

```
[1] 0.8321965
```

We can see that the phase-type result is equal to the classical formula provided above when $\rho = 0.3$.

From this multivariate phase-type representation of the ancestral recombination graph (ARG), we can simulate, for example, 1,000 samples from the joint distribution of $(T_{\text{left}}, T_{\text{right}})$ using `rMPH(1000, T_joint)` in PhaseTypeR. If the recombination rate ρ is set to a small value, then

most of the samples will result in $T_{\text{left}} = T_{\text{right}}$, and the joint density will concentrate along the diagonal, as shown in Figure 2, left (Simonsen & Churchill, 1997). If instead ρ is large, then most of the samples will result in $T_{\text{left}} \neq T_{\text{right}}$ (Figure 2, right).

```
# Simulation from the joint distribution
subintensity_matrix_09 <- ARG_subint_mat(0.166)
Tab_09 <- MPH(subintensity_matrix_09, initial_probabilities,
              matrix(c(reward_left, reward_right), nrow=5))
subintensity_matrix_01 <- ARG_subint_mat(11.316)
Tab_01 <- MPH(subintensity_matrix_01, initial_probabilities,
              matrix(c(reward_left, reward_right), nrow=5))
set.seed(3)
rTab_09 <- rMPH(1000, Tab_09)
rTab_01 <- rMPH(1000, Tab_01)
```
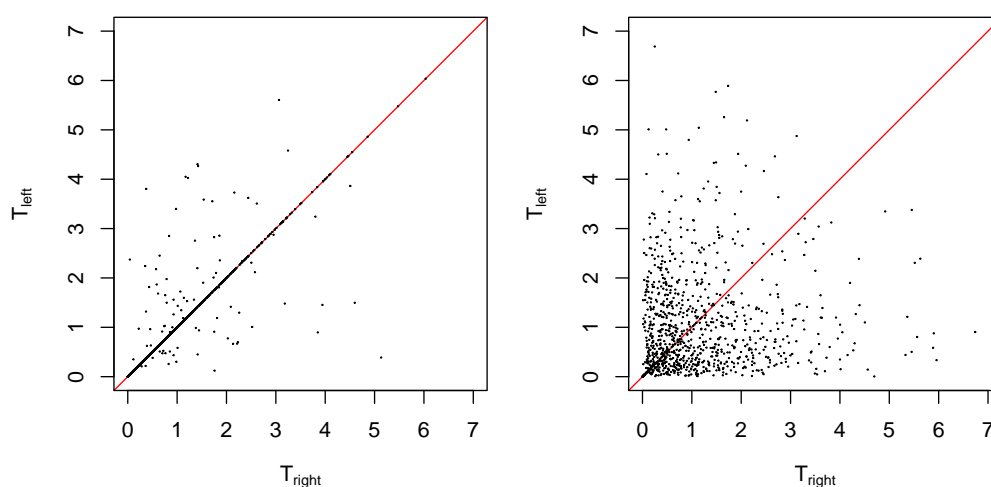

Figure 2: Random samples from the two-locus ancestral recombination graph. Left: recombination rate $\rho = 0.166$ and $P(T_{\text{left}} = T_{\text{right}}) = 0.9$. Right: recombination rate $\rho = 11.316$ and $P(T_{\text{left}} = T_{\text{right}}) = 0.1$.

Conclusion

We have described PhaseTypeR, an easy-to-use package for the analysis of time-homogeneous evolutionary models in population genetics. We have included two examples: (1) the mean and variance for the SFS of the n-coalescent with mutation, and (2) the correlation for the tree height in the two-locus coalescent with recombination. The multiple merger coalescent (Birkner & Blath, 2021), the two-island model (Legried & Terhorst, 2022) and the seed bank coalescent (Casanova et al., 2022) constitute other coalescent models where phase-type analyses have been useful. We hope that population geneticists will take advantage of PhaseTypeR in future analyses of coalescent models.

Acknowledgements

References

Albrecher, H., & Bladt, M. (2019). Inhomogeneous phase-type distributions and heavy tails. *Journal of Applied Probability*, *56*(4), 1044–1064. https://doi.org/10.1017/jpr.2019.60

Albrecher, H., Bladt, M., & Yslas, J. (2022). Fitting inhomogeneous phase-type distributions to data: The univariate and the multivariate case. *Scandinavian Journal of Statistics*, *49*(1), 44–77. https://doi.org/10.1111/sjos.12505

Birkner, M., & Blath, J. (2021). Genealogies and inference for populations with highly skewed offspring distributions. In E. Baake & A. Wakolbinger (Eds.), *Probabilistic structures in evolution* (pp. 151--178). EMS Press. https://doi.org/10.4171/ECR/17-1/8

Bladt, M., & Nielsen, B. F. (2017). *Matrix-exponential distributions in applied probability* (Vol. 81). Springer-Verlag. https://doi.org/10.1007/978-1-4939-7049-0

Campillo Navarro, A. (2018). *Order statistics and multivariate discrete phase-type distributions*. [PhD thesis]. Technical University of Denmark (Copenhagen, Denmark). Department of Applied Mathematics; Computer Science. DTU Compute.

Casanova, A. G., Peñaloza, L., & Siri-Jégousse, A. (2022). The shape of a seed bank tree. *Journal of Applied Probability*, *59*(3), 631–651. https://doi.org/10.1017/jpr.2021.79

Durrett, R. (2008). *Probability models for DNA sequence evolution* (Vol. 2). Springer. https://doi.org/10.1007/978-0-387-78168-6

Dutang, C., Goulet, V., & Pigeon, M. (2008). Actuar: An r package for actuarial science. *Journal of Statistical Software*, *25*, 1–37. https://doi.org/10.18637/jss.v025.i07

Fu, Y.-X. (1995). Statistical properties of segregating sites. *Theoretical Population Biology*, *48*(2), 172–197. https://doi.org/10.1006/tpbi.1995.1025

Hobolth, A., Bladt, M., & Andersen, L. N. (2021). Multivariate phase-type theory for the site frequency spectrum. *Journal of Mathematical Biology*, *83*(6), 1–28. https://doi.org/10.1007/s00285-021-01689-w

Hobolth, A., Siri-Jegousse, A., & Bladt, M. (2019). Phase-type distributions in population genetics. *Theoretical Population Biology*, *127*, 16–32. https://doi.org/10.1016/j.tpb.2019.02.001

Legried, B., & Terhorst, J. (2022). Rates of convergence in the two-island and isolation-with-migration models. *Theoretical Population Biology*, *147*, 16–27. https://doi.org/10.1016/j.tpb.2022.08.001

Okamura, Hiroyuki. (2015). *Mapfit: A tool for PH/MAP parameter estimation*. https://CRAN.R-project.org/package=mapfit

Okamura, Hiroyuki, & Dohi, T. (2015). Mapfit: An r-based tool for PH/MAP parameter estimation. In J. Campos & B. R. Haverkort (Eds.), *Quantitative evaluation of systems* (pp. 105–112). Springer International Publishing. https://doi.org/10.1007/978-3-319-22264-6_7

Okamura, H., & Dohi, T. (2016). PH fitting algorithm and its application to reliability engineering. *Journal of the Operations Research Society of Japan*, *59*(1), 72–109. https://doi.org/10.15807/jorsj.59.72

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Simonsen, K. L., & Churchill, G. A. (1997). A markov chain model of coalescence with recombination. *Theoretical Population Biology*, *52*(1), 43–59. https://doi.org/10.1006/tpbi.1997.1307

Wakeley, J. (2009). *Coalescent Theory: An Introduction*. W. H. Freeman. https://doi.org/10.1093/schbul/syp004

Wilton, P. R., Carmi, S., & Hobolth, A. (2015). The SMC' is a highly accurate approximation to the ancestral recombination graph. *Genetics*, *200*(1), 343–355. https://doi.org/10.1534/genetics.114.173898