





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.

DOI: [10.xxxxxx/draft](https://doi.org/10.xxxxxx/draft)

Software

- [Review](#) 
- [Repository](#) 
- [Archive](#) 

Editor: [Open Journals](#) 

Reviewers:

- [@openjournals](#)

Submitted: 01 January 1970

Published: unpublished

License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC BY 4.0](#)).

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, they do not include reward transformations, and they are tailored to actuarial sciences and queuing theory.

PhaseTypeR is particularly well suited for population genetics, and much emphasis in our software is on natural and easy-to-use R functions. 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

Table 1 provides an overview of the PhaseTypeR functions for a univariate continuous phase-type distribution $\tau \sim \text{PH}(\alpha, T)$, 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 $r(i), i = 1, \dots, p$, 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	$\tau \sim \text{PH}(a, T)$	<code>PH(T, a)</code>
Mean	$E[\tau] = a(-T)^{-1}e$	<code>mean(PH)</code>
Variance	$V[\tau] = E[\tau^2] - E[\tau]^2$	<code>var(PH)</code>
Density	$f(x) = a \exp(Tx)(-Te),$ $x \geq 0$	<code>dPH(x, PH)</code>
Cumulative distribution	$F(x) = 1 - a \exp(Tx)e,$ $x \geq 0$	<code>pPH(x, PH)</code>
Quantile function		<code>qPH(p, PH)</code>
Random sampling		<code>rPH(n, PH)</code>
Random sampling of full path		<code>rFullPH(n, PH)</code>
Reward transformation	$Y = \int_0^\tau r(X(t))dt$	<code>reward_phase_type(PH, r)</code>

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, \dots, Y_k) where $Y_j = \int_0^\tau r_j(X(t))dt$ for $j = 1, \dots, k$. 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	$\tau \sim \text{DPH}(a, T)$	<code>DPH(T, a)</code>
Mean	$E[\tau] = \pi(I - T)^{-1}e$	<code>mean(DPH)</code>
Variance	$V[\tau] = E[\tau^2] - E[\tau]^2$	<code>var(DPH)</code>
Density	$f(x) = \pi T^{x-1}t, x \geq 1$	<code>dDPH(x, DPH)</code>
Cumulative distribution	$F(x) = 1 - \pi T^x e, x \geq 1$	<code>pDPH(x, DPH)</code>
Quantile function		<code>qDPH(p, DPH)</code>
Random sampling.		<code>rDPH(n, DPH)</code>
Random sampling of full path		<code>rFullDPH(n, DPH)</code>
Reward transformation	$Y = \sum_{m=0}^{\tau-1} r(X_m)$	<code>reward_phase_type(DPH, r)</code>

Table 3: PhaseTypeR functions for multivariate continuous and multivariate discrete phase-type distributions. \mathbf{a} is the vector of initial probabilities, \mathbf{T} is the sub-intensity matrix and \mathbf{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(\mathbf{T} , \mathbf{a} , \mathbf{R})	MDPH(\mathbf{T} , \mathbf{a} , \mathbf{R})
Mean	mean(MPH)	mean(MDPH)
(Co)variance	var(MPH)	var(MDPH)
Density	dMPH(\mathbf{x} , MPH)	dMDPH(\mathbf{x} , MDPH)
Cumulative distribution	pMPH(\mathbf{x} , MPH)	pMDPH(\mathbf{x} , MDPH)
Quantile function	qMPH(\mathbf{p} , MPH)	qMDPH(\mathbf{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 $L = (L_1, \dots, L_{n-1})$ 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 $e_1 = (1, 0, \dots, 0)$. Conditionally on L , the ξ_i 's are independent and Poisson distributed, $\xi_i | L_i \sim \text{Poisson}(L_i \frac{\theta}{2})$, where θ is the underlying mutation rate (Wakeley, 2009). By the law of total variance, we have

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

where $\text{diag}(\cdot)$ is the diagonal matrix whose entries are given by $E[L]$. It is well-known that $E[L_i] = 1/i$, 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:

```
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) {
```

```

      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)
  }

```

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

```

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)

```

76 $\Sigma/4$ can now be directly calculated using var():

```

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

```

```

77      [,1] [,2] [,3] [,4] [,5] [,6] [,7]
78 [1,] 0.3211 -0.0358 -0.0210 -0.0141 -0.0103 -0.0079 0.1384
79 [2,] -0.0358 0.2495 -0.0210 -0.0141 -0.0103 0.1328 -0.0356
80 [3,] -0.0210 -0.0210 0.2076 -0.0141 0.1283 -0.0346 -0.0267
81 [4,] -0.0141 -0.0141 -0.0141 0.3173 -0.0359 -0.0275 -0.0216
82 [5,] -0.0103 -0.0103 0.1283 -0.0359 0.1394 -0.0230 -0.0183
83 [6,] -0.0079 0.1328 -0.0346 -0.0275 -0.0230 0.1310 -0.0159
84 [7,] 0.1384 -0.0356 -0.0267 -0.0216 -0.0183 -0.0159 0.1224

```

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

87 Example 2: the coalescent with recombination

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

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

96 The time τ when both loci have found their common ancestor is $\text{PH}(e_1, S)$ distributed with
97 $e_1 = (1, 0, \dots, 0)$ and

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

98 where ρ is the recombination rate.

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

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

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

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

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

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

107 and we note that for large recombination rates $\text{Cov}(T_{\text{left}}, T_{\text{right}})$ is close to zero, and for small
108 recombination rates it is close to one. Also note that T_{left} and T_{right} are both exponentially dis-
109 tributed with a rate of 1, so $\text{Var}(T_{\text{left}}) = \text{Var}(T_{\text{right}}) = 1$, and, consequently, $\text{Cor}(T_{\text{left}}, T_{\text{right}}) =$
110 $\text{Cov}(T_{\text{left}}, T_{\text{right}})$ (see also equation 3.10 in Wakeley, 2009). Moreover, as shown by a simple
111 proof in Wilton et al. (2015), we have that $P(T_{\text{left}} = T_{\text{right}}) = \text{Cov}(T_{\text{left}}, T_{\text{right}})$.

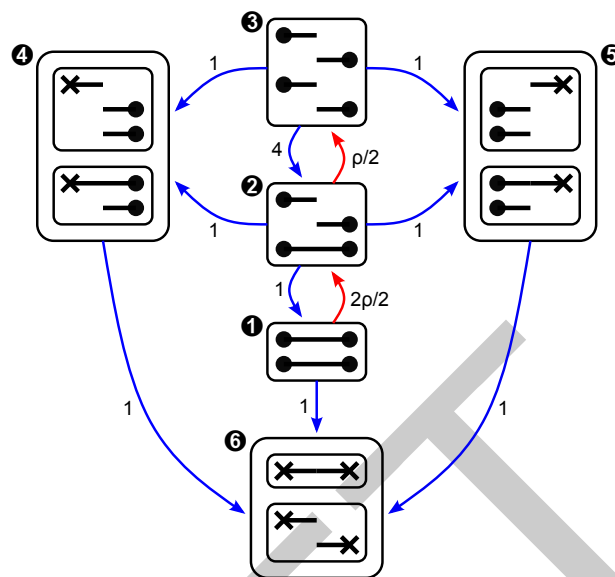


Figure 1: Two-locus ancestral recombination graph. Filled circles represent uncoalesced sites, while crosses represent coalesced sites. ρ is the recombination rate.

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

```
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]
```

115 [1] 0.8321965

116 We can see that the phase-type result is equal to the classical formula provided above when
117 $\rho = 0.3$.

118 From this multivariate phase-type representation of the ancestral recombination graph (ARG),
119 we can simulate, for example, 1,000 samples from the joint distribution of $(T_{\text{left}}, T_{\text{right}})$ using
120 `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)
```

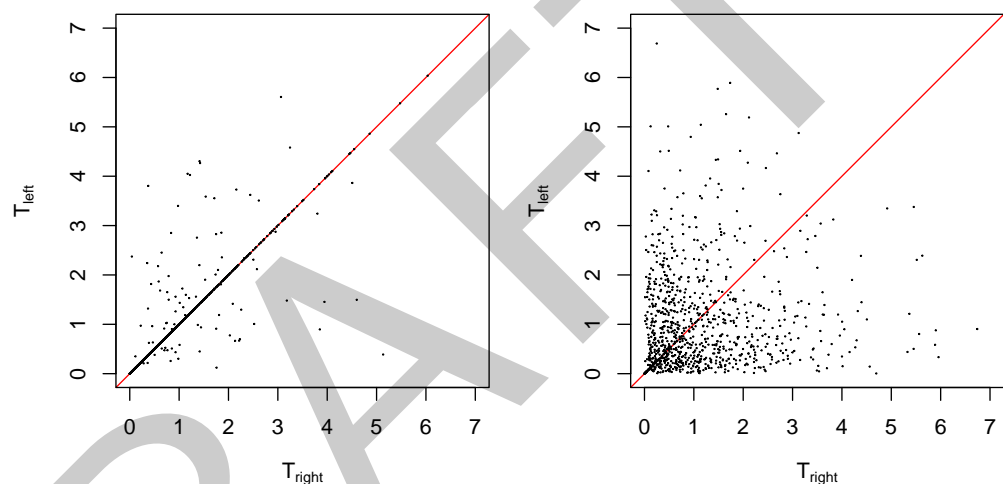


Figure 2: Random samples from the two-locus ancestral recombination graph. Left: recombination rate $\rho = 0.166$ and $P(T_{\text{left}} = T_{\text{right}}) = 0.9$. Right: recombination rate $\rho = 11.316$ and $P(T_{\text{left}} = T_{\text{right}}) = 0.1$.

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.

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*,

- 138 49(1), 44–77.
- 139 Birkner, M., & Blath, J. (2021). Genealogies and inference for populations with highly skewed
140 offspring distributions. In E. Baake & A. Wakolbinger (Eds.), *Probabilistic structures in
141 evolution* (pp. 151–178). EMS Press.
- 142 Bladt, M., & Nielsen, B. F. (2017). *Matrix-exponential distributions in applied probability*
143 (Vol. 81). Springer-Verlag. ISBN: 978-1-4939-8377-3
- 144 Campillo Navarro, A. (2018). *Order statistics and multivariate discrete phase-type distributions*.
145 [PhD thesis]. Technical University of Denmark (Copenhagen, Denmark). Department of
146 Applied Mathematics; Computer Science. DTU Compute.
- 147 Casanova, A. G., Peñaloza, L., & Siri-Jégousse, A. (2022). The shape of a seed bank tree.
148 *Journal of Applied Probability*, 59(3), 631–651.
- 149 Durrett, R. (2008). *Probability models for DNA sequence evolution* (Vol. 2). Springer.
- 150 Dutang, C., Goulet, V., & Pigeon, M. (2008). Actuar: An r package for actuarial science.
151 *Journal of Statistical Software*, 25, 1–37. <https://doi.org/10.18637/jss.v025.i07>
- 152 Fu, Y.-X. (1995). Statistical properties of segregating sites. *Theoretical Population Biology*,
153 48(2), 172–197.
- 154 Hobolth, A., Bladt, M., & Andersen, L. N. (2021). Multivariate phase-type theory for the site
155 frequency spectrum. *Journal of Mathematical Biology*, 83(6), 1–28. <https://doi.org/https://doi.org/10.1007/s00285-021-01689-w>
- 156
157 Hobolth, A., Siri-Jégousse, A., & Bladt, M. (2019). Phase-type distributions in population
158 genetics. *Theoretical Population Biology*, 127, 16–32. <https://doi.org/https://doi.org/10.1016/j.tpb.2019.02.001>
- 159
160 Legried, B., & Terhorst, J. (2022). Rates of convergence in the two-island and isolation-with-
161 migration models. *Theoretical Population Biology*, 147, 16–27.
- 162 Okamura, Hiroyuki. (2015). *Mapfit: A tool for PH/MAP parameter estimation*. <https://CRAN.R-project.org/package=mapfit>
- 163
164 Okamura, Hiroyuki, & Dohi, T. (2015). Mapfit: An r-based tool for PH/MAP parameter
165 estimation. In J. Campos & B. R. Haverkort (Eds.), *Quantitative evaluation of systems*
166 (pp. 105–112). Springer International Publishing. ISBN: 978-3-319-22263-9
- 167 Okamura, H., & Dohi, T. (2016). PH fitting algorithm and its application to reliabil-
168 ity engineering. *Journal of the Operations Research Society of Japan*, 59(1), 72–109.
169 <https://doi.org/https://doi.org/10.15807/jorsj.59.72>
- 170 R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation
171 for Statistical Computing. <https://www.R-project.org/>
- 172 Simonsen, K. L., & Churchill, G. A. (1997). A markov chain model of coalescence with
173 recombination. *Theoretical Population Biology*, 52(1), 43–59. <https://doi.org/https://doi.org/10.1006/tpbi.1997.1307>
- 174
175 Wakeley, J. (2009). *Coalescent Theory: An Introduction*. W. H. Freeman. ISBN: 078-0-
176 9747077-5-4
- 177 Wilton, P. R., Carmi, S., & Hobolth, A. (2015). The SMC' is a highly accurate approximation
178 to the ancestral recombination graph. *Genetics*, 200(1), 343–355. <https://doi.org/https://doi.org/10.1534/genetics.114.173898>
- 179