**INTRODUCTION**

Comparing phylogenetic trees provides insights into biological processes at various biological scales. These comparisons are common between the phylogenies of host and symbiont lineages or gene trees and species trees. Because of their shared histories, trees derived from associated evolutionary processes are expected to show matching patterns of diversification, yet this is often not the case. Thus, understanding the underlying causes of incongruence in phylogenies that have ecological and evolutionary links is a rich area of research (e.g. Balbuena et al., 2020; de Vienne et al., 2013; Page & Charleston, 1998; Szöllösi et al., 2013). While there are several methods for untangling cophylogenetic patterns, tools for generating simulated datasets to test the performance of tree-comparison methods are not readily available. To address this gap, we have developed treeducken, an R package for simulating datasets under host–symbiont or gene-tree–species-tree scenarios. This simulation package can be useful for quantifying the patterns generated under different simulation conditions and for evaluating the accuracy of methods used to understand codiversifying lineages.

Cophylogenetic methods are used to explain the incongruence between host and symbiont phylogenies. These methods are often divided into two categories: global-fit and event-based. Global-fit methods attempt to measure the degree of dependency between host and symbiont trees (Balbuena et al., 2020). Event-based methods use events such as host-switching or cospeciation (i.e. a host lineage and symbiont lineage speciating simultaneously) to explain the degree of congruence between host and symbiont trees (Charleston, 1998; Conow et al., 2010; Santichaivekin et al., 2020). More recently, statistical models have been used to examine host and symbiont data (Alcala et al., 2017; Baudet et al., 2015; Braga et al., 2020; Satler et al., 2019). Models for understanding gene-tree and species-tree discordance have
received much attention as genomic data have become widely available. The processes of gene duplication, gene loss, lateral gene transfer (LGT) and incomplete lineage sorting (ILS) contribute significantly to the discordance between gene and species trees (Maddison, 1997). Historically, gene duplication, gene loss, LGT, and ILS have been analyzed in host–symbiont analyses to the classical cophylogenetic events of symbiont duplication, symbiont loss, host-switching and missing-the-boat (Page & Charleston, 1998). More recently, models describing these discordance-causing processes (Heled & Drummond, 2009; Rannala & Yang, 2003; Rasmussen & Kellis, 2012) can provide more robust species-tree estimates (Kubatko et al., 2009) and a deeper understanding of the nature of genomic evolution (Szöllösi et al., 2013).

Simulation tools are essential for validating the performance of methods for comparing phylogenetic trees. There are currently few existing cophylogenetic simulators available for generating host-symbiont datasets. The simulation program, CoRE-PA (Keller-Schmidt et al., 2011), allows two types of events in host trees (cospeciation and host speciation with sorting of symbionts), and two types of events in symbiont trees (host-switching and symbiont speciation). Baudet et al. (2015) and Alcala et al. (2017) each introduced simulation methods that generate symbiont phylogenies along user-defined host trees under forward-time birth–death processes, though, neither approach simultaneously simulates the evolution of both the host and symbiont. There are a number of tools for generating discordant species trees and gene trees, including DLCoalSim (Rasmussen & Kellis, 2012) and SimPhy (Mallo et al., 2016). As noted above, these distinct biological processes of gene-tree/species-tree evolution and host–symbiont evolution produce similar patterns that warrant comparison. However, no existing tool can simulate both gene trees, species trees and cophylogenetic data while accounting for extinction.

Here we present an R package, treeducken, for simulating both cophylogenetic and gene-tree/species-tree data. treeducken currently generates data under two scenarios: the cophylogenetic birth–death model for simulating host–symbiont coevolution and a hierarchical model (Mallo et al., 2016; Rasmussen & Kellis, 2012) for simulating gene-tree/species-tree coevolution. The cophylogenetic birth–death model presented here extends the work of Keller-Schmidt et al. (2011) by allowing for extinction and host-shift or host-switch speciation and the tandem simulation of host and symbiont phylogenies. The hierarchical gene–family model used by treeducken simulates trees under processes implicated in gene-tree/species-tree discordance. The modular simulation framework provides a user-friendly and reproducible workflow allowing interoperability with existing R packages.

2 | MODEL

treeducken simulates under two distinct types of processes: host-symbiont evolution and gene–species evolution. Here, we present the cophylogenetic birth–death model for simulating the tandem evolution of interacting hosts and symbionts (Figure 1). To simulate gene–species coevolution, we use a hierarchical three-tree model consisting of three levels: species, locus and gene (Figure 2). These three levels model different causes of gene-tree and species-tree discordance including ILS, LGT, and gene duplication and loss (Rasmussen & Kellis, 2012).

2.1 | Cophylogenetic birth–death model

The cophylogenetic birth–death model simulates a pair of phylogenies, the host phylogeny and the symbiont phylogeny, and their ecological interactions or associations. The resulting two trees and extant associations are intended to mimic the data used in cophylogenetic analyses. Ecological interactions are represented by a presence–absence matrix, hereafter the association matrix, with rows representing hosts and columns representing symbionts. The association matrix determines which hosts and symbionts can evolve together; for example, the model does not allow cospeciation in unassociated hosts and symbionts. The symbiont and host lineages can undergo speciation and extinction independently, and they can speciate together (cospeciation) or go extinct together (co-extinction; Figure 1). The user can also set symbiont lineages’ maximum number of host associations at any given time. For example, with a host limit of three, a symbiont is only able to be associated with, at the most, three hosts.

The symbiont speciation and extinction rates determine the independent evolutionary history of the symbiont taxa (Figure 1a,b). Following a symbiont speciation event, both descendant lineages inherit their ancestral associations. The symbiont speciation event described here corresponds to the cophylogenetic event of duplication. The model also includes host-expansion speciation, a special case of symbiont speciation where one of the descendant lineages, chosen at random, gains a novel association in addition to the symbiont’s ancestral host repertoire (Figure 1c). The host-expansion event described here is similar to the spreading event described in Brooks et al. (1991). Users can use the hs_mode = TRUE argument to change the host-expansion rate to the host-switching rate. Here we define host-switching as a symbiont speciation where one descendant gains a randomly chosen novel association and the other descendant inherits the ancestral host repertoire. The symbiont tree has three parameters: symbiont speciation rate \( \lambda_s \), symbiont extinction rate \( \mu_s \), and host-expansion rate \( \chi \).

Cospeciation occurs when one host lineage bifurcates and a simultaneous speciation event occurs on one of the host’s (randomly selected) symbiont lineages (Figure 1d). Following cospeciation, each descendant host lineage is then associated with one of the new descendant symbiont lineages. The remainder of the associations of the ancestral lineages of host or symbiont are sorted at random among the host’s or symbiont’s descendants. The cospeciation rate, \( \lambda_C \), is a shared parameter between the host and the symbiont tree.

Host speciation—controlled by rate \( \lambda_H \)—describes events where the host speciates independently from its symbiont, that is, no cospeciation occurs (Figure 1e). This is equivalent to the failure-to-diverge event used in many event-based methods.
Following a host speciation event, ancestral associations are sorted randomly on either or both descendant lineages. Host extinction refers to events where the host goes extinct (Figure 1f). If there are symbiont lineages left without a host following a symbiont extinction event, then coextinction occurs. Host extinction occurs at rate $\mu_H$, is not independent of the symbiont phylogeny because host extinction can cause symbiont extinction. Coextinction does not occur if a host is left with no associations following a symbiont extinction event. For host-parasite systems this is adequate, but for obligate mutualisms, this may not be biologically realistic if hosts are unable to persist without their symbionts.

The generating process is a birth–death model with parameters $\lambda_S$, $\mu_S$, $\lambda_C$, $\mu_C$, $\lambda_H$ and $\mu_H$ conditioned on time. This forward-time simulation terminates after a pre-specified amount of time, resulting in two phylogenies and an association matrix. The cophylogenetic birth–death process simulation is performed using a single R function `sim_cophyloBD` that takes as input all six parameters of the model, the pre-specified amount of time and the number of cophylogenetic datasets to simulate. The function outputs a list containing a host tree, a symbiont tree, the extant association matrix with hosts in rows and symbionts in columns, and a data frame containing all events that have occurred during the simulation.

### 2.2 | Three-tree model

We implemented a hierarchical model, the three-tree model, to allow simulation from the species level down to coalescent sites (Mallo et al., 2016; Rasmussen & Kellis, 2012). The three-tree model has three levels: the species tree, the locus tree and the gene tree (Rasmussen & Kellis, 2012). At the highest level is the
species tree describing the history of speciation and extinction (Figure 2a). The next level down is the locus tree, evolving within the species tree, which models gene duplications and losses at a locus within a genome (Figure 2b). Within the locus tree, the gene-tree models the dynamics of the multi-locus coalescent process allowing for ILS (Figure 2c). treeducken extends the three-tree model of Rasmussen and Kellis (2012) to simulate gene duplication, gene loss, ILS and LGT along all lineages including those bound for extinction.

The three-tree model in treeducken first simulates the species tree (Figure 3a) under the birth–death process (Gernhard, 2008; Kendall, 1948) using the general sampling algorithm (GSA) given in Hartmann et al. (2010) for forward simulating a tree to a set number of tips with set birth and death rates. treeducken is also able to simulate species trees to a pre-specified time using the simple sampling algorithm (SSA; Stadler, 2011). For the species tree, users can use host and symbiont trees simulated from the cophylogenetic simulator, or use one of the species-tree simulation functions sim_stBD for the GSA simulation or sim_stBD_t for the SSA simulation.

Next, a locus tree is simulated within the full species tree—that is, the species tree containing records of extinct lineages (Figure 3b). This locus is simulated over the time spanned by the complete species tree under a birth–death process coupled with LGT. Within treeducken, the species tree is used as input along with a gene birth rate, gene death rate, LGT rate and a number of loci to the sim_ltBD function to simulate a set of locus trees under a birth–death process.

Finally, gene trees are simulated backwards in time along the locus tree using a multi-locus coalescent process (Figure 2c; Rasmussen & Kellis, 2012). Gene trees can also be simulated along species trees using the multispecies coalescent process (Rannala & Yang, 2003). In treeducken, locus trees or species trees are used as input for the multi-locus sim_mlc for the multispecies coalescent sim_msc functions. Users are able to set the generation time, mutation rate and effective population size. Each of these gene trees corresponds to a single coalescent independent site within its containing locus.
3 | USAGE

3.1 | Description of the R package

treeducken is implemented as an R package (R Core Team, 2020) and is available on CRAN (https://cran.r-project.org/web/packages/treeducken/index.html). Validation testing was conducted in R and is available on Github (https://github.com/wadedismukes/treeduckenValidation). This package makes extensive use of the Rcpp and RcppArmadillo packages to wrap C++ code and improve performance (Eddelbuettel & Sanderson, 2014). In addition to the simulation functions described above, the package builds on previous R phylogenetics packages to provide functions for assistance in determining various simulation parameters, calculating summary statistics and plotting host–symbiont tree sets (Harmon et al., 2007; Revell, 2012). All phylogenies are output in the format of the APE package as full phylogenies containing extinct tips (Paradis & Schliep, 2018). An example cophylogenetic plot is shown in Figure 3. Integration with R allows for straightforward simulation of parameters from statistical distributions, and intuitive integration with other macroevolutionary and cophylogenetic tools (e.g. phytools, PACo; Hutchinson et al., 2017; Revell, 2012).

3.2 | Simulating cophylogenetic datasets

To generate cophylogenetic data, we first set the cophylogenetic birth–death process parameters: host speciation and extinction rate, symbiont speciation and extinction rate, host-expansion rate, and cospeciation rate and the pre-specified simulation time. These rates are set relative to time; for example, a speciation rate of 0.1 corresponds to on average 1 speciation every 10 time units.

```
library(treeducken)
set.seed(54)
lambda_H <- rexp(n=1)
mu_H <- 0.0
lambda_C <- rexp(n=1)
time <- 1.0

lambda_S <- rexp(n=1)
mu_S <- 0.0
lambda_total_H <- lambda_H + lambda_C
lambda_total_S <- lambda_S + lambda_C
H_tips <- ave_tips_st(lambda = lambda_total_H, mu = mu_H, t = time)
S_tips <- ave_tips_st(lambda = lambda_total_S, mu = mu_S, t = time)
cophy_obj <- sim_cophyloBD(hbr = lambda_H, hdr = mu_H, sbr = lambda_S, sdr = mu_S, cosp_rate =

library(paco)
host_dist <- cophenetic(host_tree(cophy_obj[[1]]))
symb_dist <- cophenetic(symb_tree(cophy_obj[[1]]))
links <- association_mat(cophy_obj[[1]])
D <- paco::prepare_paco_data(H = host_dist, P = symb_dist, HP = links)
D <- paco::add_pcoord(D)
D <- paco::PACo(D, nperm=100, seed = 11, method="r0")
```

Cophylogenetic objects are output as a cophy object with many generic functions implemented including summary, plot and print. These objects can be used with functions within treeducken to perform the ParaFit global fit test, or with existing packages for cophylogenetics (Legendre et al., 2002). We provide an example of using treeducken with the paco package (Hutchinson et al., 2017).

3.3 | Simulating under the three-tree model

In addition to the cophylogenetic simulations, treeducken can simulate using the three-tree model to simulate gene-tree and species-tree discordance. For instance, we can use the host tree from the cophylogenetic example above to simulate a locus tree with gene birth and death rates. Then we use the locus tree and set mutation rate, generation time and population size to simulate under the multi-locus coalescent. We can perform a similar simulation with the symbiont tree.

```
host_tree <- host_tree(cophy_obj[[1]])
host_locus_trees <- sim_ltBD(host_tree, gbr = 0.2, gdr = 0.1, lgtr = 0.0, num_loci = 10)
host_gene_trees <- sim_mlc(host_locus_trees[[1]], effective_pop_size = 10000, generation_time = 1e-6, num_reps = 100)

symb_tree <- symb_tree(cophy_obj[[1]])
symb_locus_trees <- sim_ltBD(symb_tree, gbr = 0.3, gdr = 0.1, lgtr = 0.0, num_loci = 10)
symb_gene_trees <- sim_mlc(symb_locus_trees[[1]], effective_pop_size = 50000, generation_time = 1e-5, num_reps = 100)
```

4 | CONCLUSION

treeducken adds to the phylogenetic analysis and simulation toolbox available in the R programming language (a list of those existing on the CRAN R project can be found here: https://cran.rproject.org/web/views/Phylogenetics.html). As a tree simulator, treeducken outputs phylogenetic trees in the format of the APE package, allowing interoperability with many other packages in the R ecosystem. This package fills a needed gap in the available phylogenetic tools in R by providing a straightforward means of simulating phylogenetic data under a variety of models. treeducken allows for a high degree of flexibility and complexity in simulations that make it straightforward to use for testing the performance of phylogenetic methods. We intend to improve this tool and increase its usability and functionality for empiricists and theoreticians using cophylogenetic methods.

ACKNOWLEDGEMENTS

Thanks to J. Justison for helpful comments on simulation of the cophylogenetic model, and J. Barido-Sottani, J. Satler, K. Quinteros and B. Petrucci for code review. Also thank you to D. Mallo for helpful comments on an earlier version of this software and manuscript. This work was supported by an NSF GRFP Award to W.D. and NSF grant DEB-1556853 (Awarded to J.D. Nason, T.A. Heath and E.A. Herre).

AUTHORS’ CONTRIBUTIONS

W.D. and T.A.H. designed the simulator and models; W.D. implemented the methods; W.D. and T.A.H. wrote the manuscript.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1111/2041-210X.13625.

DATA AVAILABILITY STATEMENT

treeducken GPL v3 open-source license held by Wade T. Dismukes and Tracy A. Heath @ Iowa State University, 2020. treeducken is tested on previous, current and future versions of R. The gopher and lice data used, source code and documentation are freely available on CRAN (http://cran.r-project.org/web/packages/treeducken) and Github (http://github.com/wadedismukes/treeducken). The software is also archived via Zenodo (https://zenodo.org/record/4710662#.Y1GlhpBKIuk; Dismukes & Justison, 2021). Validation testing code is also available on Github (http://github.com/wadedismukes/treeduckenValidation).

ORCID

Wade Dismukes https://orcid.org/0000-0002-6662-0427

REFERENCES

Acala, N., Jenkins, T., Christie, P., & Vuilleumier, S. (2017). Host shift and cospeciation rate estimation from co-phylogenies. Ecology Letters, 20(8), 1014–1024.

Balbuena, J.A., Pérez-Escobar, Ó.A., Llopis-Belenguer, C., & Llopis-Belenguer, I. (2020). Random tanglegram partitions (Random TaPas): An Alexandrian approach to the cophylogenetic Gordian knot. Systematic Biology, 69(6), 1212–1230. https://doi.org/10.1093/sysbio/sya033

Baudet, C., Donati, B., Sinaimeri, B., Crescenzi, P., Gautier, C., Matias, C., & Sagot, M.-F. (2015). Cophylogeny reconstruction via an approximate Bayesian computation. Systematic Biology, 64(3), 416–431.

Braga, M.P., Landis, M.J., Nylin, S., Janz, N., & Ronquist, F. (2020). Bayesian inference of ancestral host–parasite interactions under a phylogenetic model of host repertoire evolution. Systematic Biology, 69(6), 1149–1162. https://doi.org/10.1093/sysbio/sya019

Brooks, D.R., McLennan, D.A., & McLennan, D.A. (1991). Phylogeny, Ecology, and Behavior: A Research Program in Comparative Biology. University of Chicago press.

Charleston, M. (1998). Jungles: A new solution to the host/parasite cospeciation rate estimation from cophylogenies. Methods in Ecology and Evolution, 25(3), 769–778.

Harmon, L.J., Weir, J.T., Brock, C.D., Glor, R.E., & Challenger, W. (2007). GEIGER: Investigating evolutionary radiations. Bioinformatics, 24(1), 129–131.

Hartmann, K., Wong, D., & Stadler, T. (2010). Sampling trees from evolutionary models. Systematic Biology, 59(4), 465–476.

Heled, J., & Drummond, A.J. (2009). Bayesian inference of species trees from multilocus data. Molecular Biology and Evolution, 27(3), 570–580.

Hutchinson, M.C., Cagua, E.F., Balbuena, J.A., Stouffer, D.B., & Poisot, T. (2017). paco: Implementing Procrustean approach to cophylogeny in R. Methods in Ecology and Evolution, 8(8), 932–940.

Keller-Schmidt, S., Wieseke, N., Klemm, K., & Middendorf, M. (2011). Evaluation of host parasite reconciliation methods using a new approach for cophylogeny generation (pp. 11–13). University of Leipzig. Technical Report.

Kendall, D.G. (1948). On the generalized ‘birth-and-death’ process. Annals of Mathematical Statistics, 19, 1–15.

Kubatko, L.S., Carstens, B.C., & Knowles, L.L. (2009). STEM: Species tree estimation using maximum likelihood for gene trees under coalescence. Bioinformatics, 25(7), 971–973.

Legendre, P., Dessevives, Y., & Bazin, E. (2002). A statistical test for host–parasite coevolution. Systematic Biology, 51(2), 217–234.

Maddison, W.P. (1997). Gene trees in species trees. Systematic Biology, 46, 523–536.

Mallo, D., De Oliveira Martins, L., & Posada, D. (2016). SimPhy: Phylogenomic simulation of gene, locus, and species trees. Systematic Biology, 65(2), 334–344.

Page, R.D.M., & Charleston, M.A. (1998). Trees within trees: Phylogenetic and historical associations. Trends in Ecology & Evolution, 13, 356–359.

Paradis, E., & Schliep, K. (2018). ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics, 35(3), 526–528.

R Core Team. (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing.
Rannala, B., & Yang, Z. (2003). Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics, 164(4), 1645–1656.

Rasmussen, M.D., & Kellis, M. (2012). Unified modeling of gene duplication, loss, and coalescence using a locus tree. Genome Research, 22(4), 755–765.

Revell, L.J. (2012). phytools: An R package for phylogenetic comparative biology (and other things). Methods in Ecology and Evolution, 3(2), 217–223.

Santichaivekin, S., Yang, Q., Liu, J., Mawhorter, R., Jiang, J., Wesley, T., Wu, Y.-C., & Libeskind-Hadas, R. (2020). eMPRes: A systematic co-phylogeny reconciliation tool. Bioinformatics, in press.

Satler, J.D., Herre, E.A., Jandér, K.C., Eaton, D.A., Machado, C.A., Heath, T.A., & Nason, J.D. (2019). Inferring processes of coevolutionary diversification in a community of Panamanian strangler figs and associated pollinating wasps. Evolution, 73(11), 2295–2311.

Stadler, T. (2011). Simulating trees with a fixed number of extant species. Systematic Biology, 60(5), 676–684.

Szöllosi, G.J., Tannier, E., Lartillot, N., & Daubin, V. (2013). Lateral gene transfer from the dead. Systematic Biology, 62(3), 386–397.

How to cite this article: Dismukes W, Heath TA. treeducken: An R package for simulating cophylogenetic systems. Methods Ecol Evol. 2021;00:1–7. https://doi.org/10.1111/2041-210X.13625