A peer-reviewed version of this preprint was published in PeerJ on 12 May 2016.

Stephens TG, Bhattacharya D, Ragan MA, Chan CX. (2016) PhySortR: a fast, flexible tool for sorting phylogenetic trees in R. PeerJ 4:e2038
https://doi.org/10.7717/peerj.2038

PhySortR: a fast, flexible tool for sorting phylogenetic trees in R

Timothy G. Stephens[1], Debashish Bhattacharya[2], Mark A. Ragan[1] and Cheong Xin Chan[1]

[1]ARC Centre of Excellence in Bioinformatics, and Institute for Molecular Bioscience, The University of Queensland, Brisbane, QLD 4072, Australia

[2]Department of Ecology, Evolution and Natural Resources, Rutgers University, New Brunswick, NJ 08901, U.S.A.

Corresponding author:
Cheong Xin Chan[1]
Institute for Molecular Bioscience, The University of Queensland, Brisbane, QLD 4072, Australia
Email address: c.chan1@uq.edu.au

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1609v1 | CC-BY 4.0 Open Access | rec: 28 Dec 2015, publ: 28 Dec 2015

Abstract

A frequent bottleneck in interpreting phylogenomic output is the need to screen often thousands of trees for features of interest, such as robust clades of specific taxa, as evidence of monophyletic relationship and/or reticulated evolution. Here we present PhySortR, a fast, flexible R package for sorting phylogenetic trees. Unlike existing utilities, PhySortR allows for identification of both exclusive and non-exclusive clades uniting the target taxa, with customisable options to assess clades within the context of the whole tree. PhySortR is a command-line tool that is freely available, highly scalable, and easily automatable.

Main Text

Phylogenomics increasingly involves screening of thousands of phylogenetic trees using specialised sorting algorithms for features of interest, e.g. strongly supported monophyletic relationships of taxa in question (i.e. the "target" taxa). Currently available tree-sorting utilities, e.g. PhyloSort (Moustafa and Bhattacharya 2008) and SICLE (Deblasio and Wisecaver 2013) screen a set of trees (in Newick format) for the presence of clades that unite a set of user-defined target taxa based on clade support (that exceeds a defined threshold), and sort these trees accordingly. However, these tools do not consider the proportion of non-target taxa and overall taxon composition in a tree during the sorting process.

Here we present PhySortR, a fast, flexible R package for screening and sorting phylogenetic trees (Cardona et al. 2008). The package provides the quick and highly flexible *sortTrees* function, allowing for screening (within a tree) for "Exclusive" clades that contain only the target taxa and/or "Non-Exclusive" clades that include a defined portion of non-target taxa. The

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1609v1 | CC-BY 4.0 Open Access | rec: 28 Dec 2015, publ: 28 Dec 2015

algorithm and all available options of PhySortR are described in detail in Supplementary Text S1 and Supplementary Figures S1 and S2.

PhySortR allows the user to specify one or more target taxa using the *target.groups* argument, providing that the taxa (tree-tip labels in the Newick files) are named consistently across all input trees; this is a simple string-matching exercise, i.e. the terms specified here determine taxon-level resolution of targets. The minimum support for a clade (*min.support*) can refer to bootstrap, Bayesian posterior probability, or any other measure of support. Existing utilities such as PhyloSort (Moustafa and Bhattacharya 2008), when assessing the relationship between two target groups in a 100-taxon tree, would identify both (a) a robust two-member clade (one from each target group), and (b) a robust 50-member clade containing multiple taxa from both groups, as "Exclusive", although (b) is likely the more-convincing evidence of a close association between the targets and of greater biological significance. To address this issue, PhySortR allows one to define *min.prop.target*, the minimum required proportion of target(s) present in a clade relative to the total number of target(s) found in a tree.

Alternatively, one might wish to screen for a robust clade that contains the target groups and a small proportion of "interrupting" non-target taxa, e.g. a 41-member clade consisting of 20 taxa each from the targets Rhodophyta and Viridiplantae, as well as a stramenopile taxon (a non-target such as a diatom). Whereas the diatom is "interrupting" the otherwise exclusive clade of Rhodophyta + Viridiplantae, the association between the targets is still of interest and the diatom presence might be readily explained by lateral gene transfer (LGT) due to plastid endosymbiosis. Composite clades such as these are considered "Non-Exclusive" (Chan et al. 2011) and are not identified by existing tree-sorting tools. The concept of exclusivity (Fig. 1A) versus non-

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1609v1 | CC-BY 4.0 Open Access | rec: 28 Dec 2015, publ: 28 Dec 2015

exclusivity (Fig. 1B) of clades in tree sorting has proven crucial in a number of genome-wide studies that have investigated the impact of LGT on the evolution of diverse algal and protist phyla (e.g., Chan et al. 2011; Curtis et al. 2012; Price et al. 2012; Bhattacharya et al. 2013). PhySortR identifies both types of clades by default. The user has the option to define a "Non-Exclusive" clade based on the proportion of target versus non-target taxa using the option *clade.exclusivity*. At the default setting (*clade.exclusivity* = 0.9), the minimum proportion of target taxa within a "Non-Exclusive" clade is 0.9, thus the maximum proportion of non-target taxa allowed in the clade is 0.1 (i.e. 1 minus 0.9). This option accepts any value <1.0, and is only applicable for sorting "Non-Exclusive" clades (see Supplementary Fig. S2); at 1.0 (no non-target taxa allowed), the clade is considered "Exclusive". PhySortR is an R implementation of the algorithm deployed in Chan et al. (2011) that has been adopted in other phylogenomic studies (Curtis et al. 2012; Price et al. 2012; Bhattacharya et al. 2013). The R platform is open source, platform-independent, and broadly accessible to researchers, with continued support by the R Core Team; functional modularity and the command-line interface enable batch automation and workflow integration.

The runtime of PhySortR is dependent on the number of trees (N) to be sorted and the number of taxa (X) within a tree. We benchmarked PhySortR using two simulated datasets to assess runtime (t): one varying N at fixed X = 100 (Supplementary Data S1), another varying X at fixed N = 1000 (Supplementary Data S2). We required 20% of X within an "Exclusive" clade in a tree. We observed that the runtime scales linearly with N (Figure 1C) and superlinearly with X (Figure 1D). In the extreme case, sorting through 10000 trees took <500 seconds (~8.3 minutes). We observed negligible differences in t with negative controls (trees containing no identifiable clades) as input, compared to the test set in Figure 1D.

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1609v1 | CC-BY 4.0 Open Access | rec: 28 Dec 2015, publ: 28 Dec 2015

PhySortR incorporates existing functionalities and data structures in the commonly used phylogenetic packages *ape* (Paradis et al. 2004) and *phytools* (Revell 2012), allowing for streamlined interoperability within the R environment. Whereas *ape* and *phytools* accept only traditional Newick as input, PhySortR accepts tree files in both traditional and extended Newick formats (Cardona et al. 2008). The function *convert.eNewick* in isolation can be used as a general-purpose tool for converting extended Newick into traditional Newick format. PhySortR is freely available from the Comprehensive R Archive Network (cran.r-project.org). See also https://cloudstor.aarnet.edu.au/plus/index.php/s/StQRSJBmcSZKd7y.

Supplementary Material

Supplementary Text, Supplementary Figures S1-S2, Supplementary Data S1-S2.

Acknowledgments

This work was supported by the Australian Research Council Discovery Project (DP150101875) grant awarded to M.A.R., C.X.C, and D.B. C.X.C. is supported by Great Barrier Reef Foundation Bioinformatics Fellowship awarded to M.A.R. D.B. acknowledges support from the National Science Foundation (1004213).

References

Bhattacharya D, Price DC, Chan CX, Qiu H, Rose N, Ball S, Weber APM, Arias MC, Henrissat B, Coutinho PM, Krishnan A, Zauner S, Morath S, Hilliou F, Egizi A, Perrineau MM, Yoon HS. 2013. Genome of the red alga *Porphyridium purpureum*. *Nat. Commun.* 4:1941.

Cardona G, Rossello F, Valiente G. 2008. Extended Newick: it is time for a standard representation of phylogenetic networks. *BMC Bioinformatics* 9:532.

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1609v1 | CC-BY 4.0 Open Access | rec: 28 Dec 2015, publ: 28 Dec 2015

100 Chan CX, Yang EC, Banerjee T, Yoon HS, Martone PT, Estevez JM, Bhattacharya D. 2011. Red
101 and green algal monophyly and extensive gene sharing found in a rich repertoire of red algal
102 genes. *Curr. Biol.* 21:328-333.

103 Curtis BA, Tanifuji G, Burki F, Gruber A, Irimia M, Maruyama S, Arias MC, Ball SG, Gile GH,
104 Hirakawa Y, Hopkins JF, Kuo A, Rensing SA, Schmutz J, Symeonidi A, Elias M, Eveleigh
105 RJM, Herman EK, Klute MJ, Nakayama T, Obornik M, Reyes-Prieto A, Armbrust EV, Aves SJ,
106 Beiko RG, Coutinho P, Dacks JB, Durnford DG, Fast NM, Green BR, Grisdale CJ, Hempel F,
107 Henrissat B, Hoppner MP, Ishida KI, Kim E, Koreny LK, Kroth PG, Liu Y, Malik SB, Maier
108 UG, McRose D, Mock T, Neilson JAD, Onodera NT, Poole AM, Pritham EJ, Richards TA,
109 Rocap G, Roy SW, Sarai C, Schaack S, Shirato S, Slamovits CH, Spencer DF, Suzuki S, Worden
110 AZ, Zauner S, Barry K, Bell C, Bharti AK, Crow JA, Grimwood J, Kramer R, Lindquist E,
111 Lucas S, Salamov A, McFadden GI, Lane CE, Keeling PJ, Gray MW, Grigoriev IV, Archibald
112 JM. 2012. Algal genomes reveal evolutionary mosaicism and the fate of nucleomorphs. *Nature*
113 492:59-65.

114 Deblasio D, Wisecaver J. 2013. SICLE: A high-throughput tool for extracting evolutionary
115 relationships from phylogenetic trees. arXiv:1303.5785.

116 Moustafa A, Bhattacharya D. 2008. PhyloSort: a user-friendly phylogenetic sorting tool and its
117 application to estimating the cyanobacterial contribution to the nuclear genome of
118 *Chlamydomonas*. *BMC Evol. Biol.* 8:6.

119 Paradis E, Claude J, Strimmer K. 2004. APE: Analyses of Phylogenetics and Evolution in R
120 language. *Bioinformatics* 20:289-290.

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1609v1 | CC-BY 4.0 Open Access | rec: 28 Dec 2015, publ: 28 Dec 2015

Price DC, Chan CX, Yoon HS, Yang EC, Qiu H, Weber APM, Schwacke R, Gross J, Blouin NA, Lane C, Reyes-Prieto A, Durnford DG, Neilson JAD, Lang BF, Burger G, Steiner JM, Loffelhardt W, Meuser JE, Posewitz MC, Ball S, Arias MC, Henrissat B, Coutinho PM, Rensing SA, Symeonidi A, Doddapaneni H, Green BR, Rajah VD, Boore J, Bhattacharya D. 2012. *Cyanophora paradoxa* genome elucidates origin of photosynthesis in algae and plants. *Science* 335:843-847.

Revell LJ. 2012. phytools: an R package for phylogenetic comparative biology (and other things). *Methods Ecol. Evol.* 3:217-223.

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1609v1 | CC-BY 4.0 Open Access | rec: 28 Dec 2015, publ: 28 Dec 2015

Figure

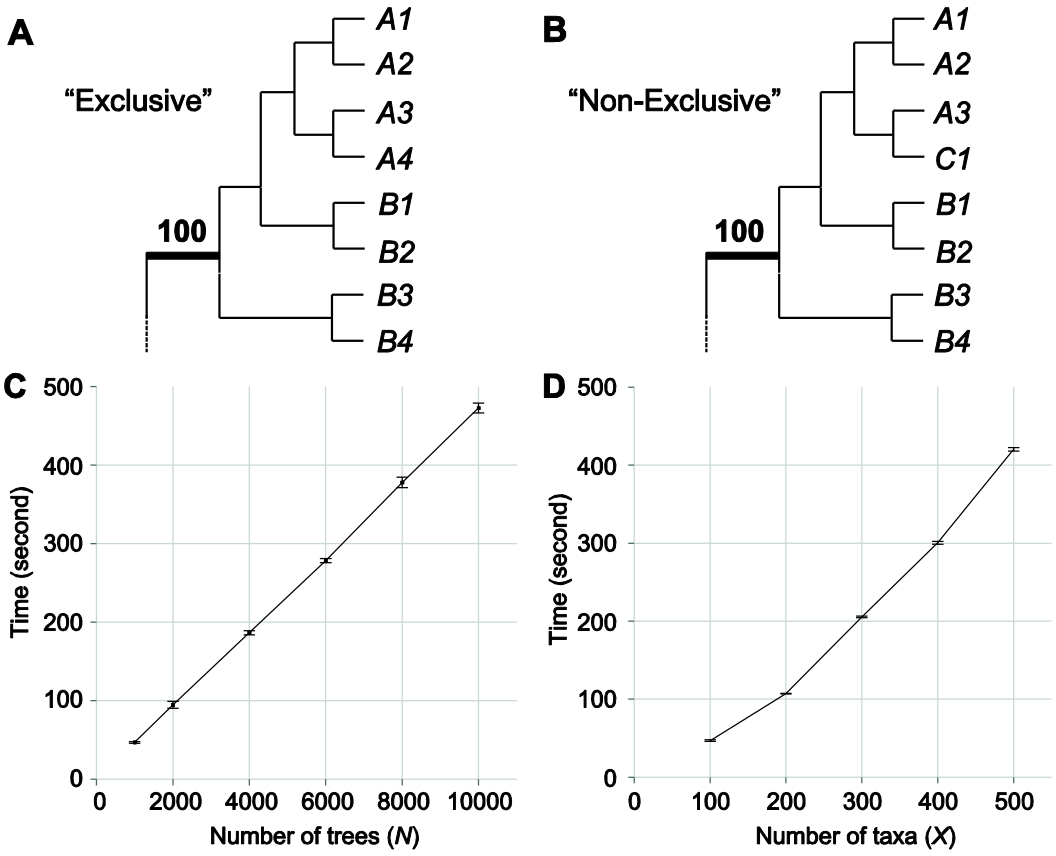

Figure 1. The concept of clade exclusivity and benchmarking results of PhySortR using simulated data. The schematic diagram of an "Exclusive" clade shown in (A) and consists solely of taxa from targets A and B, whereas the "Non-Exclusive" clade shown in (B) consists of targets A and B plus an "interrupting" taxon from group C; each clade has strong bootstrap support at 100%. The mean runtime (*t*) of PhySortR is shown for analysis across datasets (C) with different numbers of trees, *N* (Supplementary Data S1), and (D) with different numbers of taxa per tree, *X* (Supplementary Data S2). Values of *t* are mean across 100 replicates, error bars indicate the standard deviation of the mean.

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1609v1 | CC-BY 4.0 Open Access | rec: 28 Dec 2015, publ: 28 Dec 2015