2019

RASP 4: ancestral state reconstruction tool for multiple genes and characters

Yan Yu
Sichuan University

Christopher Blair
CUNY New York City College of Technology

Xingjin He
Sichuan University

RASP 4: ancestral state reconstruction tool for multiple genes and characters

Yan Yu[1, *], Christopher Blair[2,3], Xingjin He[1, *]

[1] *Key Laboratory of Bio-Resources and Eco-Environment of Ministry of Education, College of Life Sciences, Sichuan University, Chengdu, Sichuan, 610065, P. R. China.*

[2] *Department of Biological Sciences, New York City College of Technology, The City University of New York, 285 Jay Street, Brooklyn, NY 11201, USA.*

[3]*Biology PhD Program, CUNY Graduate Center, 365 5th Ave., New York, NY 10065*

* *Corresponding author: E-mail:* yyu@scu.edu.cn and xjhe@scu.edu.cn

Abstract: With the continual progress of sequencing techniques, genome-scale data are increasingly used in phylogenetic studies. With more data from throughout the genome, the relationship between genes and different kinds of characters is receiving more attention. Here, we present version 4 of RASP, a software to reconstruct ancestral states through phylogenetic trees. RASP can apply generalized statistical ancestral reconstruction methods to phylogenies, explore the phylogenetic signal of characters to particular trees, calculate distances between trees, and cluster trees into groups. RASP 4 has an improved graphic user interface and is freely available from http://mnh.scu.edu.cn/soft/blog/RASP (program) and https://github.com/sculab/RASP (source code).

Key words: Ancestral state reconstruction, Genome, Phylogeny, Phylogenetic signal

RASP (Reconstruct Ancestral State in Phylogenies) is a software to reconstruct ancestral states through phylogenetic trees. To date, the program has been used to infer biogeographic history in numerous groups of animals, plants, fungi and bacteria (Blair, et al. 2015; Yu, et al. 2015; Stucki, et al. 2016; Bourguignon, et al. 2018; Navaud, et al. 2018; Yan, et al. 2018). With the continual progress of sequencing techniques, the data from genomes, transcriptomes and proteomes have been increasingly used in phylogenetic studies (Choi and Kim 2017). Additionally, morphology, ecology, and distribution data are increasingly integrated into

research (Soltis and Soltis 2016). This motivated us to add more functionality into RASP to implement additional algorithms and tools.

The new version of RASP can analyze phylogenomic data (and other types of data), make inference on our generalized statistical method for ancestral state reconstruction (Fig 1-A and B) and summarize results under a graphical user interface (Fig 1-C). Users are also allowed to quantify phylogenetic signal of different morphological or ecological characters to particular trees (Fig 1-D), measure the fit between a tree and geography, and compute a distance matrix to cluster trees (Fig 1-E).

Methods to reconstruct ancestral geographical distributions using a combination of phylogenetic and distributional information are increasing rapidly. In RASP 4, we implement a generalized statistical method for models implemented in the R package 'BioGeoBEARS' (Matzke 2014) and 'APE' (Paradis and Schliep 2018); namely our method summarizes ancestral reconstructions across all input trees. The probability (p) of an ancestral range x at node n on the final species tree is calculated as $p(x_n) = {\sum_{t \in T}[w(x_n)_t]}/{g_n}$ where T is the set of trees, $w(x_n)_t$ is the weight of ancestral range x at node n for tree t, and g_n is the number of times node n occurs in T (see supplementary material for details). To reduce computational burden, RASP applies parallel computing to all models both by taking advantage of multiple threads and splitting trees into small groups. See Table S1 for a full comparison of the methods of ancestral reconstruction implemented in RASP.

Phylogenetic signal is the tendency of related species to resemble each other in a specific character more than species drawn at random from the same tree (Münkemüller, et al. 2012). To test for phylogenetic signal for continuous states, RASP calculates Moran's I (Moran 1948, 1950), Abouheif's C_{mean} (Abouheif 1999), Pagel's λ (Pagel 1999) and Blomberg's K (Blomberg, et al. 2003) using the R package 'adephylo' (Jombart, et al. 2010). For discrete states, RASP fits models of trait evolution using a likelihood ratio test and calculates Pagel's λ using the R package 'geiger' (Pennell, et al. 2014) (Fig 1-C). If some species have more than one state, RASP will convert them to all possible combinations of single states and calculate Pagel's λ for each of them. The largest Pagel's λ is used in the final result.

Tree distances are often used as a formal way to quantify the differences of trees inferred

from different genes and reconstruction methods (Sand, et al. 2014). In RASP, users can compute trees distances using different methods: KC distance (Kendall and Colijn 2016), triplet distance implemented in mp-est (Liu, et al. 2010), RF distance (Robinson and Foulds 1981), KF distance (Kuhner and Felsenstein 1994), path differences (Steel and Penny 1993), and SPR distance (de Oliveira Martins, et al. 2008; De Oliveira Martins, et al. 2014) implemented in the R package 'phangorn' (Schliep 2010) (Table S2). Having the distance matrix, values can be normalized using min-max normalization (Han, et al. 2006). Next, trees are clustered into groups using the R package 'adegenet' (Jombart 2008) according to the distance matrix. The end result may provide insight into the sources of heterogeneity among gene/species histories. For example, distinct clusters of genes may indicate unique phylogenetic signatures (Kendall and Colijn 2016). Additionally, the tree distance matrix and groups could be used to provide a candidate species tree under the coalescent model (Liu, et al. 2010).

In summary, the new version of RASP 4 implements several tools for multiple gene and species trees and characters while simultaneously making it easier to process trees generated from different sources. We provide six tutorials to help users select appropriate methods for different research questions on the our website (http://mnh.scu.edu.cn/soft/blog/RASP). We will continue to develop RASP with a focus on implementing new algorithms and integrating more tools. RASP for Windows and macOS are available freely from http://mnh.scu.edu.cn/soft/blog/RASP (program) and https://github.com/sculab/RASP (source code), and licensed under the terms of the MIT license.

Acknowledgments

We thank Liang Liu, Jianquan Liu and Kangshan Mao for discussion. This work was supported by the National Natural Science Foundation of China (Grant No. 31500188), the Specimen Platform of China, Teaching Specimen's sub-platform, the Science and Technology Basic Work (Grant No. 2013FY112100).

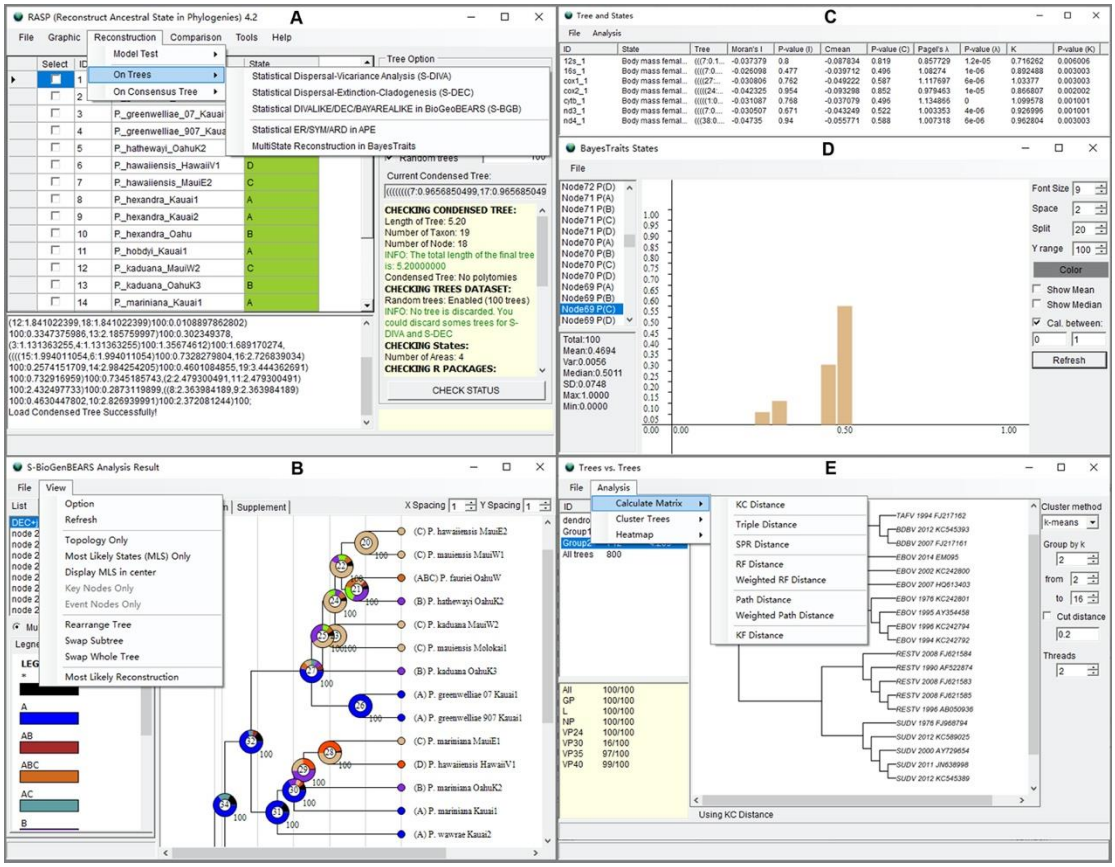

Fig. 1. Screenshots from RASP 4. The sample data and tutorials can be found on the RASP website. A) The main screen of RASP. The expanded menu shows the ancestral state reconstruction methods implemented in RASP. B) The tree view interface of RASP. The graphic shows the results of using the DIVALike model in BioGeoBears. The expanded menu shows the operations that can be performed on the results. C) The Trees and States tool. The list shows the results of phylogenetic signal for three states. D) Graphical interface showing ancestral state reconstruction results from BayesTraits (Meade and Pagel 2018). E) The Trees vs. Trees tool. The expanded menu shows the supported distance methods.

References

Abouheif E. 1999. A method for testing the assumption of phylogenetic independence in comparative data. Evolutionary Ecology Research 1:895-909.

Blair C, Noonan B, Brown J, Raselimanana A, Vences M, Yoder A. 2015. Multilocus phylogenetic and geospatial analyses illuminate diversification patterns and the biogeographic history of Malagasy endemic plated lizards (Gerrhosauridae: Zonosaurinae). Journal of Evolutionary Biology 28:481-492.

Blomberg SP, Garland Jr T, Ives AR. 2003. Testing for phylogenetic signal in comparative data: behavioral traits are more labile. Evolution 57:717-745.

Bourguignon T, Tang Q, Ho SYW, Juna F, Wang Z, Arab DA, Cameron SL, Walker J, Rentz D, Evans TA, et al. 2018. Transoceanic Dispersal and Plate Tectonics Shaped Global Cockroach Distributions: Evidence from Mitochondrial Phylogenomics. Mol Biol Evol 35:970-983.

Choi J, Kim S-H. 2017. A genome tree of life for the fungi kingdom. Proceedings of the National Academy of Sciences 114:9391-9396.

de Oliveira Martins L, Leal E, Kishino H. 2008. Phylogenetic detection of recombination with a Bayesian prior on the distance between trees. PLoS One 3:e2651.

De Oliveira Martins L, Mallo D, Posada D. 2014. A Bayesian supertree model for genome-wide species tree reconstruction. Systematic Biology 65:397-416.

Han J, Jian P, Michelin K. 2006. Data Mining, Southeast Asia Edition. In: San Francisco: Elsevier Inc.

Jombart T. 2008. adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics 24:1403-1405.

Jombart T, Balloux F, Dray S. 2010. Adephylo: new tools for investigating the phylogenetic signal in biological traits. Bioinformatics 26:1907-1909.

Kendall M, Colijn C. 2016. Mapping Phylogenetic Trees to Reveal Distinct Patterns of Evolution. Mol Biol Evol 33:2735-2743.

Kuhner MK, Felsenstein J. 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol Biol Evol 11:459-468.

Liu L, Yu L, Edwards SV. 2010. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. BMC evolutionary biology 10:302.

Matzke NJ. 2014. Model selection in historical biogeography reveals that founder-event speciation is a crucial process in island clades. Systematic Biology 63:951-970.

Meade A, Pagel M. 2018. BayesTraits: a computer package for analyses of trait evolution. In: Version.

Moran PA. 1948. The interpretation of statistical maps. Journal of the Royal Statistical Society. Series B (Methodological) 10:243-251.

Moran PA. 1950. Notes on continuous stochastic phenomena. Biometrika 37:17-23.

Münkemüller T, Lavergne S, Bzeznik B, Dray S, Jombart T, Schiffers K, Thuiller W. 2012. How to measure and test phylogenetic signal. Methods in Ecology and Evolution 3:743-756.

Navaud O, Barbacci A, Taylor A, Clarkson JP, Raffaele S. 2018. Shifts in diversification rates and host jump frequencies shaped the diversity of host range among Sclerotiniaceae fungal plant pathogens. Mol Ecol 27:1309-1323.

Pagel M. 1999. Inferring the historical patterns of biological evolution. Nature 401:877.

Paradis E, Schliep K. 2018. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics 35:526-528.

Pennell MW, Eastman JM, Slater GJ, Brown JW, Uyeda JC, FitzJohn RG, Alfaro ME, Harmon LJ. 2014. geiger v2. 0: an expanded suite of methods for fitting macroevolutionary models to phylogenetic trees.

Bioinformatics 30:2216-2218.
Robinson DF, Foulds LR. 1981. Comparison of phylogenetic trees. Mathematical biosciences 53:131-147.
Sand A, Holt MK, Johansen J, Brodal GS, Mailund T, Pedersen CN. 2014. tqDist: a library for computing the quartet and triplet distances between binary or general trees. Bioinformatics 30:2079-2080.
Schliep KP. 2010. phangorn: phylogenetic analysis in R. Bioinformatics 27:592-593.
Soltis DE, Soltis PS. 2016. Mobilizing and integrating big data in studies of spatial and phylogenetic patterns of biodiversity. Plant Diversity 38:264-270.
Steel MA, Penny D. 1993. Distributions of tree comparison metrics—some new results. Systematic Biology 42:126-141.
Stucki D, Brites D, Jeljeli L, Coscolla M, Liu Q, Trauner A, Fenner L, Rutaihwa L, Borrell S, Luo T, et al. 2016. Mycobacterium tuberculosis lineage 4 comprises globally distributed and geographically restricted sublineages. Nat Genet 48:1535-1543.
Yan H-F, Zhang C-Y, Anderberg AA, Hao G, Ge X-J, Wiens JJ. 2018. What explains high plant richness in East Asia? Time and diversification in the tribe Lysimachieae (Primulaceae). New Phytologist 219:436-448.
Yu Y, Harris AJ, Blair C, He X. 2015. RASP (Reconstruct Ancestral State in Phylogenies): a tool for historical biogeography. Molecular Phylogenetics and Evolution 87:46-49.