RCVT: a novel diagnostic to quantify compositional biases among taxa for phylogenomics

Jacob L Steenwyk[1,*] & Thomas J. Buida[2]

[1] Howards Hughes Medical Institute and the Department of Molecular and Cell Biology, University of California, Berkeley, Berkeley, CA, USA

[2] Independent Researcher, Nashville, TN 37209, USA

* Correspondence should be addressed to: jlsteenwyk@berkeley.edu

ORCiDs

J.L. Steenwyk: 0000-0002-8436-595X

T.J. Buida III: 0000-0001-9367-6189

Running title: RCVT quantifies compositional biases among taxa

Keywords: phylogeny, sequence alignments, genomics, software, tools, phylogenetics

Abstract

Phylogenomics aims to reconstruct the history of genes and genomes. However, noise or error during inference can stem from diverse sources, such as compositional biases. Here, we introduce RCVT (**R**elative **C**omposition **V**ariability among **T**axa), a novel metric to quantify compositional biases among taxa. We demonstrate the utility of RCVT using example data and quantify compositional biases in 16 empirical phylogenomic datasets, revealing variation in biases across datasets and taxa therein. RCVT may help researchers diagnose and potentially ameliorate phylogenomic noise associated with compositional biases.

Introduction

Phylogenetics aims to reconstruct the evolutionary history of genomes, genes, and other biological entities (Kapli et al. 2020; Steenwyk et al. 2023). However, sources of noise and error can cause inaccurate phylogenetic inferences (Betancur-R. et al. 2019; Steenwyk et al. 2023). Accordingly, numerous metrics have been devised to diagnose potential sources of error at the level of taxa, genes, and sites. For example, there are numerous metrics to diagnose potential errors among genes, such as alignment length, the number of parsimony informative sites, saturation, and clock-like evolution (Philippe et al. 2011; Shen et al. 2016; Liu et al. 2017; Steenwyk et al. 2020). In contrast, there are fewer metrics to evaluate compositional bias among taxa.

Here, we introduce a novel metric to quantify compositional bias among taxa termed RCVT (**R**elative **C**omposition **V**ariability among **T**axa). To demonstrate the utility of RCVT, we calculate RCVT in a fictitious four-taxon case. After that, we calculate RCVT in 16 phylogenomic data matrices from diverse lineages and uncover variation in compositional biases among taxa among datasets and organisms represented therein. For example, a dataset of Caryophyllales (an order of flowering plants) had the greatest variance in RCVT values among the datasets, and 15 out of the 16 datasets had outlier taxa, suggesting that outlier compositional biases are prevalent and may warrant further scrutiny in phylogenomic studies. To facilitate researchers calculating RCVT, we incorporate a new function to calculate RCVT in PhyKIT, a command-line toolkit for processing and analyzing phylogenomic data (i.e., multiple sequence alignments and phylogenetic trees) (https://jlsteenwyk.com/PhyKIT) (Steenwyk et al. 2021, 2024).

Methods

RCVT measures how much a particular taxon differs from the average composition of all taxa in a multiple-sequence alignment. RCVT is adapted from the metric *relative composition variability*, which measures compositional bias at the gene level (Phillips and Penny 2003). The following formula is used to calculate RCVT:

$$RCVT_j = \sum_{i=1}^{c} \frac{|c_{ij} - \overline{c_i}|}{s \times n}$$

wherein abbreviations are as follows:

- $RCVT_j$: relative composition variability for taxon *j*, which quantifies its variability compared to others;
- *c*: the number of different character states per sequence type (e.g., A, T, C, and G for DNA sequences) in an alignment;
- c_{ij}: the number of the *i*th character state (e.g., how many times A appears) for the *j*th taxon;
- $\bar{c}_i$: the average number of the *i*th *c* character state across *n* taxa, which represents the general occurrence of that character in a dataset;
- *s*: total number of sites; and
- *n*: the number of taxa.

Using this formula, higher values are associated with greater composition biases, whereas lower values are associated with less compositional bias. For example, in a toy four-taxon case, one taxon may have only one nucleotide character, but the other three taxa have an equal number of each of the four nucleotides (A, T, C, and G). In this case, the taxon with only one nucleotide character will have a much higher RCVT than the other three taxa (Figure 1a).

A function to calculate RCVT has been incorporated into the Python-based tool, PhyKIT (Steenwyk et al. 2021), as of version 1.13.0 (https://jlsteenwyk.com/PhyKIT/usage/index.html#relative-composition-variability-taxon). To ensure the efficacy of the newly added function, we wrote five integration tests that cover 100% of the code for the RCVT function. These tests, alongside a continuous integration pipeline, which leverages architecture we had built elsewhere (Steenwyk et al. 2020, 2022b,a; Steenwyk and Rokas 2021), help ensure PhyKIT properly builds, packages, and installs across multiple versions of Python. These efforts were motivated by nearly 30% of all bioinformatic tools failing to install (Mangul et al. 2019), which threatens the reproducibility of bioinformatics research. More broadly, we aim to provide the research community with a long-term dependable tool for rigorous phylogenomics research.

We also calculated RCVT for 16 empirical phylogenomic datasets (Table 1). For each dataset, single-gene multiple sequence alignments were downloaded from a previous study that examined phylogenomic methodology (Liu et al. 2024) or from the primary study, as in the case of the Anglerfishes dataset (Miller et al. 2024). The resulting multiple sequence alignments were then concatenated into a supermatrix using the *create_concat* function in PhyKIT (Steenwyk et al. 2021). Next, RCVT was calculated for the taxa in each supermatrix using the *rcvt* function in PhyKIT. To detect outlier values, we identified RCVT values that had values greater than the 75th percentile plus 1.5 times the interquartile range (the difference between the 75th and 25th percentiles).

Results and Discussion

To determine how common compositional outliers are, we calculated RCVT in 16 phylogenomic datasets (Figure 1b and Table 1). These datasets are composed of amino acid or nucleotide sequences, span a diversity of taxonomic ranks and lineages, and often include many taxa and loci. Across each dataset, the variance among RCVT values varied (Figure 1b, top panel). For example, a dataset of Caryophyllales species – including flowering plants like cacti – had the greatest variance, followed by a dataset of anglerfishes and plants from the genus *Protea*. In contrast, the dataset of birds, species from the genus *Jaltomata*, and lizards had low variance values. These results indicate that compositional biases can widely vary in some datasets, which may impact phylogenetic inferences.

Quantifying the number of taxa with high outlier RCVT values revealed variation in the number of outlier taxa per dataset (Figure 1b, middle panel). For example, the dataset of birds had the highest number of outliers. Similarly, the anglerfish dataset also had many outliers. In contrast, the dataset of Rodents and Saccharomycotina yeast had few outliers. The dataset of Lizards had no outlier taxa. Examination of the taxa that were outliers revealed, at times, that outlier taxa were outgroup taxa. For example, although *Penicillium verrucosum* had the highest RCVT value (0.0045) in the Aspergillaceae dataset, the next highest RCVT value of 0.0043 was observed in the outgroup species

Histoplasma capsulatum. This finding suggests that outgroup species may influence phylogenomic inferences due to compositional biases.

Practical considerations

The intended use of RCVT is to identify individual taxa that may have compositional biases that could negatively impact phylogenomic inference. While we demonstrate the utility of RCVT, it is a novel metric that, therefore, warrants additional independent scrutiny. We also note that RCVT is a relative metric. Thus, if a user wants to measure compositional biases for individual taxa across single genes, it may be beneficial to z-transform RCVT values for individual genes.

Data availability

RCVT has been incorporated into PhyKIT (Steenwyk et al. 2021) and comes complete with documentation (https://jlsteenwyk.com/PhyKIT/usage/index.html#relative-composition-variability-taxon). PhyKIT is freely available under the MIT license and can be downloaded from PyPi (https://pypi.org/project/phykit/) or GitHub (https://github.com/JLSteenwyk/PhyKIT).

Main figure and legend

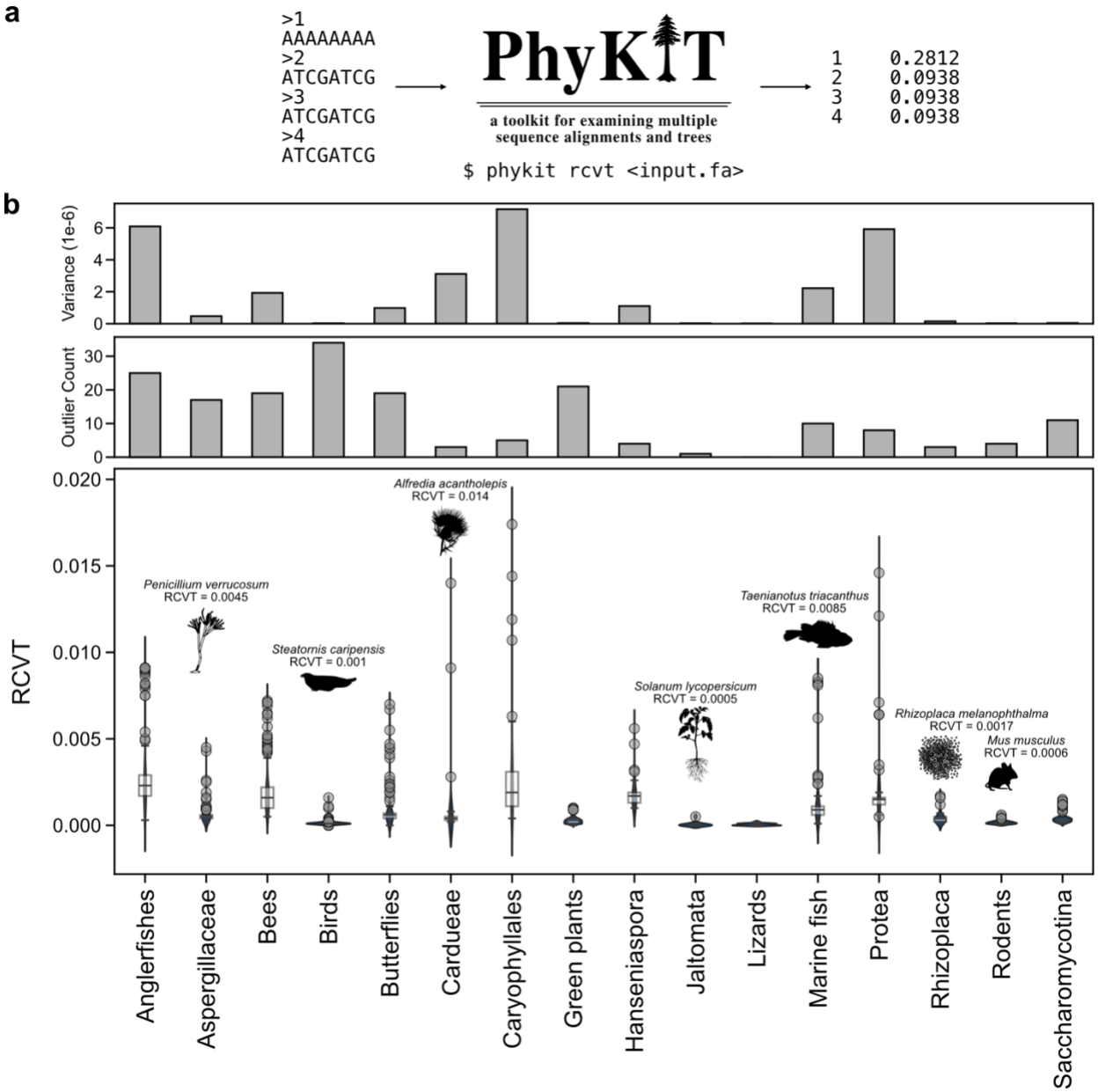

Figure 1. RCVT measures compositional bias and reveals that compositional outliers are common among phylogenomic datasets. (a) In a toy dataset wherein a taxon has only one character state in an alignment, whereas the other taxa have equal numbers of the four nucleotides, the taxon with one character state has a higher RCVT value. RCVT can easily be calculated using the *rcvt* function in the command-line tool, PhyKIT (Steenwyk et al. 2021). (b) The distribution of individual RCVT values reveals variation in compositional biases among 16 empirical datasets (Table 1). (b, top) Each dataset had different degrees of variance among RCVT values and (b, middle) different

numbers of outlier taxa. (b, bottom) Boxplots with violin plots provide additional information about the median, 75th, and 25th percentiles. For select datasets, we highlight outlier taxa and their RCVT values; for example, in the Aspergillaceae dataset, *Penicillium verrucosum* had the highest RCVT value of 0.0045.

Main table

Dataset identifier	Description	Kingdom	Taxonomic Breadth	Number of taxa	Number of loci	Data type	Reference
Anglerfishes	Fish with luminescent lures	Animals	Order	153	1092	DNA	(Miller et al. 2024)
Aspergillaceae	Primarily *Aspergillus* and *Penicillium* species	Fungi	Family	93	1668	DNA	(Steenwyk et al. 2019b)
Bees	Carpenter bees	Animals	Genus	190	753	DNA	(Blaimer et al. 2018)
Birds	Spans the diversity of Neoaves	Animals	Class	200	259	DNA	(Prum et al. 2015)
Butterflies	Spans the diversity of Rhopalocera	Animals	Order	207	352	DNA	(Espeland et al. 2018)
Cardueae	Tribe of flower plants daisies	Plants	Family	85	570	DNA	(Herrando-Moraira et al. 2019)

Caryophyllales	A clade of flowering plants like cacti	Plants	Order	95	1122	AA	(Yang et al. 2015)
Green plants	Spans the diversity of Viridiplantae	Plants	Phylum	1178	410	AA	(One Thousand Plant Transcriptomes Initiative 2019)
Hanseniaspora	Bipolar budding yeast	Fungi	Genus	29	1033	AA	(Steenwyk et al. 2019a)
Jaltomata	The sister genus to tomatoes and close relatives	Plants	Genus	15	6431	DNA	(Wu et al. 2018)
Lizards	Australian lizards from the genus *Cryptoblepharus*	Animals	Genus	29	1361	DNA	(Blom et al. 2016)

Marine fishes	Acanthomorphs	Animals	Superorder	120	1001	DNA	(Alfaro et al. 2018)
Protea	Evergreen shrubs and flowers	Plants	Genus	65	498	DNA	(Mitchell et al. 2017)
Rhizoplaca	Lichenized Fungi	Fungi	Genus	31	303	DNA	(Leavitt et al. 2016)
Saccharomycotina	Yeast	Fungi	Subphylum	343	2048	AA	(Shen et al. 2018)

Table 1. Empirical phylogenomic datasets examined.

Acknowledgements

JLS thanks Kassiani Panagiotou for the helpful discussion while at a workshop together.

Funding

JLS is a Howard Hughes Medical Institute Awardee of the Life Sciences Research Foundation.

Conflicts of Interest

JLS is an advisor to ForensisGroup Inc. JLS is a scientific consultant to FutureHouse Inc. JLS is a Bioinformatics Visiting Scholar at MantleBio Inc.

References

Alfaro, M. E., B. C. Faircloth, R. C. Harrington, L. Sorenson, M. Friedman, C. E. Thacker, C. H. Oliveros, D. Černý, and T. J. Near. 2018. Explosive diversification of marine fishes at the Cretaceous–Palaeogene boundary. Nat Ecol Evol 2:688–696.

Betancur-R., R., D. Arcila, R. P. Vari, L. C. Hughes, C. Oliveira, M. H. Sabaj, and G. Ortí. 2019. Phylogenomic incongruence, hypothesis testing, and taxonomic sampling. Evolution 73:329–345. [Society for the Study of Evolution, Wiley].

Blaimer, B. B., J. R. Mawdsley, and S. G. Brady. 2018. Multiple origins of sexual dichromatism and aposematism within large carpenter bees. Evolution 72:1874–1889.

Blom, M. P. K., J. G. Bragg, S. Potter, and C. Moritz. 2016. Accounting for Uncertainty in Gene Tree Estimation: Summary-Coalescent Species Tree Inference in a Challenging Radiation of Australian Lizards. Syst Biol syw089.

Espeland, M., J. Breinholt, K. R. Willmott, A. D. Warren, R. Vila, E. F. A. Toussaint, S. C. Maunsell, K. Aduse-Poku, G. Talavera, R. Eastwood, M. A. Jarzyna, R. Guralnick, D. J. Lohman, N. E. Pierce, and A. Y. Kawahara. 2018. A Comprehensive and Dated Phylogenomic Analysis of Butterflies. Current Biology 28:770-778.e5.

Herrando-Moraira, S., J. A. Calleja, M. Galbany-Casals, N. Garcia-Jacas, J.-Q. Liu, J. López-Alvarado, J. López-Pujol, J. R. Mandel, S. Massó, N. Montes-Moreno, C. Roquet, L. Sáez, A. Sennikov, A. Susanna, and R. Vilatersana. 2019. Nuclear and plastid DNA phylogeny of tribe Cardueae (Compositae) with Hyb-Seq data: A new subtribal classification and a temporal diversification framework. Molecular Phylogenetics and Evolution 137:313–332.

Kapli, P., Z. Yang, and M. J. Telford. 2020. Phylogenetic tree building in the genomic age. Nat Rev Genet 21:428–444.

Leavitt, S. D., F. Grewe, T. Widhelm, L. Muggia, B. Wray, and H. T. Lumbsch. 2016. Resolving evolutionary relationships in lichen-forming fungi using diverse phylogenomic datasets and analytical approaches. Sci Rep 6:22262.

Liu, C., X. Zhou, Y. Li, C. T. Hittinger, R. Pan, J. Huang, X. Chen, A. Rokas, Y. Chen, and X.-X. Shen. 2024. The Influence of the Number of Tree Searches on Maximum Likelihood Inference in Phylogenomics. Systematic Biology 73:807–822.

Liu, L., J. Zhang, F. E. Rheindt, F. Lei, Y. Qu, Y. Wang, Y. Zhang, C. Sullivan, W. Nie, J. Wang, F. Yang, J. Chen, S. V. Edwards, J. Meng, and S. Wu. 2017. Genomic evidence reveals a radiation of placental mammals uninterrupted by the KPg boundary. Proc. Natl. Acad. Sci. U.S.A. 114.

Mangul, S., T. Mosqueiro, R. J. Abdill, D. Duong, K. Mitchell, V. Sarwal, B. Hill, J. Brito, R. J. Littman, B. Statz, A. K.-M. Lam, G. Dayama, L. Grieneisen, L. S. Martin, J. Flint, E. Eskin, and R. Blekhman. 2019. Challenges and recommendations to improve the installability and archival stability of omics computational tools. PLoS Biol 17:e3000333.

Miller, E. C., R. Faucher, P. B. Hart, M. Rincón-Sandoval, A. Santaquiteria, W. T. White, C. C. Baldwin, M. Miya, R. Betancur-R, L. Tornabene, K. Evans, and D. Arcila. 2024. Reduced evolutionary constraint accompanies ongoing radiation in deep-sea anglerfishes. Nat Ecol Evol, doi: 10.1038/s41559-024-02586-3.

Mitchell, N., P. O. Lewis, E. M. Lemmon, A. R. Lemmon, and K. E. Holsinger. 2017. Anchored phylogenomics improves the resolution of evolutionary relationships in the rapid radiation of *Protea* L. American J of Botany 104:102–115.

One Thousand Plant Transcriptomes Initiative. 2019. One thousand plant transcriptomes and the phylogenomics of green plants. Nature 574:679–685.

Philippe, H., H. Brinkmann, D. V. Lavrov, D. T. J. Littlewood, M. Manuel, G. Wörheide, and D. Baurain. 2011. Resolving Difficult Phylogenetic Questions: Why More Sequences Are Not Enough. PLoS Biol 9:e1000602.

Phillips, M. J., and D. Penny. 2003. The root of the mammalian tree inferred from whole mitochondrial genomes. Molecular Phylogenetics and Evolution 28:171–185.

Prum, R. O., J. S. Berv, A. Dornburg, D. J. Field, J. P. Townsend, E. M. Lemmon, and A. R. Lemmon. 2015. A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. Nature 526:569–573.

Shen, X.-X., D. A. Opulente, J. Kominek, X. Zhou, J. L. Steenwyk, K. V. Buh, M. A. B. Haase, J. H. Wisecaver, M. Wang, D. T. Doering, J. T. Boudouris, R. M. Schneider, Q. K. Langdon, M. Ohkuma, R. Endoh, M. Takashima, R. Manabe, N. Čadež, D. Libkind, C. A. Rosa, J. DeVirgilio, A. B. Hulfachor, M. Groenewald, C. P. Kurtzman, C. T. Hittinger, and A. Rokas. 2018. Tempo and Mode of Genome Evolution in the Budding Yeast Subphylum. Cell 175:1533-1545.e20.

Shen, X.-X., L. Salichos, and A. Rokas. 2016. A Genome-Scale Investigation of How Sequence, Function, and Tree-Based Gene Properties Influence Phylogenetic Inference. Genome Biol Evol 8:2565–2580.

Steenwyk, J. L., T. J. Buida, C. Gonçalves, D. C. Goltz, G. Morales, M. E. Mead, A. L. LaBella, C. M. Chavez, J. E. Schmitz, M. Hadjifrangiskou, Y. Li, and A. Rokas. 2022a. BioKIT: a versatile toolkit for processing and analyzing diverse types of sequence data. Genetics 221:iyac079.

Steenwyk, J. L., T. J. Buida, A. L. Labella, Y. Li, X.-X. Shen, and A. Rokas. 2021. PhyKIT: a broadly applicable UNIX shell toolkit for processing and analyzing phylogenomic data. Bioinformatics 37:2325–2331.

Steenwyk, J. L., T. J. Buida, Y. Li, X.-X. Shen, and A. Rokas. 2020. ClipKIT: A multiple sequence alignment trimming software for accurate phylogenomic inference. PLoS Biol 18:e3001007.

Steenwyk, J. L., D. C. Goltz, T. J. Buida, Y. Li, X.-X. Shen, and A. Rokas. 2022b. OrthoSNAP: A tree splitting and pruning algorithm for retrieving single-copy orthologs from gene family trees. PLoS Biol 20:e3001827.

Steenwyk, J. L., Y. Li, X. Zhou, X.-X. Shen, and A. Rokas. 2023. Incongruence in the phylogenomics era. Nature Reviews Genetics, doi: 10.1038/s41576-023-00620-x.

Steenwyk, J. L., G. I. Martínez-Redondo, T. J. Buida, E. Gluck-Thaler, X. Shen, T. Gabaldón, A. Rokas, and R. Fernández. 2024. PhyKIT: A Multitool for Phylogenomics. Current Protocols 4:e70016.

Steenwyk, J. L., D. A. Opulente, J. Kominek, X.-X. Shen, X. Zhou, A. L. Labella, N. P. Bradley, B. F. Eichman, N. Čadež, D. Libkind, J. DeVirgilio, A. B. Hulfachor, C. P. Kurtzman, C. T. Hittinger, and A. Rokas. 2019a. Extensive loss of cell-cycle and

DNA repair genes in an ancient lineage of bipolar budding yeasts. PLoS Biol 17:e3000255.

Steenwyk, J. L., and A. Rokas. 2021. orthofisher: a broadly applicable tool for automated gene identification and retrieval. G3 Genes|Genomes|Genetics 11:jkab250.

Steenwyk, J. L., X.-X. Shen, A. L. Lind, G. H. Goldman, and A. Rokas. 2019b. A Robust Phylogenomic Time Tree for Biotechnologically and Medically Important Fungi in the Genera *Aspergillus* and *Penicillium*. mBio 10:e00925-19.

Wu, M., J. L. Kostyun, M. W. Hahn, and L. C. Moyle. 2018. Dissecting the basis of novel trait evolution in a radiation with widespread phylogenetic discordance. Molecular Ecology 27:3301–3316.

Yang, Y., M. J. Moore, S. F. Brockington, D. E. Soltis, G. K.-S. Wong, E. J. Carpenter, Y. Zhang, L. Chen, Z. Yan, Y. Xie, R. F. Sage, S. Covshoff, J. M. Hibberd, M. N. Nelson, and S. A. Smith. 2015. Dissecting Molecular Evolution in the Highly Diverse Plant Clade Caryophyllales Using Transcriptome Sequencing. Mol Biol Evol 32:2001–2014.