




ecodive: Parallel and Memory-Efficient R Package for Ecological Diversity Analysis

Daniel P Smith^{1,2}[✉], Sara J Javornik Cregeen^{1,2}, and Joseph F Petrosino^{1,2}

¹ The Alkek Center for Metagenomics and Microbiome Research, Department of Molecular Virology and Microbiology, Baylor College of Medicine, Houston, TX 77030, USA ² Department of Molecular Virology and Microbiology, Baylor College of Medicine, Houston, TX, USA [✉] Corresponding author

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Published: unpublished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

Characterizing the composition of biological communities is a fundamental task in ecology, but the calculations involved can be computationally prohibitive. *ecodive* is an R package that addresses this challenge by providing a highly optimized implementation of common ecological diversity metrics, including alpha-diversity (within-sample richness and evenness) and beta-diversity (between-sample dissimilarity). These metrics can incorporate species counts, relative abundances, and evolutionary relationships, providing a multi-faceted view of ecological structure. By leveraging a compiled C library with pthreads for parallelization, *ecodive* delivers substantial performance gains in both speed and memory usage, enabling researchers to analyze larger datasets more efficiently.

Statement of Need

The analysis of ecological diversity in large-scale studies is often hampered by the computational demands of calculating metrics across thousands of communities, a common requirement in modern microbiome research, where studies routinely involve analyzing thousands of samples from large cohorts. This is particularly true for phylogenetic metrics like Faith's PD (Faith, 1992) and the UniFrac distance family (Q. Chang et al., 2011; Chen et al., 2012; C. A. Lozupone et al., 2007; C. Lozupone & Knight, 2005), which integrate species abundance with evolutionary data from phylogenetic trees. The resulting high demand on processing time and memory can limit the scope and scale of scientific inquiry.

ecodive overcomes these limitations by offering a significantly faster and more memory-efficient solution. This allows researchers to analyze more samples, explore more complex questions, and obtain more robust insights from their data. By providing a high-performance, parallelized engine for these calculations, *ecodive* empowers researchers to push the boundaries of large-scale ecological analysis.

Comparison to Existing Packages

To evaluate its performance, *ecodive* was benchmarked against numerous R packages that provide their own implementations of diversity metrics, including *abdiv* (Bittinger, 2020), *adiv* (Pavoine, 2020), *ampvis2* (Andersen et al., 2018), *ecodist* (Goslee & Urban, 2007), *entropart* (Marcon & Herault, 2015), *GUniFrac* (Chen et al., 2023), *labdsv* (Roberts, 2025), *parallelDist* (Eckert, 2022), *philentropy* (HG, 2018), *phyloregion* (Daru et al., 2020), *phyloseq* (McMurdie & Holmes, 2013), *picante* (Kembel et al., 2010), *tabula* (Frerebeau, 2019), and *vegan* (Oksanen et al., 2025). The results, conducted using the *bench* package, are

summarized in Figure 1 and demonstrate ecodive's superior speed and memory efficiency for each of the metrics tested. Crucially, the benchmark suite confirms these performance gains do not come at the cost of accuracy, as ecodive produces numerically identical output to the other packages. Beyond its computational advantages, ecodive has zero external R dependencies. This makes it a lightweight, stable, and secure backend, minimizing installation conflicts and simplifying long-term maintenance for developers who build upon it. The complete benchmark code and results are available in the package vignette (`vignette('benchmark')`) and online.

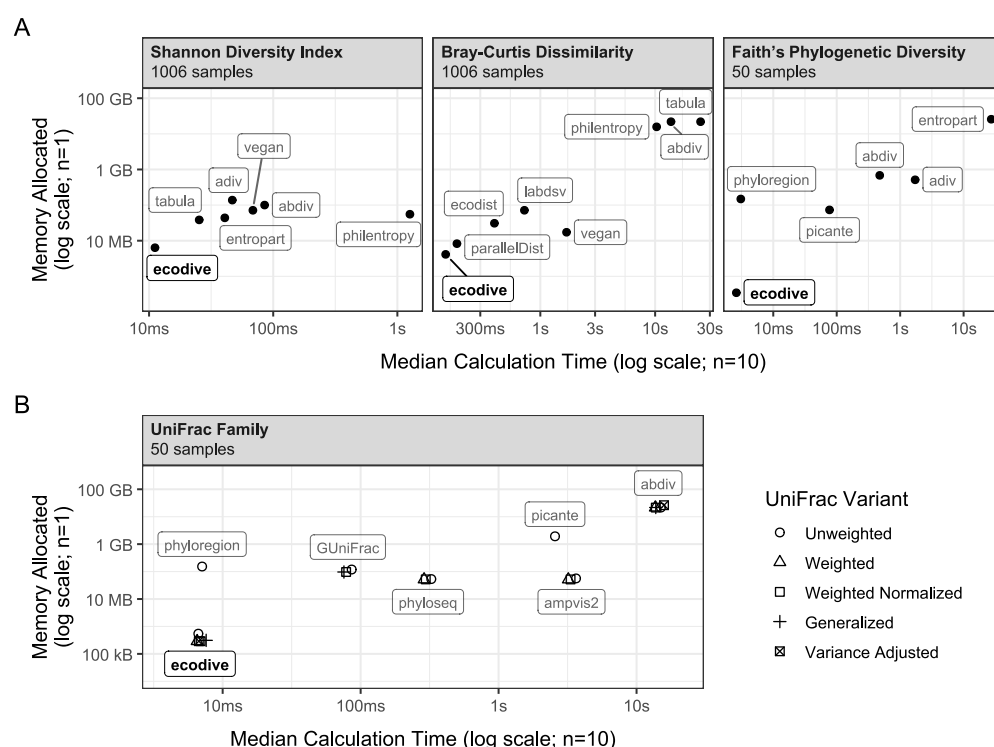


Figure 1: ecodive performance benchmarks. Each point represents an R package, plotted by median calculation time (x-axis) and memory consumption (y-axis) from ten trials. (A) Benchmarks for Shannon Diversity Index, Bray-Curtis Dissimilarity, and Faith's Phylogenetic Diversity. (B) Benchmarks for the UniFrac family of metrics, with different variants distinguished by point shape. Not all packages implement every metric, but ecodive is consistently the fastest and most memory-efficient across all tested metrics, often by several orders of magnitude.

Implemented Metrics

ecodive stands out by offering an extensive and diverse collection of over 50 metrics for both alpha and beta diversity analysis, making it a uniquely comprehensive tool. It provides researchers with a wide array of both traditional and phylogeny-aware algorithms in a single, high-performance package. The suite of alpha diversity metrics includes staples like the Shannon Diversity Index (Shannon, 1948) and Chao1 (Chao, 1984), important estimators such as the Abundance-based Coverage Estimator (ACE) (Chao & Lee, 1992) and Fisher's Alpha (Fisher et al., 1943), and key phylogeny-aware metrics like Faith's Phylogenetic Diversity, offering robust ways to assess within-sample richness and evenness. For assessing between-sample dissimilarity, ecodive implements essential beta diversity metrics widely used in microbial ecology, including Bray-Curtis Dissimilarity (Bray & Curtis, 1957; Sorenson, 1948), the complete UniFrac family, and the Aitchison distance (Aitchison, 1982) for compositional data analysis. This extensive collection allows for a thorough and multi-faceted analysis of community structure.

60 For the most up-to-date list and detailed descriptions, please refer to the official ecodive
61 documentation at <https://cmmr.github.io/ecodive>.

62 Programmatic Use and API

63 Beyond interactive analysis, ecodive is engineered for programmatic use, making it an ideal
64 backend for applications like R Shiny web apps (W. Chang et al., 2024). The package includes
65 a `list_methods()` function that allows developers to dynamically filter and present available
66 diversity metrics based on specific criteria. For instance, methods can be programmatically
67 selected if they are phylogeny-aware, abundance-weighted, capable of handling non-integer
68 counts, or are “true metrics” that satisfy the triangle inequality. This powerful API simplifies
69 the integration of ecodive into other software, enabling developers to build sophisticated tools
70 that offer users tailored diversity analysis options based on their dataset and analytical needs.

71 Example Usage

72 ecodive is designed for ease of use and integrates seamlessly with existing bioinformatics
73 workflows, such as those using phyloseq objects. For example, calculating weighted UniFrac
74 distances is straightforward:

```
data(esophagus, package = 'phyloseq')  
ecodive::weighted_unifrac(esophagus)  
#>           B           C  
#> C 0.1050480  
#> D 0.1401124 0.1422409
```

75 Acknowledgements

76 This study was supported by NIH/NIAD (Grant number U19 AI144297), and Baylor College
77 of Medicine and Alkek Foundation Seed. The authors also acknowledge the use of Google's
78 Gemini for assistance in refining this manuscript.

79 References

- 80 Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal*
81 *Statistical Society. Series B (Methodological)*, 44(2), 139–177. [https://doi.org/10.1111/j.](https://doi.org/10.1111/j.2517-6161.1982.tb01195.x)
82 [2517-6161.1982.tb01195.x](https://doi.org/10.1111/j.2517-6161.1982.tb01195.x)
- 83 Andersen, K. S., Kirkegaard, R. H., Karst, S. M., & Albertsen, M. (2018). ampvis2: An r
84 package to analyse and visualise 16S rRNA amplicon data. *bioRxiv*. [https://doi.org/10.](https://doi.org/10.1101/299537)
85 [1101/299537](https://doi.org/10.1101/299537)
- 86 Bittinger, K. (2020). *Abdiv: Alpha and beta diversity measures*. [https://doi.org/10.32614/](https://doi.org/10.32614/CRAN.package.abdiv)
87 [CRAN.package.abdiv](https://doi.org/10.32614/CRAN.package.abdiv)
- 88 Bray, J. R., & Curtis, J. T. (1957). An ordination of the upland forest communities of southern
89 wisconsin. *Ecological Monographs*, 27(4), 325–349. <https://doi.org/10.2307/1942268>
- 90 Chang, Q., Luan, Y., & Sun, F. (2011). Variance adjusted weighted UniFrac: A powerful beta
91 diversity measure for comparing communities based on phylogeny. *BMC Bioinformatics*,
92 12(1). <https://doi.org/10.1186/1471-2105-12-118>
- 93 Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson,
94 J., Dipert, A., & Borges, B. (2024). *Shiny: Web application framework for r*. <https://doi.org/10.32614/CRAN.package.shiny>

- Chao, A. (1984). Non-parametric estimation of the number of classes in a population. *Scandinavian Journal of Statistics*, 11, 265–270. <https://doi.org/10.2307/4616294>
- Chao, A., & Lee, S.-M. (1992). Estimating the number of classes via sample coverage. *Journal of the American Statistical Association*, 87(417), 210–217. <https://doi.org/10.1080/01621459.1992.10475194>
- Chen, J., Bittinger, K., Charlson, E. S., Hoffmann, C., Lewis, J., Wu, G. D., Collman, R. G., Bushman, F. D., & Li, H. (2012). Associating microbiome composition with environmental covariates using generalized UniFrac distances. *Bioinformatics*, 28(16), 2106–2113. <https://doi.org/10.1093/bioinformatics/bts342>
- Chen, J., Zhang, X., Yang, L., & Zhang, L. (2023). *GUniFrac: Generalized UniFrac distances, distance-based multivariate methods and feature-based univariate methods for microbiome data analysis*. <https://doi.org/10.32614/CRAN.package.GUniFrac>
- Daru, B. H., Karunaratne, P., & Schliep, K. (2020). Phyloregion: R package for biogeographic regionalization and macroecology. *Methods in Ecology and Evolution*, 11, 1483–1491. <https://doi.org/10.1111/2041-210X.13478>
- Eckert, A. (2022). *parallelDist: Parallel distance matrix computation using multiple threads*. <https://doi.org/10.32614/CRAN.package.parallelDist>
- Faith, D. P. (1992). Conservation evaluation and phylogenetic diversity. *Biological Conservation*, 61, 1–10. [https://doi.org/10.1016/0006-3207\(92\)91201-3](https://doi.org/10.1016/0006-3207(92)91201-3)
- Fisher, R. A., Corbet, A. S., & Williams, C. B. (1943). The relation between the number of species and the number of individuals in a random sample of an animal population. *Journal of Animal Ecology*, 12(1), 42–58. <https://doi.org/10.2307/1411>
- Frerebeau, N. (2019). Tabula: An r package for analysis, seriation, and visualization of archaeological count data. *Journal of Open Source Software*, 4(44), 1821. <https://doi.org/10.21105/joss.01821>
- Goslee, S. C., & Urban, D. L. (2007). The ecodist package for dissimilarity-based analysis of ecological data. *Journal of Statistical Software*, 22, 1–19. <https://doi.org/10.18637/jss.v022.i07>
- HG, D. (2018). Philentropy: Information theory and distance quantification with r. *Journal of Open Source Software*, 3(26), 765. <https://doi.org/10.21105/joss.00765>
- Kembel, S. W., Cowan, P. D., Helmus, M. R., Cornwell, W. K., Morlon, H., Ackerly, D. D., Blomberg, S. P., & Webb, C. O. (2010). Picante: R tools for integrating phylogenies and ecology. *Bioinformatics*, 26, 1463–1464. <https://doi.org/10.1093/bioinformatics/btq166>
- Lozupone, C. A., Hamady, M., Kelley, S. T., & Knight, R. (2007). Quantitative and qualitative beta diversity measures lead to different insights into factors that structure microbial communities. *Applied and Environmental Microbiology*, 73(5), 1576–1585. <https://doi.org/10.1128/aem.01996-06>
- Lozupone, C., & Knight, R. (2005). UniFrac: A new phylogenetic method for comparing microbial communities. *Applied and Environmental Microbiology*, 71(12), 8228–8235. <https://doi.org/10.1128/AEM.71.12.8228-8235.2005>
- Marcon, E., & Herault, B. (2015). Entropart: An r package to measure and partition diversity. *Journal of Statistical Software*, 67(8), 1–26. <https://doi.org/10.18637/jss.v067.i08>
- McMurdie, P. J., & Holmes, S. (2013). Phyloseq: An r package for reproducible interactive analysis and graphics of microbiome census data. *PloS One*, 8(4), e61217. <https://doi.org/10.1371/journal.pone.0061217>
- Oksanen, J., Simpson, G. L., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., O'Hara, R. B., Solymos, P., Stevens, M. H. H., Szoecs, E., Wagner, H., Barbour, M., Bedward, M.,

- 143 Bolker, B., Borcard, D., Carvalho, G., Chirico, M., De Caceres, M., Durand, S., ... Borman,
144 T. (2025). *Vegan: Community ecology package*. [https://doi.org/10.32614/CRAN.package.](https://doi.org/10.32614/CRAN.package.vegan)
145 [vegan](https://doi.org/10.32614/CRAN.package.vegan)
- 146 Pavoine, S. (2020). Adiv: An r package to analyse biodiversity in ecology. *Methods in Ecology*
147 *and Evolution*, 11, 1106–1112. <https://doi.org/10.1111/2041-210X.13430>
- 148 Roberts, D. W. (2025). *Labdsv: Ordination and multivariate analysis for ecology*. [https:](https://doi.org/10.32614/CRAN.package.labdsv)
149 [//doi.org/10.32614/CRAN.package.labdsv](https://doi.org/10.32614/CRAN.package.labdsv)
- 150 Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical*
151 *Journal*, 27(3), 379–423. <https://doi.org/10.1002/j.1538-7305.1948.tb01338.x>
- 152 Sorenson, T. (1948). A method of establishing groups of equal amplitude in plant sociology
153 based on similarity of species content. *Kongelige Danske Videnskabernes Selskab*, 5, 1–34.

DRAFT