






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

Daniel P Smith <sup>1,2</sup>, Sara J Javornik Cregeen <sup>1,2</sup>, and Joseph F Petrosino <sup>1,2</sup>

<sup>1</sup> The Alkek Center for Metagenomics and Microbiome Research, Department of Molecular Virology and Microbiology, Baylor College of Medicine, Houston, TX 77030, USA <sup>2</sup> 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Published: unpublished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

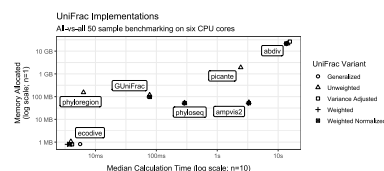
The analysis of ecological diversity in large-scale studies is often hampered by the computational demands of calculating metrics across thousands of communities, a common requirement in modern microbiome research. This is particularly true for phylogenetic metrics like Faith's PD (Faith, 1992) and the UniFrac distance family (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.

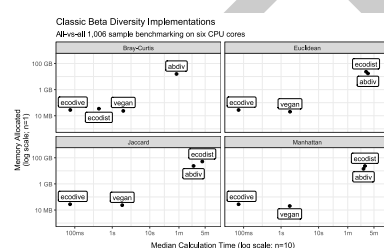
## Comparison to Existing Packages

While numerous R packages can calculate diversity metrics, our comparison focuses on those that provide their own implementations: *abdiv* (Bittinger, 2020), *adiv* (Pavoine, 2020), *ampvis2* (Andersen et al., 2018), *ecodist* (Goslee & Urban, 2007), *entropart* (Marcon & Herault, 2015), *GUniFrac* (Chen et al., 2023), *phylregion* (Daru et al., 2020), *phylseq* (McMurdie & Holmes, 2013), *picante* (Kembel et al., 2010), and *vegan* (Oksanen et al., 2025). *ecodive* sets itself apart from these packages through its superior performance. Furthermore, *ecodive* has zero external R dependencies. This makes it a lightweight, stable, and secure computational backend, minimizing installation conflicts and simplifying long-term maintenance for developers who build upon it.

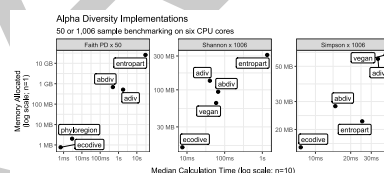
Comprehensive benchmarks, conducted using the bench package (Hester & Vaughan, 2025), demonstrate these advantages across a range of metrics (Figures 1-3). The complete benchmark code and results are available in the package vignette (vignette('benchmark')) and online.



**Figure 1:** Figure 1: UniFrac benchmarks. ecodiv demonstrates substantial performance gains for UniFrac, being 2 to 3,900x faster and using 50 - 32,000x less memory, which helps overcome computational bottlenecks in large-scale analyses.



**Figure 2:** Figure 2: Classic beta diversity benchmarks. ecodiv is 6 to 2,300x faster and uses 1 to 1,800x less memory, enabling more efficient analysis of community dissimilarities.



**Figure 3:** Figure 3: Alpha diversity benchmarks. ecodiv is 2 to 43,000x faster and uses 1 to 33,000x less memory, significantly accelerating the analysis of diversity within single samples.

## Implemented Metrics

ecodiv provides a comprehensive suite of alpha and beta diversity metrics. The current implementation includes:

### Alpha Diversity

- Classic: Shannon Index (Shannon, 1948), Simpson Index (Gini, 1912; Simpson, 1949), Inverse Simpson Index (Simpson, 1949), and Chao1 (Chao, 1984).
- Phylogenetic: Faith's Phylogenetic Diversity (Faith, 1992).

### Beta Diversity

- Classic: Bray-Curtis (Bray & Curtis, 1957; Sorenson, 1948), Canberra (Lance & Williams, 1967), Euclidean (Gower & Legendre, 1986; Legendre & Caceres, 2013), Gower (Gower, 1971; Gower & Legendre, 1986), Jaccard (Jaccard, 1908), Kulczynski (Kulczynski, 1927), and Manhattan (Kaufman & Rousseeuw, 1990).

- 55     ▪ Phylogenetic: Unweighted UniFrac (C. Lozupone & Knight, 2005), Weighted UniFrac  
56       (C. A. Lozupone et al., 2007), Normalized Weighted UniFrac (C. A. Lozupone et al.,  
57       2007), Generalized UniFrac (Chen et al., 2012), and Variance Adjusted Weighted UniFrac  
58       (Chang et al., 2011).

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

## 61     Example Usage

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

```
library(phyloseq)
data(esophagus)

ecodive::weighted_unifrac(esophagus)
#>           B           C
#> C 0.1050480
#> D 0.1401124 0.1422409
```

## 65     Acknowledgements

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

## 69     References

- 70     Andersen, K. S., Kirkegaard, R. H., Karst, S. M., & Albertsen, M. (2018). ampvis2: An r  
71     package to analyse and visualise 16S rRNA amplicon data. *bioRxiv*. <https://doi.org/10.1101/299537>
- 72     Bittinger, K. (2020). *Abdiv: Alpha and beta diversity measures*. <https://doi.org/10.32614/CRAN.package.abdiv>
- 73     Bray, J. R., & Curtis, J. T. (1957). An ordination of the upland forest communities of southern  
74     wisconsin. *Ecological Monographs*, 27(4), 325–349. <https://doi.org/10.2307/1942268>
- 75     Chang, Q., Luan, Y., & Sun, F. (2011). Variance adjusted weighted UniFrac: A powerful beta  
76     diversity measure for comparing communities based on phylogeny. *BMC Bioinformatics*,  
77     12(1). <https://doi.org/10.1186/1471-2105-12-118>
- 78     Chao, A. (1984). Non-parametric estimation of the number of classes in a population.  
79     *Scandinavian Journal of Statistics*, 11, 265–270. <https://doi.org/10.2307/4616294>
- 80     Chen, J., Bittinger, K., Charlson, E. S., Hoffmann, C., Lewis, J., Wu, G. D., Collman,  
81     R. G., Bushman, F. D., & Li, H. (2012). Associating microbiome composition with  
82     environmental covariates using generalized UniFrac distances. *Bioinformatics*, 28(16),  
83     2106–2113. <https://doi.org/10.1093/bioinformatics/bts342>
- 84     Chen, J., Zhang, X., Yang, L., & Zhang, L. (2023). *GUniFrac: Generalized UniFrac distances,*  
85     *distance-based multivariate methods and feature-based univariate methods for microbiome*  
86     *data analysis*. <https://doi.org/10.32614/CRAN.package.GUniFrac>

- 89 Daru, B. H., Karunaratne, P., & Schliep, K. (2020). Phyloregion: R package for biogeographic  
90 regionalization and macroecology. *Methods in Ecology and Evolution*, 11, 1483–1491.  
91 <https://doi.org/10.1111/2041-210X.13478>
- 92 Faith, D. P. (1992). Conservation evaluation and phylogenetic diversity. *Biological Conservation*,  
93 61, 1–10. [https://doi.org/10.1016/0006-3207\(92\)91201-3](https://doi.org/10.1016/0006-3207(92)91201-3)
- 94 Gini, C. (1912). *Variabilita e mutabilita*. Tipografia di Paolo Cuppini.
- 95 Goslee, S. C., & Urban, D. L. (2007). The ecodist package for dissimilarity-based analysis of  
96 ecological data. *Journal of Statistical Software*, 22, 1–19. <https://doi.org/10.18637/jss.v022.i07>
- 98 Gower, J. (1971). A general coefficient of similarity and some of its properties. *Biometrics*,  
99 27(4), 857–871. <https://doi.org/10.2307/2528823>
- 100 Gower, J., & Legendre, P. (1986). Metric and euclidean properties of dissimilarity coefficients.  
101 *Journal of Classification*, 3, 5–48. <https://doi.org/10.1007/BF01896809>
- 102 Hester, J., & Vaughan, D. (2025). *Bench: High precision timing of r expressions*. <https://doi.org/10.32614/CRAN.package.bench>
- 104 Jaccard, P. (1908). Nouvelles recherches sur la distribution florale. *Bulletin de La Societe Vau-*  
105 *doise Des Sciences Naturelles*, 44(163), 223–270. <https://doi.org/10.5169/seals-268384>
- 106 Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data: An introduction to cluster*  
107 *analysis*. John Wiley & Sons. <https://doi.org/10.1002/9780470316801>
- 108 Kembel, S. W., Cowan, P. D., Helmus, M. R., Cornwell, W. K., Morlon, H., Ackerly, D. D.,  
109 Blomberg, S. P., & Webb, C. O. (2010). Picante: R tools for integrating phylogenies and  
110 ecology. *Bioinformatics*, 26, 1463–1464. <https://doi.org/10.1093/bioinformatics/btq166>
- 111 Kulczynski, S. (1927). Die pflanzenassoziationen der pieninen. *Bulletin International de*  
112 *l'Académie Polonaise Des Sciences Et Des Lettres, Classe Des Sciences Mathématiques Et*  
113 *Naturelles, Série B: Sciences Naturelles*, 57–203.
- 114 Lance, G. N., & Williams, W. T. (1967). A general theory of classificatory sorting strategies II.  
115 Clustering systems. *The Computer Journal*, 10(3). <https://doi.org/10.1093/comjnl/10.3.271>
- 117 Legendre, P., & Caceres, M. (2013). Beta diversity as the variance of community data:  
118 Dissimilarity coefficients and partitioning. *Ecology Letters*, 16(8). <https://doi.org/10.1111/ele.12141>
- 120 Lozupone, C. A., Hamady, M., Kelley, S. T., & Knight, R. (2007). Quantitative and qualitative  
121 beta diversity measures lead to different insights into factors that structure microbial  
122 communities. *Applied and Environmental Microbiology*, 73(5), 1576–1585. <https://doi.org/10.1128/aem.01996-06>
- 124 Lozupone, C., & Knight, R. (2005). UniFrac: A new phylogenetic method for comparing  
125 microbial communities. *Applied and Environmental Microbiology*, 71(12), 8228–8235.  
126 <https://doi.org/10.1128/AEM.71.12.8228-8235.2005>
- 127 Marcon, E., & Herault, B. (2015). Entropart: An r package to measure and partition diversity.  
128 *Journal of Statistical Software*, 67(8), 1–26. <https://doi.org/10.18637/jss.v067.i08>
- 129 McMurdie, P. J., & Holmes, S. (2013). Phyloseq: An r package for reproducible interactive  
130 analysis and graphics of microbiome census data. *PloS One*, 8(4), e61217. <https://doi.org/10.1371/journal.pone.0061217>
- 132 Oksanen, J., Simpson, G. L., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., O'Hara,  
133 R. B., Solymos, P., Stevens, M. H. H., Szoecs, E., Wagner, H., Barbour, M., Bedward, M.,  
134 Bolker, B., Borcard, D., Carvalho, G., Chirico, M., De Caceres, M., Durand, S., ... Borman,

- 135 T. (2025). *Vegan: Community ecology package*. [https://doi.org/10.32614/CRAN.package.](https://doi.org/10.32614/CRAN.package.vegan)  
136 [vegan](https://doi.org/10.32614/CRAN.package.vegan)
- 137 Pavoine, S. (2020). Adiv: An r package to analyse biodiversity in ecology. *Methods in Ecology*  
138 *and Evolution*, 11, 1106–1112. <https://doi.org/10.1111/2041-210X.13430>
- 139 Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical*  
140 *Journal*, 27(3), 379–423. <https://doi.org/10.1002/j.1538-7305.1948.tb01338.x>
- 141 Simpson, E. H. (1949). Measurement of diversity. *Nature*, 163(4148), 688–688. [https:](https://doi.org/10.1038/163688a0)  
142 [//doi.org/10.1038/163688a0](https://doi.org/10.1038/163688a0)
- 143 Sorenson, T. (1948). A method of establishing groups of equal amplitude in plant sociology  
144 based on similarity of species content. *Kongelige Danske Videnskabernes Selskab*, 5, 1–34.

DRAFT