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 ROR ¶ Corresponding author

Summary

Characterizing the composition of biological communities is a fundamental task in ecology, but the calculations involved can be computationally prohibitive when applied to large datasets. ecodive is an R package that addresses this challenge by providing highly optimized implementations of 50 ecological metrics for 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 large datasets quickly and efficiently.

Statement of Need

A primary challenge in large-scale ecological analysis is the computational complexity of beta diversity calculations. These algorithms exhibit $O(n^2)$ complexity, meaning their computational cost scales quadratically with the number of samples (n). As microbiome and ecological studies grow to include thousands of samples, this quadratic scaling creates a significant bottleneck, demanding immense processing time and memory.

A second challenge is the fragmentation of diversity metrics across numerous R packages. Researchers often need to install and manage a suite of dependencies to access the full range of metrics required for a comprehensive analysis, leading to potential version conflicts and a disjointed workflow.

ecodive addresses both of these critical needs. First, it provides a highly optimized, parallelized C-based engine that dramatically reduces the time and memory required by these algorithms, enabling the analysis of much larger datasets. Second, it consolidates a vast collection of alpha and beta diversity metrics into a single, dependency-free package. By solving the dual problems of computational inefficiency and methodological fragmentation, 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),

39 parallelDist (Eckert, 2022), philentropy (HG, 2018), phyloregion (Daru et al., 2020),
40 phyloseq (McMurdie & Holmes, 2013), picante (Kembel et al., 2010), tabula (Frerebeau,
41 2019), and vegan (Oksanen et al., 2025). The results, conducted using the bench R package,
42 are summarized in Figure 1 and demonstrate ecodive's superior speed and memory efficiency
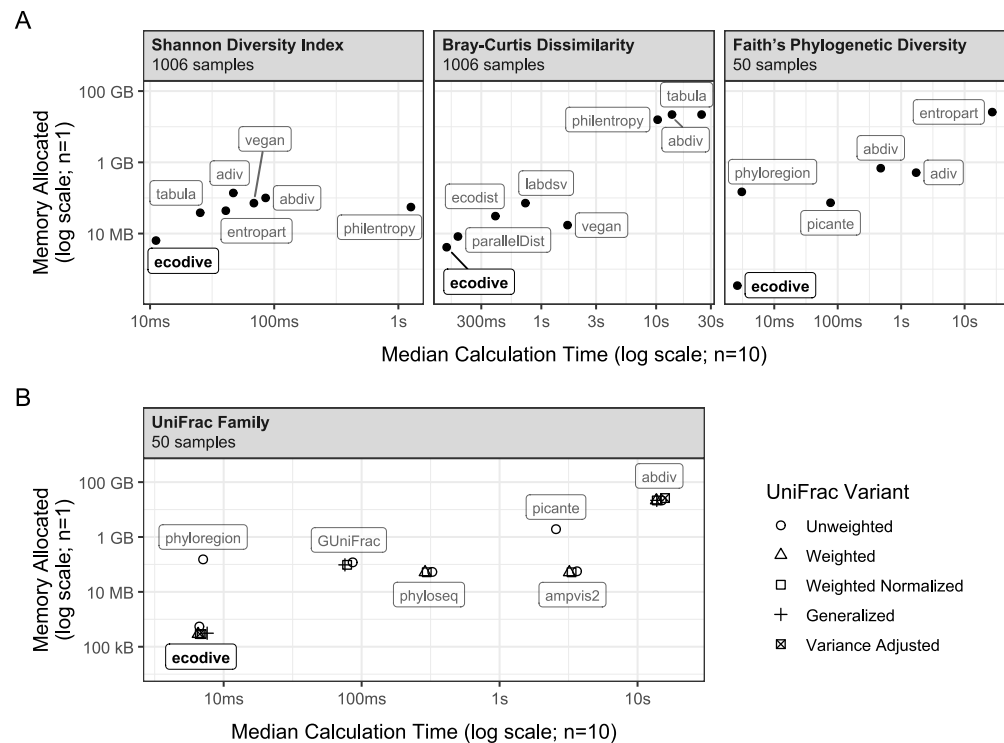43 for each of the metrics tested.

Figure 1: 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.

44 Crucially, the benchmark suite also confirms these performance gains do not come at the
45 cost of accuracy, as ecodive produces numerically identical output to the other packages.
46 Beyond its computational advantages, ecodive has zero external R dependencies. This makes
47 it a lightweight, stable, and secure backend, minimizing installation conflicts and simplifying
48 long-term maintenance for developers who build upon it. The complete benchmark code and
49 results are available in the package vignette (vignette('benchmark')) and online.

Implemented Metrics

51 ecodive stands out by offering an extensive and diverse collection of 50 metrics for both alpha
52 and beta diversity analysis, making it a uniquely comprehensive tool. It provides researchers with
53 a wide array of both traditional and phylogeny-aware algorithms in a single, high-performance
54 package. The suite of alpha diversity metrics includes staples like the Shannon Diversity Index
55 (Shannon, 1948) and Chao1 (Chao, 1984), important estimators such as the Abundance-based
56 Coverage Estimator (ACE) (Chao & Lee, 1992) and Fisher's Alpha (Fisher et al., 1943), and
57 key phylogeny-aware metrics like Faith's Phylogenetic Diversity (Faith, 1992), 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 (Q. Chang et al., 2011; Chen et al., 2012; C. A. Lozupone et al., 2007; C. Lozupone & Knight, 2005), and the Aitchison distance (Aitchison, 1982) for compositional data analysis. This extensive collection allows for a thorough and multi-faceted analysis of community structure.

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

Programmatic Use and API

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

Example Usage

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

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

Acknowledgements

References

Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society. Series B (Methodological)*, *44*(2), 139–177. https://doi.org/10.1111/j.2517-6161.1982.tb01195.x

Andersen, K. S., Kirkegaard, R. H., Karst, S. M., & Albertsen, M. (2018). ampvis2: An r package to analyse and visualise 16S rRNA amplicon data. *bioRxiv*. https://doi.org/10.1101/299537

Bittinger, K. (2020). *Abdiv: Alpha and beta diversity measures*. https://doi.org/10.32614/CRAN.package.abdiv

Bray, J. R., & Curtis, J. T. (1957). An ordination of the upland forest communities of southern wisconsin. *Ecological Monographs*, *27*(4), 325–349. https://doi.org/10.2307/1942268

94 Chang, Q., Luan, Y., & Sun, F. (2011). Variance adjusted weighted UniFrac: A powerful beta
95   diversity measure for comparing communities based on phylogeny. *BMC Bioinformatics*,
96   *12*(1). https://doi.org/10.1186/1471-2105-12-118

97 Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson,
98   J., Dipert, A., & Borges, B. (2024). *Shiny: Web application framework for r*. https:
99   //doi.org/10.32614/CRAN.package.shiny

100 Chao, A. (1984). Non-parametric estimation of the number of classes in a population.
101   *Scandinavian Journal of Statistics*, *11*, 265–270. https://www.jstor.org/stable/4616294

102 Chao, A., & Lee, S.-M. (1992). Estimating the number of classes via sample coverage. *Journal*
103   *of the American Statistical Association*, *87*(417), 210–217. https://doi.org/10.1080/
104   01621459.1992.10475194

105 Chen, J., Bittinger, K., Charlson, E. S., Hoffmann, C., Lewis, J., Wu, G. D., Collman,
106   R. G., Bushman, F. D., & Li, H. (2012). Associating microbiome composition with
107   environmental covariates using generalized UniFrac distances. *Bioinformatics*, *28*(16),
108   2106–2113. https://doi.org/10.1093/bioinformatics/bts342

109 Chen, J., Zhang, X., Yang, L., & Zhang, L. (2023). *GUniFrac: Generalized UniFrac distances,*
110   *distance-based multivariate methods and feature-based univariate methods for microbiome*
111   *data analysis*. https://doi.org/10.32614/CRAN.package.GUniFrac

112 Daru, B. H., Karunarathne, P., & Schliep, K. (2020). Phyloregion: R package for biogeographic
113   regionalization and macroecology. *Methods in Ecology and Evolution*, *11*, 1483–1491.
114   https://doi.org/10.1111/2041-210X.13478

115 Eckert, A. (2022). *parallelDist: Parallel distance matrix computation using multiple threads*.
116   https://doi.org/10.32614/CRAN.package.parallelDist

117 Faith, D. P. (1992). Conservation evaluation and phylogenetic diversity. *Biological Conservation*,
118   *61*, 1–10. https://doi.org/10.1016/0006-3207(92)91201-3

119 Fisher, R. A., Corbet, A. S., & Williams, C. B. (1943). The relation between the number of
120   species and the number of individuals in a random sample of an animal population. *Journal*
121   *of Animal Ecology*, *12*(1), 42–58. https://doi.org/10.2307/1411

122 Frerebeau, N. (2019). Tabula: An r package for analysis, seriation, and visualization of
123   archaeological count data. *Journal of Open Source Software*, *4*(44), 1821. https://doi.
124   org/10.21105/joss.01821

125 Goslee, S. C., & Urban, D. L. (2007). The ecodist package for dissimilarity-based analysis of
126   ecological data. *Journal of Statistical Software*, *22*, 1–19. https://doi.org/10.18637/jss.
127   v022.i07

128 HG, D. (2018). Philentropy: Information theory and distance quantification with r. *Journal of*
129   *Open Source Software*, *3*(26), 765. https://doi.org/10.21105/joss.00765

130 Kembel, S. W., Cowan, P. D., Helmus, M. R., Cornwell, W. K., Morlon, H., Ackerly, D. D.,
131   Blomberg, S. P., & Webb, C. O. (2010). Picante: R tools for integrating phylogenies and
132   ecology. *Bioinformatics*, *26*, 1463–1464. https://doi.org/10.1093/bioinformatics/btq166

133 Lozupone, C. A., Hamady, M., Kelley, S. T., & Knight, R. (2007). Quantitative and qualitative
134   beta diversity measures lead to different insights into factors that structure microbial
135   communities. *Applied and Environmental Microbiology*, *73*(5), 1576–1585. https://doi.
136   org/10.1128/aem.01996-06

137 Lozupone, C., & Knight, R. (2005). UniFrac: A new phylogenetic method for comparing
138   microbial communities. *Applied and Environmental Microbiology*, *71*(12), 8228–8235.
139   https://doi.org/10.1128/AEM.71.12.8228-8235.2005

140 Marcon, E., & Herault, B. (2015). Entropart: An r package to measure and partition diversity.

¹⁴¹ *Journal of Statistical Software*, *67*(8), 1–26. https://doi.org/10.18637/jss.v067.i08

¹⁴² McMurdie, P. J., & Holmes, S. (2013). Phyloseq: An r package for reproducible interactive
¹⁴³ analysis and graphics of microbiome census data. *PloS One*, *8*(4), e61217. https://doi.
¹⁴⁴ org/10.1371/journal.pone.0061217

¹⁴⁵ Oksanen, J., Simpson, G. L., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., O'Hara,
¹⁴⁶ R. B., Solymos, P., Stevens, M. H. H., Szoecs, E., Wagner, H., Barbour, M., Bedward, M.,
¹⁴⁷ Bolker, B., Borcard, D., Carvalho, G., Chirico, M., De Caceres, M., Durand, S., … Borman,
¹⁴⁸ T. (2025). *Vegan: Community ecology package*. https://doi.org/10.32614/CRAN.package.
¹⁴⁹ vegan

¹⁵⁰ Pavoine, S. (2020). Adiv: An r package to analyse biodiversity in ecology. *Methods in Ecology
¹⁵¹ and Evolution*, *11*, 1106–1112. https://doi.org/10.1111/2041-210X.13430

¹⁵² Roberts, D. W. (2025). *Labdsv: Ordination and multivariate analysis for ecology*. https:
¹⁵³ //doi.org/10.32614/CRAN.package.labdsv

¹⁵⁴ Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical
¹⁵⁵ Journal*, *27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

¹⁵⁶ Sorenson, T. (1948). A method of establishing groups of equal amplitude in plant sociology
¹⁵⁷ based on similarity of species content. *Kongelige Danske Videnskabernes Selskab*, *5*, 1–34.