The Journal of Open Source Software

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: 10.21105/joss.09777

Software

- Review
- Repository
- Archive

Editor: Nick Golding

Reviewers:

- @cstawitz
- @CosteaPaul

Submitted: 22 September 2025
Published: 29 May 2026

Summary

Modern ecology relies on quantifying the biological complexity of communities across diverse scales, from microscopic human microbiomes to vast forest ecosystems. Researchers use diversity metrics to describe the variety within a site (alpha diversity) and the compositional differences between sites (beta diversity). `ecodive` is an R package built to compute these metrics with high efficiency. It provides a comprehensive, unified toolset that allows researchers to analyze large-scale datasets that were previously computationally prohibitive. By integrating optimized C-based algorithms with a parallel processing engine, `ecodive` enables rapid, high-throughput ecological analysis through a lightweight, portable framework.

Statement of Need

A primary challenge in modern ecological analysis is the management of high-dimensional data. As sequencing technologies improve, datasets are growing to include thousands of samples and tens of thousands of unique taxa. Beta diversity calculations, which involve comparing every sample to every other sample, exhibit $O(n^2)$ complexity. This quadratic scaling creates a significant bottleneck; doubling a dataset's size quadruples the processing time and memory requirements, frequently exceeding the capacity of standard research workstations.

Furthermore, the software landscape for ecological metrics is fragmented. A researcher needing to calculate a specific set of indices - for example, Faith's Phylogenetic Diversity, Bray-Curtis dissimilarity, and UniFrac distances - often must install and manage multiple R packages (`picante`, `vegan`, `GUniFrac`), each with different dependencies, input formats, and performance limitations.

`ecodive` solves these problems by providing a centralized, high-performance library. It targets ecologists, microbiologists, and bioinformaticians who require a robust, dependency-free solution for diversity analysis. By consolidating 50 standard metrics into a single, optimized framework, it eliminates the need for "package hopping" and enables the analysis of massive datasets on standard hardware.

State of the Field

While several R packages support diversity analysis, `ecodive` offers a unique scholarly contribution by bridging the gap between specialized high-performance tools and comprehensive ecological libraries. As shown in Table 1, the current landscape is defined by three primary trade-offs: analytical breadth, computational efficiency, and infrastructure overhead.

Table 1: Comparison of community diversity metrics and architectural features across 16 R packages. Metrics were identified using a large language model protocol on package reference manuals to categorize metrics and filter for symmetry and ecological relevance. Alpha: Number of alpha diversity metrics; Beta: Number of beta diversity metrics; Uni: ✓ indicates UniFrac support; Par: Parallelized calculation of metrics; Cmp: Implementation uses compiled code (e.g. C/C++); Dep: Transitive count of hard R dependencies. Full methodology and LLM prompts are available in ecodive's benchmarking documentation.

R Package	Alpha	Beta	Uni	Par	Cmp	Dep	Citation
`ecodive`	14	36	✓	✓	✓	0	This work
`abdiv`	17	48	✓	—	—	5	Bittinger (2020)
`adiv`	42	58	—	—	—	97	Pavoine (2020)
`ampvis2`	6	25	✓	—	—	75	Andersen et al. (2018)
`ecodist`	0	9	—	—	✓	10	Goslee & Urban (2007)
`entropart`	11	9	—	—	—	82	Marcon & Herault (2015)
`GUniFrac`	0	4	✓	—	✓	47	Chen et al. (2023)
`labdsv`	2	7	—	—	✓	8	Roberts (2005)
`OmicFlow`	6	9	✓	✓	✓	90	Gusinac et al. (2025)
`parallelDist`	0	26	—	✓	✓	2	Eckert (2017)
`philentropy`	1	43	—	✓	✓	4	Drost (2018)
`phyloregion`	9	10	✓	—	—	64	Daru et al. (2020)
`phyloseq`	7	43	✓	✓	—	54	McMurdie & Holmes (2013)
`picante`	14	9	✓	—	—	11	Kembel et al. (2010)
`tabula`	16	12	—	—	—	2	Frerebeau (2019)
`vegan`	17	52	—	—	✓	7	Oksanen et al. (2001)

1. Analytical Breadth vs. Computational Scalability

The ecological software landscape often requires researchers to choose between a wide variety of metrics and the speed necessary for high-dimensional community data. As shown in Table 1, packages like `adiv` and `abdiv` offer significant analytical breadth, providing 100 and 65 total metrics, respectively. However, both lack parallelized execution (Par) and compiled backend implementations (Cmp), which limits their utility for large-scale datasets. `ecodive` overcomes this limitation by providing a comprehensive suite of 50 metrics - comparable in scope to the foundational `vegan` package (69 metrics) - while utilizing a parallelized C engine to ensure scalability.

2. Domain-Specific Functionality and UniFrac Support

General-purpose high-performance libraries often fail to address specific ecological needs, such as phylogenetic awareness. For instance, despite their high metric counts, neither `vegan` (69 metrics), `adiv` (100 metrics), nor `philentropy` (44 metrics) provide UniFrac support. While specialized tools like `picante` and `GUniFrac` do include these indices, they operate serially and carry significant dependency burdens. `ecodive` is unique in combining extensive general metrics with parallelized UniFrac support, functioning as a unified tool for both phylogenetic and non-phylogenetic analyses without sacrificing performance.

3. Infrastructure Stability and Dependency Management

A critical but often overlooked limitation in existing software is the "transitive burden" of recursive dependencies. Table 1 reveals significant infrastructure overhead across the field: `adiv` requires 97 hard dependencies, `OmicFlow` requires 90, and `phyloregion` requires 64. Such high counts increase maintenance complexity and supply chain vulnerability for research pipelines. `ecodive` addresses this by delivering its full feature set - including parallelization and

compiled performance - with zero external R dependencies. This zero-dependency architecture ensures high community readiness and long-term stability for ecological research.

The ecodive Contribution

ecodive offers a unique scholarly contribution by bridging these gaps into a single framework. It implements 50 symmetric metrics chosen specifically for their relevance to ecological ordination. Built on a parallelized C engine, the package matches or exceeds the performance of specialized tools while ensuring that results remain numerically identical to established packages. This performance is delivered within a zero-dependency architecture, which minimizes maintenance overhead and protects against the supply chain vulnerabilities common in complex bioinformatic pipelines.

Software Design

The architecture of ecodive balances the user-friendly conventions of R with the raw performance of C. A critical design trade-off centered on data representation.

Most R users work with dense matrices where samples are rows and features are columns. Standard R functions like `dist()` expect this format. However, ecological matrices are typically 90-99% zeros (sparse). Storing them as dense matrices wastes gigabytes of RAM, and processing them row-by-row is cache-inefficient for many distance algorithms.

To address this, ecodive maintains the standard R interface (samples-as-rows) but fundamentally alters the backend data structure:

1. **Transparent Conversion:** When a standard matrix is passed to ecodive, it is internally converted into a column-compressed sparse matrix (`dgCMatrix`) with samples transposed to columns. This incurs a one-time overhead but allows the C engine to skip zeros entirely and access memory in a cache-friendly, column-major pattern.
2. **Transformation Bypass:** For extremely large datasets where the overhead of this transformation is non-trivial, users can manually provide data in the native `dgCMatrix` format (samples as columns). ecodive detects this optimized state and bypasses the transformation step, operating directly on the existing C pointers. This allows for "zero-copy" analysis of massive datasets.
3. **Parallelization Strategy:** ecodive employs a direct implementation using the standard POSIX threads (`pthreads`) library, avoiding the memory duplication overhead of forking processes found in R's `parallel` package. This design enables fine-grained, dynamic load balancing, ensuring efficient execution even when calculating partial distance matrices.

Research Impact Statement

ecodive has demonstrated immediate utility in high-dimensional microbiome studies. The core C algorithms in ecodive were originally developed for and deployed in the `rbiom` package (Smith, 2020). As part of `rbiom`, these optimized metrics have already been utilized in diverse microbial ecology studies, including research on preterm infant microbiomes (Ahearn-Ford et al., 2025), dietary interventions (DiMattia et al., 2025), and relationship satisfaction (Cheng et al., 2023). ecodive extracts these proven, high-performance components into a standalone, lightweight library to make them accessible to the broader R ecosystem without `rbiom`'s specific visualization and data structure dependencies.

To evaluate ecodive under realistic research conditions, we conducted benchmarks on a workstation with a 6-core Intel i5-9600K CPU and 64GB RAM running Windows 11. While several packages leveraged all six cores, our results demonstrate that algorithmic efficiency and sparse data handling are as critical as raw multi-threading.

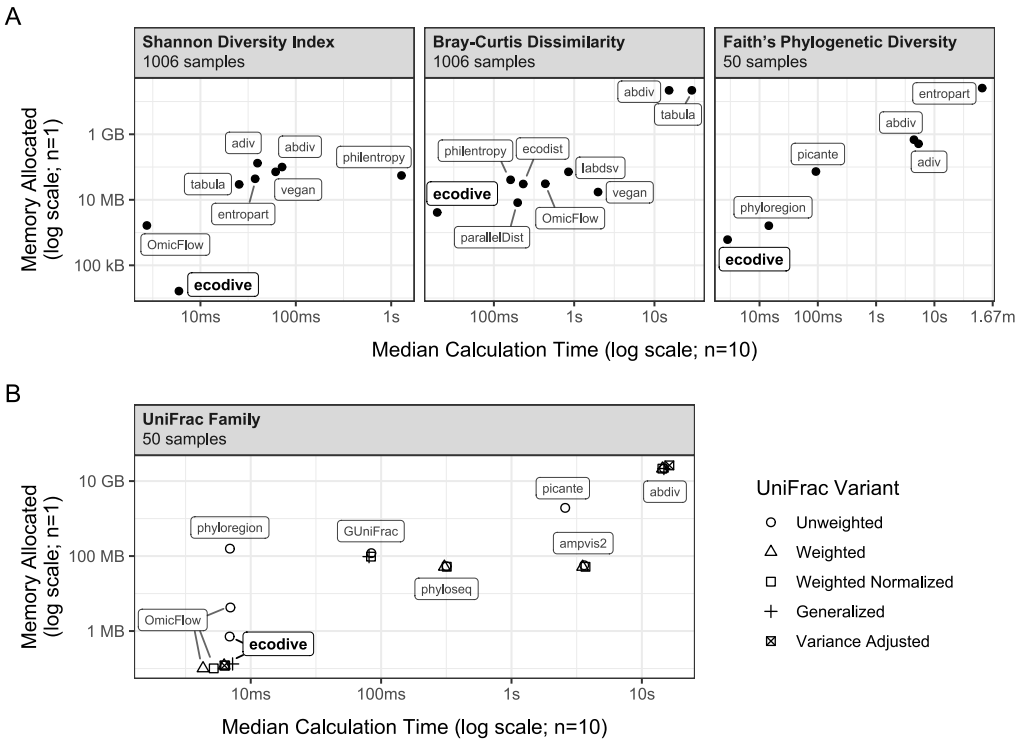

Figure 1: Benchmarking results. Execution time (x-axis) vs. peak memory usage (y-axis) for various diversity metrics across 16 R packages. `ecodive` (highlighted) consistently occupies the bottom-left quadrant, indicating high speed and low memory footprint. Note the log scale on both axes.

As illustrated in Figure 1, `ecodive` consistently occupies the bottom-left quadrant across all metrics, indicating a simultaneous optimization of execution time and memory footprint:

- **Performance:** For the widely-used Unweighted UniFrac metric ($N = 50$ samples), `ecodive` completed calculations in 6.9ms, approximately 400x faster than `picante` (2.6s) and 50x faster than `phyloseq` (320ms).
- **Scalability:** When processing Bray-Curtis dissimilarity for large datasets ($N = 1006$ samples), `ecodive` required only 19ms, whereas vegan required 2.0s (~100x faster).
- **Memory Efficiency:** The sparse architecture reduces memory allocation from gigabytes to megabytes, enabling the analysis of massive UniFrac matrices on standard laptops that would otherwise require high-performance computing clusters.

By operationalizing these performance gains within a stable, high-performance framework, `ecodive` bridges the gap between specialized high-performance computing and everyday ecological research workflows.

The package is available for installation via CRAN and Conda-Forge, supported by comprehensive vignettes that guide users through metric selection and performance tuning. This distribution model, combined with a zero-dependency architecture, ensures high community readiness and long-term stability for ecological software pipelines.

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:

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

AI Usage Disclosure

Generative AI tools (Google Gemini) were used to assist in the drafting and revision of this manuscript and the generation of documentation. No AI tools were used to write the functional source code (R or C) of the software. All AI-generated text was critically reviewed, verified for accuracy, and edited by the authors.

Acknowledgements

This study was supported by NIH/NIAD (Grant number U19 AI144297), and Baylor College of Medicine and Alkek Foundation Seed.

References

Ahearn-Ford, S., Kakaroukas, A., Young, G. R., Nelson, A., Abrahamse-Berkeveld, M., Elburg, R. M. van, Smith, D., Berrington, J. E., Embleton, N. D., & Stewart, C. J. (2025). Spatiotemporal development of late and moderate preterm infant gut and oral microbiomes and impact of gestational age on early colonization. *mSystems*, *10*(12). https://doi.org/10.1128/msystems.00667-25

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

Bittinger, K. (2020). abdiv: Alpha and beta diversity measures. In *CRAN: Contributed Packages*. The R Foundation. https://doi.org/10.32614/cran.package.abdiv

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. In *CRAN: Contributed Packages*. The R Foundation. https://doi.org/10.32614/CRAN.package.GUniFrac

Cheng, Q., Krajmalnik-Brown, R., DiBaise, J. K., Maldonado, J., Guest, M. A., Todd, M., & Langer, S. L. (2023). Relationship functioning and gut microbiota composition among older adult couples. *International Journal of Environmental Research and Public Health*, *20*(8), 5435. https://doi.org/10.3390/ijerph20085435

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

DiMattia, Z. S., Zhao, J., Hao, F., Koshkin, S., Bisanz, J. E., Patterson, A. D., Fleming, J. A., Kris-Etherton, P. M., & Petersen, K. S. (2025). Effect of varying quantities of lean beef as part of a mediterranean-style dietary pattern on gut microbiota and plasma, fecal, and urinary metabolites: A randomized crossover controlled feeding trial. *Journal of the American Heart Association*, *14*(19). https://doi.org/10.1161/jaha.125.041063

Drost, H.-G. (2018). Philentropy: Information theory and distance quantification with R. *Journal of Open Source Software*, *3*(26), 765. https://doi.org/10.21105/joss.00765

Eckert, A. (2017). parallelDist: Parallel distance matrix computation using multiple threads. In *CRAN: Contributed Packages*. The R Foundation. https://doi.org/10.32614/cran.package.

paralleldist

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

Gusinac, A., Ederveen, T., & Boleij, A. (2025). OmicFlow: Fast and efficient (automated) analysis of sparse omics data. In *CRAN: Contributed Packages*. The R Foundation. https://doi.org/10.32614/cran.package.omicflow

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

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., Bolker, B., Borcard, D., Carvalho, G., Chirico, M., De Caceres, M., Durand, S., … Borman, T. (2001). vegan: Community ecology package. In *CRAN: Contributed Packages*. The R Foundation. https://doi.org/10.32614/CRAN.package.vegan

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

Roberts, D. W. (2005). labdsv: Ordination and multivariate analysis for ecology. In *CRAN: Contributed Packages*. The R Foundation. https://doi.org/10.32614/cran.package.labdsv

Smith, D. P. (2020). rbiom: Read/write, analyze, and visualize "BIOM" data. In *CRAN: Contributed Packages*. The R Foundation. https://doi.org/10.32614/cran.package.rbiom