

¹ ecodive: Parallel and Memory-Efficient R Package for Ecological Diversity Analysis

³ Daniel P Smith  ^{1,2}¶, Sara J Javornik Cregeen  ^{1,2}, and Joseph F Petrosino  ^{1,2}

⁵ 1 The Alkek Center for Metagenomics and Microbiome Research, Department of Molecular Virology and
⁶ Microbiology, Baylor College of Medicine, Houston, TX 77030, USA ² Department of Molecular Virology
⁷ 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Published: unpublished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](#))

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; a dataset that doubles in size requires four times the processing power, often overwhelming standard desktop computers.

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.

³⁵ State of the Field

³⁶ Several R packages exist for diversity analysis, but ecodive offers a unique contribution through
³⁷ its scope and performance. The standard package for community ecology, `vegan` ([Oksanen et al., 2001](#)), provides excellent implementations of non-phylogenetic metrics (e.g., Bray-Curtis)

but lacks phylogenetic awareness (e.g., UniFrac). Conversely, packages like picante (Kembel et al., 2010) and GUniFrac (Chen et al., 2023) specialize in phylogenetic metrics but do not offer a comprehensive suite of general-purpose indices. The phyloseq (McMurdie & Holmes, 2013) package wraps many of these tools but relies on their underlying, often serial, implementations.

More generalized packages like phlentropy (Drost, 2018) offer an impressive breadth of distinct distance measures. However, phlentropy focuses on information theory and general probability distributions rather than ecology. Critically, it includes many asymmetric divergences (where distance $A \rightarrow B \neq B \rightarrow A$) which, while mathematically valuable, are unsuitable for standard ecological ordination methods like PCoA that require symmetric distance matrices. Furthermore, phlentropy lacks critical domain-specific metrics such as the UniFrac family and alpha diversity richness estimators like Chao1.

ecodive builds upon this landscape by unifying these distinct domains. It implements symmetric metrics chosen specifically for their relevance to ecological ordination and analysis. Crucially, unlike the serial R or C implementations found in most peer packages (see **Research Impact**), ecodive is built entirely on a parallelized C engine, providing orders-of-magnitude faster performance while ensuring numerical identity with established tools.

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 Transformation:** 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. **Power User 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.

86 In benchmarks comparing 15 ecological R packages, ecodive consistently ranked as the fastest
 87 and most memory-efficient solution:

- 88 ▪ **Speed:** For the widely-used Unweighted UniFrac metric ($N = 50$), ecodive completed
 89 calculations in 6.4ms, compared to 2.5s for picante (396x faster) and 297ms for phyloseq
 90 (46x faster).
- 91 ▪ **Scalability:** For standard Bray-Curtis dissimilarity ($N = 1006$), ecodive processed the
 92 matrix in \sim 20ms, whereas vegan required 1.68s.
- 93 ▪ **Memory:** ecodive's sparse architecture reduced memory allocation for large operations
 94 from gigabytes (in abdiv or tabula) to megabytes, enabling analyses on laptops that
 95 previously required clusters.

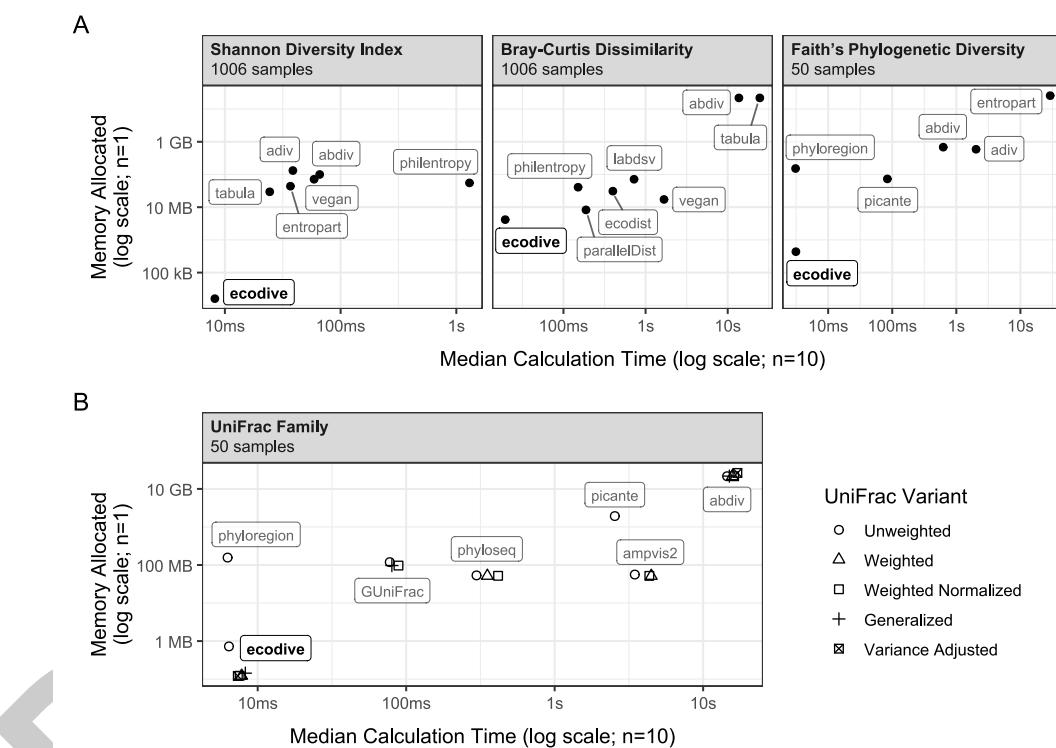


Figure 1: Benchmarking results. Execution time (x-axis) vs. peak memory usage (y-axis) for various diversity metrics across 15 R packages. ecodive (highlighted) consistently occupies the bottom-left quadrant, indicating high speed and low memory footprint. Note the log scale on both axes.

96 The package is fully documented with vignettes covering performance tuning and metric
 97 selection, and is available for installation with zero external R dependencies, ensuring high
 98 community readiness and long-term stability.

99 Example Usage

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

```
103 data(esophagus, package = 'phyloseq')
104 ecodive::weighted_unifrac(esophagus)
105 #>          B          C
```

#> C 0.1050480
#> D 0.1401124 0.1422409

103 AI Usage Disclosure

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

108 Acknowledgements

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

111 References

- 112 Ahearn-Ford, S., Kakaroukas, A., Young, G. R., Nelson, A., Abrahamse-Berkeveld, M.,
113 Elburg, R. M. van, Smith, D., Berrington, J. E., Embleton, N. D., & Stewart, C. J.
114 (2025). Spatiotemporal development of late and moderate preterm infant gut and oral
115 microbiomes and impact of gestational age on early colonization. *mSystems*, 10(12).
116 <https://doi.org/10.1128/msystems.00667-25>
- 117 Chen, J., Zhang, X., Yang, L., & Zhang, L. (2023). *GUniFrac: Generalized UniFrac distances,*
118 *distance-based multivariate methods and feature-based univariate methods for microbiome*
119 *data analysis.* <https://doi.org/10.32614/CRAN.package.GUniFrac>
- 120 Cheng, Q., Krajmalnik-Brown, R., DiBaise, J. K., Maldonado, J., Guest, M. A., Todd, M.,
121 & Langer, S. L. (2023). Relationship functioning and gut microbiota composition among
122 older adult couples. *International Journal of Environmental Research and Public Health*,
123 20(8), 5435. <https://doi.org/10.3390/ijerph20085435>
- 124 DiMattia, Z. S., Zhao, J., Hao, F., Koshkin, S., Bisanz, J. E., Patterson, A. D., Fleming, J.
125 A., Kris-Etherton, P. M., & Petersen, K. S. (2025). Effect of varying quantities of lean
126 beef as part of a mediterranean-style dietary pattern on gut microbiota and plasma, fecal,
127 and urinary metabolites: A randomized crossover controlled feeding trial. *Journal of the*
128 *American Heart Association*, 14(19). <https://doi.org/10.1161/jaha.125.041063>
- 129 Drost, H.-G. (2018). Philentropy: Information theory and distance quantification with r.
130 *Journal of Open Source Software*, 3(26), 765. <https://doi.org/10.21105/joss.00765>
- 131 Kembel, S. W., Cowan, P. D., Helmus, M. R., Cornwell, W. K., Morlon, H., Ackerly, D. D.,
132 Blomberg, S. P., & Webb, C. O. (2010). Picante: R tools for integrating phylogenies and
133 ecology. *Bioinformatics*, 26, 1463–1464. <https://doi.org/10.1093/bioinformatics/btq166>
- 134 McMurdie, P. J., & Holmes, S. (2013). Phyloseq: An r package for reproducible interactive
135 analysis and graphics of microbiome census data. *PloS One*, 8(4), e61217. <https://doi.org/10.1371/journal.pone.0061217>
- 137 Oksanen, J., Simpson, G. L., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., O'Hara,
138 R. B., Solymos, P., Stevens, M. H. H., Szoecs, E., Wagner, H., Barbour, M., Bedward, M.,
139 Bolker, B., Borcard, D., Carvalho, G., Chirico, M., De Caceres, M., Durand, S., ... Borman,
140 T. (2001). Vegan: Community ecology package. In *CRAN: Contributed Packages*. The R
141 Foundation. <https://doi.org/10.32614/CRAN.package.vegan>
- 142 Smith, D. P. (2020). Rbiom: Read/write, analyze, and visualize “BIOM” data. In *CRAN:*
143 *Contributed Packages*. The R Foundation. <https://doi.org/10.32614/cran.package.rbiom>