hilldiv: an R package for the integral analysis of diversity based on Hill numbers

Antton Alberdi[1] and M Thomas P Gilbert[1,2]

[1]Section for Evolutionary Genomics, Department of Biology, University of Copenhagen, 1350 Copenhagen, Denmark.
[2]NTNU University Museum, N-7491 Trondheim, Norway.

Correspondence:
Antton Alberdi
antton.alberdi@snm.ku.dk
Øster Farimagsgade 5, KH7, DK-1353 Copenhagen, Denmark

Running headline: R package hilldiv
Key words: biodiversity, dissimilarity, diversity index, phylodiversity, R library, Shannon, Simpson, Unifrac

Abstract

1. Hill numbers provide a powerful framework for measuring, estimating, comparing and partitioning the diversity of biological systems as characterised using high throughput DNA sequencing approaches. However, the implementation of Hill numbers in this context remains limited.
2. In order to facilitate the implementation of Hill numbers into such analyses, whether focussing on diet reconstruction, microbial community profiling or more general ecosystem characterisation analyses, we present a new R package. 'Hilldiv' provides a set of functions to assist analysis of diversity based on Hill numbers, using OTU tables and associated phylogenetic trees as inputs.
3. Multiple functionalities of the library are introduced, including (phylo)diversity measurement, (phylo)diversity profile plotting, (phylo)diversity comparison between samples and groups, (phylo)diversity partitioning and (dis)similarity measurement. All of these are grounded in abundance-based and incidence-based Hill numbers.
4. We provide a supplement that demonstrates the package's main functions using an OTU table representing the diet of eight bat species. The package can be directly installed from github (https://github.com/anttonalberdi/hilldiv).

Introduction

Tools for analysing diversity lie at the core of molecular ecology. For example, whether profiling dietary content, microorganism communities or any other bulk samples using DNA, researchers routinely need to compare diversity between samples, partition diversity between different hierarchical level, and/or compute (dis)similarity measures between samples. Although a wide repertoire of metrics have been developed to perform such operations, there is an increased awareness of the need to use general statistical frameworks to generate results that are more consistent and more easily interpretable (Jost, 2006; Chao, Chiu, & Jost, 2010).

One such general statistical framework that can enable diversity analysis is that developed around the so-called 'Hill numbers' (Hill, 1973; Jost, 2006). This framework provides a robust toolset with which to perform the most common operations researchers routinely use when analysing the diversity of biological systems, and include, among others, diversity and phylogenetic diversity measurement and estimation, diversity partitioning, and (dis)similarity computation (Alberdi & Gilbert, 2019). Despite the practicality of Hill numbers, they remain remarkably unexploited by molecular ecologists. One underlying reason might simply be the lack of easy-to-use tools with which to implement Hill numbers based diversity analysis onto the OTU tables that lie at the heart of most DNA-based diversity data sets. R packages that incorporate functions to perform some diversity analyses based on Hill numbers already exist, including *entropart* (Marcon & Hérault, 2015), *vegan* (Oksanen et al., 2013), *vegetarian* (Charney & Record, 2012), and *simba* (Jurasinski, 2007), however to the best of our knowledge, no library dedicated to DNA-derived data sets exists, that compiles functions for the integral analysis and visualisation of diversity based on Hill numbers.

To meet this need, here we present 'hilldiv', an R package that encompasses different functions with which to perform a number of diversity analyses using Hill numbers. Although it's potential uses are wider, 'hilldiv' is primarily designed for diversity analyses of molecularly (e.g. metabarcoding) characterised datasets, as reflected in the terminology used. The package includes functions for (phylo)diversity measurement, (phylo)diversity profile plotting, (phylo)diversity comparison between samples and groups, (phylo)diversity partitioning and (dis)similarity measurement (Fig. 1).

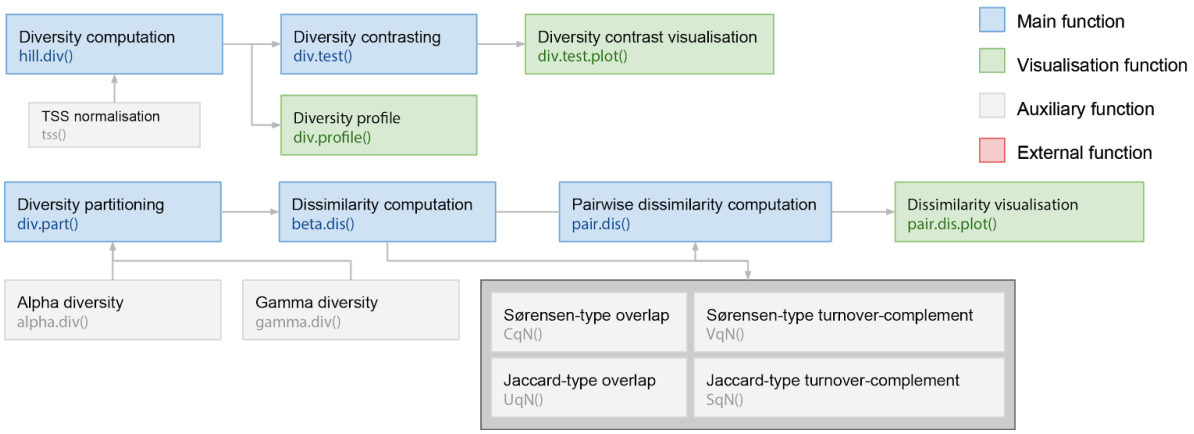

Figure 1. **Schematic representation** of the nature and relations across functions included in the package hilldiv.

Statistical background

The statistical framework developed around Hill numbers encompasses many of the most broadly employed diversity (e.g. richness, Shannon index, Simpson index), phylogenetic diversity (e.g. Faith's PD, Allen's H, Rao's quadratic entropy) and dissimilarity (e.g. Sørensen index, Unifrac distances) metrics (Chiu, Jost, & Chao, 2014). This enables the most common analyses of diversity to be performed while grounded in a single statistical framework, and provides a number of benefits in comparison to the use of each of other metrics separately.

First, Hill numbers meet the so-called doubling property, which means that when doubling the number of OTUs in a system while maintaining the rest of the parameters (e.g. evenness, phylogenetic relations between OTUs), then the diversity measured is also doubled (Chao et al., 2010). This is a property that richness owns, but Shannon and Simpson indices for instance do not (Jost, 2006).

Second, the interpretation of both the measure and measurement unit is consistent for each type of data. The basic Hill numbers expression yields a diversity measure in "effective number of OTUs", i.e. the number of equally abundant OTUs that would be needed to give the same value of diversity. This contrasts, for instance, with Shannon and Simpson indices (Shannon, 1948), which yield uncertainty and probability values, respectively (Jost, 2006).

Third, the sensitivity towards abundant and rare OTUs can be modulated with a single parameter, namely the order of diversity (q). When a diversity of order one (q=1) is used, the OTU relative abundances are weighed as their original values. However, when a diversity of order $q<1$ is used , rare OTUs are overweighed. When taken to its extreme (diversity of order set to zero, q=0), relative abundances are not considered at all, and the data simply reflects presence/absence. In contrast, when orders of diversity of $q>1$ are used, abundant OTUs are overweighed. Three q values are particularly relevant, for their close relationship to popular

diversity indices. The Hill number of q=0 yields a richness value, the Hill number of q=1 is the exponential of the Shannon index, and the Hill number of q=2 is the multiplicative inverse of the Simpson index (Jost, 2006).

Fourth, Hill numbers can also be computed while taking into account the phylogenetic or functional relationships among OTUs. When these so-called 'phylogenetic Hill numbers' are computed, the diversity is measured in effective number of equally abundant and equally distinct lineages (Chao et al., 2010). For two systems with identical number of types and relative abundances, the one with the largest phylogenetic differences across OTUs will be the one with the highest phylogenetic diversity or phylodiversity. Similar to neutral Hill numbers, phylogenetic Hill numbers are also closely related to popular phylogenetic diversity indices: Faith's PD (when q=0), Allen's H (when q=1) and Rao's Q (when q=2) (Rao, 1982; Faith, 1992; Allen, Kon, & Bar-Yam, 2009).

Fifth, although the Hill numbers framework was originally developed for abundance data, it can also be applied to incidence data, a type of information broadly employed when dealing with, for example, ecological niche-related issues. In abundance-based approaches the DNA sequence is the unit upon which diversity is computed, while in incidence-based approaches the sample is the unit upon which diversity is measured (Chao et al., 2014). When computing Hill numbers from incidence data, it is important to note that the interpretation of both the measure and the measurement unit is slightly different to that of abundance data. Abundance-based Hill numbers measure the effective number of equally abundant OTUs in the system, while incidence-based Hill numbers measure the effective number of equally frequent (across samples) OTUs in the system (Chao et al., 2014).

Sixth, the Hill numbers framework enables the diversity of a system to be partitioned following the multiplicative definition (Jost, 2007; Chao, Chiu, & Hsieh, 2012). Alpha diversity is obtained by computing the Hill numbers from the averaged basic sums of the samples, while gamma diversity is obtained by taking the average of OTU relative abundances across samples, and then computing the Hill numbers of the pooled system (Alberdi & Gilbert, 2019). The division of gamma diversity by alpha diversity yields the beta diversity, which quantifies how many times richer an entire system is in effective OTUs (gamma diversity) than its constituent samples are on average (alpha diversity). However, the Hill numbers beta diversity can also be considered an actual diversity value, as the same metric also measures the effective number of equally large and completely distinct samples in a system (Tuomisto, 2010).

Finally, it is possible to compute multiple (dis)similarity measurements derived from beta diversities, both for Hill numbers and phylogenetic Hill numbers. Four classes of similarity measures derived from Hill numbers beta diversities have been developed. The Sørensen-type overlap (CqN) quantifies the effective average proportion of a subsystem's OTUs (or lineages in the case of phylodiversities) that are shared across all subsystems. The Jaccard-type overlap (UqN) quantifies the effective proportion of OTUs or lineages in a system that are shared across all subsystems. The Sørensen-type turnover-complement (VqN) is the complement of the

Sørensen-type turnover, which quantifies the normalized OTU turnover rate with respect to the average subsystem (i.e. alpha). Finally, the Jaccard-type turnover-complement (SqN) is the complement of the Jaccard-type turnover, which quantifies the normalized OTU turnover rate with respect to the whole system (i.e. gamma). Dissimilarity measures can be obtained by calculating their one-complements (1-XqN). Interestingly, many popular dissimilarity metrics (Sørensen, Jaccar, Morisita-Horn, Unifrac) are special cases of these (dis)similarity metrics for a given type of Hill number (neutral or phylogenetic), order of diversity (q value) and sample size (N=2) (Jost, 2006; Chao et al., 2012).

Applications

The first version of the package hilldiv contains 15 functions that enable different types of diversity measurements to be performed, as well as compared and visualised. For operating some of the scripts, hilldiv calls functions from other R packages, including *ggplot2* (Wickham, 2010), *ape* (Paradis, Claude, & Strimmer, 2004), *ade4* (Dray, Dufour, & Others, 2007) and *vegan* (Oksanen et al., 2013). We introduce the main functions for performing the most relevant tasks in the following section.

(Phylo)diversity computation

The diversity of individual samples can be calculated for vectors (one sample) or matrices (OTU table containing multiple samples) using the function *hill.div()*. The function requires as input the abundance data (as a vector or matrix), and specification of the order of diversity (q-value) at which diversity will be computed. As diversity measurement based on Hill numbers requires the relative abundances per sample to sum up to 1, if such assumption is not met, the script applies the *tss()* function to normalise the data using the total sum scaling method. As default, *hill.div()* yields the diversity measure in 'effective number of OTUs'. If an ultrametric tree object that establishes the (e.g. phylogenetic) relations between OTUs is also provided, the function considers the correlation between OTUs, and yields the diversity measure in 'effective number of lineages'. If the argument 'type' is set as 'incidence', incidence-based Hill numbers are computed. It must be noted that in systems comprised of large number of samples and OTUs, incorporating the phylogenetic information considerably increases the computation time.

(Phylo)diversity profile

An effective way for representing the different components of the diversity of a system is to plot its diversity profile, as it provides information about the richness and evenness of a sample at a glance (Fig. 2a) (Chiu et al., 2014). This can be done with the *div.profile()* function, which can plot a single sample (Fig. 2b), a set of samples (Fig. 2c) or multiple sets of samples aggregated by groups (Fig. 2d). Diversity profiles can be generated both for neutral and phylogenetic (if a tree object is provided) diversity measures, although the latter requires longer computation times. The hierarchy argument enables the plotting of diversity profiles of the alpha or gamma (default) diversities of groups. The default range of orders of diversity (q) is from 0 to 5 in

intervals of 0.1, although this can be modified through specifying a different vector of sequential numbers in the *qvalues* argument. If a hierarchy table is provided, it is possible to plot either the alpha or gamma diversity profiles of the groups as specified by the argument 'level'.

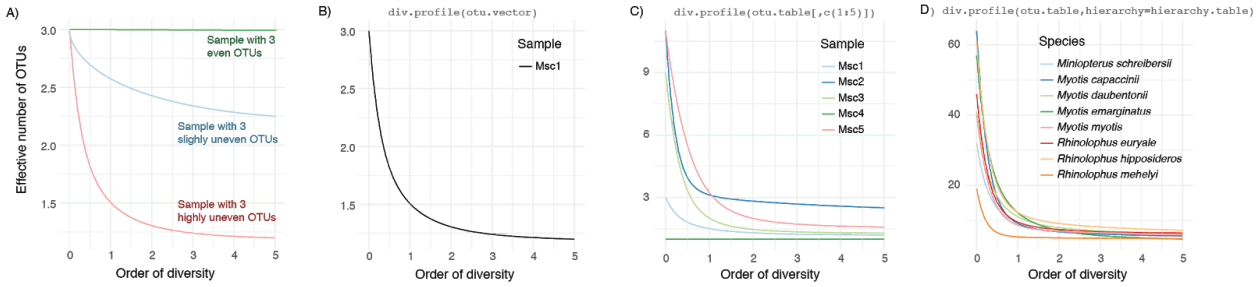

Figure 2. **Diversity profiles**. A) Three diversity profiles of hypothetical data sets with 3 OTUs with different abundance distributions, and B-D) diversity profiles generated with the function *div.profile()* for B) one sample, C) multiple individual samples, and D) multiple samples aggregated in groups.

(Phylo)diversity comparison tests

The function *div.test()* performs statistical hypothesis testing for comparing the neutral or phylogenetic (if tree is provided) diversities of a given order across groups of samples. The function requires a so-called hierarchy table (a two-column matrix in which sample names are provided in the 1st column and group names in the 2nd column) that establishes the relationship between the samples and their respective groups. The function first runs Shapiro and Barlett tests to assess normality and homogeneity of the data, and depending on the results performs either a Student's t-test or a Wilcoxon Rank Sum Test if there are 2 groups, and either an ANOVA or a Kruskal-Wallis test if there are multiple groups. The related function *div.test.plot()* uses the output object of the *div.test()* function to visually summarise the diversity values as a box, (Fig 3a), jitter (Fig 3b) or violin (Fig 3c) plots.

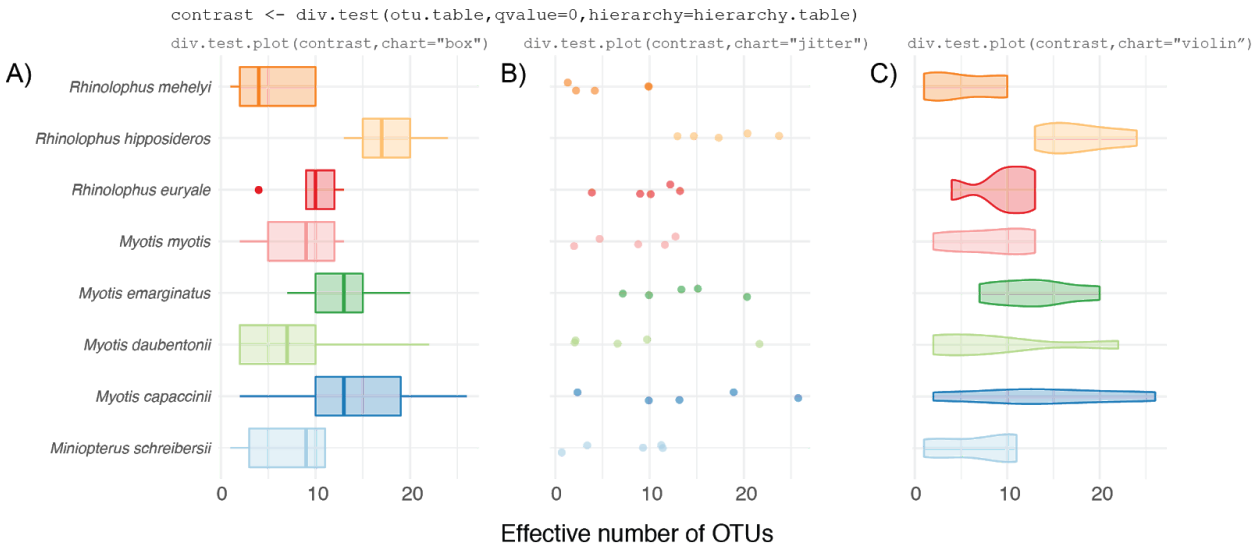

Figure 3. **Diversity comparison plots** produced using the *div.test.plot()* function. A) Box plot generated using the chart argument "box", B) Jitter plot generated using the chart argument "jitter", and C) Violin plot generated using the chart argument "violin".

(Phylo)diversity partitioning

Diversity and phylodiversity partitioning based on abundance and incidence data can be carried out using the function *div.part()*. The function assumes a 2-level hierarchy and yields alpha, gamma and beta values based on abundance data. If a hierarchy table is provided, the function yields alpha, gamma and beta values based on incidence data, i.e. alpha diversity reflects the incidence-based diversity of groups.

(Dis)similarity measurement

The function *beta.dis()* performs similarity or dissimilarity measurement based on Hill numbers beta diversity, sample size and order of diversity. The function can be run by inputting those values manually, or by using the list object outputted by the *div.part()* function, which contains all the mentioned information. As specified by the argument "metric": the function can compute the following similarity measures: the Sørensen-type overlap (C_{qN}), the Jaccard-type overlap (U_{qN}), the Sørensen-type turnover-complement (V_{qN}), and the Jaccard-type turnover-complement (S_{qN}). The argument 'type' enables either similarities or dissimilarities (one-complements of the similarity values) to be outputted.

Pairwise (dis)similarity

The function *pair.dis()* performs pairwise diversity partitioning and yields matrices containing pairwise beta diversity and (dis)similarity measures. If a hierarchy table is provided, pairwise calculations can be carried out at some or all the specified hierarchical levels. The results are outputted as a list of matrices. The related function *pair.dis.plot()* uses any of the dissimilarity matrices yielded by *pair.dis()* (e.g. 1-U_{qN}) to visualize it either as a NMDS chart, a qgraph plot or a heatmap/correlogram.

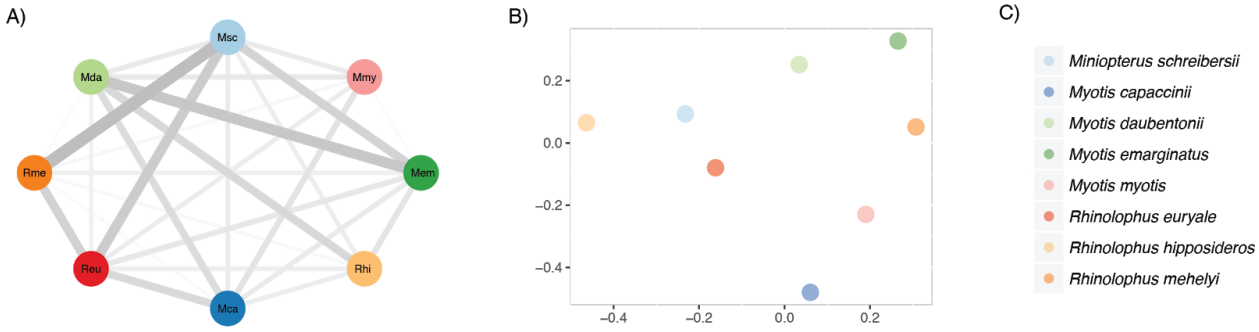

Figure 4. **Dissimilarity visualisation plots**. A) A qgraph network diagram, B) a NMDS chart and C) legend of the groups.

Conclusions

The package 'hilldiv' enables the most common statistical operations that molecular ecologists routinely require to be performed in an straightforward way based on the statistical framework developed around the Hill numbers. Although it is primarily devised for metabarcoding data containing multiple samples and OTUs, it could also be useful to perform diversity analyses on shotgun metagenomics data, or indeed any other non-molecular data that is comprised of multiple sampling units and types that enable measures of diversity to be quantified.

Data accessibility

All the data used in this paper are included in the R package. The R package can be downloaded from Github (https://github.com/anttonalberdi/hilldiv).

Supporting information

Examples of the usage of all functions are shown in the supporting document "supporting.scripts.txt".

Acknowledgements

A.A. was supported by Lundbeckfonden (grant R250-2017-1351) and M.T.P.G. acknowledges ERC Consolidator Grant (681396-Extinction Genomics).

References

Alberdi, A., & Gilbert, M. T. P. (2019). A guide to the application of Hill numbers to DNA based diversity analyses. *Molecular Ecology Resources*. Under review.

Allen, B., Kon, M., & Bar-Yam, Y. (2009). A new phylogenetic diversity measure generalizing the shannon index and its application to phyllostomid bats. *The American Naturalist*, *174*(2), 236–243.

Chao, A., Chiu, C.-H., & Hsieh, T. C. (2012). Proposing a resolution to debates on diversity partitioning. *Ecology*, *93*(9), 2037–2051.

Chao, A., Chiu, C.-H., & Jost, L. (2010). Phylogenetic diversity measures based on Hill numbers. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *365*(1558), 3599–3609.

Chao, A., Gotelli, N. J., Hsieh, T. C., Sander, E. L., Ma, K. H., Colwell, R. K., & Ellison, A. M. (2014). Rarefaction and extrapolation with Hill numbers: a framework for sampling and estimation in species diversity studies. *Ecological Monographs*, *84*(1), 45–67.

Charney, N., & Record, S. (2012, January 1). vegetarian: Jost Diversity Measures for Community Data.

Chiu, C.-H., Jost, L., & Chao, A. (2014). Phylogenetic beta diversity, similarity, and differentiation measures based on Hill numbers. *Ecological Monographs*, *84*.

Dray, S., Dufour, A.-B., & Others. (2007). The ade4 package: implementing the duality diagram for ecologists. *Journal of Statistical Software*, *22*(4), 1–20.

Faith, D. P. (1992). Conservation evaluation and phylogenetic diversity. *Biological Conservation*, *61*(1), 1–10.

Hill, M. O. (1973). Diversity and Evenness: A Unifying Notation and Its Consequences. *Ecology*, *54*(2), 427–432.

Jost, L. (2006). Entropy and diversity. *Oikos* , *113*, 363–375.

Jost, L. (2007). Partitioning diversity into independent alpha and beta components. *Ecology*, *88*(10), 2427–2439.

Jurasinski, G. (2007). Simba: a collection of functions for similarity calculation of binary data. *R Package Version 0. 2-5, URL http://CRAN. R-Project. Org/package= Simba*.

Marcon, E., & Hérault, B. (2015). entropart: An R Package to Measure and Partition Diversity. *Journal of Statistical Software, Articles*, *67*(8), 1–26.

Oksanen, J., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., O'hara, R. B., … Others. (2013). Package 'vegan'. *Community Ecology Package, Version*, *2*(9). Retrieved from http://cran.ism.ac.jp/web/packages/vegan/vegan.pdf

Paradis, E., Claude, J., & Strimmer, K. (2004). APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* , *20*(2), 289–290.

Rao, C. R. (1982). Diversity and dissimilarity coefficients: a unified approach. *Theoretical Population Biology*, *21*, 24–43.

Shannon, C. E. (1948). A Mathematical Theory of Communication. *The Bell System Technical Journal*, *27*, 379–423.

Tuomisto, H. (2010). A diversity of beta diversities: straightening up a concept gone awry. Part 1. Defining beta diversity as a function of alpha and gamma diversity. *Ecography*, *33*(1), 2–22.

Wickham, H. (2010). ggplot2: elegant graphics for data analysis. *Journal of Statistical Software*, *35*(1), 65–88.