Metadata S1

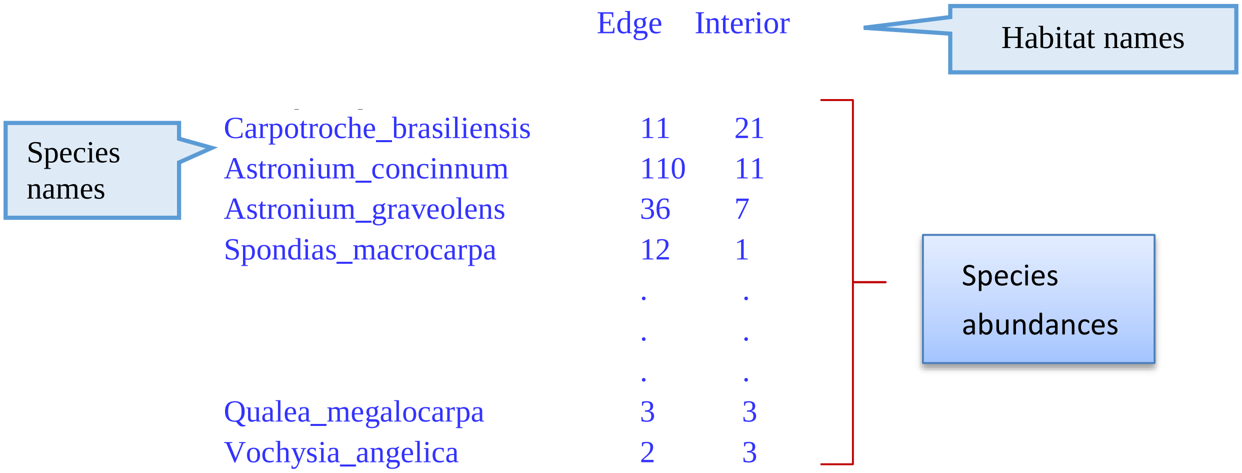
Deciphering the Enigma of Undetected Species, Phylogenetic, and Functional Diversity Based on Good-Turing Theory

Anne Chao, Chun-Huo Chiu, Robert K. Colwell, Luiz Fernando S. Magnago, Robin L. Chazdon, and Nicholas J. Gotelli

Metadata S1: Description of Data Sets and Running Procedures for the R Code "Good-Turing"

Data S1: species-by-assemblage abundance data for 425 tree species (Magnago et al. 2014)

As presented below, the first row in Data S1 shows the names of the two assemblages/habitats ("Edge" and "Interior") from the illustrative example considered in the main text. Then, beginning with the second row, there are three entries in each row: the species name followed by this species' sample abundance/frequency in the Edge and Interior habitats, respectively.



	Edge	Interior
Carpotroche_brasiliensis	11	21
Astronium_concinnum	110	11
Astronium_graveolens	36	7
Spondias_macrocarpa	12	1
	.	.
	.	.
	.	.
Qualea_megalocarpa	3	3
Vochysia_angelica	2	3

The R code "Good-Turing" for computing all the estimators discussed in the main text is available in Github (https://github.com/AnneChao). As an alternative, readers without a background in R can utilize the online software "GoodTuring", made available from https://chao.shinyapps.io/GoodTuring/ to facilitate all computations.

There are six main functions in the R code:

(1) Function Richness(data) for estimating the species richness in each individual assemblage given specified species-by-assemblage abundance matrix data. (Species names are optional).

(2) Function Shared_richess(data) for estimating the shared species richness between two assemblages given specified species-by-assemblage abundance matrix data. (Species names are optional).

(3) Function PD(data, tree) for estimating Faith's PD in each individual assemblage given specified species-by-assemblage abundance matrix data and a specified phylogenetic tree (in Newick format) spanned by all observed species. (Species names are required in species-by-assemblage abundance data and must match those names in the specified phylogenetic tree.)

(4) Function Shared_PD (data, tree) for estimating the shared PD between two assemblages given specified species-by-assemblage abundance matrix data and a specified phylogenetic tree (in Newick format) spanned by all observed species. (Species names are required in species-by-assemblage abundance data and must match those names in the specified phylogenetic tree.)

(5) Function FAD(data, dis_matrix) for estimating the FAD in each individual assemblage given specified species-by-assemblage abundance matrix data and a specified pairwise distance matrix of all observed species. (Species names are required in species-by-assemblage abundance data and must match those names in the specified distance matrix. Also, the ordering of the species in the abundance data should also be the same as that in the distance matrix.)

(6) Function Shared_FAD(data, dis_matrix) for estimating the shared FAD between two assemblages given specified species-by-assemblage abundance matrix data and a specified pairwise distance matrix of all observed species. (Species names are required in species-by-assemblage abundance data and must match those names in the specified distance matrix. Also, the ordering of the species in the abundance data should also be the same as that in the distance matrix.)

(7)

Running procedures

First, copy and paste the R code "Good-Turing", available from the Github website, into the R Console. The following steps show how to run the R function Richness(data) to obtain the Chao1 species richness estimator and 95% confidence interval. The input data include a species-by-assemblage abundance matrix (species names are optional). The number of assemblages can be any positive integer.

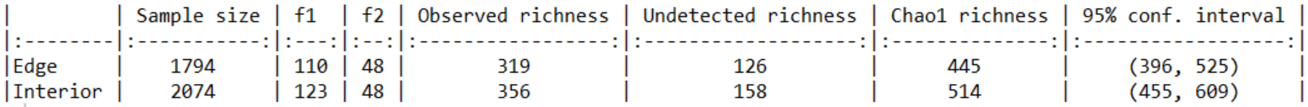
```
# The package "knitr" must be installed and loaded before running the R function Richness()
# Install the Knitr package from CRAN
install.packages(“knitr”)

# Import Knitr
library(knitr)

# Import Data S1 (species abundances)
# note the / instead of \ on MS windows systems
Brazil=read.table("C:/Users/stat_pc/Desktop/DataS1.txt")

# Run R function Richness(data) to obtain the Chao1 richness estimate and 95% confidence interval via a
# log-transformation
Richness(Brazil)
```

The output is shown below; see Table 2(a) of the main text for notation and interpretation.



```
|         | Sample size | f1  | f2 | Observed richness | Undetected richness | Chao1 richness | 95% conf. interval |
|:--------|:-----------:|:---:|:--:|:-----------------:|:-------------------:|:--------------:|:------------------:|
|Edge     |    1794     | 110 | 48 |        319        |         126         |      445       |     (396, 525)     |
|Interior |    2074     | 123 | 48 |        356        |         158         |      514       |     (455, 609)     |
```

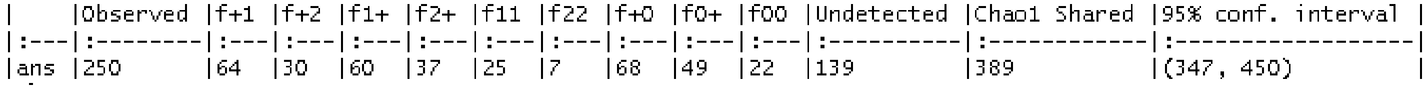
The following steps show how to run the R function `Shared_Richness(data, tree)` to obtain the Chao1-shared species richness estimator and 95% confidence interval. The input `data` include a species-by-assemblage abundance matrix (species names are optional). The number of assemblages must be two.

```
# The package "knitr" must be installed and loaded before running the R function Shared_richness().
install.packages( “knitr”)
library(knitr)

# Import Data S1 (species abundances)
# note the / instead of \ on MS windows systems
Brazil=read.table("C:/Users/stat_pc/Desktop/DataS1.txt")

# Run R function Shared_richness(data) to obtain the Chao1-shared richness estimate and 95% confidence
# interval via a log-transformation
Shared_richness (Brazil)
```

The output is shown below; see Table 2(b) of the main text for notation and interpretation.



```
|    |Observed |f+1 |f+2 |f1+ |f2+ |f11 |f22 |f+0 |f0+ |f00 |Undetected |Chao1 Shared |95% conf. interval |
|:---|:--------|:---|:---|:---|:---|:---|:---|:---|:---|:---|:----------|:------------|:------------------|
|ans |250      |64  |30  |60  |37  |25  |7   |68  |49  |22  |139        |389          |(347, 450)         |
```

Data S2: the phylogenetic tree (in Newick format) of 425 species

The data file consists of the phylogenetic tree (in Newick format) of the 425 observed species listed in Data S1. The tree was constructed using the software Phylomatic (Webb and Donoghue 2005). The following steps show how to run the R function `PD(data, tree)` to obtain the Chao1-PD estimate and 95% confidence interval. The input species-by-assemblage abundance data matrix must include species names. The number of assemblages is allowed to be any positive integer. Species names in the abundance data must match those in the specified Newick-format phylogenetic tree.

```
# The packages "ade4", "phytools", "ape" and "knitr" must be installed and loaded before running the R
# function PD()
install.packages(c(“phytools”, “ade4”, “knitr”, “ape”))
library(phytools)
library(ade4)
library(knitr)
library(ape)

# Import abundance data and tree
# note the / instead of \ on MS windows systems
Brazil=read.table("C:/Users/stat_pc/Desktop/DataS1.txt")
tree=read.newick("C:/Users/stat_pc/Desktop/DataS2.txt")

# Run R function PD(data, tree) to obtain the Chao1-PD estimate and 95% confidence interval via a
# log-transformation
PD(Brazil, tree)
```

The output is shown below; see Table 3(a) of the main text for notation and interpretation.

```
|         | Sample size |  g1  |  g2  | Observed | Undetected | Chao1-PD | 95% conf. interval |
|:--------|:-----------:|:----:|:----:|:--------:|:----------:|:--------:|:------------------:|
|Edge     |    1794     | 6578 | 2885 |  24516   |    7495    |  32011   |   (31542, 32511)   |
|Interior |    2074     | 7065 | 3656 |  27727   |    6823    |  34550   |   (34143, 34983)   |
```

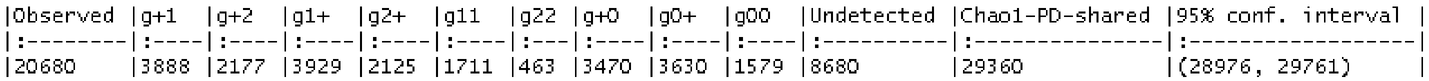
The following steps show how to run the R function Shared_PD(data, tree) to obtain the Chao1-PD-shared estimate and 95% confidence interval. The input species-by-assemblage abundance data matrix must include species names. The number of assemblages is allowed to be any positive integer. Species names in the abundance data must match those in the specified Newick-format phylogenetic tree.

```
# The packages "ade4", "phytools", "ape" and "knitr" must be installed and loaded before running the
# R function Shared_PD().
install.packages(c(“phytools”, “ade4”, “knitr”, “ape”))
library(phytools)
library(ade4)
library(knitr)
library(ape)

# Import abundance data and tree
# note the / instead of \ on MS windows systems
Brazil=read.table("C:/Users/stat_pc/Desktop/DataS1.txt")
tree=read.newick("C:/Users/stat_pc/Desktop/DataS2.txt")

# Run R function Shared_PD(data, tree) to obtain the Chao1-PD-shared estimate and 95% confidence
# interval via a log-transformation
Shared  PD(Brazil, tree)
```

The output is shown below; see the main text and Table 3(b) of the main text for notation and interpretation.



```
|Observed |g+1  |g+2  |g1+  |g2+  |g11  |g22 |g+0  |g0+  |g00  |Undetected |Chao1-PD-shared |95% conf. interval |
|:--------|:----|:----|:----|:----|:----|:---|:----|:----|:----|:----------|:---------------|:------------------|
|20680    |3888 |2177 |3929 |2125 |1711 |463 |3470 |3630 |1579 |8680       |29360           |(28976, 29761)     |
```

Data S3: Species pairwise Gower distance matrix of 425 species based on six species traits

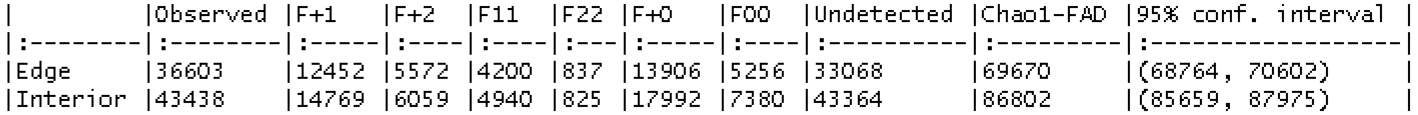
All the 425 observed species listed in Data S1 were described by a set of six functional traits, including five categorical variables: fruit size (size categories), seed size (size categories), fruit type, fruit dispersal syndrome, and successional group, together with one quantitative variable: wood density (Magnago et al. 2014). Based on these six traits, the species distance matrix in Data S3 was calculated by a Gower mixed-variables coefficient of distance with equal weights for all traits. The following steps show how to run the R function FAD(data, dis_matrix) to obtain the Chao1-FAD estimate and 95% confidence interval. The input species-by-assemblage abundance data matrix must include species names; the input dis_matrix denotes a species pairwise distance matrix. The number of assemblages is allowed to be any positive integer. Species names in the species abundance data and must match those in the specified distance matrix. Moreover, the ordering of species in the abundance data should also be exactly the same as that in the specified distance matrix.

```
# The package "knitr" must be installed and loaded before running the R function FAD().
install.packages( “knitr”)
library(knitr)

# Import abundance data and distance matrix
# note the / instead of \ on MS windows systems
Brazil=read.table("C:/Users/stat_pc/Desktop/DataS1.txt")
dis=read.table("C:/Users/stat_pc/Desktop/DataS3.txt")

# Run R function FAD(data, tree) to obtain the Chao1-FAD estimate and 95% confidence interval
# via a log-transformation
FAD(Brazil, as.matrix(dis))
```

The output is shown below; see Table 4(a) of the main text for notation and interpretation.



```
|         |Observed |F+1   |F+2  |F11  |F22 |F+0   |F00  |Undetected |Chao1-FAD |95% conf. interval |
|:--------|:--------|:-----|:----|:----|:---|:-----|:----|:----------|:---------|:------------------|
|Edge     |36603    |12452 |5572 |4200 |837 |13906 |5256 |33068      |69670     |(68764, 70602)     |
|Interior |43438    |14769 |6059 |4940 |825 |17992 |7380 |43364      |86802     |(85659, 87975)     |
```

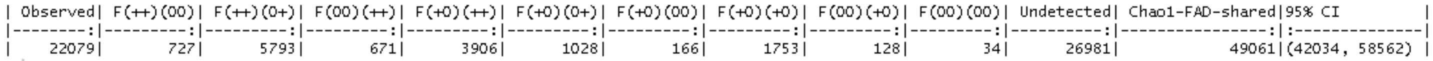
The following steps show how to run the R function Shared_FAD(data, dis_matrix) to obtain the Chao1-FAD-shared estimate and 95% confidence interval. The input species-by-assemblage abundance data matrix must include species names; the input dis_matrix denotes a species pairwise distance matrix. The number of assemblages is allowed to be any positive integer. Species names in the species abundance data must match those in the specified distance matrix. Moreover, the ordering of species in the abundance data should also be exactly the same as that in the specified distance matrix.

```
# The package "knitr" must be installed and loaded before running the following function.
install.packages( “knitr”)
library(knitr)

# Import abundance data and distance matrix
# note the / instead of \ on MS windows systems
Brazil=read.table("C:/Users/stat_pc/Desktop/DataS1.txt")
dis=read.table("C:/Users/stat_pc/Desktop/DataS3.txt")

# Run R function Shared_FAD(data, tree) to obtain the Chao1-FAD-shared estimate and 95%
# confidence interval via a log-transformation based on 200 bootstrap replications
Shared_FAD(Brazil, as.matrix(dis))
```

The output is shown below; see Table 4(b) of the main text for notation and interpretation.



```
| Observed| F(++)(00)| F(++)(0+)| F(00)(++)| F(+0)(++)| F(+0)(0+)| F(+0)(00)| F(+0)(+0)| F(00)(+0)| F(00)(00)| Undetected| Chao1-FAD-shared|95% CI         |
|--------:|---------:|---------:|---------:|---------:|---------:|---------:|---------:|---------:|---------:|----------:|----------------:|:--------------|
|    22079|       727|      5793|       671|      3906|      1028|       166|      1753|       128|        34|      26981|            49061|(42034, 58562) |
```

NOTE:

(1) A standard approximation method (i.e., delta-method) is applied to obtain a variance estimate and the associated 95% confidence interval via a log-transformation for the estimator of species richness, shared species richness, PD, shared PD, and FAD.

(2) For the shared FAD estimator, a bootstrap method with 200 replications is applied to obtain a variance estimate and the associated 95% confidence interval via a log-transformation. Due to the randomness of the bootstrapping process, the estimated 95% confidence interval will vary slightly each time the same abundance and distance matrix data are imported.

References

Magnago, L. F. S., D. P. Edwards, F. A. Edwards, A. Magrach, S. V. Martins, and W. F. Laurance. 2014. Functional attributes change but functional richness is unchanged after fragmentation of Brazilian Atlantic forests. Journal of Ecology 102:475−485.

Webb, C. O., and M. J. Donoghue. 2005. Phylomatic: tree assembly for applied phylogenetics. Molecular Ecology Notes 5:181−183.