

# Package ‘DGBD’

November 6, 2024

**Title** Discrete Generalized Beta Distribution Modeling and Plotting

**Version** 0.0.0.9006

**Description** Process abundance data and estimate the parameters of a discrete generalized beta distribution (DGBD) that fits the data. This distribution was first described in Mansilla et al. (2007).  
<[doi:10.1016/j.joi.2007.01.001](https://doi.org/10.1016/j.joi.2007.01.001)>  
and is further explored in Martinez-Mekler et al. (2009).  
<[doi:10.1371/journal.pone.0004791](https://doi.org/10.1371/journal.pone.0004791)>  
Generates linear or non-linear model reports and uses 'ggplot2' to plot rank abundance diagrams.

**License** MIT + file LICENSE

**Encoding** UTF-8

**Roxygen** list(markdown = TRUE)

**RoxygenNote** 7.3.2

**Date/Publication** 2024-09-23 21:05:00 UTC

**Imports** ggplot2,  
phyloseq,  
gslns,  
R.utils,  
utils

**Depends** R (>= 3.5.0)

**LazyData** true

**URL** <https://github.com/Fa-Moe/DGBD>, <https://fa-moe.github.io/DGBD/>

**BugReports** <https://github.com/Fa-Moe/DGBD/issues>

**Suggests** knitr,  
rmarkdown

**VignetteBuilder** knitr

## Contents

|                            |   |
|----------------------------|---|
| Author_Citations . . . . . | 2 |
| BC_compare . . . . .       | 3 |
| BC_many . . . . .          | 4 |
| BC_model . . . . .         | 5 |
| BC_multiple . . . . .      | 5 |

|                                  |    |
|----------------------------------|----|
| BC_param . . . . .               | 6  |
| BC_plot . . . . .                | 7  |
| BC_random . . . . .              | 9  |
| BC_rank . . . . .                | 10 |
| BC_report . . . . .              | 11 |
| Billionaires . . . . .           | 12 |
| Citation_ISICatalogued . . . . . | 12 |
| EC_Codon . . . . .               | 13 |
| EC_Metabolite . . . . .          | 14 |
| Flares . . . . .                 | 14 |
| hmp_wgs . . . . .                | 15 |
| MOMv3.3 . . . . .                | 16 |
| Phy_Citation . . . . .           | 17 |
| Quakes . . . . .                 | 17 |
| Tara_Data . . . . .              | 18 |
| Weblinks . . . . .               | 19 |

## Index 20

---

|                  |                          |
|------------------|--------------------------|
| Author_Citations | <i>Author_Citations.</i> |
|------------------|--------------------------|

---

### Description

How many times a given author was referenced in ISI hournals from 1981 to 1997. Authors with the same name are merged.

### Usage

Author\_Citations

### Format

Author\_Citations:

Data frame that contains.

**Name** Author name.

**Citations per article** How many citations on average does the author get on his articles.

**Articles** How many articles the author has published.

**Citations** How many times the published article has been cited.

**Rank** The ranking by how many total citations the author has gotten. ...

### Details

" 'This data covers all publications from ISI catalogued journals from January 1981 - June 1997 and was compiled by H. Small and D. Pendlebury of the Institute for Scientific Information. This file includes several corrections to the data originally supplied by the ISI: missing decimal points in several entries are now provided, and the data is correctly ordered according to total number of citations (a block of the original data that was provided by the ISI was not in correct rank order). Notice also that different authors with the same last name and first initial are erroneously lumped into one entry in the table (for example #12 or #397 on this list). Use this list with caution! ' Quoted from: Citations of the 1120 Most-Cited Physicists 1981 - June 1997. (n.d.). Retrieved September 11, 2024, from <https://physics.bu.edu/~redner/projects/citation/physics-by-person.html>"

**Source**

<https://github.com/robertoalvarezm/DataNonPowerLaws>

---

|            |                                                          |
|------------|----------------------------------------------------------|
| BC_compare | <i>Compare the linear and the nls methods parameters</i> |
|------------|----------------------------------------------------------|

---

**Description**

Creates a graph that illustrates error in the predictions of the linear and nls models for quick comparison. The y axis is the difference between the predicted value and the original abundance value. This value is then made absolute and 1 is added before taking its logarithm. The addition of 1 keeps a difference of 0 as 0 in the plot and prevents a small decimal linear difference from becoming a large difference in the plot. The x axis is the rank for that given error. This allows us to see where in the distribution each model performs better or worse.

**Usage**

```
BC_compare(
  df_abundance = NULL,
  column = NULL,
  BC_plot_list = NULL,
  c_gfx_title = "Linear vs nls model error",
  c_gfx_label = TRUE,
  ...
)
```

**Arguments**

|              |                                                                                                                                                                           |
|--------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| df_abundance | A data frame that contains abundance data.                                                                                                                                |
| column       | Either a string with the name of the column or the number of the column that stores the abundances in the data frame.                                                     |
| BC_plot_list | A list that contains 2 objects previously generated with BC_plot. The first one must use the linear parameters and the second one parameters estimated by the nls method. |
| c_gfx_title  | String. Changes the title of the graph.                                                                                                                                   |
| c_gfx_label  | Logical. Adds a label that adds the model_extra data of both models. Defaults to true.                                                                                    |
| ...          | passes arguments to BC_plot.                                                                                                                                              |

**Value**

A list with that includes a data frame with difference data between predicted and real values and

**Examples**

```
BC_compare(hmp_wgs, 2)

BC_compare(EC_Metabolite, column = 2, model_extra = "S")
```

BC\_many

*BC\_multiple for Many Objects***Description**

Analyses N objects using BC\_multiple in sequence.

**Usage**

```
BC_many(
  object_names,
  column = NULL,
  use_paths = FALSE,
  is_phyloseq = FALSE,
  supress_messages_many = TRUE,
  ...
)
```

**Arguments**

|                       |                                                                                                                                                                    |
|-----------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| object_names          | A vector containing the names of the objects to process.                                                                                                           |
| column                | Optional. A list that contains vectors that indicate which columns to select for each processed object. Order must be matched with object_names.                   |
| use_paths             | Optional. Set to true when the object_names vector uses paths instead of loaded objects.                                                                           |
| is_phyloseq           | Optional. Set to true if the objects to analyse have the phyloseq class.                                                                                           |
| supress_messages_many | Whether to hide the report progress that indicates how many objects are left to analyse. Also stops the resulting data frame from being printed. Defaults to true. |
| ...                   | Passes arguments to BC_multiple                                                                                                                                    |

**Value**

A data frame that contains BC\_multiple style information with a column appended to the right that indicates to which object the analysis belongs.

**Examples**

```
to_analyze <- c("Weblinks", "Billionaires", "MOMv3.3")
columns_to_use <- list(2, 2, c(7, 8))
BC_many(to_analyze, columns_to_use)
```

BC\_model

*Predict Rank Abundance Distribution***Description**

Uses the parameters calculated by BC\_param to model a DGB distribution (Mansilla et al. (2007) [doi:10.1016/j.joi.2007.01.001](https://doi.org/10.1016/j.joi.2007.01.001)) from the rank information in the data frame.

**Usage**

```
BC_model(df_abundance = NULL, column = NULL, BC_param_object = NULL, ...)
```

**Arguments**

df\_abundance     A data frame that contains abundance data.

column            Either a string with the name of the column or the number of the column that stores the abundances in the data frame.

BC\_param\_object   Optional. A previous object generated by BC\_param.

...                passes arguments to BC\_param.

**Value**

A list with the following elements: The input data frame with added processed ranking data, model data and confidence interval data, the adjusted parameters, the confidence interval of the parameters, the linear model, a summary of the model and a generated function for use with raw numeric data.

**Examples**

```
modelISISCatalogued <- BC_model(Citation_ISISCatalogued, column = 2, show_stats = F, rank_threshold=1)
head(modelISISCatalogued[[1]])
modelISISCatalogued[2:6]
```

BC\_multiple

*Create Plots and Reports for N Columns***Description**

Basic wrapper for BC\_plot and BC\_report intended to be used for large data frames.

**Usage**

```
BC_multiple(
  df_abundance,
  column = NULL,
  is_phyloseq = FALSE,
  supress_messages = TRUE,
  report_only = FALSE,
  file_output_mode = FALSE,
  called_from_many = NULL,
  ...
)
```

**Arguments**

|                  |                                                                                                                                                                                       |
|------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| df_abundance     | A data frame that contains abundance data.                                                                                                                                            |
| column           | Optional. A vector that contains either column name strings or numeric column indices to be analyzed. By default includes all columns.                                                |
| is_phyloseq      | Logical, optional. Use if the object to be analysed has the phyloseq class.                                                                                                           |
| supress_messages | Logical. Allows to skip messages that detail how many samples have been processed. Also stops the resulting data frame from being printed. Defaults to true.                          |
| report_only      | Logical. Use to output a data frame without the BC_plot object, intended to reduce the size of the output for large analyses. Defaults to false.                                      |
| file_output_mode | Logical. Whether to provide all analysis inside the R environment (when false) or to packet BC_report results into files (when true). Recommended when analyzing very large datasets. |
| called_from_many | Automatically handled. An object name passed from BC_many.                                                                                                                            |
| ...              | passes arguments to BC_report and BC_plot.                                                                                                                                            |

**Value**

A data frame where each row represents a fitted DGBD. The first column stores the BC\_plot objects and the second column stores the BC\_reports. Each object's output is stored inside of a list.

**Examples**

```
BC_multiple(df_abundance=DGBD::MOMv3.3, column = c(7:8))

BC_multiple(df_abundance=DGBD::EC_Codon, model_extra="S", confint_col="#448666")
```

BC\_param

*Estimate the DGB Distribution Parameters***Description**

Estimates the parameters of the DGB distribution (also known as the Beta-Cocho distribution) first defined in Mansilla et al. (2007) [doi:10.1016/j.joi.2007.01.001](https://doi.org/10.1016/j.joi.2007.01.001) and further characterized in Martinez-Mekler et al. (2009) [doi:10.1371/journal.pone.0004791](https://doi.org/10.1371/journal.pone.0004791) for a given set of data. BC\_param calculates the log of the data and estimates the abundance from the ranking using a linear model. The coefficients of the linear model are then scaled for future use.

**Usage**

```
BC_param(
  df_abundance = NULL,
  column = NULL,
  confidence_interval = 0.95,
  nls = FALSE,
  nls_loop = 1,
  nls_algo = "lm",
```

```

    nls_control = list(scale = "more"),
    BC_rank_object = NULL,
    ...
  )

```

### Arguments

|                                  |                                                                                                                                                                       |
|----------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <code>df_abundance</code>        | A data frame that contains abundance data.                                                                                                                            |
| <code>column</code>              | Either a string with the name of the column or the number of the column that stores the abundances in the data frame.                                                 |
| <code>confidence_interval</code> | Numeric. The confidence interval to calculate for the DGB distribution.                                                                                               |
| <code>nls</code>                 | Logical. Set to TRUE to use a non-linear least squares fitting method from the 'gslnls' package.                                                                      |
| <code>nls_loop</code>            | Numeric. Set to values higher than 1 to repeat the nls method as many times as needed, reseeding with the last value. This improves the fit with diminishing returns. |
| <code>nls_algo</code>            | String. The algorithm to use for the <code>gsl_nls</code> function.                                                                                                   |
| <code>nls_control</code>         | List. Provided for the control argument for the <code>gsl_nls</code> function.                                                                                        |
| <code>BC_rank_object</code>      | Optional. A previous object generated by <code>BC_rank</code> .                                                                                                       |
| <code>...</code>                 | passes arguments to <code>BC_rank</code> .                                                                                                                            |

### Value

A list with the following elements: The input data frame with added processed ranking data and predicted confidence interval values, the adjusted parameters, the confidence interval of the parameters, the linear model and a summary of the model.

### Examples

```

paramBillionaires <- BC_param(df_abundance=DGBD::Billionaires, column= 2, confidence_interval=0.99)
head(paramBillionaires[[1]])
paramBillionaires[2:5]

paramMOMv3.3 <- BC_param(df_abundance=DGBD::MOMv3.3, column=7, nls=TRUE)
head(paramMOMv3.3[[1]])
paramMOMv3.3[2:5]

```

---

BC\_plot

---

*Plot DGB Distribution and Observed Data*


---

### Description

Creates a graphical representation of the DGB Distribution (Mansilla et al. (2007) [doi:10.1016/j.joi.2007.01.001](https://doi.org/10.1016/j.joi.2007.01.001)) model. It supports both linear and nls fits done in `BC_param`. Requires a function generated by `BC_model`.

**Usage**

```
BC_plot(
  df_abundance = NULL,
  column = NULL,
  BC_model_object = NULL,
  obs = TRUE,
  obs_shape = 16,
  obs_col = "#78a7ff",
  obs_size = 1,
  model = TRUE,
  model_col = "#000000",
  model_width = 0.5,
  model_extra = "MSE",
  confint = TRUE,
  confint_col = "#ed8666",
  confint_width = 1,
  confrange = TRUE,
  confrange_col = "#ffd078",
  gfx_alpha = 0.75,
  gfx_title = "Rank-Abundance Diagram",
  gfx_label = TRUE,
  gfx_label_coords = NULL,
  gfx_xy_trans = c("identity", "log10"),
  gfx_theme = ggplot2::theme_gray(),
  plot_silent = FALSE,
  ...
)
```

**Arguments**

|                              |                                                                                                                       |
|------------------------------|-----------------------------------------------------------------------------------------------------------------------|
| <code>df_abundance</code>    | A data frame that contains abundance data.                                                                            |
| <code>column</code>          | Either a string with the name of the column or the number of the column that stores the abundances in the data frame. |
| <code>BC_model_object</code> | Optional. A previous object generated by <code>BC_model</code> .                                                      |
| <code>obs</code>             | Logical. Whether to plot the observed abundance data. Defaults to true.                                               |
| <code>obs_shape</code>       | Numerical. The shape of the plotted observed abundance data.                                                          |
| <code>obs_col</code>         | The color for the observations.                                                                                       |
| <code>obs_size</code>        | Numeric. The size for the observations.                                                                               |
| <code>model</code>           | Logical. Whether to show the models predicted data. Defaults to true.                                                 |
| <code>model_col</code>       | Specify a color for the model.                                                                                        |
| <code>model_width</code>     | Numeric. Changes the width of the lines to use for the model.                                                         |
| <code>model_extra</code>     | String. Has to be one of: "MSE" (Mean Square Error), "S" (Standard error of the Estimate), "R2". Defaults to "MSE".   |
| <code>confint</code>         | Logical. Whether to add the confidence interval lines. Defaults to true.                                              |
| <code>confint_col</code>     | Specify a color for the confidence interval lines.                                                                    |
| <code>confint_width</code>   | Numeric. Changes the width of the confidence interval lines.                                                          |
| <code>confrange</code>       | Logical. Whether to shade the area in the confidence interval. Defaults to true.                                      |



|                  |                                                                                                                                          |
|------------------|------------------------------------------------------------------------------------------------------------------------------------------|
| confrange_col    | Specify a color to use for the confidence interval shading.                                                                              |
| gfx_alpha        | Numeric. Modifies all the graphed objects alpha. Default=0.75.                                                                           |
| gfx_title        | String. Changes the title of the graph.                                                                                                  |
| gfx_label        | Logical. Whether to show the parameters used and model_extra info.                                                                       |
| gfx_label_coords | A vector that provides custom x and y values to move the label.                                                                          |
| gfx_xy_trans     | A vector with 2 strings that define the ggplot2 transformations to be applied to the x and y scales. Defaults to c("identity", "log10"). |
| gfx_theme        | Provide a ggplot2 theme function to use. Defaults to theme_gray().                                                                       |
| plot_silent      | Logical. Whether to print to console the output list and plot the graph.                                                                 |
| ...              | passes arguments to BC_model.                                                                                                            |

### Value

A list with the following elements: The input data frame with added processed ranking data, model data and confidence interval data, the adjusted parameters, the confidence interval of the parameters, the linear model, a summary of the model, a generated function for use with raw numeric data and a ggplot2 object that shows the DGBD distribution and observed data, a model\_extra vector with 2 elements, model\_extra name and value.

### Examples

```
plotWeblinks <- BC_plot(Weblinks, column=2, rank_threshold=4, confint=FALSE, confrange=FALSE, plot_silent=TRUE)
head(plotWeblinks[[1]])
plotWeblinks[2:8]

plothmp_wgs <- BC_plot(DGBD::hmp_wgs, 2, obs=FALSE, plot_silent=TRUE)
head(plothmp_wgs[[1]])
plothmp_wgs[2:8]

plotBillionaires <- BC_plot(Billionaires, column= 2, gfx_xy_trans=c("log10", "log10"), plot_silent=TRUE)
head(plotBillionaires[[1]])
plotBillionaires[2:8]
```

---

BC\_random

---

*Generate random points on a DGB Distribution.*


---

### Description

Receives a set of parameters and generates a set of "x" ranking points that are uniformly random in a linear space and then calculates their expected "y" or abundance values. Function derived from: Fontanelli, O., Miramontes, P., Mansilla, R., Cocho, G., & Li, W. (2022). Beta rank function: A smooth double-Pareto-like distribution. Communications in Statistics - Theory and Methods, 51(11), 3645–3668. <https://doi.org/10.1080/03610926.2020.1800739>

### Usage

```
BC_random(parameters)
```

**Arguments**

parameters      A numeric vector (A,a,b,max,n).

**Value**

A data frame with x,y values

**Examples**

```
BC_random(c(25,0.3,0.1,50,4))

plot(BC_random(c(25,0.3,0.1,50,100)))
```

---

|         |                                                     |
|---------|-----------------------------------------------------|
| BC_rank | <i>Rank the Abundance from a File or Data Frame</i> |
|---------|-----------------------------------------------------|

---

**Description**

Sorts the df\_abundance dataframe by the given column argument.

**Usage**

```
BC_rank(df_abundance, column, rank_threshold = 0, is_phyloseq = FALSE, ...)
```

**Arguments**

df\_abundance      The data frame to use. A string that contains a path can be used to load files.

column             Can be either a string containing the name of the column or an int that is its index

rank\_threshold    Optional. Integer. Discards rows if their column value is equal to or lower than it. Defaults to 0.

is\_phyloseq       Logical, optional. Use if the object to be analysed has the phyloseq class.

...                passes arguments to read.table.

**Value**

The input data frame with a column appended to the left that stores the generated ranking. The column used for ranking is renamed "abundance".

**Examples**

```
rankedWeblinks <- BC_rank(df_abundance=DGBD::Weblinks, column=2, rank_threshold=4)
head(rankedWeblinks)

rankedTara <- BC_rank(Tara_Data,3,is_phyloseq=TRUE)
head(rankedTara)
```

BC\_report

*Summarize the Modeled DGB Distribution***Description**

Summarize the Modeled DGB Distribution

**Usage**

```
BC_report(
  df_abundance = NULL,
  column = NULL,
  BC_plot_object = NULL,
  show_plot = TRUE,
  show_stats = TRUE,
  ...
)
```

**Arguments**

|                |                                                                                                                        |
|----------------|------------------------------------------------------------------------------------------------------------------------|
| df_abundance   | A data frame that contains abundance data.                                                                             |
| column         | Either a string with the name of the column or the number of the column that stores the abundances in the data frame.  |
| BC_plot_object | A previous object generated by BC_plot.                                                                                |
| show_plot      | Logical. Whether to plot the DGB distribution. Defaults to true.                                                       |
| show_stats     | Logical. Whether to print the parameters of the DGB distribution and its model_extra to the console. Defaults to true. |
| ...            | passes arguments to BC_plot.                                                                                           |

**Value**

A list with the following elements: a matrix that contains the parameters used and their confidence intervals, the model\_extra vector and the DGBD plot.

**Examples**

```
BC_report(EC_Metabolite,2)

BC_report(df_abundance=DGBD::MOMv3.3, column=7, show_stats=FALSE)

BC_report(Tara_Data,2,is_phyloseq=TRUE)

EC_codon_dir <- system.file("extdata", "EC_Codon.csv", package = "DGBD")
EC_codon <- read.csv(EC_codon_dir,header=TRUE,row.names=1)
BC_report(EC_codon, column=1)
```

---

Billionaires

*Billionaires.*


---

### Description

The degree distribution of links between pages in a network built from a 1997 web crawl.

### Usage

Billionaires

### Format

Billionaires:

Data frame that contains.

**rank** The position in the rank of the individual.

**finalWorth** The aggregate net worth of the individual. ...

### Details

" 'The cumulative distribution of the total wealth of the richest people in the United States. Wealth is defined as aggregate net worth, i.e., total value in dollars at current market prices of all an individual's holdings, minus their debts. ... The data are from Forbes magazine, 6 October 2003.' Quoted from: Newman, M. (2005). Power laws, Pareto distributions and Zipf's law. Contemporary Physics, 46(5), 323–351. <https://doi.org/10.1080/00107510500052444>"

### Source

<https://github.com/robertoalvarezm/DataNonPowerLaws>

---

Citation\_ISICatalogued

*Citation\_ISICatalogued.*


---

### Description

Distribution of how many times a paper was cited, taken from ISI journals.

### Usage

Citation\_ISICatalogued

### Format

Citation\_ISICatalogued:

Data frame that contains.

**N** For a given number of citations:

**Number.of.papers.with.N.citations** How many articles there are. ...

## Details

" 'This distribution covers all publications from ISI catalogued journals that were published in 1981 and cited during the period January 1981 - June 1997. This information was compiled by H. Small and D. Pendlebury of the Institute for Scientific Information.' Quoted from: ISI Citation Data. (n.d.). Retrieved September 11, 2024, from <https://physics.bu.edu/~redner/projects/citation/isi.html> "

## Source

<https://github.com/robertoalvarezm/DataNonPowerLaws>

---

|          |                  |
|----------|------------------|
| EC_Codon | <i>EC_Codon.</i> |
|----------|------------------|

---

## Description

The codon frequency of a dozen different strains of Escherichia coli.

## Usage

EC\_Codon

## Format

EC\_Codon:

Data frame that contains.

**NC\_008253** Codon freq for NC\_008253 strain.  
**NC\_008563** Codon freq for NC\_008563 strain.  
**NC\_010468** Codon freq for NC\_010468 strain.  
**NC\_004431** Codon freq for NC\_004431 strain.  
**NC\_009801** Codon freq for NC\_009801 strain.  
**NC\_009800** Codon freq for NC\_009800 strain.  
**NC\_002655** Codon freq for NC\_002655 strain.  
**NC\_002695** Codon freq for NC\_002695 strain.  
**NC\_010498** Codon freq for NC\_010498 strain.  
**NC\_007946** Codon freq for NC\_007946 strain.  
**NC\_010473** Codon freq for NC\_010473 strain.  
**NC\_000913** Codon freq for NC\_000913 strain. ...

## Details

"Pending citation"

## Source

<https://github.com/robertoalvarezm/DataNonPowerLaws>

---

|               |                       |
|---------------|-----------------------|
| EC_Metabolite | <i>EC_Metabolite.</i> |
|---------------|-----------------------|

---

### Description

The degrees of metabolites in the metabolic network of the bacterium Escherichia coli.

### Usage

EC\_Metabolite

### Format

EC\_Metabolite:

Data frame that contains.

**degree** The number of connections the metabolite has in the network.

**count** How many metabolites have this number of connections (degree distribution of the network). ...

### Details

" Mirrored at <https://aaronclauset.github.io/powerlaws/data.htm>"

### Source

<https://github.com/robertoalvarezm/DataNonPowerLaws>

---

|        |                |
|--------|----------------|
| Flares | <i>Flares.</i> |
|--------|----------------|

---

### Description

Peak gamma-ray intensity of solar flares.

### Usage

Flares

### Format

Flares:

Data frame that contains.

**intensity** The number of links pointing to the page ...

**Details**

" 'The cumulative distribution of the peak gamma-ray intensity of solar flares. The observations were made between 1980 and 1989 by the instrument known as the Hard X-Ray Burst Spectrometer aboard the Solar Maximum Mission satellite launched in 1980. The spectrometer used a CsI scintillation detector to measure gamma-rays from solar flare... The data are from the NASA Goddard Space Flight Center, [umbra.nascom.nasa.gov/smm/hxrbs.html](http://umbra.nascom.nasa.gov/smm/hxrbs.html).' Quoted from: Newman, M. (2005). Power laws, Pareto distributions and Zipf's law. Contemporary Physics, 46(5), 323–351. <https://doi.org/10.1080/00107510500052444>"

**Source**

<https://github.com/robertoalvarezm/DataNonPowerLaws>

---

|         |                                       |
|---------|---------------------------------------|
| hmp_wgs | <i>Human Microbiome Project WGS .</i> |
|---------|---------------------------------------|

---

**Description**

Whole Genome Sequencing data obtained from the Human Microbiome Project production phase I.

**Usage**

hmp\_wgs

**Format**

hmp\_wgs:

Data frame that contains.

**PID** Module Table ID that identifies organisms that contain complete KEGG modules

**Abundance** Abundance data ...

**Details**

"The overall mission of the HMP is to generate resources to facilitate characterization of the human microbiota to further our understanding of how the microbiome impacts human health and disease. We are currently in the second phase of the HMP. In this phase, the iHMP will create integrated longitudinal datasets from both the microbiome and host from three different cohort studies of microbiome-associated conditions using multiple 'omics technologies. Each of these study groups has engaged in providing new computational tools and integrative molecular perspectives on microbial activity during dysbiosis. As a result of creating these multi'omic data resources, the iHMP has opened up new opportunities for data integration in the human microbiome." <https://www.hmpdacc.org/ihmp/>

**Source**

<https://portal.hmpdacc.org/files/1670203039de370df9a35a043735e964>

MOMv3.3

MOMv3.3.

**Description**

Average body mass of late quaternary mammals.

**Usage**

MOMv3.3

**Format**

MOMv3.3:

Data frame that contains.

**Continent** Continent that a species resides on. If species resides on more than one continent, a continent specific body mass is reported when available. Thus, some mammals have multiple entries. The division between North and South America occurs at the isthmus of Panama.

**Status** Whether species is currently present in the wild (extant); extinct as of late Pleistocene (extinct), extinct within the last 300 years (historical); or an introduction (introduction); Note these do not necessarily follow CITES or IUCN categories.

**Order** Taxonomic order of species

**Family** Taxonomic family of species

**Genus** Taxonomic genus of species

**Species** Species epithet

**Log mass** Log10 transformation of Combined Mass

**Combined mass** Adult body mass averaged across males and females and geographic locations.

**Reference** Reference source for body mass information and/or status for that species; the updated electronic version of Wilson and Reeder (1993). (6 June 2002; [www.nmnh.si.edu/msw](http://www.nmnh.si.edu/msw)) serves as the status reference for all extant species. ...

**Details**

"-999 is a missing value code. 'The purpose of this data set was to compile body mass information for all mammals on Earth so that we could investigate the patterns of body mass seen across geographic and taxonomic space and evolutionary time.' Quoted from: Smith, F. A., Lyons, S. K., Ernest, S. K. M., Jones, K. E., Kaufman, D. M., Dayan, T., Marquet, P. A., Brown, J. H., & Haskell, J. P. (2003). Body Mass of Late Quaternary Mammals. *Ecology*, 84(12), 3403–3403. <https://doi.org/10.1890/02-9003> Metadata available: Ecological Archives E084-093-metadata. (n.d.). Retrieved September 11, 2024, from <https://esapubs.org/archive/ecol/E084/094/metadata.htm>"

**Source**

<https://github.com/robertoalvarezm/DataNonPowerLaws>



Phy\_Citation

*Phy\_Citation.***Description**

The degree distribution of links between pages in a network built from a 1997 web crawl.

**Usage**

Phy\_Citation

**Format**

Phy\_Citation:

Data frame that contains.

**Volume** Physical Review D Volume

**page number** Physical Review D page number being referenced.

**Number of citations** How many articles were cited in the given page ...

**Details**

" 'This data covers all citations from the SPIRES High-Energy Physics Database at SLAC as of June 1997 and was compiled by Dr. H. Galic. ... The total number of articles in this list is 24,296. Warning and Disclaimer (May 19, 1998) I just discovered that there is inaccuracy in this data which stems from authors giving incorrect page numbers for citations. This leads, for example, to citations to 524 separate articles in volume 13 while there are, in fact, only 426 articles in this volume! This gives an idea of the number of articles in the following list that are actually non-existent. SPIRES has been informed of this disparity. Use this data with caution!' Quoted From: Physical Review D Citation Data 1975-1994. (n.d.). Retrieved September 11, 2024, from <https://physics.bu.edu/~redner/projects/citation/prd.html> The corrected 2003 data is not available at this moment."

**Source**

<https://github.com/robertoalvarezm/DataNonPowerLaws>

Quakes

*Quakes.***Description**

The distribution of earthquake magnitudes in California between 1910 and 1992.

**Usage**

Quakes

**Format**

Quakes:

Data frame that contains.

**magnitude** The magnitude of an earthquake ...

**Details**

" 'Magnitude of earthquakes: The cumulative distribution of the Richter (local) magnitude of earthquakes occurring in California between January 1910 and May 1992, as recorded in the Berkeley Earthquake Catalog. The Richter magnitude is defined as the logarithm, base 10, of the maximum amplitude of motion detected in the earthquake. ... The data are from the National Geophysical Data Center, [www.ngdc.noaa.gov](http://www.ngdc.noaa.gov). ' Quoted from: Newman, M. (2005). Power laws, Pareto distributions and Zipf's law. Contemporary Physics, 46(5), 323–351. <https://doi.org/10.1080/00107510500052444>"

**Source**

<https://github.com/robertoalvarezm/DataNonPowerLaws>

---

|           |                         |
|-----------|-------------------------|
| Tara_Data | <i>Tara Oceans data</i> |
|-----------|-------------------------|

---

**Description**

A subset (3 samples) of data from Shotgun Sequencing of Tara Oceans DNA samples corresponding to size fractions for small DNA viruses.

**Usage**

Tara\_Data

**Format**

Tara\_Data:

Phyloseq-class object with 3 samples.

**otu\_table()** 1701 taxa and 3 samples

**sample\_data()** 3 samples by 61 sample variables

**tax\_table()** 1701 taxa by 7 taxonomic ranks ...

**Details**

"Tara Expeditions are global scientific voyages that probe morphological and molecular diversity, evolution and ecology of marine plankton to explore how they are impacted by changes in the Earth's climate. The first expeditions collected samples of marine plankton containing viruses, bacteria, archaea, protists and planktonic metazoans living in the photic layer of the world's oceans. These expeditions, which took place between 2009 and 2013, include Tara Oceans: a global view, and Tara Oceans Polar Circle, both of which followed the same sampling protocol". <https://www.ebi.ac.uk/services/tara-oceans-data>

**Source**

<https://www.ebi.ac.uk/metagenomics/analyses/MGYA00133816>, <https://www.ebi.ac.uk/metagenomics/analyses/MGYA00133817>, <https://www.ebi.ac.uk/metagenomics/analyses/MGYA00133818>

---

|          |                  |
|----------|------------------|
| Weblinks | <i>Weblinks.</i> |
|----------|------------------|

---

**Description**

The degree distribution of links between pages in a network built from a 1997 web crawl.

**Usage**

Weblinks

**Format**

Weblinks:

Data frame that contains.

**degree** The number of links pointing to the page.

**frequency** The number of pages that have a given degree. The degree distribution of the network.

...

**Details**

" The dataset contains 'Two AltaVista crawls each with over 200 million pages and 1.5 billion links.' Broder, A., R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener, 2000, Computer Networks 33, 309. "

**Source**

<https://github.com/robertoalvarezm/DataNonPowerLaws>

# Index

## \* datasets

- Author\_Citations, [2](#)
- Billionaires, [12](#)
- Citation\_ISICatalogued, [12](#)
- EC\_Codon, [13](#)
- EC\_Metabolite, [14](#)
- Flares, [14](#)
- hmp\_wgs, [15](#)
- MOMv3.3, [16](#)
- Phy\_Citation, [17](#)
- Quakes, [17](#)
- Tara\_Data, [18](#)
- Weblinks, [19](#)

Author\_Citations, [2](#)

BC\_compare, [3](#)  
BC\_many, [4](#)  
BC\_model, [5](#)  
BC\_multiple, [5](#)  
BC\_param, [6](#)  
BC\_plot, [7](#)  
BC\_random, [9](#)  
BC\_rank, [10](#)  
BC\_report, [11](#)  
Billionaires, [12](#)

Citation\_ISICatalogued, [12](#)

EC\_Codon, [13](#)  
EC\_Metabolite, [14](#)

Flares, [14](#)

hmp\_wgs, [15](#)

MOMv3.3, [16](#)

Phy\_Citation, [17](#)

Quakes, [17](#)

Tara\_Data, [18](#)

Weblinks, [19](#)