Package 'EvaluateCore'

June 30, 2022

Title Quality Evaluation of Core Collections

Version 0.1.3

Description Implements various quality evaluation statistics to assess the value of plant germplasm core collections using qualitative and quantitative phenotypic trait data according to Odong et al. (2015) <doi:10.1007/s00122-012-1971-y>.

Copyright 2018-2022, ICAR-NBPGR

License GPL-2 | GPL-3

Encoding UTF-8

RoxygenNote 7.2.0

URL https://github.com/aravind-j/EvaluateCore
https://CRAN.R-project.org/package=EvaluateCore
https://aravind-j.github.io/EvaluateCore/
https://doi.org/10.5281/zenodo.3875930

BugReports https://github.com/aravind-j/EvaluateCore/issues

RdMacros mathjaxr,
Rdpack

Depends R (>= 3.5.0)

Imports agricolae,
boot,
car,
cluster,
dplyr,
entropy,
ggcorrplot,
ggplot2,
grDevices,
gridExtra,
kSamples,
mathjaxr,
psych,
reshape2,
Rdpack,
stats,
vegan

Suggests corehunter,
pander,
rJava (>= 0.9-8)

LazyData true

R topics documented:

`bar.evaluate.core` *Bar Plots*

Description

Plot Bar plots to graphically compare the frequency distributions of qualitative traits between entire collection (EC) and core set (CS).

Usage

```
bar.evaluate.core(data, names, qualitative, selected)
```

Arguments

`data`	The data as a data frame object. The data frame should possess one row per individual and columns with the individual names and multiple trait/character data.
`names`	Name of column with the individual names as a character string
`qualitative`	Name of columns with the qualitative traits as a character vector.

`selected` Character vector with the names of individuals selected in core collection and present in the `names` column.

Value

A list with the `ggplot` objects of relative frequency bar plots of CS and EC for each trait specified as `qualitative`.

See Also

`barplot`, `geom_bar`

Examples

```
data("cassava_CC")
data("cassava_EC")

ec <- cbind(genotypes = rownames(cassava_EC), cassava_EC)
ec$genotypes <- as.character(ec$genotypes)
rownames(ec) <- NULL

core <- rownames(cassava_CC)

quant <- c("NMSR", "TTRN", "TFWSR", "TTRW", "TFWSS", "TTSW", "TTPW", "AVPW",
           "ARSR", "SRDM")
qual <- c("CUAL", "LNGS", "PTLC", "DSTA", "LFRT", "LBTEF", "CBTR", "NMLB",
          "ANGB", "CUAL9M", "LVC9M", "TNPR9M", "PL9M", "STRP", "STRC",
          "PSTR")

ec[, qual] <- lapply(ec[, qual],
                     function(x) factor(as.factor(x)))

bar.evaluate.core(data = ec, names = "genotypes",
                  qualitative = qual, selected = core)
```

---

box.evaluate.core *Box Plots*

Description

Plot Box-and-Whisker plots (Tukey 1970; McGill et al. 1978) to graphically compare the probability distributions of quantitative traits between entire collection (EC) and core set (CS).

Usage

```
box.evaluate.core(data, names, quantitative, selected)
```

Arguments

`data` The data as a data frame object. The data frame should possess one row per individual and columns with the individual names and multiple trait/character data.

names	Name of column with the individual names as a character string
quantitative	Name of columns with the quantitative traits as a character vector.
selected	Character vector with the names of individuals selected in core collection and present in the names column.

Value

A list with the ggplot objects of box plots of CS and EC for each trait specified as quantitative.

References

McGill R, Tukey JW, Larsen WA (1978). "Variations of box plots." *The American Statistician*, **32**(1), 12.

Tukey JW (1970). *Exploratory Data Analysis. Preliminary edition*. Addison-Wesley.

See Also

boxplot, geom_boxplot

Examples

```
data("cassava_CC")
data("cassava_EC")

ec <- cbind(genotypes = rownames(cassava_EC), cassava_EC)
ec$genotypes <- as.character(ec$genotypes)
rownames(ec) <- NULL

core <- rownames(cassava_CC)

quant <- c("NMSR", "TTRN", "TFWSR", "TTRW", "TFWSS", "TTSW", "TTPW", "AVPW",
           "ARSR", "SRDM")
qual <- c("CUAL", "LNGS", "PTLC", "DSTA", "LFRT", "LBTEF", "CBTR", "NMLB",
          "ANGB", "CUAL9M", "LVC9M", "TNPR9M", "PL9M", "STRP", "STRC",
          "PSTR")

ec[, qual] <- lapply(ec[, qual],
                     function(x) factor(as.factor(x)))

box.evaluate.core(data = ec, names = "genotypes",
                  quantitative = quant, selected = core)
```

Description

An example germplasm characterisation data of a core collection generated from 1591 accessions of IITA Cassava collection (International Institute of Tropical Agriculture et al. 2019) using 10 quantitative and 48 qualitative trait data with CoreHunter3 (`corehunter`). The core set was generated using distance based measures giving equal weightage to Average entry-to-nearest-entry distance (EN) and Average accession-to-nearest-entry distance (AN). Includes data on 26 descriptors for 168 (10 % of `cassava_EC`) accessions. It is used to demonstrate the various functions of `EvaluateCore` package.

Usage

```
cassava_CC
```

Format

A data frame with 58 columns:

CUAL Colour of unexpanded apical leaves

LNGS Length of stipules

PTLC Petiole colour

DSTA Distribution of anthocyanin

LFRT Leaf retention

LBTEF Level of branching at the end of flowering

CBTR Colour of boiled tuberous root

NMLB Number of levels of branching

ANGB Angle of branching

CUAL9M Colours of unexpanded apical leaves at 9 months

LVC9M Leaf vein colour at 9 months

TNPR9M Total number of plants remaining per accession at 9 months

PL9M Petiole length at 9 months

STRP Storage root peduncle

STRC Storage root constrictions

PSTR Position of root

NMSR Number of storage root per plant

TTRN Total root number per plant

TFWSR Total fresh weight of storage root per plant

TTRW Total root weight per plant

TFWSS Total fresh weight of storage shoot per plant

TTSW Total shoot weight per plant

TTPW Total plant weight

AVPW Average plant weight

ARSR Amount of rotted storage root per plant

SRDM Storage root dry matter

Details

Further details on how the example dataset was built from the original data is available online.

References

International Institute of Tropical Agriculture, Benjamin F, Marimagne T (2019). "Cassava morphological characterization. Version 2018.1." *www.genesys-pgr.org*.

Examples

```
data(cassava_CC)
summary(cassava_CC)

quant <- c("NMSR", "TTRN", "TFWSR", "TTRW", "TFWSS", "TTSW", "TTPW", "AVPW",
           "ARSR", "SRDM")
qual <- c("CUAL", "LNGS", "PTLC", "DSTA", "LFRT", "LBTEF", "CBTR", "NMLB",
          "ANGB", "CUAL9M", "LVC9M", "TNPR9M", "PL9M", "STRP", "STRC",
          "PSTR")

lapply(seq_along(cassava_CC[, qual]),
       function(i) barplot(table(cassava_CC[, qual][, i]),
                           xlab = names(cassava_CC[, qual])[i]))

lapply(seq_along(cassava_CC[, quant]),
       function(i) hist(table(cassava_CC[, quant][, i]),
                        xlab = names(cassava_CC[, quant])[i],
                        main = ""))
```

---

cassava_EC *IITA Cassava Germplasm Data - Entire Collection*

Description

An example germplasm characterisation data of a subset of IITA Cassava collection (International Institute of Tropical Agriculture et al. 2019). Includes data on 26 (out of 62) descriptors for 1684 (out of 2170) accessions. It is used to demonstrate the various functions of `EvaluateCore` package.

Usage

```
cassava_EC
```

Format

A data frame with 58 columns:

CUAL Colour of unexpanded apical leaves

LNGS Length of stipules

PTLC Petiole colour

DSTA Distribution of anthocyanin

LFRT Leaf retention

LBTEF Level of branching at the end of flowering
CBTR Colour of boiled tuberous root
NMLB Number of levels of branching
ANGB Angle of branching
CUAL9M Colours of unexpanded apical leaves at 9 months
LVC9M Leaf vein colour at 9 months
TNPR9M Total number of plants remaining per accession at 9 months
PL9M Petiole length at 9 months
STRP Storage root peduncle
STRC Storage root constrictions
PSTR Position of root
NMSR Number of storage root per plant
TTRN Total root number per plant
TFWSR Total fresh weight of storage root per plant
TTRW Total root weight per plant
TFWSS Total fresh weight of storage shoot per plant
TTSW Total shoot weight per plant
TTPW Total plant weight
AVPW Average plant weight
ARSR Amount of rotted storage root per plant
SRDM Storage root dry matter

Details

Further details on how the example dataset was built from the original data is available online.

References

International Institute of Tropical Agriculture, Benjamin F, Marimagne T (2019). "Cassava morphological characterization. Version 2018.1." *www.genesys-pgr.org*.

Examples

```
data(cassava_EC)
summary(cassava_EC)

quant <- c("NMSR", "TTRN", "TFWSR", "TTRW", "TFWSS", "TTSW", "TTPW", "AVPW",
           "ARSR", "SRDM")
qual <- c("CUAL", "LNGS", "PTLC", "DSTA", "LFRT", "LBTEF", "CBTR", "NMLB",
          "ANGB", "CUAL9M", "LVC9M", "TNPR9M", "PL9M", "STRP", "STRC",
          "PSTR")

lapply(seq_along(cassava_EC[, qual]),
       function(i) barplot(table(cassava_EC[, qual][, i]),
                           xlab = names(cassava_EC[, qual])[i]))

lapply(seq_along(cassava_EC[, quant]),
       function(i) hist(table(cassava_EC[, quant][, i]),
                        xlab = names(cassava_EC[, quant])[i],
                        main = ""))
```

`chisquare.evaluate.core` *Chi-squared Test for Homogeneity*

Description

Compare the distribution frequencies of qualitative traits between entire collection (EC) and core set (CS) by Chi-squared test for homogeneity (Pearson 1900; Snedecor and Irwin 1933).

Usage

```
chisquare.evaluate.core(data, names, qualitative, selected)
```

Arguments

`data`	The data as a data frame object. The data frame should possess one row per individual and columns with the individual names and multiple trait/character data.
`names`	Name of column with the individual names as a character string
`qualitative`	Name of columns with the qualitative traits as a character vector.
`selected`	Character vector with the names of individuals selected in core collection and present in the `names` column.

Value

A a data frame with the following columns.

`Trait`	The qualitative trait.
`EC_No.Classes`	The number of classes in the trait for EC.
`EC_Classes`	The frequency of the classes in the trait for EC.
`CS_No.Classes`	The number of classes in the trait for CS.
`CS_Classes`	The frequency of the classes in the trait for CS.
`chisq_statistic`	The χ^2 test statistic.
`chisq_pvalue`	The p value for the test statistic.
`chisq_significance`	The significance of the test statistic (*: $p \leq 0.01$; **: $p \leq 0.05$; ns: $p > 0.05$).

References

Pearson K (1900). "X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling." *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, **50**(302), 157–175.

Snedecor G, Irwin MR (1933). "On the chi-square test for homogeneity." *Iowa State College Journal of Science*, **8**, 75–81.

See Also

chisq.test

Examples

```
data("cassava_CC")
data("cassava_EC")

ec <- cbind(genotypes = rownames(cassava_EC), cassava_EC)
ec$genotypes <- as.character(ec$genotypes)
rownames(ec) <- NULL

core <- rownames(cassava_CC)

quant <- c("NMSR", "TTRN", "TFWSR", "TTRW", "TFWSS", "TTSW", "TTPW", "AVPW",
           "ARSR", "SRDM")
qual <- c("CUAL", "LNGS", "PTLC", "DSTA", "LFRT", "LBTEF", "CBTR", "NMLB",
          "ANGB", "CUAL9M", "LVC9M", "TNPR9M", "PL9M", "STRP", "STRC",
          "PSTR")

ec[, qual] <- lapply(ec[, qual],
                     function(x) factor(as.factor(x)))

chisquare.evaluate.core(data = ec, names = "genotypes",
                        qualitative = qual, selected = core)
```

---

corr.evaluate.core *Phenotypic Correlations*

Description

Compute phenotypic correlations (Pearson 1895) between traits, plot correlation matrices as correlograms (Friendly 2002) and calculate mantel correlation (Legendre and Legendre 2012) between them to compare entire collection (EC) and core set (CS).

Usage

```
corr.evaluate.core(data, names, quantitative, qualitative, selected)
```

Arguments

data	The data as a data frame object. The data frame should possess one row per individual and columns with the individual names and multiple trait/character data.
names	Name of column with the individual names as a character string
quantitative	Name of columns with the quantitative traits as a character vector.
qualitative	Name of columns with the qualitative traits as a character vector.
selected	Character vector with the names of individuals selected in core collection and present in the names column.

Value

A list with the following components.

`Correlation Matrix`
: The matrix with phenotypic correlations between traits in EC (below diagonal) and CS (above diagonal).

`Correologram`
: A correlogram of phenotypic correlations between traits in EC (below diagonal) and CS (above diagonal) as a `ggplot` object.

`Mantel Correlation`
: A data frame with Mantel correlation coefficient (r) between EC and CS phenotypic correlation matrices, it's p value and significance (*: $p \leq 0.01$; **: $p \leq 0.05$; ns: $p > 0.05$).

References

Friendly M (2002). "Corrgrams." *The American Statistician*, **56**(4), 316–324.

Legendre P, Legendre L (2012). "Interpretation of ecological structures." In *Developments in Environmental Modelling*, volume 24, 521–624. Elsevier.

Pearson K (1895). "Note on regression and inheritance in the case of two parents." *Proceedings of the Royal Society of London*, **58**, 240–242.

See Also

`cor`, `cor_pmat` `ggcorrplot`, `mantel`

Examples

```
data("cassava_CC")
data("cassava_EC")

ec <- cbind(genotypes = rownames(cassava_EC), cassava_EC)
ec$genotypes <- as.character(ec$genotypes)
rownames(ec) <- NULL

core <- rownames(cassava_CC)

quant <- c("NMSR", "TTRN", "TFWSR", "TTRW", "TFWSS", "TTSW", "TTPW", "AVPW",
           "ARSR", "SRDM")
qual <- c("CUAL", "LNGS", "PTLC", "DSTA", "LFRT", "LBTEF", "CBTR", "NMLB",
          "ANGB", "CUAL9M", "LVC9M", "TNPR9M", "PL9M", "STRP", "STRC",
          "PSTR")

ec[, qual] <- lapply(ec[, qual],
                     function(x) factor(as.factor(x)))

corr.evaluate.core(data = ec, names = "genotypes", quantitative = quant,
                   qualitative = qual, selected = core)
```

`coverage.evaluate.core` *Class Coverage*

Description

Compute the Class Coverage (Kim et al. 2007) to compare the distribution frequencies of qualitative traits between entire collection (EC) and core set (CS).

Usage

```
coverage.evaluate.core(data, names, qualitative, selected)
```

Arguments

`data`	The data as a data frame object. The data frame should possess one row per individual and columns with the individual names and multiple trait/character data.
`names`	Name of column with the individual names as a character string
`qualitative`	Name of columns with the qualitative traits as a character vector.
`selected`	Character vector with the names of individuals selected in core collection and present in the `names` column.

Details

Class Coverage (Kim et al. 2007) is computed as follows.

$$Class\,Coverage = \left(\frac{1}{n} \sum_{i=1}^{n} \frac{A_{CS_i}}{A_{EC_i}} \right) \times 100$$

Where, A_{CS_i} is the sets of categories in the CS for the ith trait, A_{EC_i} is the sets of categories in the EC for the ith trait and n is the total number of traits.

Value

The Class Coverage value.

References

Kim K, Chung H, Cho G, Ma K, Chandrabalan D, Gwag J, Kim T, Cho E, Park Y (2007). "PowerCore: A program applying the advanced M strategy with a heuristic search for establishing core sets." *Bioinformatics*, **23**(16), 2155–2162.

Examples

```
data("cassava_CC")
data("cassava_EC")

ec <- cbind(genotypes = rownames(cassava_EC), cassava_EC)
ec$genotypes <- as.character(ec$genotypes)
rownames(ec) <- NULL
```

```
core <- rownames(cassava_CC)

quant <- c("NMSR", "TTRN", "TFWSR", "TTRW", "TFWSS", "TTSW", "TTPW", "AVPW",
           "ARSR", "SRDM")
qual <- c("CUAL", "LNGS", "PTLC", "DSTA", "LFRT", "LBTEF", "CBTR", "NMLB",
          "ANGB", "CUAL9M", "LVC9M", "TNPR9M", "PL9M", "STRP", "STRC",
          "PSTR")

ec[, qual] <- lapply(ec[, qual],
                     function(x) factor(as.factor(x)))

coverage.evaluate.core(data = ec, names = "genotypes",
                       qualitative = qual, selected = core)
```

`cr.evaluate.core` *Coincidence Rate of Range*

Description

Compute the Coincidence Rate of Range (CR) (Hu et al. 2000) (originally described by (Diwan et al. 1995) as Mean range ratio) to compare quantitative traits of the entire collection (EC) and core set (CS).

Usage

```
cr.evaluate.core(data, names, quantitative, selected)
```

Arguments

`data`	The data as a data frame object. The data frame should possess one row per individual and columns with the individual names and multiple trait/character data.
`names`	Name of column with the individual names as a character string
`quantitative`	Name of columns with the quantitative traits as a character vector.
`selected`	Character vector with the names of individuals selected in core collection and present in the `names` column.

Details

The Coincidence Rate of Range (CR) is computed as follows.

$$CR = \left(\frac{1}{n} \sum_{i=1}^{n} \frac{R_{CS_i}}{R_{EC_i}} \right) \times 100$$

Where, R_{CS_i} is the range of the ith trait in the CS, R_{EC_i} is the range of the ith trait in the EC and n is the total number of traits.

A representative CS should have a CR value no less than 70% (Diwan et al. 1995) or 80% (Hu et al. 2000).

Value

The CR value.

References

Diwan N, McIntosh MS, Bauchan GR (1995). "Methods of developing a core collection of annual *Medicago* species." *Theoretical and Applied Genetics*, **90**(6), 755–761.

Hu J, Zhu J, Xu HM (2000). "Methods of constructing core collections by stepwise clustering with three sampling strategies based on the genotypic values of crops." *Theoretical and Applied Genetics*, **101**(1), 264–268.

See Also

`wilcox.test`

Examples

```
data("cassava_CC")
data("cassava_EC")

ec <- cbind(genotypes = rownames(cassava_EC), cassava_EC)
ec$genotypes <- as.character(ec$genotypes)
rownames(ec) <- NULL

core <- rownames(cassava_CC)

quant <- c("NMSR", "TTRN", "TFWSR", "TTRW", "TFWSS", "TTSW", "TTPW", "AVPW",
           "ARSR", "SRDM")
qual <- c("CUAL", "LNGS", "PTLC", "DSTA", "LFRT", "LBTEF", "CBTR", "NMLB",
          "ANGB", "CUAL9M", "LVC9M", "TNPR9M", "PL9M", "STRP", "STRC",
          "PSTR")

ec[, qual] <- lapply(ec[, qual],
                     function(x) factor(as.factor(x)))

cr.evaluate.core(data = ec, names = "genotypes",
                 quantitative = quant, selected = core)
```

---

dist.evaluate.core *Distance Measures*

Description

Compute average Entry-to-nearest-entry distance (E-EN), Accession-to-nearest-entry distance (A-EN) and Entry-to-entry distance (E-E) (Odong et al. 2013) to evaluate a core set (CS) selected from an entire collection (EC).

Usage

```
dist.evaluate.core(data, names, quantitative, qualitative, selected, d = NULL)
```

Arguments

data	The data as a data frame object. The data frame should possess one row per individual and columns with the individual names and multiple trait/character data.
names	Name of column with the individual names as a character string
quantitative	Name of columns with the quantitative traits as a character vector.
qualitative	Name of columns with the qualitative traits as a character vector.
selected	Character vector with the names of individuals selected in core collection and present in the names column.
d	A distance matrix of class "dist" with individual names in the names column in data as labels. If NULL (default), then a distance matrix is computed using Gower's metric. (Gower 1971).

Value

A data frame with the average values of E-EN, A-EN and E-E.

References

Gower JC (1971). "A general coefficient of similarity and some of its properties." *Biometrics*, **27**(4), 857–871.

Odong TL, Jansen J, van Eeuwijk FA, van Hintum TJL (2013). "Quality of core collections for effective utilisation of genetic resources review, discussion and interpretation." *Theoretical and Applied Genetics*, **126**(2), 289–305.

See Also

evaluateCore

Examples

```
data("cassava_CC")
data("cassava_EC")

ec <- cbind(genotypes = rownames(cassava_EC), cassava_EC)
ec$genotypes <- as.character(ec$genotypes)
rownames(ec) <- NULL

core <- rownames(cassava_CC)

quant <- c("NMSR", "TTRN", "TFWSR", "TTRW", "TFWSS", "TTSW", "TTPW", "AVPW",
           "ARSR", "SRDM")
qual <- c("CUAL", "LNGS", "PTLC", "DSTA", "LFRT", "LBTEF", "CBTR", "NMLB",
          "ANGB", "CUAL9M", "LVC9M", "TNPR9M", "PL9M", "STRP", "STRC",
          "PSTR")

ec[, qual] <- lapply(ec[, qual],
                     function(x) factor(as.factor(x)))

dist.evaluate.core(data = ec, names = "genotypes", quantitative = quant,
                   qualitative = qual, selected = core)
```

```
###################################
# Compare with corehunter
###################################

library(corehunter)
# Prepare phenotype dataset
dtype <- c(rep("RD", length(quant)),
           rep("NS", length(qual)))
rownames(ec) <- ec[, "genotypes"]
ecdata <- corehunter::phenotypes(data = ec[, c(quant, qual)],
                                 types = dtype)

# Compute average distances
EN <- evaluateCore(core = rownames(cassava_CC), data = ecdata,
                   objective = objective("EN", "GD"))
AN <- evaluateCore(core = rownames(cassava_CC), data = ecdata,
                   objective = objective("AN", "GD"))
EE <- evaluateCore(core = rownames(cassava_CC), data = ecdata,
                   objective = objective("EE", "GD"))
EN
AN
EE
```

`diversity.evaluate.core` *Diversity Indices*

Description

Compute the following diversity indices and perform corresponding statistical tests to compare the phenotypic diversity for qualitative traits between entire collection (EC) and core set (CS).

- Simpson's and related indices
  - Simpson's Index (d) (Simpson 1949; Peet 1974)
  - Simpson's Index of Diversity or Gini's Diversity Index or Gini-Simpson Index or Nei's Diversity Index or Nei's Variation Index (D) (Gini 1912, 1912; Greenberg 1956; Berger and Parker 1970; Nei 1973; Peet 1974)
  - Maximum Simpson's Index of Diversity or Maximum Nei's Diversity/Variation Index (D_{max}) (Hennink and Zeven 1990)
  - Simpson's Reciprocal Index or Hill's N_2 (D_R) (Williams 1964; Hill 1973)
  - Relative Simpson's Index of Diversity or Relative Nei's Diversity/Variation Index (D') (Hennink and Zeven 1990)
- Shannon-Weaver and related indices
  - Shannon or Shannon-Weaver or Shannon-Weiner Diversity Index (H) (Shannon and Weaver 1949; Peet 1974)
  - Maximum Shannon-Weaver Diversity Index (H_{max}) (Hennink and Zeven 1990)
  - Relative Shannon-Weaver Diversity Index or Shannon Equitability Index (H') (Hennink and Zeven 1990)
- McIntosh Diversity Index
  - McIntosh Diversity Index (D_{Mc}) (McIntosh 1967; Peet 1974)

Usage

```
diversity.evaluate.core(data, names, qualitative, selected, base = 2, R = 1000)
```

Arguments

`data`	The data as a data frame object. The data frame should possess one row per individual and columns with the individual names and multiple trait/character data.
`names`	Name of column with the individual names as a character string
`qualitative`	Name of columns with the qualitative traits as a character vector.
`selected`	Character vector with the names of individuals selected in core collection and present in the `names` column.
`base`	The logarithm base to be used for computation of Shannon-Weaver Diversity Index (I). Default is 2.
`R`	The number of bootstrap replicates. Default is 1000.

Value

A list with three data frames as follows.

`simpson`

Trait The qualitative trait.
EC_No.Classes The number of classes in the trait for EC.
CS_No.Classes The number of classes in the trait for CS.
EC_d The Simpson's Index (d) for EC.
EC_D The Simpson's Index of Diversity (D) for EC.
EC_D.max The Maximum Simpson's Index of Diversity (D_{max}) for EC.
EC_D.inv The Simpson's Reciprocal Index (D_R) for EC.
EC_D.rel The Relative Reciprocal Index (D') for EC.
EC_d.V The variance of d for EC according to (Simpson 1949).
EC_d.boot.V The bootstrap variance of d for EC.
CS_d The Simpson's Index (d) for CS.
CS_D The Simpson's Index of Diversity (D) for CS.
CS_D.max The Maximum Simpson's Index of Diversity (D_{max}) for CS.
CS_D.inv The Simpson's Reciprocal Index (D_R) for CS.
CS_D.rel The Relative Reciprocal Index (D') for CS.
CS_d.V The variance of d for CS according to (Simpson 1949).
CS_d.boot.V The bootstrap variance of d for CS.
d.t.df The degrees of freedom for t test.
d.t.stat The t statistic.
d.t.pvalue The p value for t test.
d.t.significance The significance of t test for t-test
d.boot.z.df The degrees of freedom for bootstrap z score.
d.boot.z.stat The bootstrap z score.
d.boot.z.pvalue The p value of z score.
d.boot.z.significance The significance of z score.

`shannon`

Trait The qualitative trait.
EC_No.Classes The number of classes in the trait for EC.

CS_No.Classes The number of classes in the trait for CS.
EC_I The Shannon-Weaver Diversity Index (I) for EC.
EC_I.max The Maximum Shannon-Weaver Diversity Index (I_{max}) for EC.
EC_I.rel The Relative Shannon-Weaver Diversity Index (I') for EC.
EC_I.V The variance of I for EC according to (Hutcheson 1970).
EC_I.boot.V The bootstrap variance of I for EC.
CS_I The Shannon-Weaver Diversity Index (I) for CS.
CS_I.max The Maximum Shannon-Weaver Diversity Index (I_{max}) for CS.
CS_I.rel The Relative Shannon-Weaver Diversity Index (I') for CS.
CS_I.V The variance of I for CS according to (Hutcheson 1970).
CS_I.boot.V The bootstrap variance of I for CS.
I.t.stat The t statistic.
I.t.df The degrees of freedom for t test.
I.t.pvalue The p value for t test.
I.t.significance The significance of t test for t-test
I.boot.z.df The degrees of freedom for bootstrap z score.
I.boot.z.stat The bootstrap z score.
I.boot.z.pvalue The p value of z score.
I.boot.z.significance The significance of z score.

`mcintosh`
EC_No.Classes The number of classes in the trait for EC.
CS_No.Classes The number of classes in the trait for CS.
EC_D.Mc The McIntosh Index (D_{Mc}) for EC.
CS_D.Mc The McIntosh Index (D_{Mc}) for CS.
M.boot.z.stat The bootstrap z score.
M.boot.z.df The degrees of freedom for bootstrap z score.
M.boot.z.pvalue The p value of z score.
M.boot.z.significance The significance of z score.

Details

The diversity indices and the corresponding statistical tests implemented in `diversity.evaluate.core` are as follows.

Simpson's and related indices: Simpson's index (d) which estimates the probability that two accessions randomly selected will belong to the same phenotypic class of a trait, is computed as follows (Simpson 1949; Peet 1974).

$$d = \sum_{i=1}^{k} p_i^2$$

Where, p_i denotes the proportion/fraction/frequency of accessions in the ith phenotypic class for a trait and k is the number of phenotypic classes for the trait.

The value of d can range from 0 to 1 with 0 representing maximum diversity and 1, no diversity.

d is subtracted from 1 to give Simpson's index of diversity (D) (Greenberg 1956; Berger and Parker 1970; Peet 1974; Hennink and Zeven 1990) originally suggested by Gini (1912, 1912) and described in literature as Gini's diversity index or Gini-Simpson index. It is the same as Nei's diversity index or Nei's variation index (Nei 1973; Hennink and Zeven 1990). Greater the value of D, greater the diversity with a range from 0 to 1.

$$D = 1 - d$$

The maximum value of D, D_{max} occurs when accessions are uniformly distributed across the phenotypic classes and is computed as follows (Hennink and Zeven 1990).

$$D_{max} = 1 - \frac{1}{k}$$

Reciprocal of d gives the Simpson's reciprocal index (D_R) (Williams 1964; Hennink and Zeven 1990) and can range from 1 to k. This was also described in Hill (1973) as (N_2).

$$D_R = \frac{1}{d}$$

Relative Simpson's index of diversity or Relative Nei's diversity/variation index (H') (Hennink and Zeven 1990) is defined as follows (Peet 1974).

$$D' = \frac{D}{D_{max}}$$

Differences in Simpson's diversity index for qualitative traits of EC and CS can be tested by a t-test using the associated variance estimate described in Simpson (1949) (Lyons and Hutcheson 1978).
The t statistic is computed as follows.

$$t = \frac{d_{EC} - d_{CS}}{\sqrt{V_{d_{EC}} + V_{d_{CS}}}}$$

Where, the variance of d (V_d) is,

$$V_d = \frac{4N(N-1)(N-2)\sum_{i=1}^{k}(p_i)^3 + 2N(N-1)\sum_{i=1}^{k}(p_i)^2 - 2N(N-1)(2N-3)\left(\sum_{i=1}^{k}(p_i)^2\right)^2}{[N(N-1)]^2}$$

The associated degrees of freedom is computed as follows.

$$df = (k_{EC} - 1) + (k_{CS} - 1)$$

Where, k_{EC} and k_{CS} are the number of phenotypic classes in the trait for EC and CS respectively.

Shannon-Weaver and related indices: An index of information H, was described by Shannon and Weaver (1949) as follows.

$$H = -\sum_{i=1}^{k} p_i \log_2(p_i)$$

H is described as Shannon or Shannon-Weaver or Shannon-Weiner diversity index in literature. Alternatively, H is also computed using natural logarithm instead of logarithm to base 2.

$$H = -\sum_{i=1}^{k} p_i \ln(p_i)$$

The maximum value of H (H_{max}) is $\ln(k)$. This value occurs when each phenotypic class for a trait has the same proportion of accessions.

$$H_{max} = \log_2(k) \text{ OR } H_{max} = \ln(k)$$

The relative Shannon-Weaver diversity index or Shannon equitability index (H') is the Shannon diversity index (I) divided by the maximum diversity (H_{max}).

$$H' = \frac{H}{H_{max}}$$

Differences in Shannon-Weaver diversity index for qualitative traits of EC and CS can be tested by Hutcheson t-test (Hutcheson 1970).
The Hutcheson t statistic is computed as follows.

$$t = \frac{H_{EC} - H_{CS}}{\sqrt{V_{H_{EC}} + V_{H_{CS}}}}$$

Where, the variance of H (V_H) is,

$$V_H = \frac{\sum_{i=1}^{k} n_i (\log_2 n_i)^2 \frac{(\sum_{i=1}^{k} \log_2 n_i)^2}{N}}{N^2}$$

OR

$$V_H = \frac{\sum_{i=1}^{k} n_i (\ln n_i)^2 \frac{(\sum_{i=1}^{k} \ln n_i)^2}{N}}{N^2}$$

The associated degrees of freedom is approximated as follows.

$$df = \frac{(V_{H_{EC}} + V_{H_{CS}})^2}{\frac{V_{H_{EC}}^2}{N_{EC}} + \frac{V_{H_{CS}}^2}{N_{CS}}}$$

McIntosh Diversity Index: A similar index of diversity was described by McIntosh (1967) as follows (D_{Mc}) (Peet 1974).

$$D_{Mc} = \frac{N - \sqrt{\sum_{i=1}^{k} n_i^2}}{N - \sqrt{N}}$$

Where, n_i denotes the number of accessions in the ith phenotypic class for a trait and N is the total number of accessions so that $p_i = n_i/N$.

Testing for difference with bootstrapping: Bootstrap statistics are employed to test the difference between the Simpson, Shannon-Weaver and McIntosh indices for qualitative traits of EC and CS (Solow 1993).
If I_{EC} and I_{CS} are the diversity indices with the original number of accessions, then random samples of the same size as the original are repeatedly generated (with replacement) R times and the corresponding diversity index is computed for each sample.

$$I_{EC}^{*} = \{H_{EC_1}, H_{EC}, \cdots, H_{EC_R}\}$$

$$I^*_{CS} = \{H_{CS_1}, H_{CS}, \cdots, H_{CS_R}\}$$

Then the bootstrap null sample I_0 is computed as follows.

$$\Delta^* = I^*_{EC} - I^*_{CS}$$

$$I_0 = \Delta^* - \overline{\Delta^*}$$

Where, $\overline{\Delta^*}$ is the mean of Δ^*.
Now the original difference in diversity indices ($\Delta_0 = I_{EC} - I_{CS}$) is tested against mean of bootstrap null sample (I_0) by a z test. The z score test statistic is computed as follows.

$$z = \frac{\Delta_0 - \overline{H_0}}{\sqrt{V_{H_0}}}$$

Where, $\overline{H_0}$ and V_{H_0} are the mean and variance of the bootstrap null sample H_0.
The corresponding degrees of freedom is estimated as follows.

$$df = (k_{EC} - 1) + (k_{CS} - 1)$$

References

Berger WH, Parker FL (1970). "Diversity of planktonic foraminifera in deep-sea sediments." *Science*, **168**(3937), 1345–1347.

Gini C (1912). *Variabilita e Mutabilita. Contributo allo Studio delle Distribuzioni e delle Relazioni Statistiche. [Fasc. I.]*. Tipogr. di P. Cuppini, Bologna.

Gini C (1912). "Variabilita e mutabilita." In Pizetti E, Salvemini T (eds.), *Memorie di Metodologica Statistica*. Liberia Eredi Virgilio Veschi, Roma, Italy.

Greenberg JH (1956). "The measurement of linguistic diversity." *Language*, **32**(1), 109.

Hennink S, Zeven AC (1990). "The interpretation of Nei and Shannon-Weaver within population variation indices." *Euphytica*, **51**(3), 235–240.

Hill MO (1973). "Diversity and evenness: A unifying notation and its consequences." *Ecology*, **54**(2), 427–432.

Hutcheson K (1970). "A test for comparing diversities based on the Shannon formula." *Journal of Theoretical Biology*, **29**(1), 151–154.

Lyons NI, Hutcheson K (1978). "C20. Comparing diversities: Gini's index." *Journal of Statistical Computation and Simulation*, **8**(1), 75–78.

McIntosh RP (1967). "An index of diversity and the relation of certain concepts to diversity." *Ecology*, **48**(3), 392–404.

Nei M (1973). "Analysis of gene diversity in subdivided populations." *Proceedings of the National Academy of Sciences*, **70**(12), 3321–3323.

Peet RK (1974). "The measurement of species diversity." *Annual Review of Ecology and Systematics*, **5**(1), 285–307.

Shannon CE, Weaver W (1949). *The Mathematical Theory of Communication*, number v. 2 in The Mathematical Theory of Communication. University of Illinois Press.

Simpson EH (1949). "Measurement of diversity." *Nature*, **163**(4148), 688–688.

Solow AR (1993). "A simple test for change in community structure." *The Journal of Animal Ecology*, **62**(1), 191.

Williams CB (1964). *Patterns in the Balance of Nature and Related Problems in Quantitative Ecology*. Academic Press.

See Also

shannon, diversity, boot

Examples

```
data("cassava_CC")
data("cassava_EC")

ec <- cbind(genotypes = rownames(cassava_EC), cassava_EC)
ec$genotypes <- as.character(ec$genotypes)
rownames(ec) <- NULL

core <- rownames(cassava_CC)

quant <- c("NMSR", "TTRN", "TFWSR", "TTRW", "TFWSS", "TTSW", "TTPW", "AVPW",
           "ARSR", "SRDM")
qual <- c("CUAL", "LNGS", "PTLC", "DSTA", "LFRT", "LBTEF", "CBTR", "NMLB",
          "ANGB", "CUAL9M", "LVC9M", "TNPR9M", "PL9M", "STRP", "STRC",
          "PSTR")

ec[, qual] <- lapply(ec[, qual],
                     function(x) factor(as.factor(x)))


diversity.evaluate.core(data = ec, names = "genotypes",
                        qualitative = qual, selected = core)
```

---

freqdist.evaluate.core

Frequency Distribution Histogram

Description

Plot stacked frequency distribution histogram to graphically compare the probability distributions of traits between entire collection (EC) and core set (CS).

Usage

```
freqdist.evaluate.core(
  data,
  names,
  quantitative,
  qualitative,
  selected,
  highlight = NULL,
  include.highlight = TRUE,
  highlight.se = NULL,
  highlight.col = "red"
)
```

Arguments

data	The data as a data frame object. The data frame should possess one row per individual and columns with the individual names and multiple trait/character data.
names	Name of column with the individual names as a character string
quantitative	Name of columns with the quantitative traits as a character vector.
qualitative	Name of columns with the qualitative traits as a character vector.
selected	Character vector with the names of individuals selected in core collection and present in the names column.
highlight	Individual names to be highlighted as a character vector.
include.highlight	If TRUE, the highlighted individuals are included in the frequency distribution histogram. Default is TRUE.
highlight.se	Optional data frame of standard errors for the individuals specified in highlight. It should have the same column names as in data.
highlight.col	The colour(s) to be used to highlighting individuals in the plot as a character vector of the same length as highlight. Must be valid colour values in R (named colours, hexadecimal representation, index of colours [1:8] in default R palette() etc.).

Value

A list with the ggplot objects of stacked frequency distribution histograms plots for each trait specified as quantitative and qualitative.

See Also

hist, geom_histogram

Examples

```
data("cassava_CC")
data("cassava_EC")

ec <- cbind(genotypes = rownames(cassava_EC), cassava_EC)
ec$genotypes <- as.character(ec$genotypes)
rownames(ec) <- NULL
```

```
core <- rownames(cassava_CC)

quant <- c("NMSR", "TTRN", "TFWSR", "TTRW", "TFWSS", "TTSW", "TTPW", "AVPW",
           "ARSR", "SRDM")
qual <- c("CUAL", "LNGS", "PTLC", "DSTA", "LFRT", "LBTEF", "CBTR", "NMLB",
          "ANGB", "CUAL9M", "LVC9M", "TNPR9M", "PL9M", "STRP", "STRC",
          "PSTR")

ec[, qual] <- lapply(ec[, qual],
                     function(x) factor(as.factor(x)))


freqdist.evaluate.core(data = ec, names = "genotypes",
                       quantitative = quant, qualitative = qual,
                       selected = core)

checks <- c("TMe-1199", "TMe-1957", "TMe-3596", "TMe-3392")

freqdist.evaluate.core(data = ec, names = "genotypes",
                       quantitative = quant, qualitative = qual,
                       selected = core,
                       highlight = checks, highlight.col = "red")

quant.se <- data.frame(genotypes = checks,
                       NMSR = c(0.107, 0.099, 0.106, 0.062),
                       TTRN = c(0.081, 0.072, 0.057, 0.049),
                       TFWSR = c(0.089, 0.031, 0.092, 0.097),
                       TTRW = c(0.064, 0.031, 0.071, 0.071),
                       TFWSS = c(0.106, 0.071, 0.121, 0.066),
                       TTSW = c(0.084, 0.045, 0.066, 0.054),
                       TTPW = c(0.098, 0.052, 0.111, 0.082),
                       AVPW = c(0.074, 0.038, 0.054, 0.061),
                       ARSR = c(0.104, 0.019, 0.204, 0.044),
                       SRDM = c(0.078, 0.138, 0.076, 0.079))

freqdist.evaluate.core(data = ec, names = "genotypes",
                       quantitative = quant,
                       selected = core,
                       highlight = checks, highlight.col = "red",
                       highlight.se = quant.se)
```

`iqr.evaluate.core` *Interquartile Range*

Description

Compute the Interquartile Range (IQR) (Upton and Cook 1996) to compare quantitative traits of the entire collection (EC) and core set (CS).

Usage

```
iqr.evaluate.core(data, names, quantitative, selected)
```

Arguments

data	The data as a data frame object. The data frame should possess one row per individual and columns with the individual names and multiple trait/character data.
names	Name of column with the individual names as a character string
quantitative	Name of columns with the quantitative traits as a character vector.
selected	Character vector with the names of individuals selected in core collection and present in the names column.

Value

A data frame with the IQR values of the EC and CS for the traits specified as quantitative.

References

Upton G, Cook I (1996). "General summary statistics." In *Understanding statistics*. Oxford University Press.

See Also

IQR

Examples

```
data("cassava_CC")
data("cassava_EC")

ec <- cbind(genotypes = rownames(cassava_EC), cassava_EC)
ec$genotypes <- as.character(ec$genotypes)
rownames(ec) <- NULL

core <- rownames(cassava_CC)

quant <- c("NMSR", "TTRN", "TFWSR", "TTRW", "TFWSS", "TTSW", "TTPW", "AVPW",
           "ARSR", "SRDM")
qual <- c("CUAL", "LNGS", "PTLC", "DSTA", "LFRT", "LBTEF", "CBTR", "NMLB",
          "ANGB", "CUAL9M", "LVC9M", "TNPR9M", "PL9M", "STRP", "STRC",
          "PSTR")

ec[, qual] <- lapply(ec[, qual],
                     function(x) factor(as.factor(x)))

iqr.evaluate.core(data = ec, names = "genotypes",
                  quantitative = quant, selected = core)
```

`levene.evaluate.core` *Levene's Test*

Description

Test for of variances of the entire collection (EC) and core set (CS) for quantitative traits by Levene's test (Levene 1960).

Usage

```
levene.evaluate.core(data, names, quantitative, selected)
```

Arguments

`data`	The data as a data frame object. The data frame should possess one row per individual and columns with the individual names and multiple trait/character data.
`names`	Name of column with the individual names as a character string
`quantitative`	Name of columns with the quantitative traits as a character vector.
`selected`	Character vector with the names of individuals selected in core collection and present in the `names` column.

Value

A data frame with the following columns

`Trait`	The quantitative trait.
`EC_V`	The variance of the EC.
`CS_V`	The variance of the CS.
`EC_CV`	The coefficient of variance of the EC.
`CS_CV`	The coefficient of variance of the CS.
`Levene_Fvalue`	The test statistic.
`Levene_pvalue`	The p value for the test statistic.
`Levene_significance`	The significance of the test statistic (*: $p \leq 0.01$; **: $p \leq 0.05$; ns: $p > 0.05$).

References

Levene H (1960). "Robust tests for equality of variances." In Olkin I, Ghurye SG, Hoeffding W, Madow WG, Mann HB (eds.), *Contribution to Probability and Statistics: Essays in Honor of Harold Hotelling*, 278–292. Stanford University Press, Palo Alto, CA.

See Also

`leveneTest`

Examples

```
data("cassava_CC")
data("cassava_EC")

ec <- cbind(genotypes = rownames(cassava_EC), cassava_EC)
ec$genotypes <- as.character(ec$genotypes)
rownames(ec) <- NULL

core <- rownames(cassava_CC)

quant <- c("NMSR", "TTRN", "TFWSR", "TTRW", "TFWSS", "TTSW", "TTPW", "AVPW",
           "ARSR", "SRDM")
qual <- c("CUAL", "LNGS", "PTLC", "DSTA", "LFRT", "LBTEF", "CBTR", "NMLB",
          "ANGB", "CUAL9M", "LVC9M", "TNPR9M", "PL9M", "STRP", "STRC",
          "PSTR")

ec[, qual] <- lapply(ec[, qual],
                     function(x) factor(as.factor(x)))

levene.evaluate.core(data = ec, names = "genotypes",
                     quantitative = quant, selected = core)
```

---

pca.evaluate.core *Principal Component Analysis*

Description

Compute Principal Component Analysis Statistics (Mardia et al. 1979) to compare the probability distributions of quantitative traits between entire collection (EC) and core set (CS).

Usage

```
pca.evaluate.core(
  data,
  names,
  quantitative,
  selected,
  center = TRUE,
  scale = TRUE,
  npc.plot = 6
)
```

Arguments

data	The data as a data frame object. The data frame should possess one row per individual and columns with the individual names and multiple trait/character data.
names	Name of column with the individual names as a character string
quantitative	Name of columns with the quantitative traits as a character vector.
selected	Character vector with the names of individuals selected in core collection and present in the names column.

center	either a logical value or numeric-alike vector of length equal to the number of columns of x, where 'numeric-alike' means that as.numeric(.) will be applied successfully if is.numeric(.) is not true.
scale	either a logical value or a numeric-alike vector of length equal to the number of columns of x.
npc.plot	The number of principal components for which eigen values are to be plotted. The default value is 6.

Value

A list with the following components.

EC PC Importance	A data frame of importance of principal components for EC
EC PC Loadings	A data frame with eigen vectors of principal components for EC
CS PC Importance	A data frame of importance of principal components for CS
CS PC Loadings	A data frame with eigen vectors of principal components for CS
Scree Plot	The scree plot of principal components for EC and CS as a ggplot object.
PC Loadings Plot	A plot of the eigen vector values of principal components for EC and CS as specified by npc.plot as a ggplot2 object.

References

Mardia KV, Kent JT, Bibby JM (1979). *Multivariate analysis*. Academic Press, London; New York. ISBN 0-12-471250-9 978-0-12-471250-8 0-12-471252-5 978-0-12-471252-2.

See Also

prcomp

Examples

```
data("cassava_CC")
data("cassava_EC")

ec <- cbind(genotypes = rownames(cassava_EC), cassava_EC)
ec$genotypes <- as.character(ec$genotypes)
rownames(ec) <- NULL

core <- rownames(cassava_CC)

quant <- c("NMSR", "TTRN", "TFWSR", "TTRW", "TFWSS", "TTSW", "TTPW", "AVPW",
           "ARSR", "SRDM")
qual <- c("CUAL", "LNGS", "PTLC", "DSTA", "LFRT", "LBTEF", "CBTR", "NMLB",
          "ANGB", "CUAL9M", "LVC9M", "TNPR9M", "PL9M", "STRP", "STRC",
          "PSTR")

ec[, qual] <- lapply(ec[, qual],
                     function(x) factor(as.factor(x)))

pca.evaluate.core(data = ec, names = "genotypes",
                  quantitative = quant, selected = core,
```

```
center = TRUE, scale = TRUE, npc.plot = 4)
```

`pdfdist.evaluate.core` *Distance Between Probability Distributions*

Description

Compute Kullback-Leibler (Kullback and Leibler 1951), Kolmogorov-Smirnov (Kolmogorov 1933; Smirnov 1948) and Anderson-Darling distances (Anderson and Darling 1952) between the probability distributions of collection (EC) and core set (CS) for quantitative traits.

Usage

```
pdfdist.evaluate.core(data, names, quantitative, selected)
```

Arguments

`data`	The data as a data frame object. The data frame should possess one row per individual and columns with the individual names and multiple trait/character data.
`names`	Name of column with the individual names as a character string
`quantitative`	Name of columns with the quantitative traits as a character vector.
`selected`	Character vector with the names of individuals selected in core collection and present in the `names` column.

Value

A data frame with the following columns.

`Trait`	The quantitative trait.
`KL_Distance`	The Kullback-Leibler distance (Kullback and Leibler 1951) between EC and CS.
`KS_Distance`	The Kolmogorov-Smirnov distance (Kolmogorov 1933; Smirnov 1948) between EC and CS.
`KS_pvalue`	The p value of the Kolmogorov-Smirnov distance.
`AD_Distance`	Anderson-Darling distance (Anderson and Darling 1952) between EC and CS.
`AD_pvalue`	The p value of the Anderson-Darling distance.
`KS_significance`	The significance of the Kolmogorov-Smirnov distance (*: $p \leq 0.01$; **: $p \leq 0.05$; ns: $p > 0.05$).
`AD_pvalue`	The significance of the Anderson-Darling distance (*: $p \leq 0.01$; **: $p \leq 0.05$; ns: $p > 0.05$).

See Also

`KL.plugin`, `ks.test`, `ad.test`

Examples

```
data("cassava_CC")
data("cassava_EC")

ec <- cbind(genotypes = rownames(cassava_EC), cassava_EC)
ec$genotypes <- as.character(ec$genotypes)
rownames(ec) <- NULL

core <- rownames(cassava_CC)

quant <- c("NMSR", "TTRN", "TFWSR", "TTRW", "TFWSS", "TTSW", "TTPW", "AVPW",
           "ARSR", "SRDM")
qual <- c("CUAL", "LNGS", "PTLC", "DSTA", "LFRT", "LBTEF", "CBTR", "NMLB",
          "ANGB", "CUAL9M", "LVC9M", "TNPR9M", "PL9M", "STRP", "STRC",
          "PSTR")

ec[, qual] <- lapply(ec[, qual],
                     function(x) factor(as.factor(x)))

pdfdist.evaluate.core(data = ec, names = "genotypes",
                      quantitative = quant, selected = core)
```

`percentdiff.evaluate.core`

Percentage Difference of Means and Variances

Description

Compute the following differences between the entire collection (EC) and core set (CS).

- Percentage of significant differences of mean ($MD\%_{Hu}$) (Hu et al. 2000)
- Percentage of significant differences of variance ($VD\%_{Hu}$) (Hu et al. 2000)
- Average of absolute differences between means ($MD\%_{Kim}$) (Kim et al. 2007)
- Average of absolute differences between variances ($VD\%_{Kim}$) (Kim et al. 2007)
- Percentage difference between the mean squared Euclidean distance among accessions ($\bar{d}D\%$) (Studnicki et al. 2013)

Usage

```
percentdiff.evaluate.core(data, names, quantitative, selected, alpha = 0.05)
```

Arguments

`data`	The data as a data frame object. The data frame should possess one row per individual and columns with the individual names and multiple trait/character data.
`names`	Name of column with the individual names as a character string
`quantitative`	Name of columns with the quantitative traits as a character vector.
`selected`	Character vector with the names of individuals selected in core collection and present in the `names` column.
`alpha`	Type I error probability (Significance level) of difference.

Details

The differences are computed as follows.

$$MD\%_{Hu} = \left(\frac{S_t}{n}\right) \times 100$$

Where, S_t is the number of traits with a significant difference between the means of the EC and the CS and n is the total number of traits. A representative core should have $MD\%_{Hu}$ < 20 % and CR > 80 % (Hu et al. 2000).

$$VD\%_{Hu} = \left(\frac{S_F}{n}\right) \times 100$$

Where, S_F is the number of traits with a significant difference between the variances of the EC and the CS and n is the total number of traits. Larger $VD\%_{Hu}$ value indicates a more diverse core set.

$$MD\%_{Kim} = \left(\frac{1}{n}\sum_{i=1}^{n}\frac{|M_{EC_i} - M_{CS_i}|}{M_{CS_i}}\right) \times 100$$

Where, M_{EC_i} is the mean of the EC for the ith trait, M_{CS_i} is the mean of the CS for the ith trait and n is the total number of traits.

$$VD\%_{Kim} = \left(\frac{1}{n}\sum_{i=1}^{n}\frac{|V_{EC_i} - V_{CS_i}|}{V_{CS_i}}\right) \times 100$$

Where, V_{EC_i} is the variance of the EC for the ith trait, V_{CS_i} is the variance of the CS for the ith trait and n is the total number of traits.

$$\overline{d}D\% = \frac{\overline{d}_{CS} - \overline{d}_{EC}}{\overline{d}_{EC}} \times 100$$

Where, $\overline{d}_{CS}$ is the mean squared Euclidean distance among accessions in the CS and $\overline{d}_{EC}$ is the mean squared Euclidean distance among accessions in the EC.

Value

A data frame with the values of $MD\%_{Hu}$, $VD\%_{Hu}$, $MD\%_{Kim}$, $VD\%_{Kim}$ and $\overline{d}D\%$.

References

Hu J, Zhu J, Xu HM (2000). "Methods of constructing core collections by stepwise clustering with three sampling strategies based on the genotypic values of crops." *Theoretical and Applied Genetics*, **101**(1), 264–268.

Kim K, Chung H, Cho G, Ma K, Chandrabalan D, Gwag J, Kim T, Cho E, Park Y (2007). "PowerCore: A program applying the advanced M strategy with a heuristic search for establishing core sets." *Bioinformatics*, **23**(16), 2155–2162.

Studnicki M, Madry W, Schmidt J (2013). "Comparing the efficiency of sampling strategies to establish a representative in the phenotypic-based genetic diversity core collection of orchardgrass (*Dactylis glomerata* L.)." *Czech Journal of Genetics and Plant Breeding*, **49**(1), 36–47.

See Also

snk.evaluate.core, snk.evaluate.core

Examples

```
data("cassava_CC")
data("cassava_EC")

ec <- cbind(genotypes = rownames(cassava_EC), cassava_EC)
ec$genotypes <- as.character(ec$genotypes)
rownames(ec) <- NULL

core <- rownames(cassava_CC)

quant <- c("NMSR", "TTRN", "TFWSR", "TTRW", "TFWSS", "TTSW", "TTPW", "AVPW",
           "ARSR", "SRDM")
qual <- c("CUAL", "LNGS", "PTLC", "DSTA", "LFRT", "LBTEF", "CBTR", "NMLB",
          "ANGB", "CUAL9M", "LVC9M", "TNPR9M", "PL9M", "STRP", "STRC",
          "PSTR")

ec[, qual] <- lapply(ec[, qual],
                     function(x) factor(as.factor(x)))

percentdiff.evaluate.core(data = ec, names = "genotypes",
                          quantitative = quant, selected = core)
```

---

qq.evaluate.core *Quantile-Quantile Plots*

Description

Plot Quantile-Quantile (QQ) plots (Wilk and Gnanadesikan 1968) to graphically compare the probability distributions of quantitative traits between entire collection (EC) and core set (CS).

Usage

```
qq.evaluate.core(data, names, quantitative, selected)
```

Arguments

data	The data as a data frame object. The data frame should possess one row per individual and columns with the individual names and multiple trait/character data.
names	Name of column with the individual names as a character string
quantitative	Name of columns with the quantitative traits as a character vector.
selected	Character vector with the names of individuals selected in core collection and present in the names column.

Value

A list with the ggplot objects of QQ plots of CS vs EC for each trait specified as quantitative.

References

Wilk MB, Gnanadesikan R (1968). "Probability plotting methods for the analysis for the analysis of data." *Biometrika*, **55**(1), 1–17.

See Also

qqplot

Examples

```
data("cassava_CC")
data("cassava_EC")

ec <- cbind(genotypes = rownames(cassava_EC), cassava_EC)
ec$genotypes <- as.character(ec$genotypes)
rownames(ec) <- NULL

core <- rownames(cassava_CC)

quant <- c("NMSR", "TTRN", "TFWSR", "TTRW", "TFWSS", "TTSW", "TTPW", "AVPW",
           "ARSR", "SRDM")
qual <- c("CUAL", "LNGS", "PTLC", "DSTA", "LFRT", "LBTEF", "CBTR", "NMLB",
          "ANGB", "CUAL9M", "LVC9M", "TNPR9M", "PL9M", "STRP", "STRC",
          "PSTR")

ec[, qual] <- lapply(ec[, qual],
                     function(x) factor(as.factor(x)))

qq.evaluate.core(data = ec, names = "genotypes",
                 quantitative = quant, selected = core)
```

---

signtest.evaluate.core

Sign Test

Description

Test difference between means and variances of entire collection (EC) and core set (CS) for quantitative traits by Sign test (+ versus −) (Basigalup et al. 1995; Tai and Miller 2001).

Usage

```
signtest.evaluate.core(data, names, quantitative, selected)
```

Arguments

data	The data as a data frame object. The data frame should possess one row per individual and columns with the individual names and multiple trait/character data.
names	Name of column with the individual names as a character string

quantitative	Name of columns with the quantitative traits as a character vector.
selected	Character vector with the names of individuals selected in core collection and present in the names column.

Details

The test statistic for Sign test (χ^2) is computed as follows.

$$\chi^2 = \frac{(N_1 - N_2)^2}{N_1 + N_2}$$

Where, where N_1 is the number of variables for which the mean or variance of the CS is greater than the mean or variance of the EC (number of $+$ signs); N_2 is the number of variables for which the mean or variance of the CS is less than the mean or variance of the EC (number of $-$ signs). The value of χ^2 is compared with a Chi-square distribution with 1 degree of freedom.

Value

A data frame with the following components.

Comparison	The comparison measure.
ChiSq	The test statistic (χ^2).
p.value	The p value for the test statistic.
significance	The significance of the test statistic (*: $p \leq 0.01$; **: $p \leq 0.05$; ns: $p > 0.05$).

References

Basigalup DH, Barnes DK, Stucker RE (1995). "Development of a core collection for perennial *Medicago* plant introductions." *Crop Science*, **35**(4), 1163–1168.

Tai PYP, Miller JD (2001). "A Core Collection for *Saccharum spontaneum* L. from the World Collection of Sugarcane." *Crop Science*, **41**(3), 879–885.

Examples

```
data("cassava_CC")
data("cassava_EC")

ec <- cbind(genotypes = rownames(cassava_EC), cassava_EC)
ec$genotypes <- as.character(ec$genotypes)
rownames(ec) <- NULL

core <- rownames(cassava_CC)

quant <- c("NMSR", "TTRN", "TFWSR", "TTRW", "TFWSS", "TTSW", "TTPW", "AVPW",
           "ARSR", "SRDM")
qual <- c("CUAL", "LNGS", "PTLC", "DSTA", "LFRT", "LBTEF", "CBTR", "NMLB",
          "ANGB", "CUAL9M", "LVC9M", "TNPR9M", "PL9M", "STRP", "STRC",
          "PSTR")

ec[, qual] <- lapply(ec[, qual],
                     function(x) factor(as.factor(x)))

signtest.evaluate.core(data = ec, names = "genotypes",
```

```
quantitative = quant, selected = core)
```

`snk.evaluate.core` *Student-Newman-Keuls Test*

Description

Test difference between means of entire collection (EC) and core set (CS) for quantitative traits by Newman-Keuls or Student-Newman-Keuls test (Newman 1939; Keuls 1952).

Usage

```
snk.evaluate.core(data, names, quantitative, selected)
```

Arguments

`data`	The data as a data frame object. The data frame should possess one row per individual and columns with the individual names and multiple trait/character data.
`names`	Name of column with the individual names as a character string
`quantitative`	Name of columns with the quantitative traits as a character vector.
`selected`	Character vector with the names of individuals selected in core collection and present in the `names` column.

Value

A data frame with the following components.

`Trait`	The quantitative trait.
`EC_Min`	The minimum value of the trait in EC.
`EC_Max`	The maximum value of the trait in EC.
`EC_Mean`	The mean value of the trait in EC.
`EC_SE`	The standard error of the trait in EC.
`CS_Min`	The minimum value of the trait in CS.
`CS_Max`	The maximum value of the trait in CS.
`CS_Mean`	The mean value of the trait in CS.
`CS_SE`	The standard error of the trait in CS.
`SNK_pvalue`	The p value of the Student-Newman-Keuls test for equality of means of EC and CS.
`SNK_significance`	The significance of the Student-Newman-Keuls test for equality of means of EC and CS.

References

Keuls M (1952). "The use of the „studentized range" in connection with an analysis of variance." *Euphytica*, **1**(2), 112–122.

Newman D (1939). "The distribution of range in samples from a normal population, expressed in terms of an independent estimate of standard deviation." *Biometrika*, **31**(1-2), 20–30.

See Also

SNK.test

Examples

```
data("cassava_CC")
data("cassava_EC")

ec <- cbind(genotypes = rownames(cassava_EC), cassava_EC)
ec$genotypes <- as.character(ec$genotypes)
rownames(ec) <- NULL

core <- rownames(cassava_CC)

quant <- c("NMSR", "TTRN", "TFWSR", "TTRW", "TFWSS", "TTSW", "TTPW", "AVPW",
           "ARSR", "SRDM")
qual <- c("CUAL", "LNGS", "PTLC", "DSTA", "LFRT", "LBTEF", "CBTR", "NMLB",
          "ANGB", "CUAL9M", "LVC9M", "TNPR9M", "PL9M", "STRP", "STRC",
          "PSTR")

ec[, qual] <- lapply(ec[, qual],
                     function(x) factor(as.factor(x)))

snk.evaluate.core(data = ec, names = "genotypes",
                  quantitative = quant, selected = core)
```

---

ttest.evaluate.core *Student's t Test*

Description

Test difference between means of entire collection (EC) and core set (CS) for quantitative traits by Student's t test (Student 1908).

Usage

```
ttest.evaluate.core(data, names, quantitative, selected)
```

Arguments

data	The data as a data frame object. The data frame should possess one row per individual and columns with the individual names and multiple trait/character data.
names	Name of column with the individual names as a character string
quantitative	Name of columns with the quantitative traits as a character vector.
selected	Character vector with the names of individuals selected in core collection and present in the names column.

Value

`Trait`	The quantitative trait.
`EC_Min`	The minimum value of the trait in EC.
`EC_Max`	The maximum value of the trait in EC.
`EC_Mean`	The mean value of the trait in EC.
`EC_SE`	The standard error of the trait in EC.
`CS_Min`	The minimum value of the trait in CS.
`CS_Max`	The maximum value of the trait in CS.
`CS_Mean`	The mean value of the trait in CS.
`CS_SE`	The standard error of the trait in CS.
`ttest_pvalue`	The p value of the Student's t test for equality of means of EC and CS.
`ttest_significance`	The significance of the Student's t test for equality of means of EC and CS.

References

Student (1908). "The probable error of a mean." *Biometrika*, **6**(1), 1–25.

See Also

`t.test`

Examples

```
data("cassava_CC")
data("cassava_EC")

ec <- cbind(genotypes = rownames(cassava_EC), cassava_EC)
ec$genotypes <- as.character(ec$genotypes)
rownames(ec) <- NULL

core <- rownames(cassava_CC)

quant <- c("NMSR", "TTRN", "TFWSR", "TTRW", "TFWSS", "TTSW", "TTPW", "AVPW",
           "ARSR", "SRDM")
qual <- c("CUAL", "LNGS", "PTLC", "DSTA", "LFRT", "LBTEF", "CBTR", "NMLB",
          "ANGB", "CUAL9M", "LVC9M", "TNPR9M", "PL9M", "STRP", "STRC",
          "PSTR")

ec[, qual] <- lapply(ec[, qual],
                     function(x) factor(as.factor(x)))

ttest.evaluate.core(data = ec, names = "genotypes",
                    quantitative = quant, selected = core)
```

`vr.evaluate.core` *Variable Rate of Coefficient of Variation*

Description

Compute the Variable Rate of Coefficient of Variation (VR) (Hu et al. 2000) to compare quantitative traits of the entire collection (EC) and core set (CS).

Usage

```
vr.evaluate.core(data, names, quantitative, selected)
```

Arguments

`data`	The data as a data frame object. The data frame should possess one row per individual and columns with the individual names and multiple trait/character data.
`names`	Name of column with the individual names as a character string
`quantitative`	Name of columns with the quantitative traits as a character vector.
`selected`	Character vector with the names of individuals selected in core collection and present in the `names` column.

Details

The Variable Rate of Coefficient of Variation (VR) is computed as follows.

$$VR = \left(\frac{1}{n} \sum_{i=1}^{n} \frac{CV_{CS_i}}{CV_{EC_i}} \right) \times 100$$

Where, CV_{CS_i} is the coefficients of variation for the ith trait in the CS, CV_{EC_i} is the coefficients of variation for the ith trait in the EC and n is the total number of traits

Value

The VR value.

References

Hu J, Zhu J, Xu HM (2000). "Methods of constructing core collections by stepwise clustering with three sampling strategies based on the genotypic values of crops." *Theoretical and Applied Genetics*, **101**(1), 264–268.

Examples

```
data("cassava_CC")
data("cassava_EC")

ec <- cbind(genotypes = rownames(cassava_EC), cassava_EC)
ec$genotypes <- as.character(ec$genotypes)
rownames(ec) <- NULL
```

```
core <- rownames(cassava_CC)

quant <- c("NMSR", "TTRN", "TFWSR", "TTRW", "TFWSS", "TTSW", "TTPW", "AVPW",
           "ARSR", "SRDM")
qual <- c("CUAL", "LNGS", "PTLC", "DSTA", "LFRT", "LBTEF", "CBTR", "NMLB",
          "ANGB", "CUAL9M", "LVC9M", "TNPR9M", "PL9M", "STRP", "STRC",
          "PSTR")

ec[, qual] <- lapply(ec[, qual],
                     function(x) factor(as.factor(x)))

vr.evaluate.core(data = ec, names = "genotypes",
                 quantitative = quant, selected = core)
```

`wilcox.evaluate.core` *Wilcoxon Rank Sum Test*

Description

Compare the medians of quantitative traits between entire collection (EC) and core set (CS) by Wilcoxon rank sum test or Mann-Whitney-Wilcoxon test or Mann-Whitney U test (Wilcoxon 1945; Mann and Whitney 1947).

Usage

```
wilcox.evaluate.core(data, names, quantitative, selected)
```

Arguments

`data`	The data as a data frame object. The data frame should possess one row per individual and columns with the individual names and multiple trait/character data.
`names`	Name of column with the individual names as a character string
`quantitative`	Name of columns with the quantitative traits as a character vector.
`selected`	Character vector with the names of individuals selected in core collection and present in the `names` column.

Value

`Trait`	The quantitative trait.
`EC_Med`	The median value of the trait in EC.
`CS_Med`	The median value of the trait in CS.
`Wilcox_pvalue`	The p value of the Wilcoxon test for equality of medians of EC and CS.
`Wilcox_significance`	The significance of the Wilcoxon test for equality of medians of EC and CS.

References

Mann HB, Whitney DR (1947). "On a test of whether one of two random variables is stochastically larger than the other." *The Annals of Mathematical Statistics*, **18**(1), 50–60.

Wilcoxon F (1945). "Individual comparisons by ranking methods." *Biometrics Bulletin*, **1**(6), 80.

See Also

wilcox.test

Examples

```
data("cassava_CC")
data("cassava_EC")

ec <- cbind(genotypes = rownames(cassava_EC), cassava_EC)
ec$genotypes <- as.character(ec$genotypes)
rownames(ec) <- NULL

core <- rownames(cassava_CC)

quant <- c("NMSR", "TTRN", "TFWSR", "TTRW", "TFWSS", "TTSW", "TTPW", "AVPW",
           "ARSR", "SRDM")
qual <- c("CUAL", "LNGS", "PTLC", "DSTA", "LFRT", "LBTEF", "CBTR", "NMLB",
          "ANGB", "CUAL9M", "LVC9M", "TNPR9M", "PL9M", "STRP", "STRC",
          "PSTR")

ec[, qual] <- lapply(ec[, qual],
                     function(x) factor(as.factor(x)))

wilcox.evaluate.core(data = ec, names = "genotypes",
                     quantitative = quant, selected = core)
```

Index