Package 'BKQualit'

August 8, 2026

Type Package

Title Analysis of Qualitative Traits, Segregation and Genetic Linkage

Version 0.1.1

Description A colour-first toolkit for the analysis of qualitative (categorical) traits in plant breeding and genetics. It tests observed segregation against the classical Mendelian expectations, fitting every standard mono-, di- and trihybrid ratio automatically and ranking them by goodness of fit, with Yates continuity correction and Monte Carlo exact tests for sparse tables. It partitions chi-square across families into pooled and heterogeneity components, so that a poor overall fit can be attributed either to the hypothesised ratio or to variation between families. Genetic linkage is estimated by maximum likelihood from two-locus second filial generation and backcross data, with logarithm of odds scores and likelihood-ratio confidence intervals for the recombination fraction. The package further computes Shannon-Weaver and Simpson diversity for descriptor states used in distinctness, uniformity and stability testing, and performs multiple correspondence analysis and Gower-distance clustering of mixed categorical and quantitative descriptors. Every analysis returns a tidy result object and a publication-ready 'ggplot2' figure. Methods follow Mather (1951, <ISBN:9780416470406>), Allard (1956) <doi:10.3733/hilg.v24n10p235> and Gower (1971) <doi:10.2307/2528823>.

License GPL-3

Encoding UTF-8

Depends R (>= 4.0.0)

Imports stats, utils, grDevices, ggplot2

Suggests ggrepel, patchwork, knitr, rmarkdown, testthat (>= 3.0.0)

RoxygenNote 7.3.1

Config/testthat/edition 3

URL https://github.com/bkpraveenars-del/BKQualit

BugReports https://github.com/bkpraveenars-del/BKQualit/issues

NeedsCompilation no

Author Praveen Kumar B. K. [aut, cre]

Maintainer Praveen Kumar B. K. <bkpraveenars@gmail.com>

Repository CRAN

Date/Publication 2026-08-08 13:20:02 UTC

Contents

`BKQualit-package` *BKQualit: Analysis of Qualitative Traits in Plant Breeding*

Description

Tools for the genetic analysis of qualitative (discretely inherited) characters: testing observed segregation against Mendelian expectations, partitioning chi-square across families into pooled and heterogeneity components, estimating linkage by maximum likelihood, and summarising and clustering qualitative descriptors of the kind recorded on a distinctness, uniformity and stability (DUS) form.

Analysis functions

`bq_segregate` Fit observed counts to Mendelian ratios; ranks a library of classical ratios and names the gene action implied.

`bq_heterogeneity` Partition the total chi-square over families into pooled and heterogeneity components.

`bq_linkage` Estimate the recombination fraction by maximum likelihood with a logarithm of odds (LOD) score and a likelihood-ratio confidence interval.

`bq_diversity` Shannon, Simpson and evenness indices for qualitative descriptors, with discriminating power.

`bq_mca` Multiple correspondence analysis with the Benzecri adjustment, plus Gower-distance clustering of accessions.

Support

bq_data loads the bundled example datasets, bq_plot draws every result, bq_save writes all figures for a result to disk, and bq_palette, theme_bq, scale_colour_bq and scale_fill_bq provide the shared visual style.

Design

Every analysis function returns a classed list with a print method that states the conclusion in words as well as numbers, and a bq_plot method. Nothing is written to the user's file space unless bq_save is called explicitly.

Author(s)

Praveen Kumar B. K. <bkpraveenars@gmail.com>

References

Mather K (1951). *The Measurement of Linkage in Heredity*, 2nd edition. Methuen, London.

Allard RW (1956). Formulas and tables to facilitate the calculation of recombination values in heredity. *Hilgardia* **24**, 235–278. doi:10.3733/hilg.v24n10p235

Gower JC (1971). A general coefficient of similarity and some of its properties. *Biometrics* **27**, 857–871. doi:10.2307/2528823

---

bq_data *Load a bundled example dataset*

Description

Reads one of the demonstration datasets shipped with the package. The datasets are simulated from known parameters and exist for documentation, teaching and regression testing.

Usage

```
bq_data(name = c("f2", "families", "linkage", "dus"))
```

Arguments

name	One of "f2" (single-cross segregation counts), "families" (segregation counts for many families), "linkage" (two-locus counts) or "dus" (qualitative descriptors scored on a set of accessions).

Value

A data frame.

Examples

```
str(bq_data("f2"))
head(bq_data("dus"))
```

`bq_diversity` *Diversity of qualitative descriptors*

Description

Summarises a set of categorical descriptors, of the kind recorded on a distinctness, uniformity and stability (DUS) form, into per-descriptor diversity indices, and identifies which descriptors actually discriminate among accessions.

Usage

```
bq_diversity(data, descriptors, accession = NULL, group = NULL)
```

Arguments

`data`	A data frame with one row per accession.
`descriptors`	Character vector naming the categorical columns.
`accession`	Character or `NULL`; the accession identifier column.
`group`	Character or `NULL`; an optional grouping column, giving diversity within each group as well as overall.

Details

Three complementary indices are reported for each descriptor. The Shannon-Weaver index $H' = -\sum p_i \log p_i$ is sensitive to rare states; the Simpson index $1 - \sum p_i^2$ is dominated by common states; and evenness $J' = H' / \log s$ rescales Shannon by the number of observed states s so that descriptors with different numbers of states can be compared directly. A descriptor with a single observed state is monomorphic, contributes nothing to discrimination, and is flagged.

A descriptor is useful only if it separates accessions. The discriminating power reported here is the probability that two randomly chosen accessions differ in that descriptor, which equals the Simpson index; descriptors are ranked on it so that the least informative can be dropped from a characterisation form.

Value

An object of class `bq_diversity` containing per-descriptor diversity indices, the frequency of every state, and, if `group` is supplied, the same indices within each group.

See Also

`bq_mca`

Examples

```
d <- bq_data("dus")
res <- bq_diversity(d,
  descriptors = c("growth_habit", "leaf_pubescence", "flower_colour",
                  "seed_colour", "seed_shape"),
  accession = "accession")
res
bq_plot(res)
bq_plot(res, type = "frequencies")
```

---

bq_heterogeneity *Pooled and heterogeneity chi-square across families*

Description

Partitions the total chi-square computed over several segregating families into a pooled component and a heterogeneity component. This is the standard way to decide why a segregation hypothesis fails.

Usage

```
bq_heterogeneity(data, family, classes, trait = NULL, ratio = c(3, 1))
```

Arguments

`data`	A data frame with one row per family.
`family`	Character; the family identifier column.
`classes`	Character vector; the columns holding counts of each phenotype class.
`trait`	Character or `NULL`; an optional column identifying several traits, each partitioned separately.
`ratio`	Numeric vector giving the expected ratio, default `c(3, 1)`.

Details

Writing χ^2_i for the goodness-of-fit statistic of family i, the total is $\sum_i \chi^2_i$ on $k(c-1)$ degrees of freedom, the pooled statistic is computed from the summed counts on $c-1$ degrees of freedom, and the heterogeneity statistic is the difference on $(k-1)(c-1)$ degrees of freedom, where k is the number of families and c the number of phenotype classes. The partition is exact.

The two components answer different questions. A significant pooled chi-square with a non-significant heterogeneity means the families behave alike but the hypothesised ratio is wrong. A significant heterogeneity means the families do not behave alike, so pooling them is not justified and the pooled test should not be interpreted.

Value

An object of class `bq_heterogeneity` containing the per-family goodness of fit and the pooled, heterogeneity and total chi-square components.

See Also

bq_segregate

Examples

```
d <- bq_data("families")
res <- bq_heterogeneity(d, family = "family",
                        classes = c("dominant", "recessive"),
                        trait = "trait", ratio = c(3, 1))
res
bq_plot(res)
```

---

bq_linkage *Estimate genetic linkage by maximum likelihood*

Description

Estimates the recombination fraction between two loci from two-locus second filial generation (F2) or backcross counts, and tests it against independent assortment.

Usage

```
bq_linkage(data, pair, classes, design = NULL, phase = NULL, conf = 0.95)
```

Arguments

data	A data frame with one row per locus pair.
pair	Character; a column naming the locus pair.
classes	Character vector of length four giving the count columns, in the order A_B_, A_bb, aaB_, aabb.
design	Character or NULL; a column holding "F2" or "backcross" for each pair. If NULL, "F2" is assumed.
phase	Character or NULL; a column holding "coupling" or "repulsion". If NULL, coupling is assumed.
conf	Confidence level (default 0.95).

Details

For a backcross the recombination fraction is the proportion of recombinant progeny, a binomial parameter. For an F2 in coupling phase the four phenotype classes occur in the proportions $(2+\theta)/4$, $(1-\theta)/4$, $(1-\theta)/4$ and $\theta/4$ with $\theta = (1-r)^2$; in repulsion phase the outer and inner proportions are exchanged. The recombination fraction is obtained by maximising the multinomial likelihood numerically rather than from the older product-ratio approximation.

The LOD score compares the fitted model with independent assortment, and a value above 3 is the conventional threshold for declaring linkage. The confidence interval is obtained by likelihood-ratio inversion, retaining every value of the recombination fraction whose log-likelihood lies within 1.921

of the maximum; this is preferred to a symmetric standard-error interval because the likelihood is skewed near the boundary. Map distance is reported as 100 times the recombination fraction and is not corrected for double crossovers.

Value

An object of class `bq_linkage` containing the estimated recombination fraction, its confidence interval, the logarithm of odds (LOD) score, the map distance and a test against independent assortment, together with the likelihood profile for plotting.

References

Allard RW (1956). Formulas and tables to facilitate the calculation of recombination values in heredity. *Hilgardia* **24**, 235–278. doi:10.3733/hilg.v24n10p235

Examples

```
d <- bq_data("linkage")
res <- bq_linkage(d, pair = "pair",
                  classes = c("AB", "Ab", "aB", "ab"),
                  design = "design", phase = "phase")
res
bq_plot(res)
bq_plot(res, type = "estimates")
```

---

bq_mca *Multiple correspondence analysis and Gower clustering of accessions*

Description

Places accessions and descriptor states in a common low-dimensional space by multiple correspondence analysis (MCA), and groups the accessions by hierarchical clustering of the Gower distance.

Usage

```
bq_mca(data, descriptors, accession = NULL, numeric_vars = NULL,
       k = 3, ndim = 2)
```

Arguments

`data`	A data frame with one row per accession.
`descriptors`	Character vector naming the categorical columns.
`accession`	Character or `NULL`; the accession identifier column.
`numeric_vars`	Character vector or `NULL`; numeric columns to add to the Gower distance. They do not enter the correspondence analysis.
`k`	Number of clusters to cut the dendrogram into (default 3).
`ndim`	Number of dimensions to retain (default 2).

Details

MCA is computed from the singular value decomposition of the standardised indicator matrix. Writing Z for the $n \times J$ indicator of Q descriptors, $P = Z/(nQ)$, and r, c for its row and column margins, the decomposition is applied to $D_r^{-1/2}(P - rc^\top)D_c^{-1/2}$. Row and column principal coordinates then satisfy the transition relation $F = (Z/Q)G_{std}$, so an accession sits at the centroid of the descriptor states it carries, which is what makes an MCA biplot directly readable.

The raw principal inertias of an indicator MCA understate the structure badly, because coding a descriptor into several columns injects inertia that carries no information. The Benzecri adjustment is therefore also reported: axes with inertia above $1/Q$ are rescaled to $((Q/(Q-1))(\lambda - 1/Q))^2$. Interpret the adjusted percentages, not the raw ones.

The Gower distance handles nominal and numeric descriptors together: nominal descriptors contribute 0 when two accessions match and 1 when they differ, numeric descriptors contribute the absolute difference divided by the range, and the contributions are averaged. Clustering uses Ward's minimum-variance criterion on that distance.

Value

An object of class `bq_mca` containing the principal inertias with the Benzecri adjustment, row and column coordinates, the Gower distance matrix, the clustering, and a modal profile of each cluster.

References

Gower JC (1971). A general coefficient of similarity and some of its properties. *Biometrics* **27**, 857–871. doi:10.2307/2528823

See Also

`bq_diversity`

Examples

```
d <- bq_data("dus")
res <- bq_mca(d,
  descriptors = c("growth_habit", "leaf_pubescence", "flower_colour",
                  "seed_colour", "seed_shape", "pod_curvature"),
  accession = "accession", numeric_vars = "plant_height_cm", k = 3)
res
bq_plot(res)
bq_plot(res, type = "dendrogram")
bq_plot(res, type = "scree")
```

bq_palette *Colour palettes for qualitative-trait figures*

Description

Returns the colours used throughout the package. Categorical palettes are returned unchanged when enough colours are available and interpolated otherwise; sequential palettes are always interpolated.

Usage

```
bq_palette(name = c("phenotype", "fit", "linkage", "descriptor", "contrast"),
           n = NULL, reverse = FALSE)
```

Arguments

name	Palette name. "phenotype" is a vivid categorical set for phenotype classes; "fit" runs red to green for goodness of fit; "linkage" is a sequential ramp for recombination fraction; "descriptor" is a soft categorical set for descriptor states; "contrast" is a high-separation set for clusters.
n	Number of colours. If NULL, the whole palette is returned.
reverse	Logical; reverse the colour order.

Value

A character vector of hexadecimal colours.

Examples

```
bq_palette("phenotype")
bq_palette("linkage", 6)
```

bq_plot *Plot a BKQualit result*

Description

A single plotting verb for every analysis in the package. The method chosen depends on the class of x.

Usage

```
bq_plot(x, ...)
```

Arguments

x	An object returned by one of the bq_* analysis functions.
...	Arguments passed to the method, typically type.

Details

Several results accept a type argument selecting among alternative views: bq_linkage offers "profile" and "estimates"; bq_diversity offers "indices" and "frequencies"; bq_mca offers "biplot", "dendrogram" and "scree".

Value

A ggplot object.

See Also

bq_save

Examples

```
res <- bq_segregate(bq_data("f2"), phenotype = "phenotype",
                    count = "count", trait = "trait")
bq_plot(res)
```

---

bq_save *Save every figure from a BKQualit result*

Description

Convenience wrapper that writes all available views of a result to disk as PNG files. Nothing is written unless this function is called.

Usage

```
bq_save(x, path = tempdir(), prefix = NULL, width = 8, height = 5.5,
        dpi = 300)
```

Arguments

x	An object returned by one of the bq_* analysis functions.
path	Directory to write into. Defaults to a temporary directory.
prefix	File-name prefix. Defaults to the class of x.
width, height, dpi	Passed to ggplot2::ggsave.

Value

Character vector of the files written, invisibly.

Examples

```
res <- bq_segregate(bq_data("f2"), phenotype = "phenotype",
                    count = "count", trait = "trait")
bq_save(res, path = tempdir())
```

`bq_segregate` *Test observed segregation against Mendelian ratios*

Description

Fits observed phenotype counts to the classical Mendelian expectations. When no ratio is stated, a library of standard mono-, di- and trihybrid ratios with the matching number of classes is fitted and ranked, and the gene action implied by the best fit is named.

Usage

```
bq_segregate(data = NULL, phenotype = NULL, count = NULL,
             trait = NULL, ratio = NULL, counts = NULL,
             simulate = TRUE, B = 10000, seed = 1L)
```

Arguments

`data`	A data frame in long form, one row per phenotype class.
`phenotype`	Character; the phenotype-class column.
`count`	Character; the observed-count column.
`trait`	Character or `NULL`; a column identifying several traits, each tested separately.
`ratio`	Numeric vector giving a specific expected ratio. If `NULL`, every ratio in the library with the right number of classes is fitted and ranked.
`counts`	Named numeric vector, an alternative to supplying `data`.
`simulate`	Logical; also run a Monte Carlo exact test (default `TRUE`).
`B`	Monte Carlo replicates (default 10000).
`seed`	Random seed for the Monte Carlo test.

Details

The Pearson chi-square is reported with the Yates continuity correction applied only when there is a single degree of freedom, which is where the correction is defined; applying it more widely is a common error that makes tests conservative. A Monte Carlo exact test is also run, because the chi-square approximation is unreliable when an expected count falls below five, as it often does in the rarest class of a 63:1 or 9:3:3:1 table.

A high p value does not prove the hypothesis. It means the data are compatible with it. Where two ratios fit almost equally well, the printed output says so rather than declaring a winner.

Value

An object of class `bq_segregate` holding the fitted ratios ranked by goodness of fit, the observed and expected counts, and the gene action implied by the best-fitting ratio.

References

Mather K (1951). *The Measurement of Linkage in Heredity*, 2nd edition. Methuen, London.

See Also

`bq_heterogeneity`, `bq_plot`

Examples

```
d <- bq_data("f2")
res <- bq_segregate(d, phenotype = "phenotype", count = "count",
                    trait = "trait")
res
bq_plot(res)

## a single trait against a stated hypothesis
bq_segregate(counts = c(purple = 312, white = 104), ratio = c(3, 1))
```

---

scale_colour_bq *Discrete colour and fill scales for BKQualit palettes*

Description

Discrete scales built on the package palettes. `scale_color_bq` is an alias of `scale_colour_bq`.

Usage

```
scale_colour_bq(pal_name = "phenotype", reverse = FALSE, ...)

scale_fill_bq(pal_name = "phenotype", reverse = FALSE, ...)
```

Arguments

`pal_name`	Palette name; see `bq_palette`. This argument is called `pal_name` rather than `name` so that the `ggplot2` legend title, which is also called `name`, can still be passed through `....`
`reverse`	Logical; reverse the colour order.
`...`	Passed to `ggplot2::discrete_scale`, for example `name` to set the legend title.

Value

A `ggplot2` scale object.

See Also

bq_palette, theme_bq

Examples

```
library(ggplot2)
ggplot(bq_data("dus"), aes(growth_habit, fill = seed_shape)) +
  geom_bar() +
  scale_fill_bq("descriptor", name = "Seed shape") +
  theme_bq()
```

---

theme_bq *The BKQualit ggplot2 theme*

Description

A clean theme with a soft background, muted grid and a coloured title, applied by every plotting method in the package. It can be added to any ggplot of your own.

Usage

```
theme_bq(base_size = 12, base_family = "", grid = TRUE)
```

Arguments

base_size	Base font size in points.
base_family	Base font family.
grid	Logical; draw the background grid.

Value

A ggplot2 theme object.

Examples

```
library(ggplot2)
ggplot(bq_data("dus"), aes(growth_habit)) +
  geom_bar(fill = bq_palette("phenotype", 1)) +
  theme_bq()
```

Index