Package: MADPop (via r-universe)

October 29, 2026

Type Package

Title MHC Allele-Based Differencing Between Populations

Version 1.1.7

Date 2024-02-26

Description Tools for the analysis of population differences using the Major Histocompatibility Complex (MHC) genotypes of samples having a variable number of alleles (1-4) recorded for each individual. A hierarchical Dirichlet-Multinomial model on the genotype counts is used to pool small samples from multiple populations for pairwise tests of equality. Bayesian inference is implemented via the 'rstan' package. Bootstrapped and posterior p-values are provided for chi-squared and likelihood ratio tests of equal genotype probabilities.

URL https://github.com/mlysy/MADPop

BugReports https://github.com/mlysy/MADPop/issues

License GPL-3

Depends R (>= 3.4.0), rstan (>= 2.26.0)

Imports methods, Rcpp (>= 0.12.0), stats, rstantools

Suggests knitr, rmarkdown, testthat

LinkingTo BH (>= 1.66.0), Rcpp (>= 0.12.0), RcppEigen (>= 0.3.3.3.0), rstan (>= 2.26.0), StanHeaders (>= 2.26.0), RcppParallel (>= 5.0.1)

VignetteBuilder knitr

SystemRequirements GNU make

Encoding UTF-8

LazyData true

RoxygenNote 7.2.3

NeedsCompilation yes

Author Martin Lysy [cre, aut], Peter W.J. Kim [aut], Terin Robinson [ctb]

Maintainer Martin Lysy <mlysy@uwaterloo.ca>

Config/pak/sysreqs make

Repository https://cran.r-universe.dev

Date/Publication 2024-02-27 02:29:17 UTC

RemoteUrl https://github.com/cran/MADPop

RemoteRef HEAD

RemoteSha f39285a5c57bca1379778c107683f8a33af4b97d

Contents

MADPop-package 2
chi2.stat 3
fish215 4
hUM.post 5
LRT.stat 6
UM.eqtest 7
UM.suff 8

Index **10**

`MADPop-package` *(M)HC (A)llele-Based (D)ifferencing between (Pop)ulations*

Description

Tools for the analysis of population differences using the Major Histocompatibility Complex (MHC) genotypes of samples having a variable number of alleles (1-4) recorded for each individual.

Details

For a full tutorial see package vignette: `vignette("MADPop-quicktut")`.

Author(s)

Maintainer: Martin Lysy <mlysy@uwaterloo.ca>

Authors:

- Peter W.J. Kim

Other contributors:

- Terin Robinson [contributor]

See Also

Useful links:

- https://github.com/mlysy/MADPop
- Report bugs at https://github.com/mlysy/MADPop/issues

Examples

```
# typical dataset
head(fish215[sample(nrow(fish215)),])
table(fish215$Lake) # number of samples per lake

# contingency table on two lakes
iLakes <- c("Michipicoten", "Simcoe")
Xsuff <- UM.suff(X = fish215[fish215$Lake %in% iLakes,])
ctab <- Xsuff$tab
ctab

# bootstrapped p-value calculation for chi^2 and LR tests
p.MLE <- colSums(ctab)/sum(ctab)
N1 <- sum(ctab[1,])
N2 <- sum(ctab[2,])
# bootstrapped test statistics (chi^2 and LRT)
T.boot <- UM.eqtest(N1 = N1, N2 = N2, p0 = p.MLE, nreps = 1e3)

# observed test statistics
T.obs <- c(chi2 = chi2.stat(ctab), LRT = LRT.stat(ctab))
# p-values
rowMeans(t(T.boot) > T.obs)

# posterior sampler for hierarchical model

# output posterior probability for each genotype in lake Simcoe
rhoId <- "Simcoe"
nsamples <- 500
hUM.fit <- hUM.post(nsamples = nsamples, X = fish215,
                    rhoId = rhoId, chains = 1)

# first 20 genotype posterior probabilities in lake Simcoe
rho.post <- hUM.fit$rho[,1,]
boxplot(rho.post[,1:20], las = 2,
        xlab = "Genotype", ylab = "Posterior Probability",
        pch = ".", col = "grey")
```

`chi2.stat` *Chi-squared test statistic for contingency tables*

Description

Calculates the chi-squared test statistic for a two-way contingency table.

Usage

```
chi2.stat(tab)
```

Arguments

tab	A K x C matrix (contingency table) of counts. See details.

Details

Suppose that tab consists of counts from K populations (rows) in C categories. The chi-squared test statistic is computed as

$$\sum_{i=1}^{K}\sum_{j=1}^{C}(E_{ij} - O_{ij})^2/E_{ij},$$

where O_{ij} is the observed number of counts in the ith row and jth column of tab, and E_{ij} is the expected number of counts under H_0 that the populations have indentical proportions in each category:

$$E_{ij} = \frac{1}{N}\sum_{i=1}^{K} O_{ij} \times \sum_{j=1}^{C} O_{ij}.$$

where N is the total number of counts in tab.

Value

The calculated value of the chi-squared statistic.

Examples

```
# simple contingency table
ctab <- rbind(pop1 = c(5, 3, 0, 3),
              pop2 = c(4, 10, 2, 5))
colnames(ctab) <- LETTERS[1:4]
ctab
chi2.stat(ctab) # chi^2 test statistic
```

---

fish215 *Genotypes of lake trout from Ontario, Canada*

Description

Observable genotypes (up to possibly duplicated alleles) of 215 lake trout (*Salvelinus namaycush*) collected from 11 lakes in Ontario, Canada.

Format

A data.frame with 215 rows and 5 columns. The first column is an (optional) vector of population identifiers. The next four columns contain the recorded genotype for each observation (row). Each row contains up to four distinct allele identifiers in any order. Missing alleles should be denoted by NA, or "", but not both.

Details

This `data.frame` is how a typical spreadsheet of genotype data gets imported into **R**. Data must adhere to this format to be correctly processed by the functions in **MADPop**.

Source

Kuntz, S. (2014). *Population Differentiation of Ontario Lake trout (Salvelinus namaycush) using the Major Histocompatibility Complex class II β gene* (URL).

Examples

```
head(fish215)
```

`hUM.post` *Posterior sampling from a hierarchical Unconstrained-Multinomial model*

Description

MCMC sampling from a Dirichlet-Multinomial model using `stan`.

Usage

```
hUM.post(nsamples, X, popId, rhoId, full.stan.out = FALSE, ...)
```

Arguments

`nsamples`	Number of posterior samples
`X`	4-column or 5-column matrix of observations in the correct format. See `UM.suff`.
`popId`	Optional vector of population identifiers. See `UM.suff`.
`rhoId`	Populations for which posterior samples of the genotype probability vector `rho` are desired. Defaults to all populations. Set `rhoId = NULL` not to output these for any populations.
`full.stan.out`	Logical. Whether or not to return the full `stan` output. For monitoring convergence of the MCMC sampling.
`...`	Further arguments to be passed to the `sampling` function in **rstan**.

Details

The hierarchical Dirichlet-Multinomial model is given by

$$Y_k \mid \rho_k \sim_{\text{ind}} \text{Multinomial}(\rho_k, N_k),$$

$$\rho_k \sim_{\text{iid}} \text{Dirichlet}(\alpha).$$

where $\alpha_0 = \sum_{i=1}^{C} \alpha_i$ and $\bar{\alpha} = \alpha/\alpha_0$. MCMC sampling is achieved with the **rstan** package, which is listed as a dependency for **MADPop** so as to expose **rstan**'s sophisticated tuning mechanism and convergence diagnostics.

Value

A list with elements

- A: The unique allele names.
- G: The 4-column matrix Package libcurl was not found in the pkg-config search path.of unique genotype combinations.
- rho: A matrix with ncol(rho) == nrow(G), where each row is a draw from the posterior distribution of inheritance probabilities.
- sfit: If full.stan.out = TRUE, the fitted stan object.

Examples

```
# fit hierarchical model to fish215 data

# only output posterior samples for lake Simcoe
rhoId <- "Simcoe"
nsamples <- 500
hUM.fit <- hUM.post(nsamples = nsamples, X = fish215,
                    rhoId = rhoId,
                    chains = 1) # number of MCMC chains

# plot first 20 posterior probabilities in lake Simcoe
rho.post <- hUM.fit$rho[,1,]
boxplot(rho.post[,1:20], las = 2,
        xlab = "Genotype", ylab = "Posterior Probability",
        pch = ".", col = "grey")
```

LRT.stat *Likelihood ratio test statistic for contingency tables*

Description

Calculate the likelihood ratio test statistic for general two-way contingency tables.

Usage

```
LRT.stat(tab)
```

Arguments

tab	A K x C matrix (contingency table) of counts. See details.

Details

Suppose that tab consists of counts from K populations (rows) in C categories. The likelihood ratio test statistic is computed as

$$2\sum_{i=1}^{K}\sum_{j=1}^{N} O_{ij}\log(p_{ij}^{A}/p_{j}^{0}),$$

where O_{ij} is the observed number of counts in the ith row and jth column of tab, $p_{ij}^{A} = O_{ij}/\sum_{j=1}^{C} O_{ij}$ is the unconstrained estimate of the proportion of category j in population i, and $p_{j}^{0} = \sum_{i=1}^{K} O_{ij}/\sum_{i=1}^{K}\sum_{j=1}^{C} O_{ij}$ is the estimate of this proportion under H_0 that the populations have indentical proportions in each category. If any column has only zeros it is removed before calculating the LRT statistic.

Value

The calculated value of the LRT statistic.

Examples

```
# simple contingency table
ctab <- rbind(pop1 = c(5, 3, 0, 3),
               pop2 = c(4, 10, 2, 5))
colnames(ctab) <- LETTERS[1:4]
ctab
LRT.stat(ctab) # likelihood ratio statistic
```

UM.eqtest *Equality tests for two multinomial samples*

Description

Generate multinomial samples from a common probability vector and calculate the Chi-square and Likelihood Ratio test statistics.

Usage

```
UM.eqtest(N1, N2, p0, nreps, verbose = TRUE)
```

Arguments

N1	Size of sample 1.
N2	Size of sample 2.
p0	Common probability vector from which to draw the multinomial samples. Can also be a matrix, in which case each simulation randomly draws with replacement from the rows of p0.
nreps	Number of replications of the simulation.
verbose	Logical. If TRUE prints message every 5000 replications.

Details

The chi-squared and likelihood ratio test statistics are calculated from multinomial samples $(Y_1^1, Y_2^1), \ldots, (Y_1^M, Y_2^M)$, where

$$Y_k^m \overset{\text{ind}}{\sim} \text{Multinomial}(N_k, p_0^m),$$

where p_0^m is the mth row of `p0`.

Value

An `nreps` x 2 matrix with the simulated chi-squared and LR values.

Examples

```
# bootstrapped p-value calculation against equal genotype proportions
# in lakes Michipicoten and Simcoe

# contingency table
popId <- c("Michipicoten", "Simcoe")
ctab <- UM.suff(fish215[fish215$Lake %in% popId,])$tab
ctab

# MLE of probability vector
p.MLE <- colSums(ctab)/sum(ctab)
# sample sizes
N1 <- sum(ctab[1,]) # Michipicoten
N2 <- sum(ctab[2,]) # Simcoe

# bootstrapped test statistics (chi^2 and LRT)
T.boot <- UM.eqtest(N1 = N1, N2 = N2, p0 = p.MLE, nreps = 1e3)

# observed test statistics
T.obs <- c(chi2 = chi2.stat(ctab), LRT = LRT.stat(ctab))
# p-values
rowMeans(t(T.boot) > T.obs)
```

`UM.suff` *Sufficient statistics for the Unconstrained-Multinomial model*

Description

Converts a matrix or data.frame of genotype data into the sufficient statistics required to fit a Dirichlet-Multinomial hierarchical model.

Usage

```
UM.suff(X, popId)
```

Arguments

X	Genotype adata. Either a N x 4 matrix with NA's indicating duplicates or a N x 5 column data.frame with the first column being the popId.
popId	grouping variable for X. Must be supplied if X has 4 columns.

Value

A list with elements:

- A: Vector of unique alleles names. The allele numbers in the following quantities correspond to the indices of A.
- G: 4-column matrix of unique genotype combinations. The presence of 0's indicates that less than four alleles were amplified indicating that a given genotype either has less than 4 distinct alleles or that some alleles are duplicated.
- tab: Observed data in a simplified numerical format. This is a contingency table with rows given by the unique elements of popId and columns given by each row of G.

Examples

```
# sufficient statistics in 3 lakes

X <- fish215[fish215$Lake %in% c("Hogan", "Manitou", "Simcoe"),]
suff <- UM.suff(X)

suff$A # allele names
suff$G # unique genotypes
suff$tab # contingency table
```

Index

chi2.stat, 3

fish215, 4

hUM.post, 5

LRT.stat, 6

MADPop *(*MADPop-package*)*, 2
MADPop-package, 2

sampling, *5*
stan, *5*

UM.eqtest, 7
UM.suff, *5*, 8