fitTetra documentation

Script fitTetra contains three functions that can be used to assign genotypes to a collection of tetraploid samples based on bi-allelic marker assays.
Functions fitTetra (to fit several models for one marker from the data and select the best fitting) or saveMarkerModels (calls fitTetra for multiple markers and saves the results to files) will probably be the most convenient to use. Function CodomMarker offers more control and fits one specified model for a given marker.

WARNING

The Windows 32-bit version of R2.12.0 and possibly 2.11.x and 2.10.x (but not 2.9.x) has a bug in the nls function that occasionally causes the functions in this script to "hang", i.e. they enter a perpetual loop and R does not respond any more. The bug and its fix are reported at http://bugs.r-project.org/bugzilla3/show_bug.cgi?id=14427 (R bugs report 14427) and the first patch to solve the problem is R version 2.12.0-Patched (2010-11-01 r53513). This problem does not occur in the Windows 64-bit and Linux versions of R2.12.0.

CodomMarker

Description

Function to fit a mixture model to a set of signal ratios of multiple samples for a single bi-allelic marker.

Usage

```
CodomMarker(y, ng=5, mutype=0, sdtype="sd.const", ptype="p.free",
           clus=TRUE, mustart=NA, sd.fixed=0.05, p=NA, maxiter=500,
           plothist=TRUE, nbreaks=40, maintitle=NULL, subtitle=NULL,
           xlabel=NULL, xaxis="s")
```

Arguments

y	the vector of signal ratios (each value is from one sample, vector y contains the values for 1 marker)
ng	the number of possible genotypes (mixture components) to be fitted: one more than the ploidy of the samples
mutype	an integer in 0:10. Describes how to fit the means of the components of the mixture model: with mutype=0 the means are not constrained, requiring ng degrees of freedom. With mutype in 1:10 the means are constrained based on the ng possible allele ratios according to one or 10 models; see Details.
sdtype	one of "sd.const", "sd.free", "sd.fixed". Describes how to fit the standard deviations of the components of the mixture model: with "sd.const" all standard deviations (on the transformed scale) are equal (requiring 1 degree of freedom); with "sd.free" all standard deviations are fitted separately (ng d.f.); with "sd.fixed" all sd's on the transformed scale are equal to parameter sd.fixed (0 d.f.).

ptype	one of "p.free", "p.fixed" or "p.HW". Describes how to fit the mixing proportions of the components of the mixture model: with "p.free", the proportions are not constrained (and require ng-1 degrees of freedom); with "p.fixed" the proportions given in parameter p are fixed; with "p.HW" the proportions are calculated from the overall allele frequency, requiring only 1 degree of freedom.
clus	boolean. If TRUE, the initial means and standard deviations are based on a hierarchical clustering into ng groups. If false, the initial means are equally spaced on the transformed scale between the values corresponding to 0.02 and 0.98 on the original scale and the initial standard deviations are 0.075 on the transformed scale.
mustart	vector of ng values. If present, gives the start values of mu on the original (untransformed) scale, must be strictly ascending (mu[i]>mu[i-1]). Overrides the start values determined by clus TRUE or FALSE.
sd.fixed	vector, recycled if less than ng values: if argument sdtype is "sd.fixed", argument sd.fixed specifies the fixed standard deviations.
p	a vector of ng elements with the initial (or fixed, of parameter ptype is "p.fixed") mixing proportions of the mixture model components.
maxiter	the maximum number of iterations (0 = no limit, default=500)
plothist	If TRUE a histogram of y is plotted with the fitted distributions superimposed
nbreaks	number of breaks for plotting the histogram; does not have an effect on fitting the mixture model
maintitle	string, used for plotting
subtitle	string, used for plotting
xlabel	string, used for plotting
xaxis	string, used for plotting: if "n" no x-axis is plotted

Details

This function takes as input a vector of ratios of the signals of two alleles at a genetic marker locus (ratios as a/(a+b)), one for each sample, and fits a mixture model with ng components (e.g. for a tetraploid species: ng=5 components representing the nulliplex, simplex, duplex, triplex and quadruplex genotypes). Ideally these signal ratios should reflect the possible allele ratios (for a tetraploid: 0, 0.25, 0.5, 0.75, 1) but in real life they show a continuous distribution with a number of more or less clearly defined peaks.

The arguments specify what model to fit and with what values the iterative fitting process should start. If the argument mutype is set to a value in 1:10 the means of the mixture model components are constrained based on the possible allele ratios. This constraint takes the form of one of 10 possible models, specified by mutype, as follows:
1: a basic model assuming that both allele signals have a linear response to the allele dosage; one parameter for the ratio of the slopes of the two signal responses, and two parameters for the background levels (intercepts) of both signals (total 3 parameters).
2: as 1, but with the same background level for both signals (2 parameters)

3: as 1, with two parameters for a quadratic effect in the signal responses (5 parameters)
4: as 3, but with the same background level for both signals (4 parameters)
5: as 3, but with the same quadratic parameter for both signal responses (4 parameters)
6: as 5, but with the same background level for both signals (3 parameters)

Value

A list; if an error occurs the only list component is

```
message the error message
```

If no error occurs the list has the following components:

```
loglik  the optimized log-likelihood
npar    the number of fitted parameters
AIC     Akaike's Information Criterion
BIC     Bayesian Information Criterion
psi     a list with components mu, sigma and p: each a vector of
        length ng with the means, standard deviations and mixing
        proportions of the components of the fitted mixture model;
        the means and standard deviations are on the transformed
        scale
post    a matrix of ng columns and length(y) rows; each row r gives
        the ng probabilities that the y[r] belongs to the ng
        components
nobs    the number of observations in y (including NA's and possibly
        removed outliers)
iter    the number of iterations
message an error message, "" if no error
back    a list with components mu.back and sigma.back: each a vector
        of length ng with the means and standard deviations of the
        mixture model components back-transformed to the original
        scale.
```

fitTetra

Description

This function takes a data frame with allele signals for multiple markers and samples, and finds a fitting model for one specified marker

Usage

```
fitTetra(marker, data, diplo=NA, select=TRUE, diploselect=TRUE,
           maxiter=40, try.HW=TRUE, sd.threshold=0.1,
           p.threshold=0.99, call.threshold=0.6, peak.threshold=0.85,
           dip.filter=T, plot="none", plot.type="emf")
```

Arguments

```
marker     integer: specifies the marker number to analyze. "marker"
           is the index to the alphabetically sorted MarkerNames (see
           argument "data")
```

data data frame for tetraploid samples, with (at least) columns "MarkerName", "SampleName", and "ratio", where ratio is the X allele signal divided by the sum of the X and Y allele signals.

diplo data frame like "data" with diploid samples. Facultative, only used for plotting, does not affect model fitting.

select boolean vector, recycled if shorter than the columns in data: indicates which rows are to be used (default: select=TRUE, i.e. keep all rows)

diploselect as select, for diplo instead of data

maxiter integer: the maximum number of times the nls function is called in CodomMarker

try.HW boolean: if TRUE (default), try models with and without a constraint on the mixing proportions according to Hardy-Weinberg equilibrium ratios. If FALSE, only try models without this constraint

sd.threshold the maximum value allowed for the (constant) standard deviation on the arcsine – square root transformed scale, default 0.1. If the optimal model has a larger standard deviation the marker is rejected.

p.threshold the minimum P-value required to assign a genotype to a sample; default 0.99. If the P-value for all 5 possible genotypes is less than p.threshold the sample is assigned genotype NA.

call.threshold the minimum fraction of samples to have genotypes assigned ("called"); default 0.6. If under the optimal model the fraction of "called" samples is less than call.threshold the marker is rejected.

peak.threshold the maximum allowed fraction of the scored samples that are in one peak; default 0.85. If any of the possible genotypes (peaks in the ratio histogram) contains more than peak.threshold of the samples the marker is rejected (because the remaining samples offers too little information for reliable model fitting)

dip.filter boolean: if TRUE (default), select only from models that do not have a dip (a lower peak surrounded by higher peaks: these are not expected under Hardy-Weinberg equilibrium or in cross progenies). Note: if all fitted models have a dip still the best of these is selected

plot string, "none" (default), "fitted" or "all". If "fitted" a plot of the best fitting model and the assigned genotypes is generated and saved to a file named <marker number><marker name>.<plot.type>; if "all" additionally small images of all models are saved to files (8 per file) with filename <"plots"><marker number><A/B/C><marker name>.<plot.type>

plot.type string, "emf" (default), "png" or "pdf". Indicates format of saved plot files. On non-Windows platforms the default "emf" is not available and "png" is used instead

Details

fitTetra fits a series of mixture models for the given marker by repeatedly calling CodomMarker and selects the optimal one. The models tested have four different models for the means of the mixture components: mutype 1, 2, 5 and 6 as described for CodomMarker, and one or two (depending on argument try.HW) models for the mixing proportions. These four or eight models are run using 2, 3 or 4 different start configurations. The model with the smallest Bayesian Information Criterion is selected, within the constraints specified by p.threshold, call.threshold, peak.threshold and dip.filter.

Value

a list with components:

log — a character vector with the lines of the log text

modeldata — a data frame with one row with the marker number, marker name, number of samples and (if the marker is not rejected) data of the fitted model (see below)

allmodeldata — a data frame with 16, 24 or 32 rows with data of all attempted model fits, including error messages if applicable (see below)

scores — a data frame with the name and data for all samples (including NA's for the samples that were not selected, see parameter select): marker (same as argument marker), MarkerName, SampleName, model (a string describing the model),select (value of argument select for this data point),ratio (the given or calculated ratio from argument data), P0,P1,P2,P3,P4 (the probabilities that this sample belongs to each of the five mixture components),maxgeno (the genotype = mixture component with the highest P value), maxP (the P value for this genotype) and geno (the assigned genotype number: same as maxgeno, or NA if maxP<p.threshold)

The modeldata and allmodeldata data frames present data on a fitted model. modeldata presents data on the selected model; allmodeldata lists all attempted models and gives additional information that can be used to assess the differences between these models. Both data frames contain the following columns:

marker — the sequential number of the marker (marker names are ordered alphabetically)

markername — the name of the marker

model — the fitted model. Possible values are "b1", "b2", "b1,q", "b2,q", "b1 HW", "b2 HW", "b1,q HW" and "b2,q HW" where b1 and b2 indicate whether 1 or two parameters for signal background were fitted, q indicates that a quadratic term in the signal response was fitted, and HW indicates that the mixing proportions were constrained according to Hardy-Weinberg equilibrium ratios. For more details see Voorrips et al (2011)

nsamp — the number of samples for this marker in data

nsel — the number of these samples for which select==TRUE

dip — 0 or 1 for FALSE or TRUE. If 1, at least one of the three central mixture components has a smaller mixing proportion

```
           and/or less samples than components on both sides. A “dip”
           is unexpected both under HW equilibrium and in cross
           progenies.
P80, P90, P95, P975, P99      the fraction of selected samples that
           have a probability of at least 0.8, 0.9, 0.95, 0.975 or
           0.99 to belong to one of the five mixture components (by
           default a level of 0.99 is required to assign a genotype
           score to a sample)
mu0, mu1, mu2, mu3, mu4       the means of the five mixture
           components on the original scale
P0, P1, P2, P3, P4      the mixing proportions of the five components
```

In allmodeldata additional columns are present that allow comparisons between models for the same marker and/or may be used as quality indicators:

```
m          the number of the attempted fit. The 8 (or 4 if try.HW is
           FALSE) models are tried with 2, 3 or 4 start
           configurations , so m can range from 1 to 16, 24 or 32
npar       the number of free parameters to be fitted
iter       the number of iterations to reach convergence
LL         the log-likelihood of the fitted model
AIC        Akaike’s Information Criterion
BIC        Bayesian Information Criterion
minsepar   a measure of the minimum peak separation. each difference
           of the means of two successive mixture components is
           divided by the average of the standard deviations of the
           two components. The minimum of the four values is
           reported. All calculations are on the arcsine-squere root
           transformed scale.
meanP      For each sample the maximum probability of belonging to
           any mixture component is calculated. The average of these
           P values is reported in meanP
mutrans0, mutrans1, mutrans2, mutrans3 and mutrans4: the means of the
           mixture components on the arcsine-square root transformed
           scale
sdtrans    the standard deviations of the mixture components on the
           arcsine-square root transformed scale
message    if no model was selected the reason is reported here. The
           most common case is iter>maxiter; increasing maxiter may
           solve some of these errors (but usually a high number of
           iterations indicates that the data are too noisy). Other
           error messages usually reflect numerical computation
           issues that have no obvious solution.
```

saveMarkerModels

Description

This is a convenience function that calls fitTetra for a series of markers and saves the tabular, graphical and log output to files.

Usage

```
saveMarkerModels(markers=NA, data, diplo=NA, select=T, diploselect=T,
          maxiter=40, try.HW=T, sigma.threshold=0.1,
          p.threshold=0.99, call.threshold=0.6, peak.threshold=0.85,
          dip.filter=TRUE, logfile="", modelfile, allmodelsfile="",
          scorefile, plot="none", plot.type="emf")
```

Arguments

Most of the arguments are identical to those of fitTetra and are directly passed through. Arguments specific to saveMarkerModels are:

markers	a vector listing the markers to be analyzed. The numbers refer to the levels of data$MarkerName. If "" (default) all markers are analyzed.
logfile	string, name of a text file. This file will contain several text lines per marker corresponding to component "log" in the result of fitTetra. If "" (default) no file is created.
modelfile	string, name of a text file. This file will contain one line per marker corresponding to component "modeldata" in the result of fitTetra. modelfile can be read using read.table.
allmodelsfile	string, name of a text file. This file will contain 16, 24 or 32 lines per marker, corresponding to component " allmodeldata " in the result of fitTetra. allmodelsfile can be read using read.table. If "" (default) no file is created.
scorefile	string, name of a text file. This file will contain one line per sample for every marker that could be fitted, corresponding to component "scores" in the result of fitTetra. scorefile can later be read using read.table

Value

This function does not return a value.