Package ‘radmixture’

March 13, 2017

Title    Calculate Population Stratification
Version  0.0.1
Date     2017-02-20
Maintainer Beilei Bian <bianbeilei@wegene.com>
Description Implementation of ADMIXTURE for individual ancestry inference in R. Specifically, ADMIXTURE is a software tool for maximum likelihood estimation of individual ancestries from multilocus SNP genotype datasets, see <https://www.genetics.ucla.edu/software/admixture/>. Users can use 'radmixture' to calculate ancestry components with different public datasets. It is very convenient and fast for personal genotype data. For more details, see <https://github.com/wegene-llc/radmixture/blob/master/README.md>.

Depends  R (>= 3.1.0)
Importsquadprog, plyr, magrittr, MCMCpack
Suggests rmarkdown, knitr, testthat
License  MIT + file LICENSE
URL      https://github.com/wegene-llc/radmixture
BugReports https://github.com/wegene-llc/radmixture/issues
Encoding UTF-8
LazyData true
RoxygenNote 6.0.1
NeedsCompilation yes
VignetteBuilder knitr
ByteCompile true
Author   Beilei Bian [aut, cre],
          Dajun Luo [ctb] (R’s C API),
          Gang Chen [ctb],
          Senwei Tang [ctb],
          WeGene [cph]
Repository CRAN
Date/Publication 2017-03-13 19:25:45
**R topics documented:**

- br .................................................. 2
- em .................................................. 3
- fFixBr ............................................. 3
- fFixEm ............................................. 4
- fFixQN ............................................. 5
- generateG ....................................... 5
- initQF ........................................... 6
- qn .................................................. 6
- radmixture ..................................... 7
- tfrdpub ......................................... 7

---

**Description**

This function is also used for estimating Q and F but faster than EM.

**Usage**

```r
br(g, q, f, acc, max.iter, tol, model)
```

**Arguments**

- `g` Genotype matrix with dimensions `np`, where `n` is sample size and `p` is the number of SNPs.
- `q` Ancestry coefficient matrix with dimensions `nK`, where `n` is sample size and `K` is the number of populations.
- `f` Minor allele frequency matrix with dimensions `Kp`, where `K` is the number of populations and `p` is the number of SNPs.
- `acc` a logical value indicating whether use quasi-Newton accelerated BR or not.
- `max.iter` If `acc = T`, `max.iter` must be set, the default is 3. `max.iter` should greater than 1.
- `tol` Tolerance, if `acc = F`, tolerance must be set, the default is 1e-4.
- `model` Choose which model you want to use. Supervised learning or unsupervised learning.

**Value**

Estimation results of q, f and the loglikelihood value of each iteration.
em

Do ancestry analysis with EM algorithm

Description
The EM algorithm could be used for estimating the Q and F matrix.

Usage
em(g, q, f, acc, max.iter, tol, model)

Arguments
- g: Genotype matrix with dimensions np, where n is sample size and p is the number of SNPs.
- q: Ancestry coefficient matrix with dimensions nK, where n is sample size and K is the number of populations.
- f: Minor allele frequency matrix with dimensions Kp, where K is the number of populations and p is the number of SNPs.
- acc: a logical value indicating whether use accelerated EM or not.
- max.iter: an integer. If acc is TRUE, the number of iterations must be set. max.iter should be greater than 1.
- tol: Tolerance. If acc is FALSE, tol must be set. The default is 1e-4.
- model: Choose which model you want to use. Supervised learning or unsupervised learning.

Value
Estimation results of q, f and the loglikelihood value of each iteration.

ffixbr

Block relaxation when f is fixed

Description
This function can be used for ancestry analysis when frequency matrix is fixed.

Usage
ffixbr(gnew, qnew, f, acc, max.iter, tol, pubdata)
**Arguments**

- `gnew`: Genotype matrix. The number of row present in `gnew` is 1 and the number of column is the number of SNPs.
- `qnew`: Initial q used in calculation. A vector. Sum(q) must be 1.
- `f`: Allele frequencies matrix learned from the reference panels.
- `acc`: a logical value indicating whether use quasi-Newton accelerated BR or not.
- `max.iter`: If acc = T, max.iter must be set, the default is 3. max.iter should greater than 1.
- `tol`: If acc = F, tolerance must be set, the default is 1e-4.
- `pubdata`: You can choose a public dataset here, E11, K13, K4, K12b, K7b, World9. You also can use other public dataset which is not in this package.

**Value**

Estimation results of q and the loglikelihood value of each iteration.

---

**Description**

This function can be used for ancestry analysis when frequency matrix is fixed.

**Usage**

```r
fFixEm(gnew, qnew, f, acc, max.iter, tol = 1e-4, pubdata)
```

**Arguments**

- `gnew`: Genotype matrix. The number of row present in `gnew` is 1 and the number of column is the number of SNPs.
- `qnew`: Initial q used in calculation. A vector. Sum(q) must be 1.
- `f`: Allele frequencies matrix learned from the reference panels.
- `acc`: a logical value indicating whether use quasi-Newton accelerated EM or not.
- `max.iter`: an integer. If acc is TRUE, the number of iterations must be set. max.iter should greater than 1.
- `tol`: Tolerance. If acc is FALSE, tol must be set. The default is 1e-4.
- `pubdata`: You can choose a public dataset here, E11, K13, K4, K12b, K7b, World9. You also can use other public dataset which is not in this package.

**Value**

Estimation results of q and the loglikelihood value of each iteration.
**fFixQN**

*quasi-Newton when f is fixed*

**Description**

quasi-Newton for ancestry analysis when F is fixed

**Usage**

`fFixQN(gnew, qnew, f, tol, method, pubdata)`

**Arguments**

- `gnew`: Integer which length is the number of SNPs used in calculation.
- `qnew`: Initial q used in calculation. A vector. `sum(q)` must be 1.
- `f`: Allele frequencies learned from the reference panels.
- `tol`: Tolerance, the default value is 1e-4.
- `method`: Choose which algorithm you want to use. EM or BR.
- `pubdata`: You can choose a public dataset here, E11, K13, K4, K12b, K7b, World9. You also can use other public dataset which is not in this package.

**Value**

Estimation results of q and the loglikelihood value of each iteration.

**Examples**

```r
## res <- tfrdpub(genotype, 4, globe4.alleles, globe4.4.F)
## ances <- fFixQN(res$g, res$q, res$f, tol = 1e-4, method = 'BR', pubdata = 'K4')
```

---

**generateG**

*Transfer ped file to genotype matrix*

**Description**

This function can be used to transfer a ped file to g matrix

**Usage**

`generateG(rawped)`

**Arguments**

- `rawped`: A data.frame. Standard ped format. Genotype should be transferred to 1,2,3,4 from A,C,G,T. 0 represents missing. ‘-’,’_’,’I’,’D’ should be replaced by 0 by yourself.
initQF  

Initialize $Q$ and $F$

Description

This function could help you initialize $Q$ and $F$ matrix conveniently especially when you intend to use supervised learning.

Usage

initQF(g, pop = NULL, alpha = NULL, k = NULL, model)

Arguments

g  genotype matrix

pop  A data.frame. If you intend to do supervised learning, you must specify the ancestries of the reference individuals.

alpha  Parameter for dirichlet distribution. Vector of shape parameters, or matrix of shape parameters corresponding to the number of draw.

k  If you intend to do unsupervised learning, set the number of populations you will use.

model  Choose supervised or unsupervised learning.

Value

A list contains $q$ and $f$ matrix.

---

qn  

quasi-Newton algorithm for ancestry analysis

Description

Use quasi-Newton algorithm to accelerate EM or block relaxation.

Usage

qn(g, q, f, tol = 1e-4, method, model)
Arguments

- **g**: Genotype matrix with dimensions $np$, where $n$ is sample size and $p$ is the number of SNPs.
- **q**: Ancestry coefficient matrix with dimensions $nK$, where $n$ is sample size and $K$ is the number of populations.
- **f**: Minor allele frequency matrix with dimensions $Kp$, where $K$ is the number of populations and $p$ is the number of SNPs.
- **tol**: Tolerance, the default value is $1e-4$.
- **method**: Choose which algorithm you want to use. EM or BR.
- **model**: Choose which model you want to use. Supervised learning or unsupervised learning.

Value

Estimation results of $q$, $f$ and the loglikelihood value of each iteration.

Examples

```r
## qn(g, q, f, tol = 1e-4, method = 'BR', model = 'supervised')
```

Description

`radmixture` is an R package for ancestry calculation. It provides both supervised and unsupervised learning with several algorithms for researchers and DNA customers. see README on GitHub

```
radmixture
```

```
tfrdpub
```

Transfer personal genotype raw data according public dataset

Description

Transfer personal genotype raw data to g matrix which the number of row is 1 and the number of column is the number of SNPs used here.

Usage

tfrdpub(genotype, K, map, f)
Arguments

- **genotype**: A data.frame contains your genotype information.
- **K**: The number of populations
- **map**: A data.frame, it should contain rsid, major allele and minor allele information for both plus and minus strands. You should download datasets from GitHub.
- **f**: Frequency matrix learned from reference panel. You should download datasets from GitHub.

Details

Please download datasets from GitHub See README.

Value

A list contains g, q, f which can be used for calculation.

Examples

```r
## download.file(url = 'https://github.com/wegene-llc/radmixture/raw/master/data/globe4.alleles.RData', destfile = 'K4.RData')
## download.file(url = 'https://github.com/wegene-llc/radmixture/raw/master/data/globe4.4.F.RData', destfile = 'K4f.RData')
## load('K4.RData')
## load('K4f.RData')
## res <- tfrdpub(genotype, 4, globe4.alleles, globe4.4.F)
```
Index

br, 2
em, 3
fFixBr, 3
fFixEm, 4
fFixQN, 5
generateG, 5
initQF, 6
qn, 6
radmixture, 7
radmixture-package (radmixture), 7
tfrdpub, 7