Package 'Summix'

Version 2.4.0

LazyData false

Title Summix: A method to estimate and adjust for population structure
in genetic summary data

Description This package contains the Summix method for estimating and adjusting for ancestry in genetic summary allele frequency data. The function summix estimates reference ancestry proportions using a mixture model. The adjAF function produces ancestry adjusted allele frequencies for an observed sample with ancestry proportions matching a target person or sample.

License MIT + file LICENSE

RoxygenNote 7.1.1

Suggests rmarkdown, markdown, knitr

biocViews StatisticalMethod, WholeGenome, Genetics

VignetteBuilder knitr

Encoding UTF-8

Depends R (>= 4.1)

Imports nloptr, methods

BugReports https://github.com/Bioconductor/Summix/issues

git_url https://git.bioconductor.org/packages/Summix

git_branch RELEASE_3_16

git_last_commit b5a6d27

git_last_commit_date 2022-11-01

Date/Publication 2023-04-10

Author Audrey Hendricks [cre],
Stoneman Haley [aut]

Maintainer Audrey Hendricks <audrey.hendricks@ucdenver.edu>

R topics documented:

---

adjAF   *Ancestry adjusted allele frequencies*

Description

Adjusts allele frequencies for heterogeneous populations in genetic data given proportion of reference ancestry groups

Usage

```
adjAF(data, reference, observed, pi.target, pi.observed)
```

Arguments

data	dataframe of unadjusted allele frequency for observed group, K-1 reference ancestry allele frequencies for N SNPs
reference	character vector of the column names for K-1 reference ancestry groups. The name of the last reference ancestry group is not included as that group is not used to estimate the adjusted allele frequencies.
observed	character value for the column name of observed data group
pi.target	numeric vector of the mixture proportions for K reference ancestry groups in the target sample or subject. The order must match the order of the reference columns with the last entry matching the missing reference group.
pi.observed	numeric vector of the mixture proportions for K reference ancestry groups for the observed group. The order must match the order of the reference columns with the last entry matching the missing reference group.

Value

pi: table of input reference ancestry groups, pi.observed, and pi.target

observed.data: name of the data column for the observed group from which adjusted ancestry allele frequency is estimated

Nsnps: number of SNPs for which adjusted AF is estimated

adjusted.AF: data frame of original data with an appended column of adjusted allele frequencies

Author(s)

Gregory Matesi, <gregory.matesi@ucdenver.edu>

Audrey Hendricks, <audrey.hendricks@ucdenver.edu>

See Also

summix for estimating the proportion of reference ancestry groups and https://github.com/hendriau/Summix for further documentation

Examples

```
data(ancestryData)
tmp.aa<-adjAF(data     = ancestryData,
    reference  = c("ref_AF_eur_1000G"),
    observed   = "gnomad_AF_afr",
    pi.target  = c(0, 1),
    pi.observed = c(.15, .85))
tmp.aa$adjusted.AF[1:5,]
```

---

ancestryData   *Example allele frequency data*

Description

reference data is 1000 Genomes and NAM. 1000 Genomes data was downloaded from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ft
on May 31, 2018 The IAM Affymetrix 6.0 data were downloaded from ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working
(data accessed October 2018) and had been previously harmonized with the 1000 Genomes data.
Observed data is from gnomAD. gnomAD v2 data was downloaded from https://gnomad.broadinstitute.org/downloads
on Oct. 11, 2018

Usage

```
data(ancestryData)
```

Format

Chromosome, SNP, base pair, reference and alternate alleles, reference allele frequencies, observed
allele frequencies

Examples

```
data("ancestryData")
```

---

summix   *Summix: estimating mixture proportions of reference group*

Description

Summix: estimating mixture proportions of reference groups from large (N SNPs>10,000) genetic
AF data

Usage

```
summix(data, reference, observed, pi.start = c())
```

Arguments

`data`	a dataframe of the observed and reference allele frequencies for N genetic variants. See data formatting document at https://github.com/hendriau/Summix for more information.
`reference`	a character vector of the column names for the reference ancestries.
`observed`	a character value that is the column name for the observed group.
`pi.start`	length K numeric vector of the starting guess for the ancestry proportions. If not specified, this defaults to 1/K where K is the number of reference ancestry groups.

Value

data frame with the following columns

objective: least square value at solution

iterations: number of iterations for SLSQP algorithm

time: time in seconds of SLSQP algorithm

filtered: number of SNPs not used in estimation due to missing values

K columns of mixture proportions of reference ancestry groups input into the function

Author(s)

Gregory Matesi, <gregory.matesi@ucdenver.edu>

Audrey Hendricks, <audrey.hendricks@ucdenver.edu>

See Also

adjAF for adjusting allele frequencies and https://github.com/hendriau/Summix for further documentation. slsqp function in the nloptr package for further details on Sequential Quadratic Programming https://www.rdocumentation.org/packages/nloptr/versions/1.2.2.2/topics/slsqp

Examples

```
# load the data
data("ancestryData")

# Estimate 5 reference ancestry proportion values for the gnomAD African/African American group
# using a starting guess of .2 for each ancestry proportion.
summix( data = ancestryData,
    reference=c("ref_AF_afr_1000G",
        "ref_AF_eur_1000G",
        "ref_AF_sas_1000G",
        "ref_AF_iam_1000G",
        "ref_AF_eas_1000G"),
    observed="gnomad_AF_afr",
    pi.start = c(.2, .2, .2, .2, .2) )
```

Index