Package 'gbs2ploidy'

December 1, 2016

Type Package

Title Inference of Ploidy from (Genotyping-by-Sequencing) GBS Data

Version 1.0

Date 2016-12-01

Author Zachariah Gompert

Maintainer Zachariah Gompert <zach.gompert@usu.edu>

Depends R (>= 2.10), MASS, rjags

Description Functions for inference of ploidy from (Genotyping-by-sequencing) GBS data, including a function to infer allelic ratios and allelic proportions in a Bayesian framework.

License GPL-3

NeedsCompilation no

Repository CRAN

Date/Publication 2016-12-01 21:32:40

R topics documented:

`gbs2ploidy-package` *Inference of Ploidy from (Genotyping-by-Sequencing) GBS Data*

Description

Functions for inference of ploidy from (Genotyping-by-sequencing) GBS data, including a function to infer allelic ratios and allelic proportions in a Bayesian framework.

Details

The DESCRIPTION file:

Package:	gbs2ploidy
Type:	Package
Title:	Inference of Ploidy from (Genotyping-by-Sequencing) GBS Data
Version:	1.0
Date:	2016-12-01
Author:	Zachariah Gompert
Maintainer:	Zachariah Gompert <zach.gompert@usu.edu>
Depends:	R (>= 2.10), MASS, rjags
Description:	Functions for inference of ploidy from (Genotyping-by-sequencing) GBS data, including a function to infer all
License:	GPL-3

Index of help topics:

A typical analysis will begin by estimating allelic proportions using the `estprops` function. This is done in a Bayesian framework and is the most computationally intensive part of the analysis (i.e., depending on the size of the data set, this might take a day or more). This function depends on `rjags`, which means the user needs to install the stand-alone program `JAGS` as well. Principal component analysis and discriminant analysis are then used to obtain cytotype assignment probabilities via the `estploidy` function. This can be done with or without a training set of individuals with known ploidies.

Author(s)

Zachariah Gompert

Maintainer: Zachariah Gompert <zach.gompert@usu.edu>

References

Gompert Z. and Mock K. (XXXX) Detection of individual ploidy levels with genotyping-by-sequencing (GBS) analysis. Molecular Ecology Resources, submitted.

Examples

```
## load a simulated data set
data(dat)
## Not run:
## obtain posterior estimates of allelic proportions; short chains are used for
## the example, we recommend increasing this to at least 1000 MCMC steps with a
## 500 step burnin
props<-estprops(cov1=t(dat[[1]]),cov2=t(dat[[2]]),mcmc.steps=20,mcmc.burnin=5,
    mcmc.thin=2)
```

```
## calculate observed heterozygosity and depth of coverage from the allele count
## data
hx<-apply(is.na(dat[[1]]+dat[[2]])==FALSE,1,mean)
dx<-apply(dat[[1]]+dat[[2]],1,mean,na.rm=TRUE)

## run estploidy without using known ploidy data
pl<-estploidy(alphas=props,het=hx,depth=dx,train=FALSE,pl=NA,set=NA,nclasses=2,
    ids=dat[[3]],pcs=1:2)

## boxplots to visualize posterior assignment probabilities by true ploidy
## (which is known because these are simulated data)
boxplot(pl$pp[,1] ~ dat[[3]],ylab="assignment probability",xlab="ploidy")

## run estploidy with a training data set with known ploidy; the data set is
## split into 100 individuals with known ploidy and 100 that are used for
## inference
truep<-dat[[3]]
trn<-sort(sample(1:200,100,replace=FALSE))
truep[-trn]<-NA
plt<-estploidy(alphas=props,het=hx,depth=dx,train=TRUE,pl=truep,set=trn,
    nclasses=2,ids=dat[[3]],pcs=1:2)

## boxplots to visualize posterior assignment probabilities for individuals that
## were not part of the training set by true ploidy (which is known because
## these are simulated data)
boxplot(plt$pp[,1] ~ dat[[3]][-trn],ylab="assignment probability",xlab="ploidy")

## End(Not run)
```

`dat` *Simulated allele counts*

Description

`dat` is a simulated data set meant to mimic GBS data. It is a list with three components. The first two components are N (number of individuals) by P (number of SNPs) matrixes with allele counts for the first and second allele at each locus, respectively. The third component is a numeric vector that gives the true ploidy for each individual (2 = diploid, 4 = tetraploid).

Usage

```
data("dat")
```

Format

The format is: List of 3 $: int [1:200, 1:10000] NA NA 7 4 NA NA NA NA NA NA ... $: int [1:200, 1:10000] NA NA 2 5 NA NA NA NA NA NA ... $: num [1:200] 4 2 4 2 2 2 2 2 2 2 ...

Examples

```
data(dat)
str(dat)
```

`estploidy` *Discriminate cytotypes using GBS data*

Description

Use principal component analysis (PCA) and discriminant analysis (DA) to assign individuals to different cytotype groups based on GBS data. This function can be run with or without a training set of individuals with known cytotypes.

Usage

```
estploidy(alphas = NA, het = NA, depth = NA, train = FALSE, pl = NA, set = NA,
    nclasses = 2, ids = NA, pcs = 1:2)
```

Arguments

`alphas`	a list generated by the `estprops` function that contains posterior estimates of allelic proportions.
`het`	a numeric vector with the observed heterozygosity for each individual, that is the proportion of SNPs at which the individual was heterozygous.
`depth`	a numeric vector with the mean number of reads per SNP for each individual (all SNPs or just heterozygous SNPs).
`train`	a boolean specifying whether or not a training set with known ploidy should be used.
`pl`	a vector of known ploidies with one entry per individual (use 'NA' for individuals with unknown ploidy); only used if `train == TRUE`.
`set`	indixes for the training set; only used if `train == TRUE`.
`nclasses`	the number of cyotypes expected.
`ids`	names or other IDs for individuals.
`pcs`	a vector giving the PC to use for DA.

Details

Assignment probabilities to different cytotype groups are obtained using PCA and DA, as described in Gompert & Mock (XXXX). Residual hetorzygosity (from regressing `het` on `depth`) and point estimates (posterior medians) for allelic proportions are first ordinated using PCA (on the centered covariance matrix). The first `pcs` PCs are retained for DA. If a training set is not provided, k-means clustering is used to obtain initial classifications (with `nclasses` groups), which are then used for leave-one-out cross-validation with DA. In this case, PC loadings and discriminant scores should be evaluated to associate groups from k-means clustering with specific cytotypes (Gompert & Mock, XXXX). If a training set is provided, it is used to define groups for DA and to estimate assignment probabilities without k-means clustering.

Value

estploidy returns a list with three components:

pp	A matrix with assignment probabilities for each individual (rows) to each group (columns); the first column gives the ids provided by the user. Only individuals that were not part of the training set are included.
pcwghts	A matrix with the variable loadings (PC weights) from the ordination of residual heterozygosity and allelic proportions. Columns correspond with PCs in ascending order (i.e., the PC with the largest eigenvalue is first).
pcscrs	A matrix of PC scores from the ordination of residual heterozygosity and allelic proportions. Columns correspond with PCs in ascending order (i.e., the PC with the largest eigenvalue is first).

Author(s)

Zachariah Gompert

References

Gompert Z. and Mock K. (XXXX) Detection of individual ploidy levels with genotyping-by-sequencing (GBS) analysis. Molecular Ecology Resources, submitted.

See Also

estprops

Examples

```
## load a simulated data set
data(dat)
## Not run:
## obtain posterior estimates of allelic proportions; short chains are used for
## the example, we recommend increasing this to at least 1000 MCMC steps with a
## 500 step burnin
props<-estprops(cov1=t(dat[[1]]),cov2=t(dat[[2]]),mcmc.steps=20,mcmc.burnin=5,
    mcmc.thin=1)

## calculate observed heterozygosity and depth of coverage from the allele count
## data
hx<-apply(is.na(dat[[1]]+dat[[2]])==FALSE,1,mean)
dx<-apply(dat[[1]]+dat[[2]],1,mean,na.rm=TRUE)

## run estploidy without using known ploidy data
pl<-estploidy(alphas=props,het=hx,depth=dx,train=FALSE,pl=NA,set=NA,nclasses=2,
    ids=dat[[3]],pcs=1:2)

## boxplots to visualize posterior assignment probabilities by true ploidy
## (which is known because these are simulated data)
boxplot(pl$pp[,1] ~ dat[[3]],ylab="assignment probability",xlab="ploidy")
```

```
## run estploidy with a training data set with known ploidy; the data set is
## split into 100 individuals with known ploidy and 100 that are used for
## inference
truep<-dat[[3]]
trn<-sort(sample(1:200,100,replace=FALSE))
truep[-trn]<-NA
plt<-estploidy(alphas=props,het=hx,depth=dx,train=TRUE,pl=truep,set=trn,
    nclasses=2,ids=dat[[3]],pcs=1:2)

## boxplots to visualize posterior assignment probabilities for individuals that
## were not part of the training set by true ploidy (which is known because
## these are simulated data)
boxplot(plt$pp[,1] ~ dat[[3]][-trn],ylab="assignment probability",xlab="ploidy")

## End(Not run)
```

`estprops` *Estimate allelic proportions*

Description

This functions uses Markov chain Monte Carlo to obtain Bayesian estimates of allelic proportions, which denote that proportion of heterozygous GBS SNPs with different allelic ratios.

Usage

```
estprops(cov1 = NA, cov2 = NA, props = c(0.25, 0.33, 0.5, 0.66, 0.75),
    mcmc.nchain = 2, mcmc.steps = 10000, mcmc.burnin = 1000, mcmc.thin = 2)
```

Arguments

`cov1`	a P (number of SNPs) by N (number of individuals) matrix with read counts for the first allele (e.g., the non-reference allele). Numeric values should be provided for heterygous SNPs only, homozygous SNPs should be coded as missing data (i.e., 'NA').
`cov2`	a P (number of SNPs) by N (number of individuals) matrix with read counts for second allele (e.g., the reference allele). Numeric values should be provided for heterygous SNPs only, homozygous SNPs should be coded as missing data (i.e., 'NA').
`props`	a vector containing valid allelic proportions given the expected cyotypes present in the sample.
`mcmc.nchain`	number of chains for MCMC.
`mcmc.steps`	number of post burnin iterations for each chain.
`mcmc.burnin`	number of iterations to discard from each chain as a burnin.
`mcmc.thin`	thinning interval for MCMC.

Details

Allelic proportions are inferred from the allele counts based on the Bayesian model described in Gompert \& Mock (XXXX). Please consult this publication for a detailed description of the model. Users can modify the vector of possible allelic proportions based on expectations for their data set. For example, true allelic proportions for diploids, triploids and tetraploids are 1:1 (0.5), 1:2 or 2:1 (0.33 or 0.66), and 1:3, 2:2, or 3:1 (0.25, 0.5, or 0.75), respectively.

Value

`estprops` returns a list with one component per individual. Components summarize the posterior distributions for allelic proportions. Rows correspond to different allelic proportions (as defined by 'props') and columns give the 2.5th, 25th, 50th, 75th, and 97.5th quantiles of the posterior distribution for each parameter.

Author(s)

Zachariah Gompert

References

Gompert Z. and Mock K. (XXXX) Detection of individual ploidy levels with genotyping-by-sequencing (GBS) analysis. Molecular Ecology Resources, submitted.

Examples

```
## load a simulated data set
data(dat)
## Not run:
## obtain posterior estimates of allelic proportions; short chains are used for
## the example, we recommend increasing this to at least 1000 MCMC steps with a
## 500 step burnin
props<-estprops(cov1=t(dat[[1]]),cov2=t(dat[[2]]),mcmc.steps=20,mcmc.burnin=5,
    mcmc.thin=1)

## plot point estimates and 95
## allelic proportions for the first nine individuals
par(mfrow=c(3,3))
for(i in 1:9){
    plot(props[[i]][,3],ylim=c(0,1),axes=FALSE,xlab="ratios",ylab="proportions")
axis(1,at=1:5,c("1:3","1:2","1:1","2:1","3:1"))
axis(2)
box()
segments(1:5,props[[i]][,1],1:5,props[[i]][,5])
title(main=paste("true ploidy =",dat[[3]][i]))
}

## End(Not run)
```

Index