Package 'sva'

October 9, 2015

Title Surrogate Variable Analysis

Version 3.14.0

Author Jeffrey T. Leek <jtleek@gmail.com>, W. Evan Johnson <wej@bu.edu>, Hilary S. Parker <hiparker@jhsph.edu>, Elana J. Fertig <ejfertig@jhmi.edu>, Andrew E. Jaffe <ajaffe@jhsph.edu>, John D. Storey <jstorey@princeton.edu>

Description The sva package contains functions for removing batch effects and other unwanted variation in high-throughput experiment. Specifically, the sva package contains functions for the identifying and building surrogate variables for high-dimensional data sets. Surrogate variables are covariates constructed directly from high-dimensional data (like gene expression/RNA sequencing/methylation/brain imaging data) that can be used in subsequent analyses to adjust for unknown, unmodeled, or latent sources of noise. The sva package can be used to remove artifacts in three ways: (1) identifying and estimating surrogate variables for unknown sources of variation in high-throughput experiments (Leek and Storey 2007 PLoS Genetics,2008 PNAS), (2) directly removing known batch effects using ComBat (Johnson et al. 2007 Biostatistics) and (3) removing batch effects with known control probes (Leek 2014 biorXiv). Removing batch effects and using surrogate variables in differential expression analysis have been shown to reduce dependence, stabilize error rate estimates, and improve reproducibility, see (Leek and Storey 2007 PLoS Genetics, 2008 PNAS or Leek et al. 2011 Nat. Reviews Genetics).

Maintainer Jeffrey T. Leek <jtleek@gmail.com>, John D. Storey <jstorey@princeton.edu>, W. Evan Johnson <wej@bu.edu>

Depends R (>= 2.8), mgcv, genefilter

Suggests limma, pamr, bladderbatch, BiocStyle, zebrafishRNASeq, testthat

License Artistic-2.0

biocViews Microarray, StatisticalMethod, Preprocessing, MultipleComparison, Sequencing, RNASeq, BatchEffect, Normalization

NeedsCompilation yes

R topics documented:

ComBat 2
empirical.controls 3
f.pvalue 3
fstats 4
fsva 4
irwsva.build 5
num.sv 6
psva 7
ssva 7
sva 8
sva.check 10
svaseq 10
twostepsva.build 11

Index **13**

`ComBat` *Adjust for batch effects using an empirical Bayes framework*

Description

ComBat allows users to adjust for batch effects in datasets where the batch covariate is known, using methodology described in Johnson et al. 2007. It uses either parametric or non-parametric empirical Bayes frameworks for adjusting data for batch effects. Users are returned an expression matrix that has been corrected for batch effects. The input data are assumed to be cleaned and normalized before batch effect removal.

Usage

```
ComBat(dat, batch, mod=NULL, par.prior = TRUE,
  prior.plots = FALSE)
```

Arguments

`dat`	Genomic measure matrix (dimensions probe x sample) - for example, expression matrix
`batch`	Batch covariate (multiple batches are not allowed)
`mod`	Model matrix for outcome of interest and other covariates besides batch
`par.prior`	(Optional) TRUE indicates parametric adjustments will be used, FALSE indicates non-parametric adjustments will be used
`prior.plots`	(Optional)TRUE give prior plots with black as a kernel estimate of the empirical batch effect density and red as the parametric

Value

data A probe x sample genomic measure matrix, adjusted for batch effects.

`empirical.controls` *A function for estimating the probability that each gene is an empirical control*

Description

This function uses the iteratively reweighted surrogate variable analysis approach to estimate the probability that each gene is an empirical control.

Usage

```
empirical.controls(dat, mod, mod0 = NULL, n.sv, B = 5, type = c("norm",
  "counts"))
```

Arguments

`dat`	The transformed data matrix with the variables in rows and samples in columns
`mod`	The model matrix being used to fit the data
`mod0`	The null model being compared when fitting the data
`n.sv`	The number of surogate variables to estimate
`B`	The number of iterations of the irwsva algorithm to perform
`type`	If type is norm then standard irwsva is applied, if type is counts, then the moderated log transform is applied first

Value

pcontrol A vector of probabilites that each gene is a control.

`f.pvalue` *A function for quickly calculating f statistic p-values for use in sva*

Description

This function does simple linear algebra to calculate f-statistics for each row of a data matrix comparing the nested models defined by the design matrices for the alternative (mod) and and null (mod0) cases. The columns of mod0 must be a subset of the columns of mod.

Usage

```
f.pvalue(dat, mod, mod0)
```

Arguments

dat	The transformed data matrix with the variables in rows and samples in columns
mod	The model matrix being used to fit the data
mod0	The null model being compared when fitting the data

Value

p A vector of F-statistic p-values one for each row of dat.

---

fstats *A function for quickly calculating f statistics for use in sva*

Description

This function does simple linear algebra to calculate f-statistics for each row of a data matrix comparing the nested models defined by the design matrices for the alternative (mod) and and null (mod0) cases. The columns of mod0 must be a subset of the columns of mod.

Usage

```
fstats(dat, mod, mod0)
```

Arguments

dat	The transformed data matrix with the variables in rows and samples in columns
mod	The model matrix being used to fit the data
mod0	The null model being compared when fitting the data

Value

fstats A vector of F-statistics one for each row of dat.

---

fsva *A function for performing frozen surrogate variable analysis as proposed in Parker, Corrada Bravo and Leek 2013*

Description

This function performs frozen surrogate variable analysis as described in Parker, Corrada Bravo and Leek 2013. The approach uses a training database to create surrogate variables which are then used to remove batch effects both from the training database and a new data set for prediction purposes. For inferential analysis see sva, svaseq, with low level functionality available through irwsva.build and ssva.

Usage

```
fsva(dbdat, mod, sv, newdat = NULL, method = c("fast", "exact"))
```

Arguments

dbdat	A m genes by n arrays matrix of expression data from the database/training data
mod	The model matrix for the terms included in the analysis for the training data
sv	The surrogate variable object created by running sva on dbdat using mod.
newdat	(optional) A set of test samples to be adjusted using the training database
method	If method ="fast" then the SVD is calculated using an online approach, this may introduce slight bias. If method="exact" the exact SVD is calculated, but will be slower

Value

db An adjusted version of the training database where the effect of batch/expression heterogeneity has been removed

new An adjusted version of the new samples, adjusted one at a time using the fsva methodology.

newsv Surrogate variables for the new samples

---

irwsva.build *A function for estimating surrogate variables by estimating empirical control probes*

Description

This function is the implementation of the iteratively re-weighted least squares approach for estimating surrogate variables. As a buy product, this function produces estimates of the probability of being an empirical control. See the function empirical.controls for a direct estimate of the empirical controls.

Usage

```
irwsva.build(dat, mod, mod0 = NULL, n.sv, B = 5)
```

Arguments

dat	The transformed data matrix with the variables in rows and samples in columns
mod	The model matrix being used to fit the data
mod0	The null model being compared when fitting the data
n.sv	The number of surogate variables to estimate
B	The number of iterations of the irwsva algorithm to perform

Value

sv The estimated surrogate variables, one in each column

pprob.gam: A vector of the posterior probabilities each gene is affected by heterogeneity

pprob.b A vector of the posterior probabilities each gene is affected by mod

n.sv The number of significant surrogate variables

`num.sv` *A function for calculating the number of surrogate variables to estimate in a model*

Description

This function estimates the number of surrogate variables that should be included in a differential expression model. The default approach is based on a permutation procedure originally prooposed by Buja and Eyuboglu 1992. The function also provides an interface to the asymptotic approach proposed by Leek 2011 Biometrics.

Usage

```
num.sv(dat, mod, method = c("be", "leek"), vfilter = NULL, B = 20,
  seed = NULL)
```

Arguments

`dat`	The transformed data matrix with the variables in rows and samples in columns
`mod`	The model matrix being used to fit the data
`method`	One of "be" or "leek" as described in the details section
`vfilter`	You may choose to filter to the vfilter most variable rows before performing the analysis
`B`	The number of permutaitons to use if method = "be"
`seed`	Set a seed when using the permutation approach

Value

n.sv The number of surrogate variables to use in the sva software

psva *A function for estimating surrogate variables with the two step approach of Leek and Storey 2007*

Description

This function is the implementation of the two step approach for estimating surrogate variables proposed by Leek and Storey 2007 PLoS Genetics. This function is primarily included for backwards compatibility. Newer versions of the sva algorithm are available through sva, svaseq, with low level functionality available through irwsva.build and ssva.

Usage

```
psva(dat, batch, ...)
```

Arguments

dat	The transformed data matrix with the variables in rows and samples in columns
batch	A factor variable giving the known batch levels
...	Other arguments to the sva function.

Value

psva.D Data with batch effect removed but biological heterogeneity preserved

Author(s)

Elana J. Fertig

ssva *A function for estimating surrogate variables using a supervised approach*

Description

This function implements a supervised surrogate variable analysis approach where genes/probes known to be affected by artifacts but not by the biological variables of interest are assumed to be known in advance. This supervised sva approach can be called through the sva and svaseq functions by specifying controls.

Usage

```
ssva(dat, controls, n.sv)
```

Arguments

dat	The transformed data matrix with the variables in rows and samples in columns
controls	A vector of probabilities (between 0 and 1, inclusive) that each gene is a control. A value of 1 means the gene is certainly a control and a value of 0 means the gene is certainly not a control.
n.sv	The number of surogate variables to estimate

Value

sv The estimated surrogate variables, one in each column

pprob.gam: A vector of the posterior probabilities each gene is affected by heterogeneity (exactly equal to controls for ssva)

pprob.b A vector of the posterior probabilities each gene is affected by mod (always null for ssva)

n.sv The number of significant surrogate variables

---

sva *sva: a package for removing artifacts from microarray and sequencing data*

Description

sva has functionality to estimate and remove artifacts from high dimensional data the sva function can be used to estimate artifacts from microarray data the svaseq function can be used to estimate artifacts from count-based RNA-sequencing (and other sequencing) data. The ComBat function can be used to remove known batch effecs from microarray data. The fsva function can be used to remove batch effects for prediction problems.

This function is the implementation of the iteratively re-weighted least squares approach for estimating surrogate variables. As a by product, this function produces estimates of the probability of being an empirical control. See the function empirical.controls for a direct estimate of the empirical controls.

Usage

```
sva(dat, mod, mod0 = NULL, n.sv = NULL, controls = NULL,
  method = c("irw", "two-step", "supervised"), vfilter = NULL, B = 5,
  numSVmethod = "be")
```

Arguments

dat	The transformed data matrix with the variables in rows and samples in columns
mod	The model matrix being used to fit the data
mod0	The null model being compared when fitting the data
n.sv	The number of surogate variables to estimate

controls	A vector of probabilities (between 0 and 1, inclusive) that each gene is a control. A value of 1 means the gene is certainly a control and a value of 0 means the gene is certainly not a control.
method	For empirical estimation of control probes use "irw". If control probes are known use "supervised"
vfilter	You may choose to filter to the vfilter most variable rows before performing the analysis. vfilter must be NULL if method is "supervised"
B	The number of iterations of the irwsva algorithm to perform
numSVmethod	If n.sv is NULL, sva will attempt to estimate the number of needed surrogate variables. This should not be adapted by the user unless they are an expert.

Details

A vignette is available by typing `browseVignettes("sva")` in the R prompt.

Value

sv The estimated surrogate variables, one in each column

pprob.gam: A vector of the posterior probabilities each gene is affected by heterogeneity

pprob.b A vector of the posterior probabilities each gene is affected by mod

n.sv The number of significant surrogate variables

Author(s)

Jeffrey T. Leek, W. Evan Johnson, Hilary S. Parker, Andrew E. Jaffe, John D. Storey

References

For the package: Leek JT, Johnson WE, Parker HS, Jaffe AE, and Storey JD. (2012) The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics DOI:10.1093/bioinformatics/bts034

For sva: Leek JT and Storey JD. (2008) A general framework for multiple testing dependence. Proceedings of the National Academy of Sciences , 105: 18718-18723.

For sva: Leek JT and Storey JD. (2007) Capturing heterogeneity in gene expression studies by 'Surrogate Variable Analysis'. PLoS Genetics, 3: e161.

For Combat: Johnson WE, Li C, Rabinovic A (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics, 8 (1), 118-127

For svaseq: Leek JT (2014) svaseq: removing batch and other artifacts from count-based sequencing data. bioRxiv doi: TBD

For fsva: Parker HS, Bravo HC, Leek JT (2013) Removing batch effects for prediction problems with frozen surrogate variable analysis arXiv:1301.3947

For psva: Parker HS, Leek JT, Favorov AV, Considine M, Xia X, Chavan S, Chung CH, Fertig EJ (2014) Preserving biological heterogeneity with a permuted surrogate variable analysis for genomics batch correction Bioinformatics doi: 10.1093/bioinformatics/btu375

sva.check — *A function for post-hoc checking of an sva object to check for degenerate cases.*

Description

This function is designed to check for degenerate cases in the sva fit and fix the sva object where possible.

Usage

```
sva.check(svaobj, dat, mod, mod0)
```

Arguments

svaobj	The transformed data matrix with the variables in rows and samples in columns
dat	The data set that was used to build the surrogate variables
mod	The model matrix being used to fit the data
mod0	The null model matrix being used to fit the data

Details

empirical.controls for a direct estimate of the empirical controls.

Value

sv The estimated surrogate variables, one in each column

pprob.gam: A vector of the posterior probabilities each gene is affected by heterogeneity

pprob.b A vector of the posterior probabilities each gene is affected by mod

n.sv The number of significant surrogate variables

svaseq — *A function for estimating surrogate variables for count based RNA-seq data.*

Description

This function is the implementation of the iteratively re-weighted least squares approach for estimating surrogate variables. As a by product, this function produces estimates of the probability of being an empirical control. This function first applies a moderated log transform as described in Leek 2014 before calculating the surrogate variables. See the function empirical.controls for a direct estimate of the empirical controls.

Usage

```
svaseq(dat, mod, mod0 = NULL, n.sv = NULL, controls = NULL,
  method = c("irw", "two-step", "supervised"), vfilter = NULL, B = 5,
  numSVmethod = "be", constant = 1)
```

Arguments

dat	The transformed data matrix with the variables in rows and samples in columns
mod	The model matrix being used to fit the data
mod0	The null model being compared when fitting the data
n.sv	The number of surogate variables to estimate
controls	A vector of probabilities (between 0 and 1, inclusive) that each gene is a control. A value of 1 means the gene is certainly a control and a value of 0 means the gene is certainly not a control.
method	For empirical estimation of control probes use "irw". If control probes are known use "supervised"
vfilter	You may choose to filter to the vfilter most variable rows before performing the analysis. vfilter must be NULL if method is "supervised"
B	The number of iterations of the irwsva algorithm to perform
numSVmethod	If n.sv is NULL, sva will attempt to estimate the number of needed surrogate variables. This should not be adapted by the user unless they are an expert.
constant	The function takes log(dat + constant) before performing sva. By default constant = 1, all values of dat + constant should be positive.

Value

sv The estimated surrogate variables, one in each column

pprob.gam: A vector of the posterior probabilities each gene is affected by heterogeneity

pprob.b A vector of the posterior probabilities each gene is affected by mod

n.sv The number of significant surrogate variables

`twostepsva.build` *A function for estimating surrogate variables with the two step approach of Leek and Storey 2007*

Description

This function is the implementation of the two step approach for estimating surrogate variables proposed by Leek and Storey 2007 PLoS Genetics. This function is primarily included for backwards compatibility. Newer versions of the sva algorithm are available through `sva`, `svaseq`, with low level functionality available through `irwsva.build` and `ssva`.

Usage

```
twostepsva.build(dat, mod, n.sv)
```

Arguments

dat	The transformed data matrix with the variables in rows and samples in columns
mod	The model matrix being used to fit the data
n.sv	The number of surogate variables to estimate

Value

sv The estimated surrogate variables, one in each column

pprob.gam: A vector of the posterior probabilities each gene is affected by heterogeneity

pprob.b A vector of the posterior probabilities each gene is affected by mod (this is always null for the two-step approach)

n.sv The number of significant surrogate variables

Index

ComBat, 2, *8*

empirical.controls, 3, *5*, *8*, *10*

f.pvalue, 3
fstats, 4
fsva, 4, *8*

irwsva.build, *4*, 5, *7*, *11*

num.sv, 6

psva, 7

ssva, *4*, *7*, 7, *11*
sva, *4*, *7*, *8*, 8, *11*
sva-package *(*sva*)*, 8
sva.check, 10
svaseq, *4*, *7*, *8*, 10, *11*

twostepsva.build, 11