Package 'lymphclon'

April 25, 2013

Version 0.5-1

Date 2013-04-01

Title Estimating the Clonality Score, a Measure of Conincidences with Replicated Abundances, with Applications in the Study of Lymphocytes

Author Yi Liu, Richard Olshen, Andrew Fire, Scott Boyd

Maintainer Yi Liu <liuyipei@stanford.edu>

Depends R (>= 2.12.0), VGAM, MASS, expm, Matrix

Imports VGAM, MASS, expm, Matrix

Suggests VGAM, MASS, expm, Matrix

Description We provide a clonality score estimator that takes full advantage of the multi-biological-replicate structure of modern sequencing experiments; it specifically takes into account the reality that, typically, the clonal coverage is well below 0.1%.

License LGPL-2

URL http://www.r-project.org, http://www.another.url

BugReports http://pkgname.bugtracker.url

R topics documented:

`lymphclon-package` *Estimates the clonality score from replicate of abundances data*

Description

There are an enormous number of clones(species) distributed in a highly unequal distribution. The experimenters collect a small number of very coarse samplings, and perform noisy measurements on the samplings, leading to abundance estimations for the indiviudal biological replicates. The goal of this package is to estimate the probability that two random cells belong to the same clone.

It provides two functions; one computes the estimate, given replicate of abundances; the other function generates reasonable simulation data (for B or T cell lymphocyte setting), for evaluation purposes.

This package was written for clonality quantification in the lymphocyte repertoires, as measured by blood-draw and PCR/sequencing; however it can be directly reused for broader ecological applications.

Details

Package:	lymphclon
Type:	Package
Version:	1.0
Date:	2013-04-24
License:	LGPL-2

~~ An overview of how to use the package, including the most important ~~ ~~ functions ~~ This package provides two functions; one computes the estimate, given replicate of abundances; the other function generates reasonable simulation data (for B or T cell lymphocyte setting), for evaluation purposes. infer.clonality: infers the clonality score, given replicate of abundances data simulate.clonality.data: generates simulated data, primarily for evaluation purposes

Author(s)

Author and Maintainer: Yi Liu <liuyipei@stanford.edu>

References

~~ Literature or other references for background information ~~

See Also

<vegan>

Examples

```
my.data<-simulate.clonality.data(n=2e3)
```

```
# n ~ 2e7 is more appropriate for a realistic B cell repertoire
infer.clonality(my.data$read.count.matrix)
```

`infer.clonality` *infer.clonality (part of lymphclon package)*

Description

Clonality score, a useful metric used in immunology, refers to the probability that two random lymphocyte receptor chain reads drawn with replacement (makes no difference in the immunology context) from an individual corresponded to the same clone, within some given repertoire of either B cells or T cells. This package implements ane estimator which understands the multi-replicate-with-PCR structure of thes sequencing experiments.

Usage

```
infer.clonality(
  read.count.matrix,
  regularization.clones = matrix(0, 0, ncol(read.count.matrix)))
```

Arguments

`read.count.matrix`
: A matrix of read frequences, where each row corresponds to a distinct clone, and each column corresponds to a particular biological (rather than technical) replicate. All biological replicates should be drawn from the same person. Reads from technical replicates of the same underlying templates should be merged into a single column.

`regularization.clones`
: A matrix of regularzation (i.e. "virtual") data to be appended to the bottom of the read.count.matrix. For example, this could be the vertical concatentation of an indentiy matrix the size of the number of replicates, with a row of 1s. Deafault value is no regularization (i.e. empty matrix). This is recommended if the data sampling is exceptionally sparse.

Value

`simple.precision.clonality`
: This is a base line estimator of the clonality score based on simple modeling. Briefly, for each pairing of reads possible, where the two reads arise from different biological replicates – this pairing is considered as an observation from a single share undelrying bernoulli distribution with parameter equal to the clonality score. This estimator computes maximum likelihood estimate of the underlying clonality score. This estimator can be thought of as a baseline estimate. Unlike the Gini-Simpson-Estimator, this baseline does not suffer from any convexity-based bias.

`rao.blackwell.mvg.clonality`
: This is an estimator is a weighted average of n-choose-2 individually unbiased estimators that arise from comparing the provided n biological replicates pairwise. The weighting is determined by the covariance between these estimators. This has much lower variance than the simple.precision.clonality.

Examples

```
my.data<-simulate.clonality.data(n=2e3)
# n ~ 2e7 is more appropriate for a realistic B cell repertoire
infer.clonality(my.data$read.count.matrix)
```

`simulate.clonality.data`
simulate.clonality.data (part of lymphclon package)

Description

This function generates simulated data, for evaluation purposes. We start with an underlying multinomial population with entries proportional to rank^-power distribution, where power is fixed. Next, we draw multinomially from this distribution, a fixed number of cells, to generate each desired replicate. Then, this distribution is subject to log-normal error, and subsequently scaled up to the expected number of reads. To round into integers, the expected number of reads for each clone is finally pushed through a poisson process to generate integer read counts. The poisson "rounding" process is why the resulting read counts are not exactly as specified.

Usage

```
simulate.clonality.data(
  n = 2e+07,
  num.cells.taken.vector = c(2000, 5000, 10000, 20000, 50000, 50000),
  read.count.per.replicate.vector = rep(20000, length(num.cells.taken.vector)),
  clonal.distribution.power = -sqrt(2))
```

Arguments

`n`
: The true number of distinct clones in the underlying assemblage

`num.cells.taken.vector`
: A vector specifying the number of cells taken in each independent biological replicate

`read.count.per.replicate.vector`
: A vector of the same length as num.cells.taken.vector, specifying the number of reads generated from each biological replicate, of the same corresponding indices

`clonal.distribution.power`
: The true underlying clonal multinomial distribution is proportional to (1:n)^-clonal.distribution.power

Value

`read.count.matrix`
: This is a matrix of simulated counts, with rows corresponding to clones (classes, or species), and columns corresponding to biological replicates

`true.clone.prob`
: This is the underlying simulated assemblage multinomial distribution used to generate read.count.matrix

`true.clonality`
: This is the true clonality score of the underlying simulated assemblage

Examples

```
my.data<-simulate.clonality.data(n=2e3)
# n ~ 2e7 is more appropriate for a realistic B cell repertoire
infer.clonality(my.data$read.count.matrix)
```

Index