

# Package ‘lymphclon’

May 31, 2013

**Version** 1.0.1

**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 . . . . .	2
infer.clonality . . . . .	3
simulate.clonality.data . . . . .	4

---

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 individual biological replicates. The goal of this package is to estimate the probability that two random cells belong to the same clone. Clonality estimation arises as a naturally interesting question when trying to understand the diversity of B cell populations, T cell populations, microbial populations, cancer cell subclones, and artificial biological libraries, and also broader ecological settings.

This package provides two primary functions; one computes the clonality score estimate (probability that two random individuals belong to the same clone), given replicate of abundances; it also infers the underlying clonal distribution, which may be of interest in diagnostic and scientific settings.

The other function generates reasonable simulation data, for evaluation purposes. The default behavior of this function generates a very large number of classes, which is suitable for B and T cell settings. It can be easily repurposed for other applications. The simulation assumes random sampling, amplifying, and sequencing of individual cells; this is implemented by introducing sparse sampling of individual cells, followed by introduction of log-normal error, and finally unbiased "rounding" by use of a poisson distribution.

## Details

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

## 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 an estimator which understands the multi-replicate-with-PCR structure of these sequencing experiments.

## Usage

```
infer.clonality(
  read.count.matrix,
  regularization.clones = matrix(0, 0, ncol(read.count.matrix)),
  estimate.abundances = F,
  loo.squared.err.est = T,
  use.squared.err.est = c())
```

## Arguments

`read.count.matrix`

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

`estimate.abundances`

A boolean value defaulting to false. If set to true, then the return value will be in the form of a list, and include an additional item named "estimated.abundances". The estimated abundances are computed as the (conditional) precision weighted average of the `read.count.matrix`.

`loo.squared.err.est`

A boolean value defaulting to true. When estimating the squared errors and precisions associated with each replicate, we can use the maximum likelihood estimate (i.e. set `loo.squared.err.est` to false), or use an averaged leave-one-out estimate (i.e. set `loo.squared.err.est` to true). The leave-one-out estimate is slightly biased, and slightly slower to compute, but much less variance. Especially if there are at least three replicates, this value probably should be set to true.

`use.squared.err.est`

A vector defaulting to the empty vector. If a non-empty vector is provided, then this forces `infer.clonality` to use these values as the conditional variances of the respective replicates. If non-empty, then this vector should be positive, and of length equal to the number of replicates.

## 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 underlying 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  $\text{rank}^{\text{power}}$  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

<code>n</code>	The true number of distinct clones in the underlying assemblage
<code>num.cells.taken.vector</code>	A vector specifying the number of cells taken in each independent biological replicate
<code>read.count.per.replicate.vector</code>	A vector of the same length as <code>num.cells.taken.vector</code> , specifying the number of reads generated from each biological replicate, of the same corresponding indices
<code>clonal.distribution.power</code>	The true underlying clonal multinomial distribution is proportional to $(1:n)^{-\text{clonal.distribution.power}}$

### Value

<code>read.count.matrix</code>	This is a matrix of simulated counts, with rows corresponding to clones (classes, or species), and columns corresponding to biological replicates
<code>true.clone.prob</code>	This is the underlying simulated assemblage multinomial distribution used to generate <code>read.count.matrix</code>
<code>true.clonality</code>	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)
```