RRHO package

Jonathan Rosenblatt
Dept. of Mathematics and Computer Science,
The Weizmann Institute of Science
Israel

October 17, 2016

1 Introduction

Consider the problem of comparing the degree and significance of overlap between two lists of same length. In the following, we will assume that one list consists of differential expression statistics between two conditions for each gene. A second list consists of differential expression statistics between two different conditions for the same genes. A possible way of comparison would be to test the hypothesis that the ordering of lists by their differential expression statistics is arbitrary. The test can be performed against a one sided hypothesis (an over-enrichment hypothesis), or a two sided hypothesis (looking for under- or over-enrichment). This is the purpose of this package, based on the work of Plaisier et al. [2010].

The proposed approach is to count the number of common genes in the first $i \times stepsize$ and $j \times stepsize$ elements of the first and second list respectively, where *stepsize* is an arbitrary user inputted number. As the count of common elements could be driven by chance, the significance of the observed count is computed assuming the hypothesis of completely random list orderings. As this is performed for all $i \times stepsize$ and $j \times stepsize$, correction for multiple comparisons is necessary.

The package offers both FWER control[1] using permutation testing and FDR control using the B-Y procedure [Benjamini and Yekutieli, 2001] as proposed in the original work by Plaisier et al. [2010].

Remark 1.1. *FDR or FWER?*
For brevity, i and j will denote $i \times stepsize$ and $j \times stepsize$ respectively.

[1] For a general introduction to multiple testing error rates, see [Rosenblatt, 2013].

Plaisier et al. [2010] recommend the control of the FDR over the different is and js. Each i, j combination tests the null hypothesis of "arbitrary rankings of the two lists", versus an alternative of "non-arbitrary ranking in the first i and j elements of the first and second list respectively". FDR control is thus appropriate if concerned with the number of false i, j statements made.

If only concerned with the existence of any non-arbitrariness, without claiming at which part of the lists it resides, than FWER control is more appropriate.

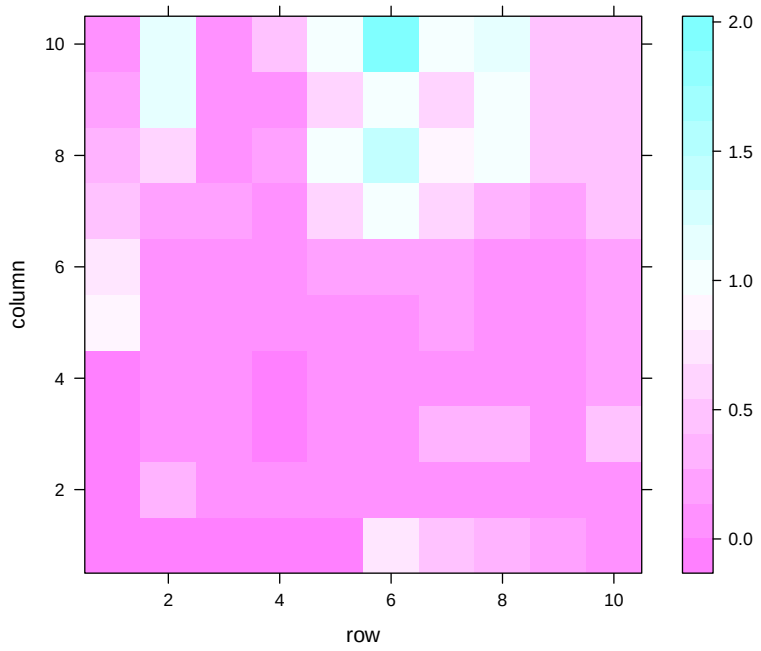
2 Comparing Two Lists

We start with a sketch of the workflow. The details follow.

- Compute the marginal significance of the gene overlap for all i and j first elements of the two lists.
- Correct the marginal significance levels for the multiple is and js.
- Report findings using the exported significance matrices and accompanying Venn diagrams.

```
> library(RRHO)
> # Create "gene" lists:
>   list.length <- 100
>         list.names <- paste('Gene',1:list.length, sep='')
>         gene.list1<- data.frame(list.names, sample(100))
>         gene.list2<- data.frame(list.names, sample(100))
> # Compute overlap and significance
>         RRHO.example <-  RRHO(gene.list1, gene.list2,
+                         BY=TRUE, alternative='enrichment')

> # Examine Nominal (-log) pvalues
>   lattice::levelplot(RRHO.example$hypermat)
> # Note: If lattice is available try:
> # levelplot(RRHO.example$hypermat)
```

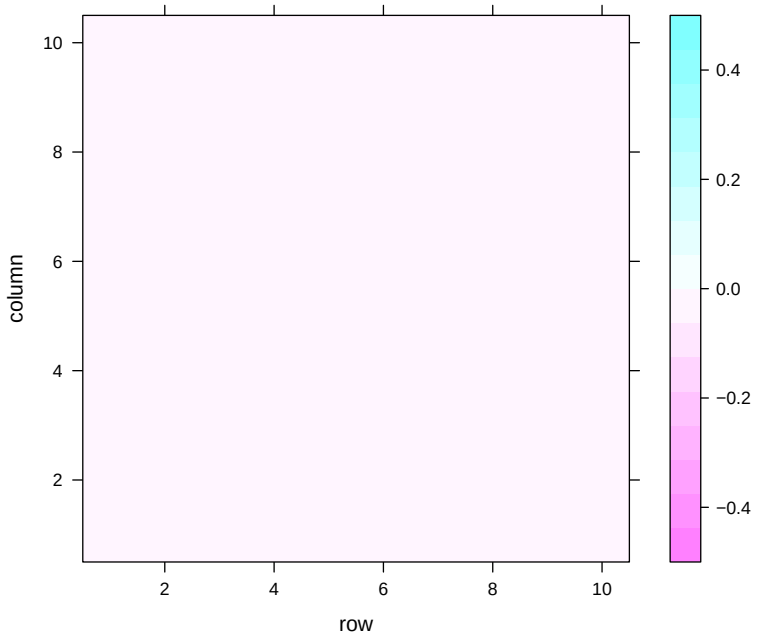



```
> # FWER corrected pvalues using 50 random permutations:
> pval.testing <- pvalRRHO(RRHO.example, 50)
> pval.testing$pval

[1] 0.92

> # The sampling distribution of the minimum
> # of the (-log) nominal p-values:
> xs<- seq(0, 10, length=100)
> plot(Vectorize(pval.testing$FUN.ecdf)(xs)~xs,
+      xlab='-log(pvalue)', ylab='ECDF', type='S')

> # Examine B-Y corrected pvalues
> # Note: probably nothing will be rejected in this
> # example as the data is generated from the null.
>   lattice::levelplot(RRHO.example$hypermat.by)
```



Remark 2.1. *As of version* 1.4.0 *a two-sided hypothesis test is now possible. The computation of the p-values differ from that described in Plaisier et al. [2010]. The algorithm propsed in [Plaisier et al., 2010, Section Hypergeometric probability distributions] does not control the type I error as demonstrated in the following simulation:*

```
> m<- 100 ; n<- 100; k<- 50
> data<- rhyper(1000, m, n, k)
> pvals<- pmin(phyper(data,m,n,k, lower.tail=TRUE),
+              phyper(data,m,n,k, lower.tail=FALSE))
> alpha<- 0.05
> prop.table(table(pvals<alpha))

FALSE  TRUE
0.894 0.106
```

We thus replace the proposed algorithm by the simple summation of the two tails of the distribution:

```
> getPval<- function(count,m,n,k){
+   the.mean<- k*m/(m+n)
+   if(count<the.mean){
+     lower<- count
+     upper<- 2*the.mean-count
+   } else{
+     lower<- 2*the.mean-count
+     upper<- count
+   }
+   phyper(q=lower, m=m, n=n, k=k, lower.tail=TRUE) +
+     phyper(q= upper, m=m, n=n, k=k, lower.tail=FALSE)
+ }
> pvals<- sapply(data, getPval, m,n,k)
> prop.table(table(pvals<alpha))

FALSE  TRUE
0.976 0.024
```

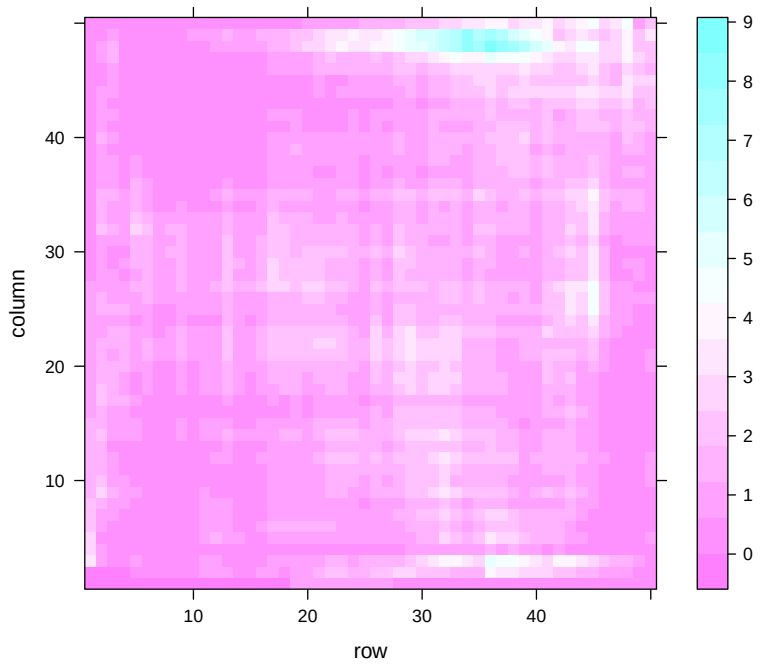
3 Comparing Three Lists

As of version 1.4.0, a comparison of three lists is possible as described by JL Stein et al. [2014]. This comprison tests whether the difference between lists 1 and 3 is different than the differences between 2 and 3. Rejecting this hypothesis implies that that the difference between 1 and 2 are non arbitrary.

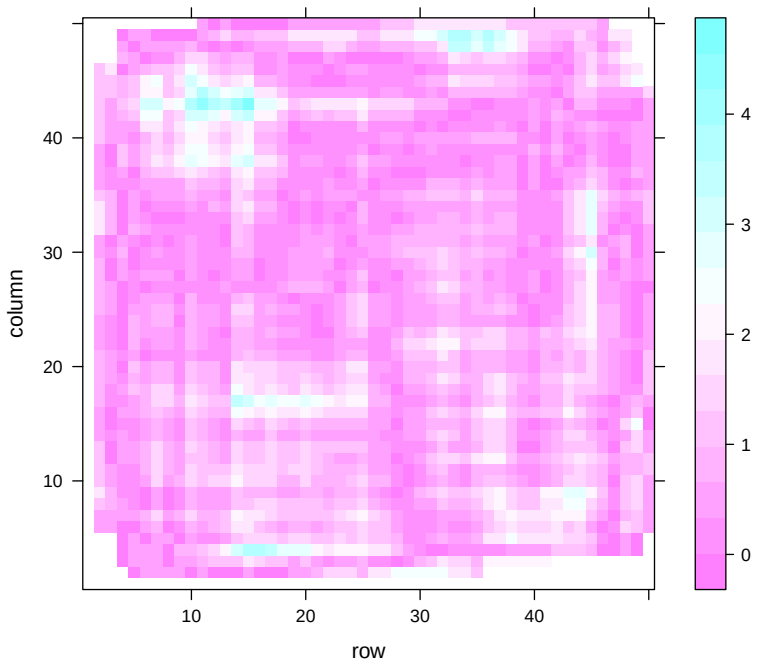
```
> size<- 500
> list1<- data.frame(
+   GeneIdentifier=paste('gen',1:size, sep=''),
+   RankingVal=-log(runif(size)))
> list2<- data.frame(
+   GeneIdentifier=paste('gen',1:size, sep=''),
+   RankingVal=-log(runif(size)))
> list3<- data.frame(
+   GeneIdentifier=paste('gen',1:size, sep=''),
+   RankingVal=-log(runif(size)))
> rrho.comparison<- RRHOComparison(list1,list2,list3,
+                 stepsize=10,
+                 labels=c("list1",
+                          "list2",
```

```
+                             "list3"),
+                   plots=FALSE,
+                   outputdir=temp.dir);

> ## The standard RRHO map between list1 and list 3.
> lattice::levelplot(rrho.comparison$hypermat1)
```




```
> ## The p-value of the difference between
> # (list1-list3)-(list2-list3).
> lattice::levelplot(rrho.comparison$Pdiff)
```



References

Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *ANNALS OF STATISTICS*, 29(4): 1165–1188, 2001.

JL Stein, L de la Torre Ubieta, Y Tian, NN Parikshak, IA Hernandez, MC Marchetto, DK Baker, D Lu, JK Lowe, EM Wexler, AR Muotri, FH Gage, KS Kosik, and DH Geschwind. A quantitative framework to evaluate modeling of cortical development by neural stem cells. *Neuron (in press)*, 2014.

Seema B. Plaisier, Richard Taschereau, Justin A. Wong, and Thomas G. Graeber. Rank–rank hypergeometric overlap: identification of statistically significant overlap between gene-expression signatures. *Nucleic Acids Research*, 38(17):e169–e169, September 2010. ISSN 0305-1048, 1362-4962.

Jonathan Rosenblatt. A practitioner's guide to multiple test-

ing error rates. arXiv e-print 1304.4920, April 2013. URL http://arxiv.org/abs/1304.4920.