OGSA Users Manual

Michael Ochs
email: ochsm@tcnj.edu

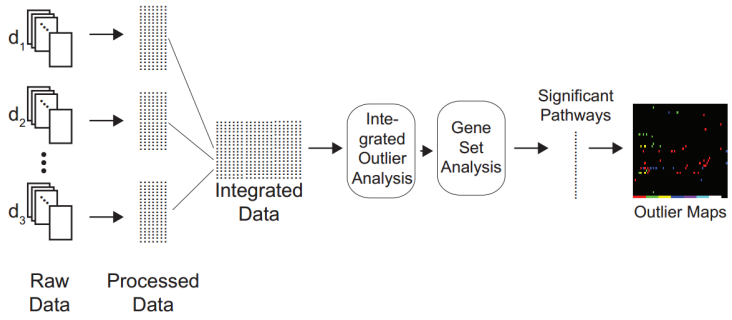
October 30, 2017

Outlier Gene Set Analysis (OGSA) provides a global estimate of pathway deregulation in cancer subtypes by integrating the estimates of significance for individual pathway members that have been identified by outlier analysis. OGSA integrates data from different molecular domains using different molecular data types simultaneously. Two methods for performing outlier analysis are included in this package: the method of Tibshirani and Hastie and a rank sum outlier approach, modified from Ghosh. The latter method sets minimum change levels for the calling of an outlier, effectively eliminating many outliers where the change was not biologically meaningful.

In this example we apply the outlier analysis to a set of measurements of expression, promoter methylation, and copy number variation. The data is processed prior to application of OGSA by any method to produce gene level summaries of the data.

Basic Concept

The figure below illustrates how the analysis works. Different molecular data measurements, d_i, are preprocessed into gene level summaries. Analysis then follows by determining outlier counts for each data type and each gene, which are then used in rank-based statistical tests on pathways of interest. Pathways can then be refined as desired (e.g., removing pathways that are not of interest either due to lack of features of interest such as druggable targets or due to being well-known already).



Example

1. Load and format the data and set initial settings. Three data sets are provided in this example. Here, they're called 'expr', 'cnv', and 'meth', for the gene expression, copy number variation, and promoter methylation measurements. Each data set is provided as a matrix with the genes in the rows and different samples in different columns. The 'pheno' in the code below is a vector of numbers with as many elements as there are samples (in this example, 69) with "1" indicating a case and "0" indicating a control. In this analysis, the cases are individual tumors samples and the controls normal tissue samples.

   In order to encode reasonable molecular relationships, we look for right-tail outliers for expression and copy number with left-tail outliers in methylation (tailRLR). This provides the proper biological relationships of amplification or loss of promoter methylation leading to increased expression. Likewise we will also look for the reverse with left tail outliers in copy number and expression tied to right tail outliers in methylation (tailLRL).

   When using the rank method to identify outliers, we will also set thresholds on the minimum acceptable change to quality as an outlier, with a log expression change of 1, a beta methylation level change of 0.1, and a copy number change of 0.5. Only when an outlier exceeds a normal sample by these amounts does it get counted into the statistic.

```
> library('OGSA')
> data('ExampleData')
> data('KEGG_BC_GS')
> phenotype <- pheno
> names(phenotype) <- colnames(cnv)
> tailLRL <- c('left','right','left')
> tailRLR <- c('right','left','right')
> offsets <- c(1.0, 0.1, 0.5)
> dataSet <- list(expr,meth,cnv)
```

```
>
>
```

2. Next, run the analysis using the functions copaInt() and testGScogps() with both tail settings. The function copaInt is the wrapper function for using any of the three outlier methods, while testGScogps generates gene set statistics from the gene lists with outlier counts. The default method in copaInt is the Tibshirani-Hastie method, so run copaInt again with method='Rank'. 'pathGS' is the list of significant pathways. Setting the variable corr to TRUE corrects the count for normal outliers in the Tibshirani-Hastie case and applies the threshold in the rank case.

```
> tibLRL <- copaInt(dataSet,phenotype,tails=tailLRL)
> gsTibLRL <- testGScogps(tibLRL,pathGS)
> tibLRLcorr <- copaInt(dataSet,phenotype,tails=tailLRL,corr=TRUE)
> gsTibLRLcorr <- testGScogps(tibLRLcorr,pathGS)
> tibRLR <- copaInt(dataSet,phenotype,tails=tailRLR)
> tibRLRcorr <- copaInt(dataSet,phenotype,tails=tailRLR,corr=TRUE)
> gsTibRLR <- testGScogps(tibRLR,pathGS)
> gsTibRLRcorr <- testGScogps(tibRLRcorr,pathGS)
> rankLRL <- copaInt(dataSet,phenotype,tails=tailLRL,method='Rank')

[1] 2000   69
[1] 2000   69
[1] 2000

> rankLRLcorr <- copaInt(dataSet,phenotype,tails=tailLRL,method='Rank',corr=TRUE,
+                        offsets=offsets)

[1] 2000   69
[1] 2000   69
[1] 2000

> gsRankLRL <- testGScogps(rankLRL,pathGS)
> gsRankLRLcorr <- testGScogps(rankLRLcorr,pathGS)
> rankRLR <- copaInt(dataSet,phenotype,tails=tailRLR,method='Rank')

[1] 2000   69
[1] 2000   69
[1] 2000

> rankRLRcorr <- copaInt(dataSet,phenotype,tails=tailRLR,method='Rank',corr=TRUE,
+                        offsets=offsets)

[1] 2000   69
[1] 2000   69
[1] 2000
```

```
> gsRankRLR <- testGScogps(rankRLR,pathGS)
> gsRankRLRcorr <- testGScogps(rankRLRcorr,pathGS)
```

3. Process and create the outlier maps to indicate where the tumor-specific outliers have been called in pathways of interest. Here, we use KEGG ECM Receptor Interaction and Biocarta PDGFB pathways to demonstrate creating and plotting the outlier maps.

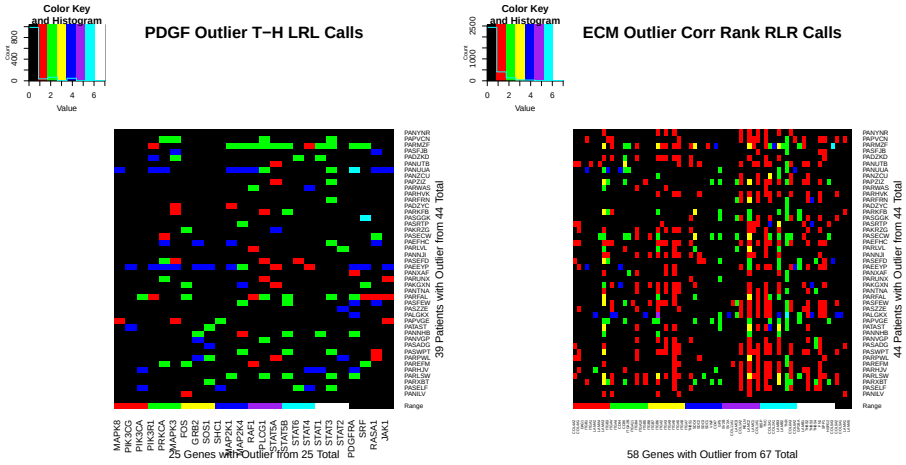
```
> outRankLRLcorr <- outCallRank(dataSet, phenotype, names=c('Expr','Meth','CNV'),tail=t
+                               corr=TRUE,offsets=offsets)
> outRankRLRcorr <- outCallRank(dataSet, phenotype, names=c('Expr','Meth','CNV'),tail=t
+                               corr=TRUE,offsets=offsets)
> print("Corrected Rank Outliers Calculated")

[1] "Corrected Rank Outliers Calculated"

> outTibLRL <- outCallTib(dataSet, phenotype, names=c('Expr','Meth','CNV'),tail=tailLRL
> outTibRLR <- outCallTib(dataSet, phenotype, names=c('Expr','Meth','CNV'),tail=tailRLR
> print("Tibshirani-Hastie Outliers Calculated")

[1] "Tibshirani-Hastie Outliers Calculated"

> pdgfB <- pathGS$'BIOCARTA_PDGF_PATHWAY'
> map1 <- outMap(outTibLRL,pdgfB,hmName='BC_PDGF_TIB.pdf',
+                plotName='PDGF Outlier T-H LRL Calls')
> ecmK <- pathGS$'KEGG_ECM_RECEPTOR_INTERACTION'
> map4 <- outMap(outRankRLRcorr,ecmK,hmName='KEGG_ECM_RANKcorr.pdf',
+                plotName='ECM Outlier Corr Rank RLR Calls')
>
```



References

Ochs, M. F., Farrar, J. E., Considine, M., Wei, Y., Meshinchi, S., & Arceci, R. J. (n.d.). Outlier Analysis and Top Scoring Pair for Integrated Data Analysis and Biomarker Discovery. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 1–1. doi:10.1109/tcbb.2013.153

R. Tibshirani and T. Hastie. (2007) Outlier Sums for Differential Gene Expression Analysis. Biostatistics, 8(1), 2-8.