sSeq: A Simple and Shrinkage Approach of Differential Expression Analysis for RNA-Seq experiments.

Danni Yu, Wolfgang Huber and Olga Vitek

April 16, 2015

The sSeq package introduced in this manual provides a simple and efficient approach to discover differentially expressed (DE) genes based on the counts of transcripts from RNA-seq experiments. It regularizes the per-gene dispersion estimates with the common information across genes so that the bias and the variability in variance estimation are maintained at the low level.

1 Simple comparison between two conditions.

In this section, we will use the Hammer *et. al.* data [1] to illustrate how to use the functions in the sSeq package. The two conditions are control Sprague Dawley after 2 months (Condition A) and L5 SNL Sprague Dawley after 2 months (Condition B). There are two samples within each condition. This data is included in the sSeq package as an example, and can be imported as follows. "countsTable" is a matrix or data frame in which a column represents a sample and a row represents a gene. "conds.Hammer" is a characteristic vector, and used to define the conditions corresponding to the samples in columns. After defining the input counts table and the groups for conditions, the function "nbTestSH" can be utilized to obtain the regularized dispersion estimates and perform the exact tests. The output is a data frame in which the "pval" column includes the p-values of the exact tests.

```
> library(sSeq);
> data(Hammer2months);
> head(countsTable);

                    A 1 A 2 B 1 B 2
ENSRNOG00000000001   2    4   18   24
ENSRNOG00000000007   4    1    3    1
ENSRNOG00000000008   0    1    4    2
ENSRNOG00000000009   0    0    0    0
ENSRNOG00000000010  19   10   19   13
ENSRNOG00000000012   7    5    1    0

> conds.Hammer=c("A","A","B","B");

> #exact test to get differential expressed genes
> res1 = nbTestSH( countsTable, conds.Hammer, "A", "B");

[1] "Shrinkage estimates on dispersion are used for the tests."
[1] "The shrink target is 0.543981062963739"
[1] "The shrink quantile is 0.963"
[1] "30% processed."
[1] "50% processed."
[1] "70% processed."
[1] "90% processed."
[1] "100% processed."
Time difference of 40.50218 secs

> head(res1);
```

```
                        Mean rawMeanA rawMeanB rawLog2FoldChange      dispMM
ENSRNOG00000000001 11.611882      3.0     21.0        -2.8073549 0.655085689
ENSRNOG00000000007  2.207431      2.5      2.0         0.3219281 0.000000000
ENSRNOG00000000008  1.722050      0.5      3.0        -2.5849625 0.329805247
ENSRNOG00000000009  0.000000      0.0      0.0               NaN 0.000000000
ENSRNOG00000000010 15.041505     14.5     16.0        -0.1420190 0.001213686
ENSRNOG00000000012  3.323526      6.0      0.5         3.5849625 0.707503560
                      dispSH       pval
ENSRNOG00000000001 0.6294031 0.01235650
ENSRNOG00000000007 0.1257451 0.76135252
ENSRNOG00000000008 0.3793135 0.18544609
ENSRNOG00000000009 0.1257451 1.00000000
ENSRNOG00000000010 0.1266782 1.00000000
ENSRNOG00000000012 0.6697042 0.01880579
```

1.1 ASD plot and Dispersion plot

In the sSeq package, the testing is based on the shrinkage estimator $\hat{\phi}^{sSeq} = (1 - \delta)\hat{\phi}^{MM} + \delta\xi$ that regularizes the method of moment estimates $\hat{\phi}^{MM}$ to a shrinkage target ξ for the dispersion parameter. The averaged squared difference (ASD) between the method of moment estimates and the shrinkage estimates is used to estimate the shrinkage target. The smallest target value that minimizes the ASD value is selected as the estimate.

If "plotASD=TRUE" is specified in the function "nbTestSH", a plot (Fig.1) of ASD values when varying the shrinkage targets is generated. In Fig.1, the dotted vertical and horizontal lines represent the estimated shrinkage target $\hat{\xi}$ and the corresponding ASD value. The argument "SHonly=TRUE" is used to only calculate the dispersion estimates without running the exact tests.

```
> disp1 <- nbTestSH( countsTable, conds.Hammer, "A", "B", SHonly=TRUE, plotASD=TRUE);

[1] "Shrinkage estimates on dispersion are used for the tests."
[1] "The shrink target is 0.543981062963739"
[1] "The shrink quantile is 0.963"
```

After running the function "nbTestSH" with the argument "SHonly=TRUE", we obtain an object (named as "disp1" in the following R scripts) that includes the dispersion estimates and the mean estimates. Using this object, a scatter plot (Fig.2) visualizing the relationship between the dispersion estimates and the mean estimates can be generated with the function "plot.dispersion".

```
> head(disp1);

         SH         raw       mus
1 0.6294031 0.655085689 11.611882
2 0.1257451 0.000000000  2.207431
3 0.3793135 0.329805247  1.722050
4 0.1257451 0.000000000  0.000000
5 0.1266782 0.001213686 15.041505
6 0.6697042 0.707503560  3.323526

> plotDispersion(disp1, legPos="bottomright")
```

Sometimes, a user may like to define the shrinkage target instead of letting the package automatically find an estimate. The sSeq package is flexible for the requirement. For example, the method of moment estimates will be shrunk toward the target 1 when the argument "shrinkTarget=1" is added in the function "nbTestSH". If the target needs to be defined as a quantile (e.g. 0.975) of the method of moment estimates across genes, then "shrinkQuantile=0.975" should be only added in the function "nbTestSH". When both the arguments are added, the sSeq package uses the pre-defined target value, not the quantile, and shrinks the method of moment estimates toward the target 1.

1.2 Variance plot

To visualize the dependence between the variance estimates and the mean estimates, the following R scripts are used to generate a scatter plot (Fig.3) of log variance estimates versus log mean estimates. The black dots are the
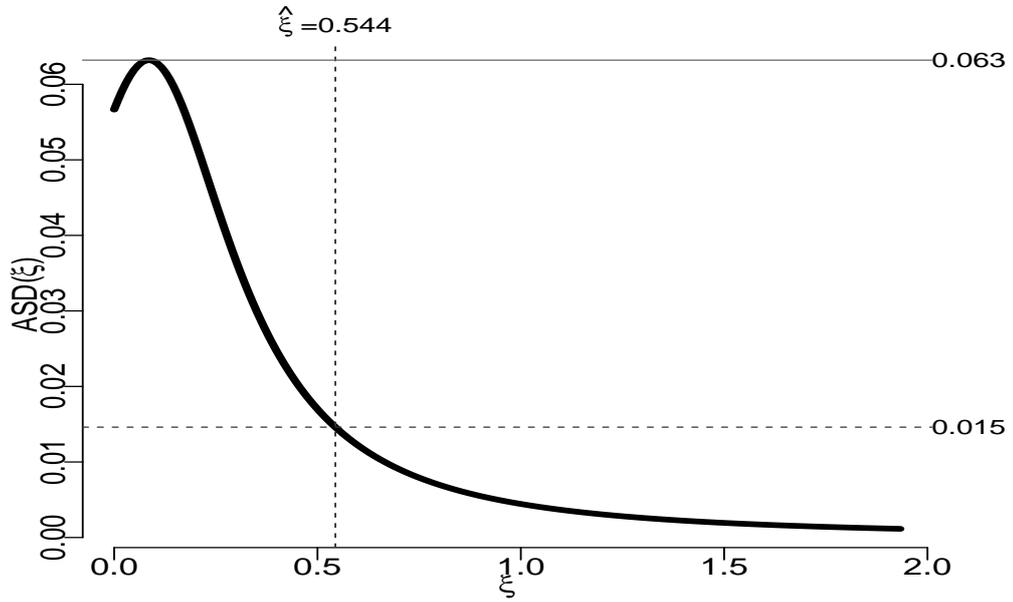
Figure 1: The plot of ASD varying the shrinkage target. The smallest target value that minimizes the ASD value is represented by the vertical dotted line, and the corresponding ASD value is represented by the horizontal dotted line.
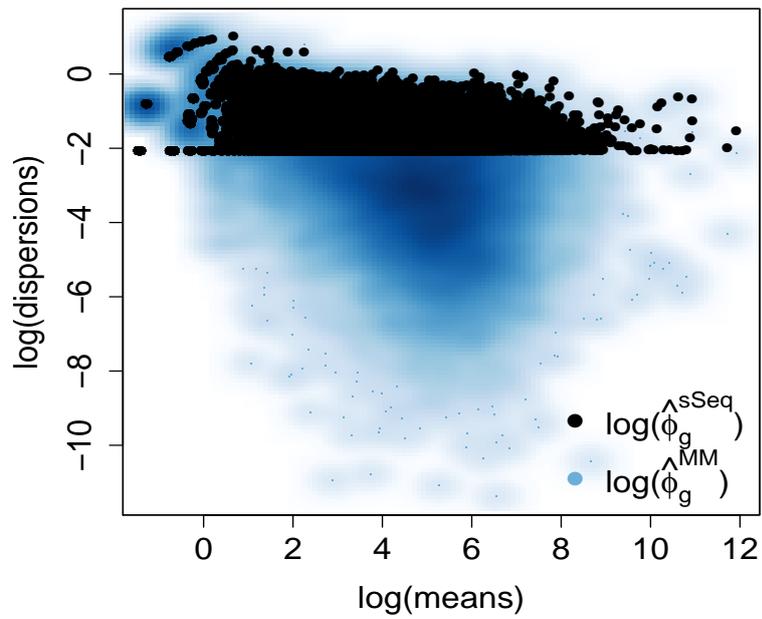

Figure 2: Dispersion plot.

variance estimates based on the shrinkage (or regularized) estimates of the dispersion. The blue smooth dots are the variance estimates directly obtained from the samples without any regularization. The variability among black dots are much lower than the variability among the blue smooth dots. Fig.3 clearly indicates that the mean-variance dependence is improved by the regularized variance estimates.

```
>   rV = rowVars(countsTable);
>   mu = rowMeans(countsTable);
>   SH.var =  (disp1$SH * mu^2 + mu)
>   smoothScatter(log(rV)~log(mu), main="Variance Plot", ylab='log(variance)',
+       xlab='log(mean)', col=blues9[5], cex.axis=1.8)
>   points( log(SH.var)~log(mu), col="black", pch=16)
>   leg1 =  expression(paste("log(", hat("V")[g]^"sSeq", ")", sep=''));
>   leg2 =  expression(paste("log(", hat("V")[g]^"MM", ")", sep=''));
>   legend("bottomright", legend=c(leg1,leg2), col=c("black",blues9[5]),
+       pch=c(16, 1), cex=2)
```
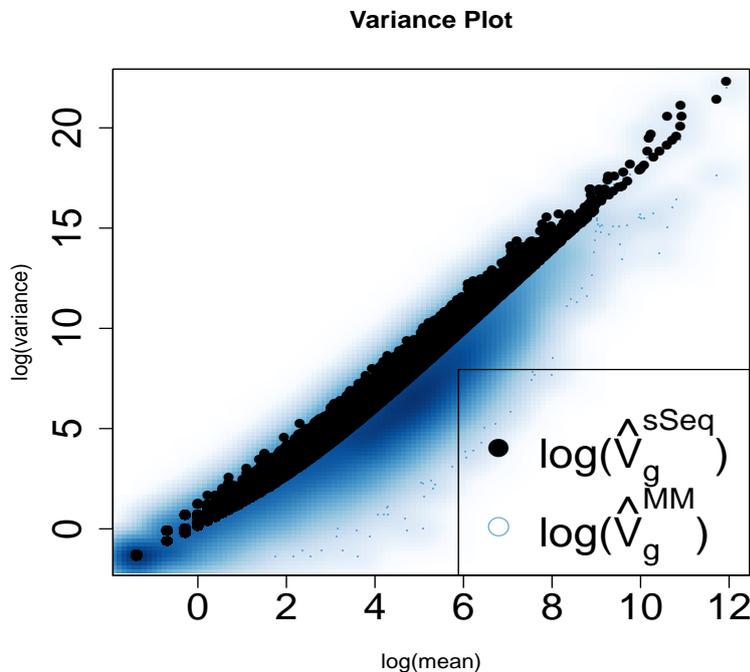

Figure 3: The plot of the variance estimates and the mean estimates.

1.3   ECDF plot

The empirical cumulative distribution function (ECDF) is an estimator of the true cumulative distribution function (CDF). It asymptotically converges to the true CDF for large number of points. In RNA-seq experiments, we typically have more than 20,000 p-values, and thus the ECDF of the p-values are very close to the true CDF.

The specificity and the sensitivity can be visualized by drawing the ECDF curves of the p-values for the within-condition comparison and the p-values for the between-condition comparison. When comparing the replicates under the same condition for the specificity, we expect to see that the genes are differentially expressed only by chance. The p-values should follow a uniform distribution (equivalent to the 45 degree line), or most p-values should be large and close to 1. On the other hand, when comparing the samples under two different conditions for the sensitivity, we expect to see that many genes are differentially expressed due to the changes of environment. The p-values should be small and close to 0. When a statistical method is robust for testing, we expect to see that the ECDF curve for the between-condition comparison is toward to the top left corner, and that the ECDF curve for the within-condition

comparison is toward to the 45 degree line or the bottom right corner. An example of this ECDF plot is shown in Fig.4. "AvsA" is for the within-condition comparison and "AvsB" is for the between-condition comparison.

```
> #obtain the p-values for the comparison AvsA.
> conds2.Hammer = c("A","B");
> res1.2 =  nbTestSH( countsTable[,1:2], conds2.Hammer, "A", "B");

[1] "Shrinkage estimates on dispersion are used for the tests."
[1] "The shrink target is 0.249603542694111"
[1] "The shrink quantile is 0.972"
[1] "30% processed."
[1] "50% processed."
[1] "70% processed."
[1] "90% processed."
[1] "100% processed."
Time difference of 22.44194 secs

> #draw the ECDF plot;
> dd = data.frame(AvsA=res1.2$pval, AvsB=res1$pval);
> ecdfAUC(dd, col.line=c("green", "red"), main = "ECDF, Hammer", drawRef = TRUE, rm1=TRUE)

     AvsA      AvsB
0.4202079 0.6283182
```
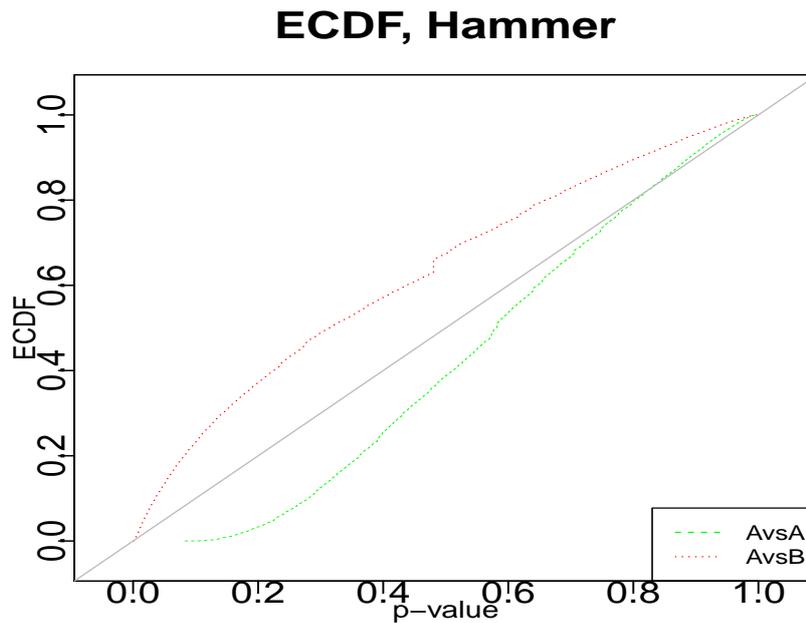

Figure 4: The ECDF plot for the profiles of p-values for the comparison between conditions ($AvsB$) and for the comparison within a condition ($AvsA$).

1.4 MV plot and volcano plot.

A MV plot is a scatter plot between the means (M=$[log_2(A)+log_2(B)]/2$) and the differences (V=$log_2(A)$ - $log_2(B)$). It helps to detect any dependent structures between the means and the differences in condition A and B. In a MV plot, we expect to see that the dots are roughly distributed on the two sides of the zero horizontal line without any

dependent patter between M and V. A volcano plot is a scatter plot that visualizes the linear dependence between the statistical changes (e.g. $-log_2$(p-value)) and the biological changes (e.g. log_2(fold change)). We expect to see that the dots are linearly and evenly distributed on the two sides of the zero vertical line. Both types of the plots are useful for visual inspection on the test results among genes. The function 'drawMA_vol' in the sSeq package can be used to draw the MV plot and the volcano plot. An example is shown in Fig.5. The red dots are the genes that have p-values less than 0.05.

```
> drawMA_vol(countsTable, conds.Hammer, res1$pval, cutoff=0.05);
```
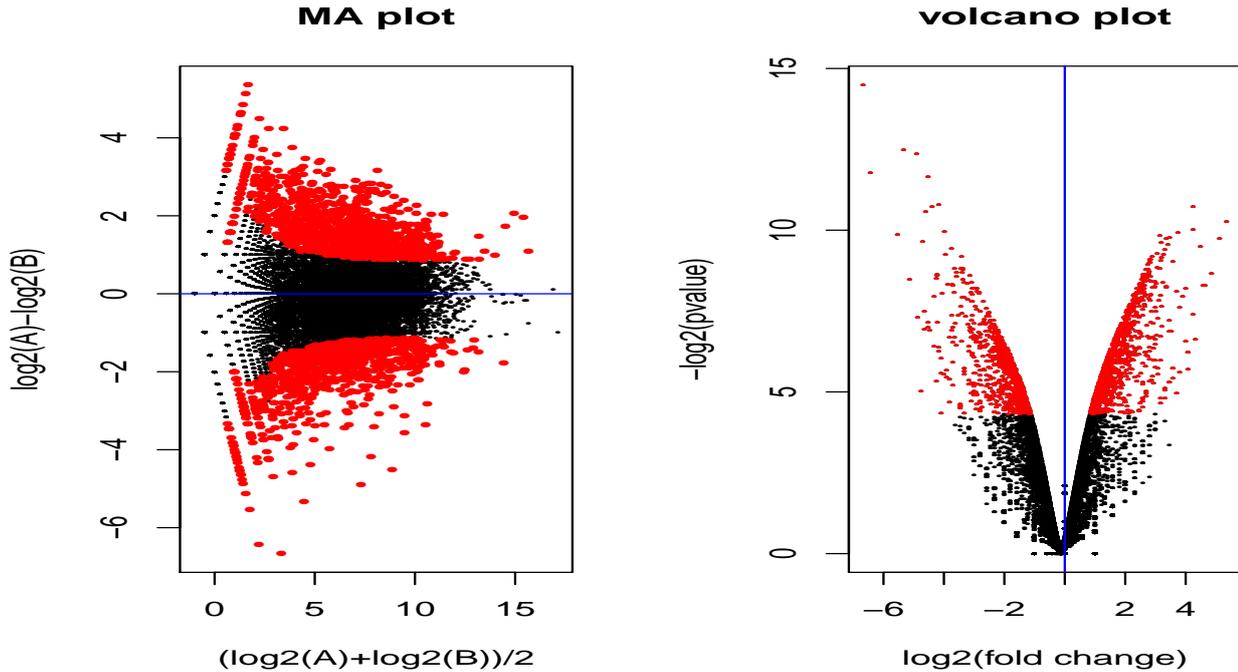

Figure 5: MA plot and volcano plot.

2 Comparison between two conditions for paired experimental design.

The sSeq package is also available to perform exact tests for complex designed experiments, such as paired design. The Tuch *et. al.* data [2] is used as an example. In the experiment, there were three patients who had oral squamous cell carcinoma, which is one of the most common cancers in humans. The paired samples from the tumor tissue and the normal tissue for each patient were collected and sequenced with the RNA-seq technology. This data set is included in the sSeq package and shown as follows. We use "normal" and "tumor" to represent the two conditions, and use 1, 2, 3 to represent the three patients. We will simultaneously compare the gene expression between the normal tissue and the tumor tissue within each of the three patients. After specifying the paired samples for each patient by the arguement 'coLevels', the exact tests for the paired-design experiment are performed. The counts of the genes that have the 25 smallest p-values are also shown as follows.

```
> data(Tuch);
> head(countsTable);

             N8    N33    N51   T8   T33   T51
NM_000014  2242   2285  15121  261   597  1991
NM_144670 11731  13308   6944  912  3071  1160
NM_017436   162    111    751  296   362   182
```

```
NM_015665    199    215    512   81   344   342
NM_023928    470    573    690  710  1112   728
NM_024666    298    332    856  203   790   909

> conds2 = c("normal", "normal", "normal", "tumor", "tumor", "tumor");
> coLevels=data.frame(subjects=c(1,2,3,1,2,3));

> res2 = nbTestSH(countsTable, conds2, "normal", "tumor",
+     coLevels= coLevels, pairedDesign=TRUE, pairedDesign.dispMethod="pooled");

[1] "Get shrinkage target at level 1"
[1] "For paired design, the aveaged dispersion estimates across paires are used."
[1] "Shrinkage estimates on dispersion are used for the tests."
[1] "The shrink target is 0.82375090107507"
[1] "The shrink quantile is 0.944"
[1] "30% processed."
[1] "50% processed."
[1] "70% processed."
[1] "90% processed."
[1] "100% processed."
Time difference of 1.448209 mins

> head(res2)

                Mean    rawMeanA   rawMeanB rawLog2FoldChange      dispMM     dispSH
NM_000014 2809.2787   6549.3333   949.6667         2.7858549 1.02097798  0.9805542
NM_144670 6895.8007 10661.0000  1714.3333         2.6366232 1.07008843  1.0195989
NM_017436  286.7446    341.3333   280.0000         0.2857545 0.38780403  0.4771561
NM_015665  252.2928    308.6667   255.6667         0.2717856 0.09804907  0.2467895
NM_023928  734.0457    577.6667   850.0000        -0.5572256 0.08411026  0.2357076
NM_024666  498.6980    495.3333   634.0000        -0.3560831 0.08028615  0.2326673
                pval
NM_000014 0.0031707506
NM_144670 0.0004503366
NM_017436 0.0097131515
NM_015665 0.0813849771
NM_023928 0.0713288404
NM_024666 0.0379905629

> countsTable[order(res2$pval),][1:25,]

                N8    N33    N51   T8   T33   T51
NM_001100112  4389   7944   9262    7    16  1818
NM_002272    76461  99082  47411  353 20651    31
NM_005181     1840   4180    552    1    35    72
NM_003280     1684   1787   4894    0     7   559
NM_152381     9915  10396  23309   15    48  7181
NM_182502     2592   7805   3372    3   321     9
NM_016190    24146  22026  12480   49  2353    26
NM_001231      519    857    833    0     3    83
NM_002371     2697   3941   1750    3   265     8
NM_001010909  4160   3425   1720    7   516     5
NM_002016      343   3180    713   91   134     0
NM_000257     3118   3767   4001    8    21   724
NM_203378     4114   6544  11279    6    59  1487
NM_145244      432    935   1005    1     6   227
NM_002465     4809   4146  15623   10    14  1311
NM_003063      532   1022   1384    1     6   200
NM_014332      406    688   2347    0     3   318
```

```
NM_198060      3741  1990 12531     4    17 1829
NM_207163       449  1032  2138     1     4  272
NM_198271       399   396  1138     2     0  140
NM_014440       367  1824   802    10    45    1
NM_032578       679   656  3256     1     3  321
NM_057088      1069  3774   885     7   358    5
NM_001122853    257   223   559     0     1   83
NM_005416     41809 49781 31185  1155  8036   45
```

References

[1] Hammer, P. and Banck, M.S. and Amberg, R. and Wang, C. and Petznick, G. and Luo, S. and Khrebtukova, I. and Schroth, G.P. and Beyerlein, P. and Beutler, A.S. mRNA-seq with agnostic splice site discovery for nervous system transcriptomics tested in chronic pain. *Genome research*, 20(6):847–860, 2010.

[2] B.B. Tuch, R.R. Laborde, X. Xu, J. Gu, C.B. Chung, C.K. Monighetti, S.J. Stanley, K.D. Olsen, J.L. Kasperbauer, E.J. Moore, et al. Tumor transcriptome sequencing reveals allelic expression imbalances associated with copy number alterations. *PloS One*, 5:e9317, 2010.