erccdashboard Package Vignette

Sarah A. Munro

November 18, 2013

This vignette describes the use of the erccdashboard R package to analyze External RNA Control Consortium (ERCC) spike-in control ratio mixtures in gene expression experiments. Two types of data from the SEQC/MAQC III project were analyzed.

1. Rat toxicogenomics treatment and control samples for different drug treatments
2. Human reference RNA samples from the MAQC I project, Universal Human Reference RNA (UHRR) and Human Brain Reference RNA (HBRR)

1 Rat Toxicogenomics Example: MET treatment

1.1 Load data and define input parameters

Load the testerccdashboard package.

```
> library( "testerccdashboard" )
```

Load the Rat Toxicogenomics Data set.

```
> load(file = system.file("data/SEQC.RatTox.Example.RData",
                          package = "testerccdashboard"))
```

The R workspace should now contain 5 count tables and for each count table a corresponding total reads vector. Take a look at the data for the MET experiment.

```
> head(COH.RatTox.ILM.MET.CTL.countTable)
     Feature MET_1 MET_2 MET_3 CTL_1 CTL_2 CTL_3
1 ERCC-00002 16629 18798 26568 36600 45436 25163
2 ERCC-00003  1347  1565  1983  3048  3447  2195
3 ERCC-00004  4569  5570  6755  1240  1484   902
4 ERCC-00009   811   869  1123   909  1073   537
5 ERCC-00013     3     1     2     1     5     1
6 ERCC-00019    24    32    43     5    13     4
> COH.RatTox.ILM.MET.CTL.totalReads
[1] 41423502 46016148 44320280 38400362 47511484 33910098
```

The first column of the count table, Feature, contains unique names for all the transcripts that were quantified in this experiment. The remaining columns represent replicates of the pair of samples, in this count table the control sample is labeled CTL and the treatment sample is labeled MET. An underscore is included to separate the sample names from the replicate numbers during analysis. This naming convention Sample_Rep is needed for the columns of any input count table.

The total reads vectors will be used for library size normalization of the count tables. Total reads can either represent the total number of reads in FASTQ files or total mapped reads. In the examples provided with this package FASTQ file total reads are used.

For our analysis of the MET-CTL experiment start by assigning the MET-CTL data to the input data variables countTable and totalReads.

```
> countTable <- COH.RatTox.ILM.MET.CTL.countTable
> totalReads <- COH.RatTox.ILM.MET.CTL.totalReads
```

In addition to countTable and totalReads, there are 7 additional variables that must be defined by the user. First the filename prefix for results files, filenameRoot, needs to be defined. Here we choose to use the lab abbreviation COH and the platform abbreviation ILM as our identifiers, but this is flexible for the user.

```
> filenameRoot = "COH.ILM"
```

Next, 5 parameters associated with the ERCC control ratio mixtures need to be defined, sample1Name, sample2Name, ERCCdilution, spikeVol, and totalRNAmass.

The sample spiked with ERCC Mix 1 is sample1Name and the sample spiked with ERCC Mix 2 is sample2Name. In this experiment sample1Name = MET and sample2Name = CTL. For a more robust experimental design the reverse spike-in design could be created using additional replicates of the treatment and control samples. ERCC Mix 2 would be spiked into MET samples and ERCC Mix 1 would be spiked into CTL control replicates.

ERCCdilution is the dilution factor of the pure Ambion ERCC mixes prior to spiking into total RNA samples. Here a 1/100 dilution was made from the Ambion ERCC mixes according to the protocol. The amount of diluted ERCC mix spiked into the total RNA sample is spikeVol (units are μL). The mass of total RNA spiked with the diluted ERCC mix is totalRNAmass (units are μg).

```
> sample1Name = "MET"
> sample2Name = "CTL"
> ERCCdilution = 1/100
> spikeVol = 1
> totalRNAmass = 0.500
```

The final required input parameter, choseFDR, is the False Discovery Rate (FDR) for differential expression testing. A typical choice would be 0.05 (5% FDR), for the rat data sets a more liberal FDR was used, choseFDR = 0.1.

```
> choseFDR = 0.1
```

In addition to the required input variables the user can also choose whether to print the results directly to a PDF file (the default is TRUE) with the variable printPDF.

1.2 Initialize the expDat list for analysis

The expDat list is created with the initDat function:

```
> expDat <- initDat(countTable, totalReads, filenameRoot, sample1Name,
                    sample2Name, ERCCdilution, spikeVol, totalRNAmass, choseFDR)
Filename root is: COH.ILM.MET.CTL
Library sizes:
41.4235 46.01615 44.32028 38.40036 47.51148 33.9101
Using total sequencing reads,
 mean library size factor = 41.93031
```

Look at the structure of expDat

```
> summary(expDat)
```

```
                Length Class      Mode
sampleInfo      16     -none-     list
totalReads       6     -none-     numeric
Transcripts      7     data.frame list
designMat        3     data.frame list
sampleNames      2     -none-     character
idCols           6     data.frame list
totalReads       6     -none-     numeric
expressDat       7     data.frame list
libeSize         6     -none-     numeric
ERCCxlabelIndiv  1     -none-     expression
ERCCxlabelAve    1     -none-     expression
spikeFraction    1     -none-     numeric
mnLibeFactor     1     -none-     numeric
sampleLibeSums   6     -none-     numeric
```

The expDat list will be passed to the erccdashboard functions for analysis of technical performance.

1.3 Estimate the mRNA fraction difference, r_m for the pair of samples

Estimate r_m for the sample pair using a negative binomial glm. The r_m results will be added to the expDat structure and are necessary for the remaining analysis.

```
> expDat <- est_r_m(expDat, cnt = expDat$Transcripts, printPlot = F)
Check for sample mRNA fraction differences(r_m)...

log.offset
17.53936 17.6445 17.60695 17.46358 17.67648 17.33922

Number of ERCC Controls Used in r_m estimate
63

Outlier ERCCs for GLM r_m Estimate:
None

GLM log(r_m) estimate:
-0.0472291

GLM log(r_m) estimate standard deviation:
0.02061546

GLM r_m estimate:
0.9538688

GLM r_m upper limit
0.9599503

GLM r_m lower limit
0.9478259
```

1.4 Test for differential expression

Test for differential expression with the geneExprTest function. This function wraps the QuasiSeq differential expression testing package. If a correctly formatted csv file is provided with the necessary DE test results, then geneExprTest will bypass DE testing (with reduced runtime). The function will look for a csv file with the name ”filenameRoot.quasiSeq.res.csv” and the first 3 column headers of the file must be ”Feature”, ”pvals”, and ”qvals”.

```
> expDat <- geneExprTest(expDat, cnt = expDat$Transcripts,
                          designMat = expDat$designMat )
 Differential expression test results exist, will use
 existing results for analysis. Delete COH.ILM.MET.CTL.quasiSeq.res.csv if
 you want to repeat differential expression testing.

Threshold P-value
0.006689177
```

1.5 Diagnostic Performance: ROC curves and AUC statistics

Generate ROC curves for the differential ratios and the corresponding Area Under the Curve (AUC) statistics.

```
> expDat = erccROC(expDat)
Area Under the Curve (AUC) Results:
  Ratio   AUC Measured Spiked
1   4:1 1.000       16     23
2 1:1.5 0.950       16     23
3   1:2 0.967       16     23
```

1.6 Diagnostic Performance: Limit of Detection of Ratios (LODR)

Find LODR estimates using the ERCC data p-values.

```
> expDat = estLODR(expDat,kind = "ERCC", prob=0.9)
  Ratio LODR Estimate 90% CI Lower Bound 90% CI Upper Bound
1   4:1            26                 19                 31
3 1:1.5           Inf               <NA>               <NA>
4   1:2           240                120                340
```

One can also obtain LODR estimates using p-values simulated from endogenous transcripts

```
> expDat = estLODR(expDat, kind = "Sim", prob = 0.9)
  Ratio LODR Estimate 90% CI Lower Bound 90% CI Upper Bound
1   4:1            17                 13                 22
3 1:1.5           Inf               <NA>               <NA>
4   1:2           Inf               <NA>               <NA>
```

1.7 Use dynRangePlot function to evaluate dynamic range data

Evaluate the dynamic range of the experiment using the ERCC controls.

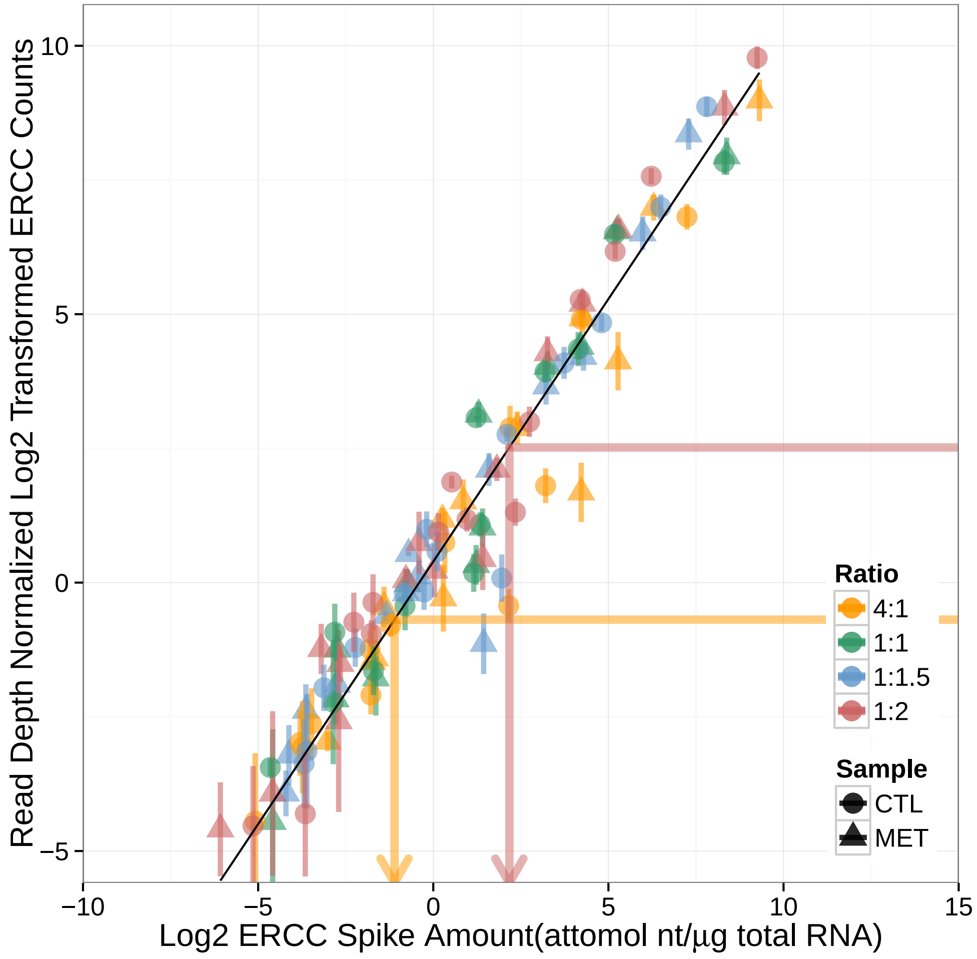
```
> expDat <- dynRangePlot(expDat, expressDat = expDat$expressDat,
                              designMat = expDat$designMat, noErrorBars = F)
[1] "Number of ERCCs in Mix 1 dyn range:"
[1] 63
[1] "Number of ERCCs in Mix 2 dyn range:"
[1] 63
These ERCCs were not included in the signal-abundance plot,
because not enough non-zero replicate measurements of these
controls were obtained for both samples:
 [1] "ERCC-00058" "ERCC-00067" "ERCC-00077" "ERCC-00168"
 [5] "ERCC-00028" "ERCC-00033" "ERCC-00040" "ERCC-00109"
 [9] "ERCC-00154" "ERCC-00158"
```

1.8 Use LODR estimates to Annotate Signal-Abundance and Ratio-Abundance Plots

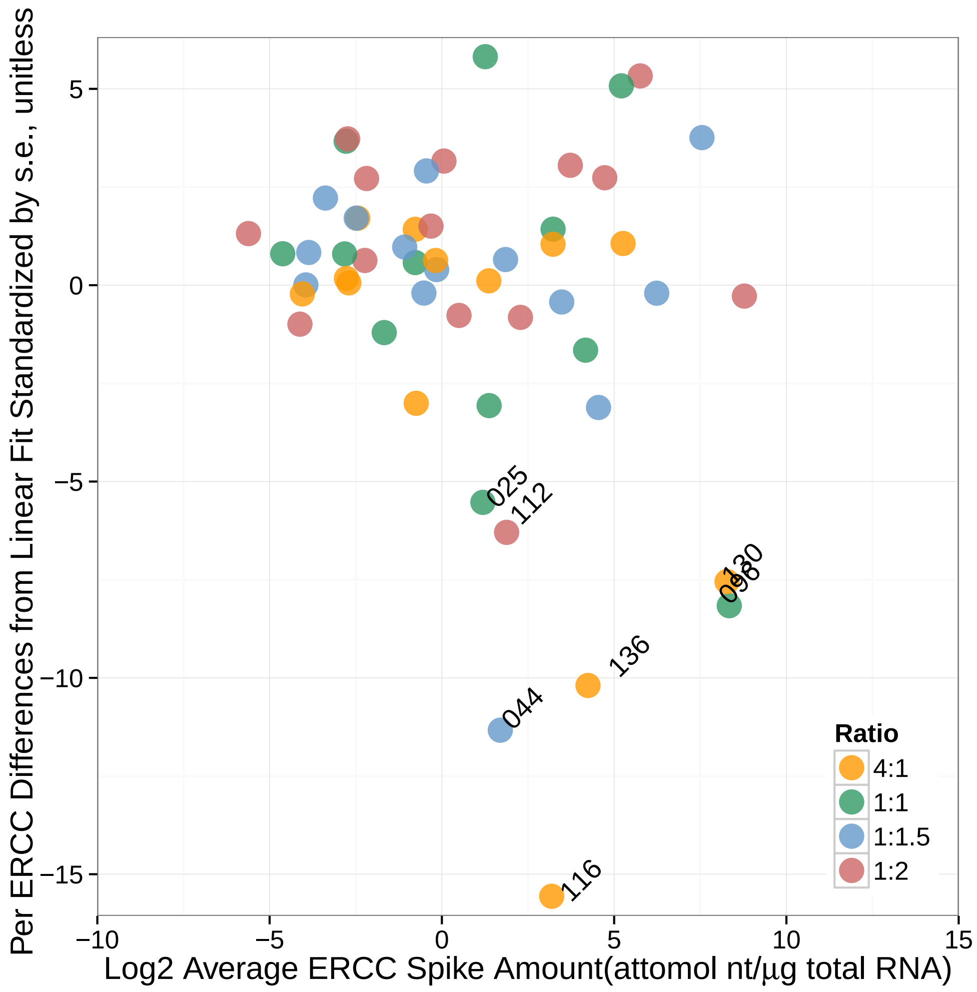
Get LODR annotations for adding to plots and then annotate the signal-abundance and ratio-abundance plots using LODR estimate information

```
> expDat <- annotLODR(expDat)
   Fold Ratio Count Log2Count_normalized  Log2Conc
1   4:1   4:1    26            -0.689482 -1.106341
2   1:1   1:1    NA                   NA        NA
3 1:1.5 1:1.5   Inf                  Inf       Inf
4   1:2   1:2   240             2.516969  2.172251
[1] "LODR estimates will be used to code ratio-abundance plot"
[1] "These ERCCs were not included in the ratio-abundance plot,  because not enough non-zero replicate
 [1] "ERCC-00028" "ERCC-00033" "ERCC-00040" "ERCC-00058"
 [5] "ERCC-00067" "ERCC-00077" "ERCC-00109" "ERCC-00154"
 [9] "ERCC-00158" "ERCC-00168"
[1] "Global Ratio SD for this sample pair is:"
[1] 0.7785041
                     Estimate Std. Error   t value
Minimum SD Estimate 0.4787852 0.03264151 14.667986
Maximum SD Estimate 1.6195817 0.20936506  7.735683
Lambda              0.4123437 0.05767326  7.149652
                        Pr(>|t|)
Minimum SD Estimate 4.049430e-20
Maximum SD Estimate 3.770026e-10
Lambda              3.158191e-09
[1] "Printing MA plot with LODR coding"
```

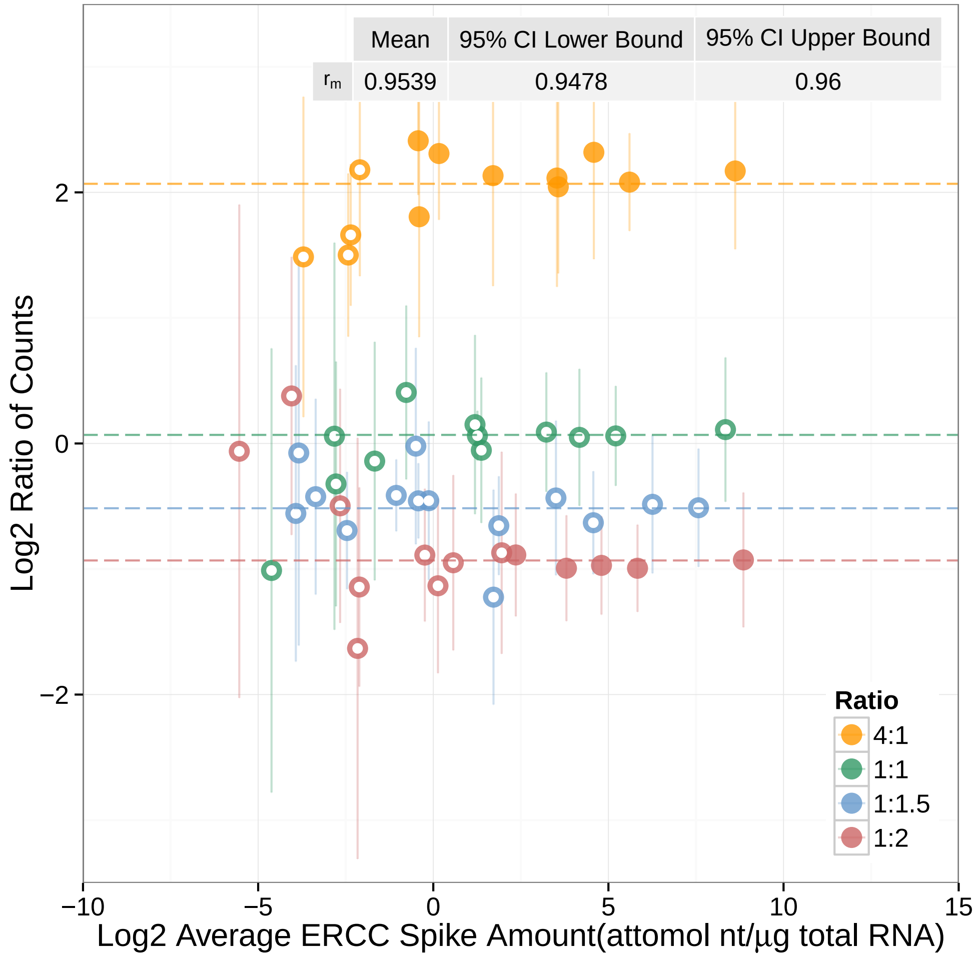
```
> expDat$Figures$plotdynRangeAnnot
```



```
> expDat$Figures$plotERCCeffects
```



```
> expDat$Figures$plotRatioAnnot
```



	Mean	95% CI Lower Bound	95% CI Upper Bound
r_m	0.9539	0.9478	0.96

1.9 Output Results for Comparisons Across Experiments or Laboratories

If you wish, save your results to an Rdata file that can be reused.

```
> saveResults(expDat)
```

You can also save any of the figures to pdf for additional use. The helper function savePlots will output selected figures to individual pages or place multiple figures per page in a pdf file

```
> savePlots(expDat)
```

2 SEQC Reference RNA Examples: UHRR vs. HBRR

2.1 Load data and define input parameters