erccdashboard Package Vignette

Sarah A. Munro

January 22, 2014

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 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( "erccdashboard" )
```

Load the Rat Toxicogenomics Data set.

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

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. Please note that the current version of the package automatically uses a library size normalization, therefore count data is expected. A future version of the package will be flexible to also accept data that has already been normalized.

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, 6 parameters associated with the ERCC control ratio mixtures need to be defined, sample1Name, sample2Name, ERCCMixes, 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"
> ERCCmixes = "RatioPair"
> 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, ERCCmixes, 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    10     -none-     list
plotInfo       8     -none-     list
erccInfo       4     -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
normERCCDat    7     data.frame list
libeSize       6     -none-     numeric
spikeFraction  1     -none-     numeric
mnLibeFactor   1     -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)
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
```

An r_m of 1 indicates that the two sample types under comparison have similar mRNA fractions of total RNA. The r_m estimate is used to adjusted the expected ERCC mixture ratios in this analysis and may indicate a need for a different sample normalization approach.

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)
```

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 Detected 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)
Estimating LODR
.............................................
  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)
Estimating LODR
.............................................
  Ratio LODR Estimate 90% CI Lower Bound 90% CI Upper Bound
1   4:1            33                 21                 44
3 1:1.5           Inf               <NA>               <NA>
4   1:2           Inf               <NA>               <NA>
```

1.7 Use dynRangePlot function to evaluate dynamic range of the data

This function will add a plot to `expDat$Figures` of the signal vs. abundance of the spiked ERCC controls.

```
> expDat <- dynRangePlot(expDat, errorBars = T)
Number of ERCCs in Mix 1 dyn range:  63

Number of ERCCs in Mix 2 dyn range:  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:

ERCC-00058
ERCC-00067
ERCC-00077
ERCC-00168
ERCC-00028
ERCC-00033
ERCC-00040
ERCC-00109
ERCC-00154
ERCC-00158
```

This figure shows that in this experiment the expected signal-abundance relationship spans a 2^{15} dynamic range. To capture the full 2^{20} dynamic range design of the control mixtures additional sequencing depth may be needed.

1.8 Use LODR estimates to Annotate MA Plot

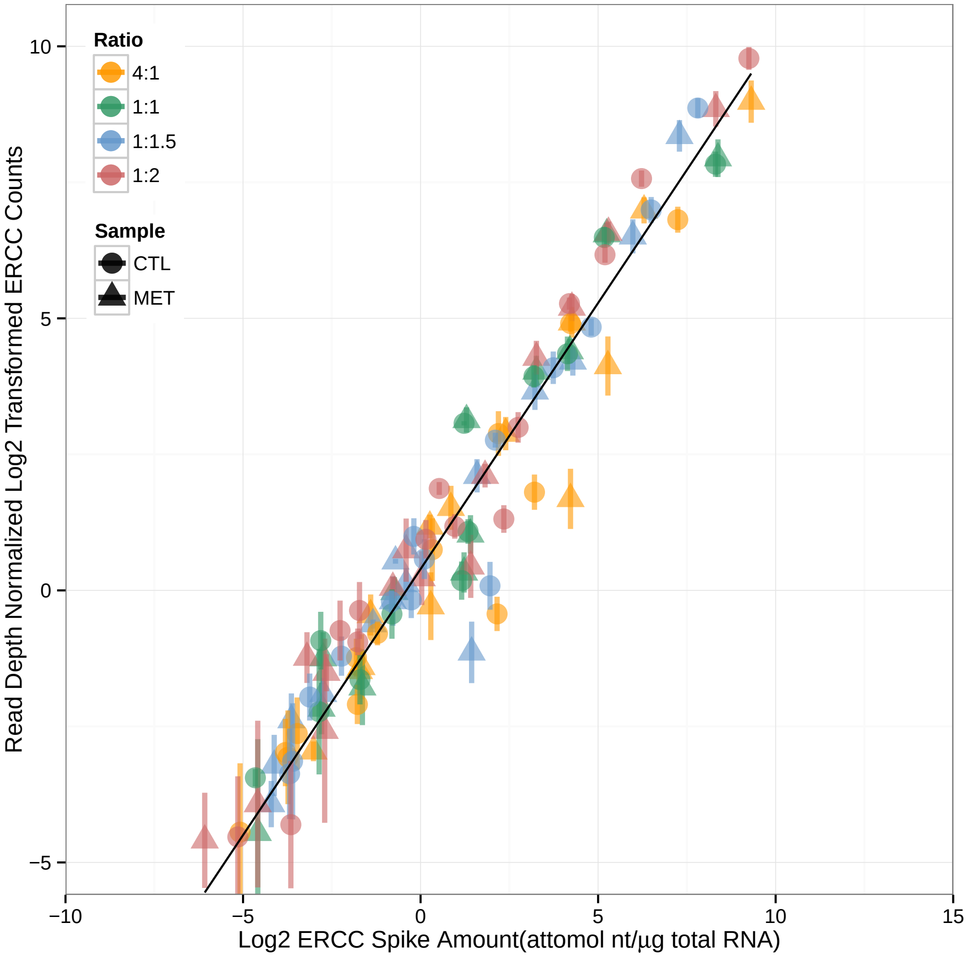
Get LODR estimates and annotate the MA plot using LODR estimate information

```
> 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

LODR estimates are available to code ratio-abundance plot
```

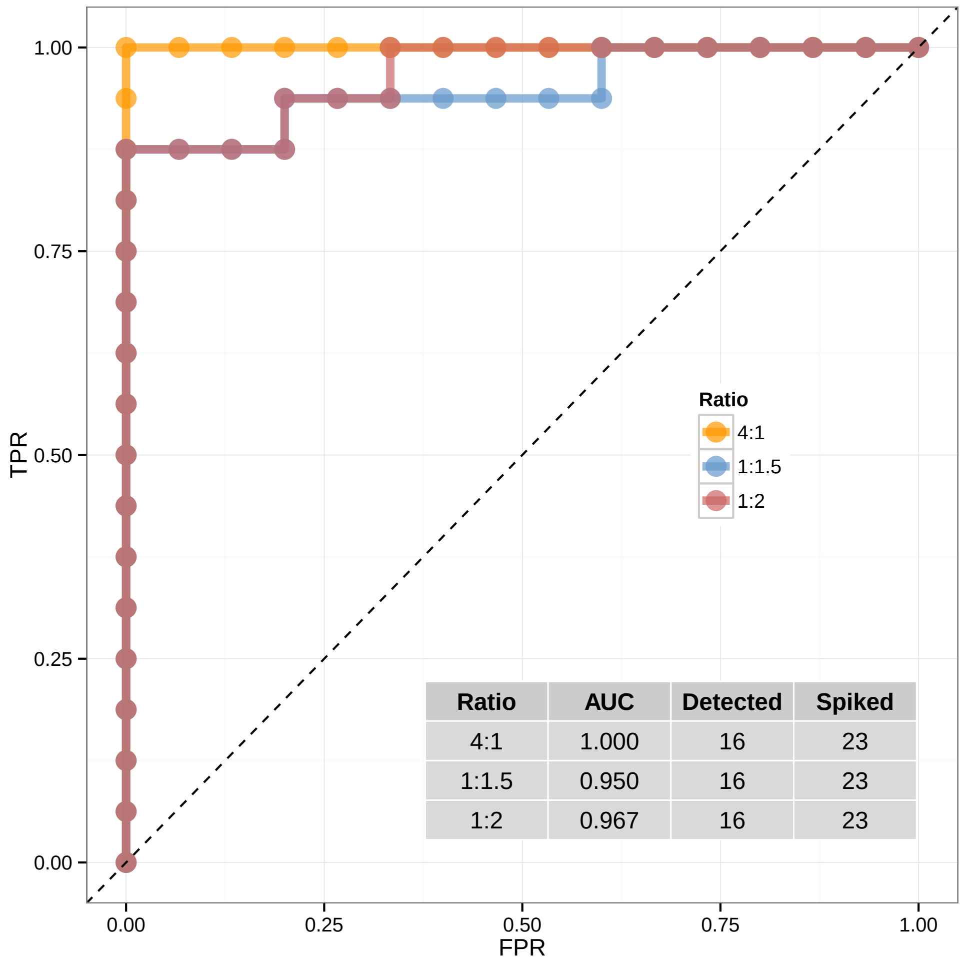
1.9 Viewing Diagnostic Plots

All dashboard plots are stored in the expDat$Figures list. You can call any figure for viewing directly and you can also save the figures to a pdf file.

```
> expDat$Figures$dynRangePlot
```

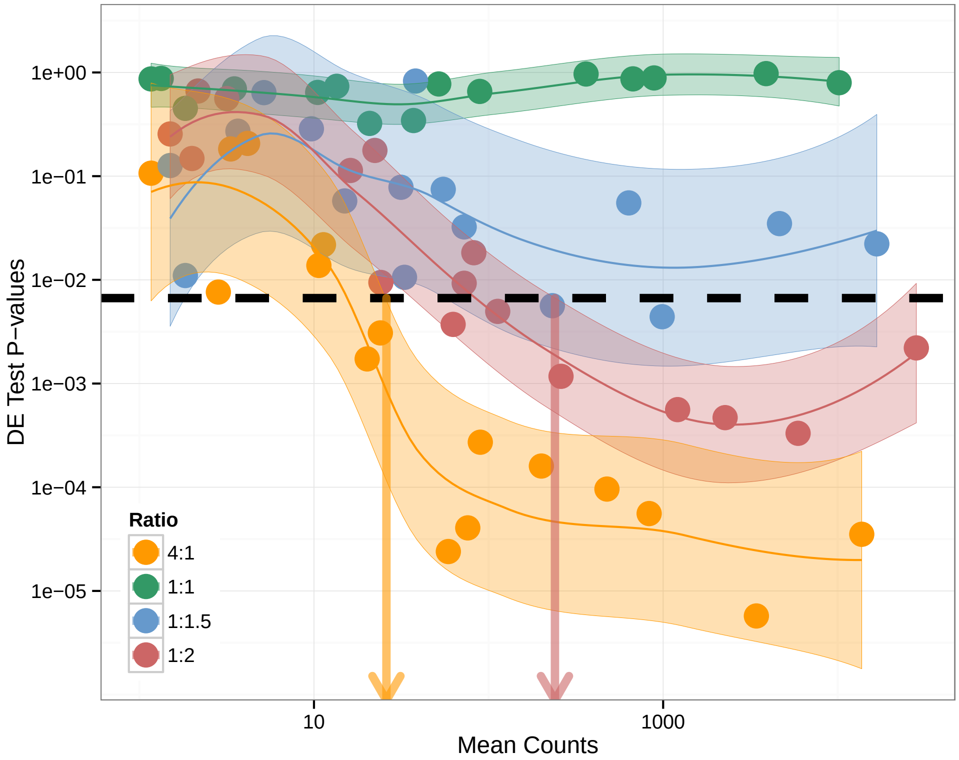


```
> expDat$Figures$rocPlot
```



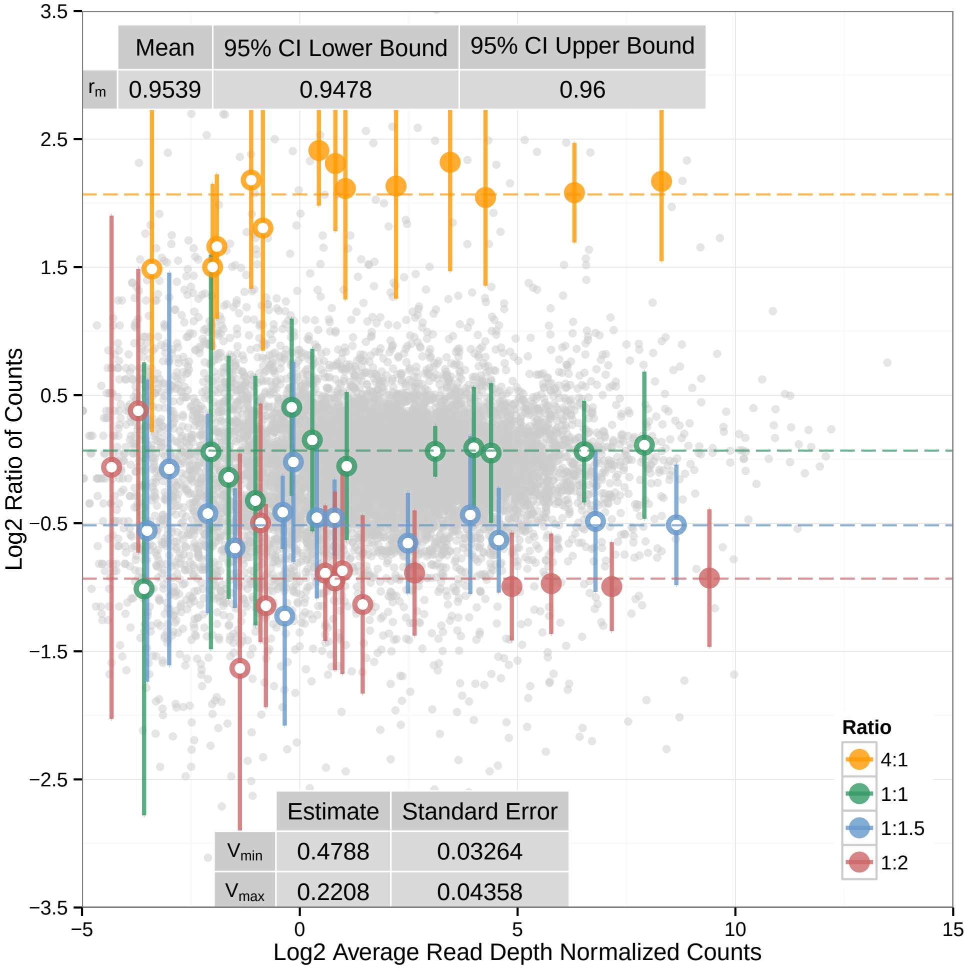
Ratio	AUC	Detected	Spiked
4:1	1.000	16	23
1:1.5	0.950	16	23
1:2	0.967	16	23

```
> expDat$Figures$lodrERCCPlot
```



Ratio	LODR Estimate	90% CI Lower Bound	90% CI Upper Bound
4:1	26	19	31
1:1.5	Inf	NA	NA
1:2	240	120	340

```
> expDat$Figures$maPlot
```



The function savePlots will save selected figures to a pdf file. The default is the 4 manuscript figures to a single page (plotsPerPg = “”manuscript“”). If plotsPerPg = “”single“” then each plot is placed on an individual page in one pdf file. If plotlist is not defined (plotlist = NULL) then all plots in `expDat$Figures` are printed to the file.

```
> savePlots(expDat)
```

1.10 Output Results for Comparisons Across Experiments or Between Laboratories

If you wish, save your expDat list to an Rdata file that can be reused for comparisons across experiments or between laboratories.

```
> save(expDat,file=paste0(expDat$sampleInfo$filenameRoot,".RData"))
```

2 SEQC Reference RNA Examples: UHRR vs. HBRR

2.1 Load data and define input parameters

Note that due to file size limitations only a subset of the data shown in the manuscript are provided with the package (data is from 6 out of 9 laboratories).

```
> load(file = system.file("data/SEQC.Main.Example.Simple.RData",
                          package = "erccdashboard"))
> countTable <- Lab5.ILM.UHRR.HBRR.countTable
> totalReads <- Lab5.ILM.UHRR.HBRR.totalReads
> filenameRoot = "Lab5"
> sample1Name = "UHRR"
> sample2Name = "HBRR"
> ERCCMixes = "RatioPair"
> ERCCdilution = 1
> spikeVol = 50
> totalRNAmass = 2.5*10^(3)
> choseFDR = 0.01
> expDat <- initDat(countTable, totalReads, filenameRoot, sample1Name,
                    sample2Name, ERCCMixes, ERCCdilution, spikeVol, totalRNAmass,
                    choseFDR)
Filename root is: Lab5.UHRR.HBRR
Library sizes:
138.7869 256.0065 199.4683 431.9338 247.9856 219.3833 251.2658 257.5082
Using total sequencing reads,
 mean library size factor = 250.2923
```

2.2 Dashboard analysis of SEQC Lab 5 data

The commands used in the Rat toxicogenomics sample can be repeated on the SEQC data for Lab 5.

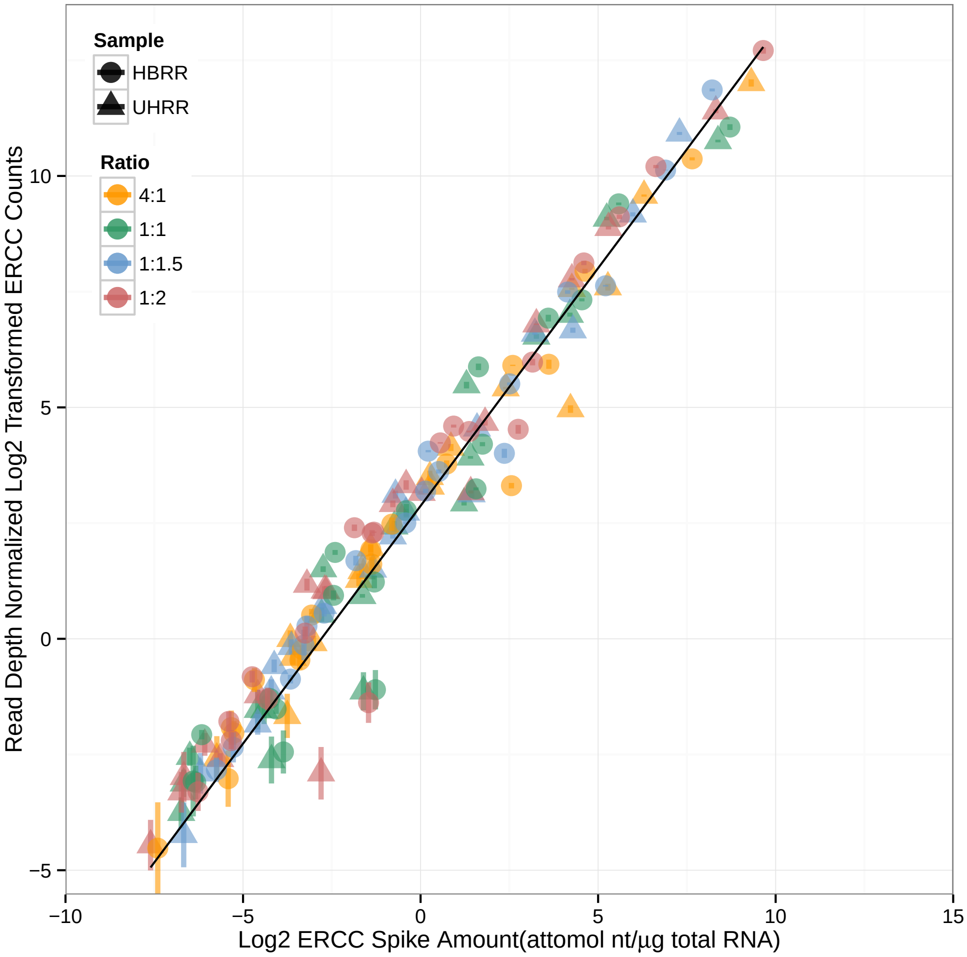
```
> expDat <- est_r_m(expDat)
Check for sample mRNA fraction differences(r_m)...

log.offset
18.74845 19.36071 19.11117 19.88378 19.32888 19.20633 19.34202 19.36656

Number of ERCC Controls Used in r_m estimate
71

Outlier ERCCs for GLM r_m Estimate:
ERCC-00137 ERCC-00085 ERCC-00054 ERCC-00019

GLM log(r_m) estimate:
0.2335326

GLM log(r_m) estimate standard deviation:
0.002941665

GLM r_m estimate:
```

```
1.263054

GLM r_m upper limit
1.264133

GLM r_m lower limit
1.261975


> expDat <- geneExprTest(expDat)


> expDat <- erccROC(expDat)
Area Under the Curve (AUC) Results:
  Ratio   AUC Detected Spiked
1   4:1 1.000       17     23
2 1:1.5 0.971       17     23
3   1:2 0.994       19     23
> expDat <- estLODR(expDat,kind = "ERCC", prob=0.9)
Estimating LODR
.............................................
  Ratio LODR Estimate 90% CI Lower Bound 90% CI Upper Bound
1   4:1           <27                <27                 36
3 1:1.5           160                <25                170
4   1:2            29                <19                 34
> expDat <- estLODR(expDat, kind = "Sim", prob = 0.9)
Estimating LODR
.............................................
  Ratio LODR Estimate 90% CI Lower Bound 90% CI Upper Bound
1   4:1           <30                <30                <30
3 1:1.5            50                <25                 69
4   1:2            45                 38                 52
> expDat <- dynRangePlot(expDat,errorBars=T)
Number of ERCCs in Mix 1 dyn range:  71

Number of ERCCs in Mix 2 dyn range:  71
> expDat <- annotLODR(expDat)
   Fold Ratio Count Log2Count_normalized  Log2Conc
1   4:1   4:1    27            -3.212583 -5.922734
2   1:1   1:1    NA                   NA        NA
3 1:1.5 1:1.5   160            -0.645542 -3.423910
4   1:2   1:2    29            -3.109489 -5.822381

LODR estimates are available to code ratio-abundance plot
> #expDat <- maSignal(expDat, alphaPoint = 0.8,  r_mAdjust = T, replicate = T)
```

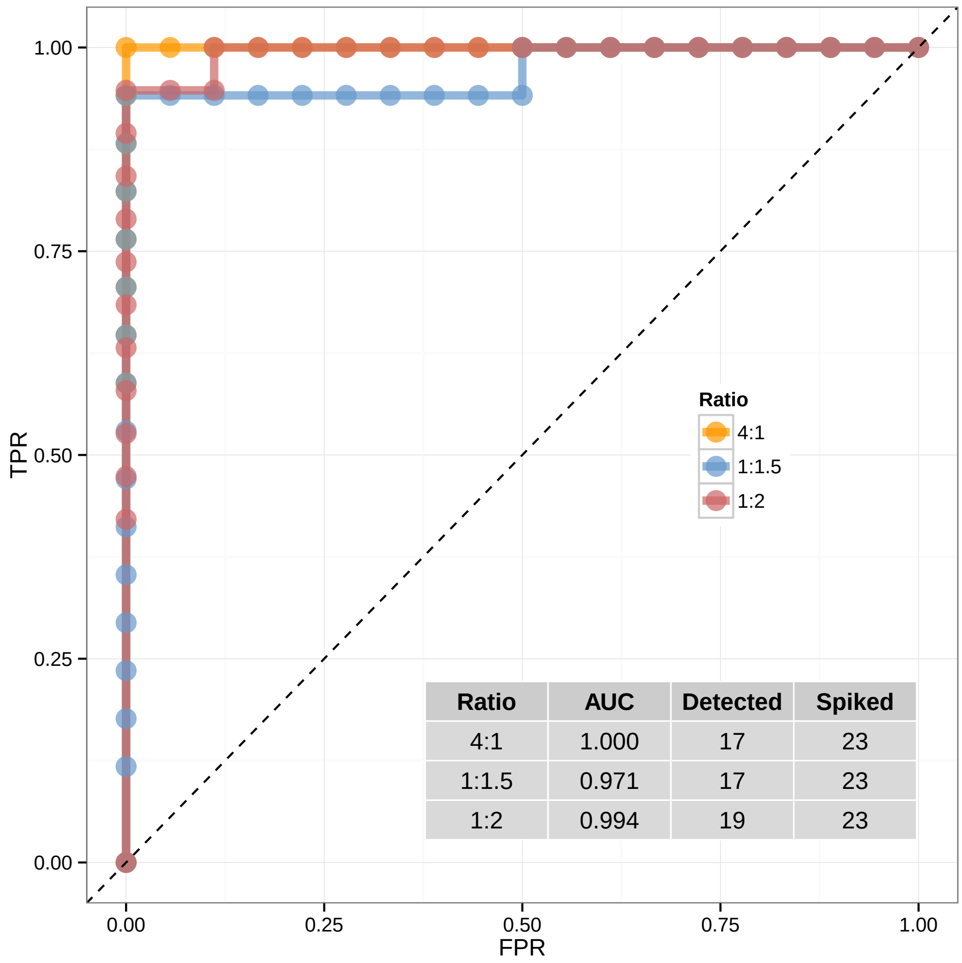
2.3 Viewing Diagnostic Plots

All dashboard plots are stored in the expDat$Figures list. You can call any figure for viewing directly and you can also save the figures to a pdf file. The six plots presented in the erccdashboard publication can be generated with the following commands.

```
> expDat$Figures$dynRangePlot
```

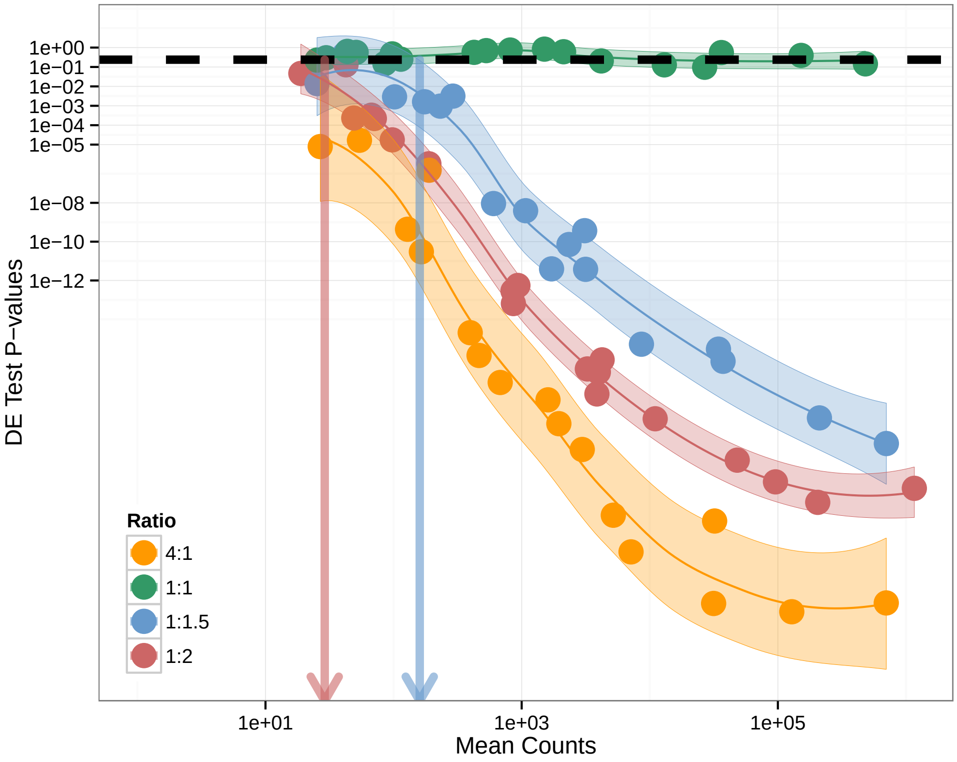


```
> expDat$Figures$rocPlot
```



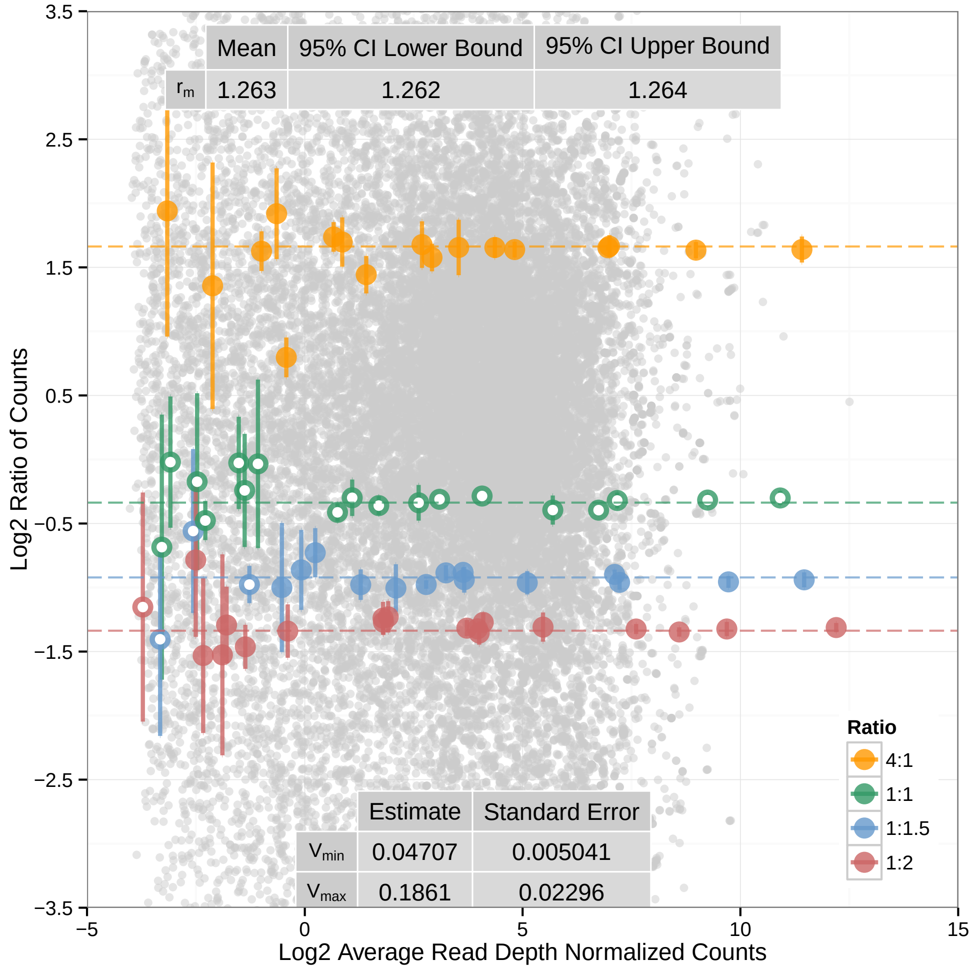
Ratio	AUC	Detected	Spiked
4:1	1.000	17	23
1:1.5	0.971	17	23
1:2	0.994	19	23

```
> expDat$Figures$lodrERCCPlot
```



Ratio	LODR Estimate	90% CI Lower Bound	90% CI Upper Bound
4:1	<27	<27	36
1:1.5	160	<25	170
1:2	29	<19	34

```
> expDat$Figures$maPlot
```



	Mean	95% CI Lower Bound	95% CI Upper Bound
r_m	1.263	1.262	1.264

	Estimate	Standard Error
V_{min}	0.04707	0.005041
V_{max}	0.1861	0.02296

As with the Rat Toxicogenomic data figures can be saved to pdf with savePlots.

```
> savePlots(expDat)
```

And if you wish, save your analysis results to an Rdata file that can be reused for comparisons across experiments or between laboratories.

```
> save(expDat,file=paste0(expDat$sampleInfo$filenameRoot,".RData"))
```

3 Alternative Spike-in Designs

By default the package is configured to analyze the ERCC ratio mixtures produced by Ambion (ERCC ExFold RNA Spike-In Mixes, Catalog Number 4456739). This pair of control ratio mixtures were designed to have 1:1, 4:1, 1:1.5, and 1:2 ratios of 92 distinct RNA transcripts (23 different RNA control sequences are in each of these four ratio subpools). Alternative ERCC RNA control ratio mixture designs can be produced using the NIST DNA Plasmid Library for External Spike-in Controls (NIST Standard Reference Material 2374, https://www-s.nist.gov/srmors/certificates/2374.pdf). For example, a pair of RNA contol mixtures could be created with a ternary ratio design, three subpools of RNA controls with either no change (1:1) or 2-fold increased (2:1) and 2-fold decreased (1:2) relative abundances between the pair of mixtures (Mix 1/Mix 2). To use alternative spike-in mixture designs with the dashboard a csv file must be provided to the package with the argument userMixFile for the initDat function.

If all samples from both conditions were only spiked with a single ERCC mixture (e.g. Ambion Catalog Number 4456740, ERCC RNA Spike-In Mix) a limited subset of the package functions can be used. For initDat use ERCCMixes=”Single” and `est_r_m` and `dynRangePlot` functions can then be used to examine the mRNA fraction differences for the pair of samples and evaluate the dynamic range of the experiment.

4 Notes on R version and session information

The results shown in this R vignette are the same as the results shown in our manuscript and were obtained in R version 2.15.3. Slightly different results will be obtained if R version 3.0.2 and associated packages are used due to differences in the differential expression testing packages. Here is the session information for this analysis in R 2.15.3.

```
> sessionInfo()
R version 2.15.3 (2013-03-01)
Platform: x86_64-apple-darwin9.8.0/x86_64 (64-bit)

locale:
[1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8

attached base packages:
[1] grid      stats     graphics  grDevices utils
[6] datasets  methods   base

other attached packages:
 [1] reshape2_1.2.2      gridExtra_0.9.1
 [3] ROCR_1.0-4          gplots_2.11.0
 [5] MASS_7.3-23         KernSmooth_2.23-10
 [7] caTools_1.14        gdata_2.12.0
 [9] gtools_2.7.1        QuasiSeq_1.0-2
[11] edgeR_3.0.8         limma_3.14.4
```

```
[13] fields_6.7          spam_0.29-2
[15] erccdashboard_0.9.3 ggplot2_0.9.3.1

loaded via a namespace (and not attached):
 [1] bitops_1.0-4.2     colorspace_1.2-1
 [3] dichromat_2.0-0    digest_0.6.3
 [5] gtable_0.1.2       labeling_0.1
 [7] lattice_0.20-15    locfit_1.5-9
 [9] munsell_0.4        plyr_1.8
[11] proto_0.3-10       RColorBrewer_1.0-5
[13] scales_0.2.3       stringr_0.6.2
[15] tools_2.15.3
```