`erccdashboard` Package Vignette

Sarah A. Munro

March 28, 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. Analysis is shown for two types of samples spiked with ERCC control ratio mixtures from the SEQC project

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

A subset of the large data set produced in the SEQC study is provided here as examples. The three sets of example data are:

1. Rat toxicogenomics RNA-Seq gene expression count data
2. UHRR/HBRR RNA-Seq gene expression count data
3. UHRR/HBRR Microarray gene expression fluorescent intensity data

1 Rat Toxicogenomics Example: MET (methimazole treatment) and CTL (control) Experiment

1.1 Load data and define input parameters

Load the testerccdashboard package.

```
> library( "erccdashboard" )
```

Load the package gene expression data.

```
> data(SEQC.Example)
```

The R workspace should now contain 5 objects

- COH.array.UHRR.HBRR - Fluorescent signal data from an Illumina beadarray microarray experiment
- COH.RatTox.ILM.MET.CTL.countTable - RNA-Seq count data from a rat toxicogenomics experiment
- COH.RatTox.ILM.MET.CTL.totalReads - total sequenced reads factors for each column in the corresponding rat experiment count table
- Lab5.ILM.UHRR.HBRR.countTable - RNA-Seq count data from the SEQC interlaboratory study with UHRR and HBRR
- Lab5.ILM.UHRR.HBRR.totalReads - total sequenced reads factors for each column in the corresponding UHRR/HBRR count table

Take a look at the count table for the MET experiment.

```
> head(COH.RatTox.ILM.MET.CTL.countTable)
         Feature MET_1 MET_2 MET_3 CTL_1 CTL_2 CTL_3
16499 ERCC-00002 16629 18798 26568 36600 45436 25163
16500 ERCC-00003  1347  1565  1983  3048  3447  2195
16501 ERCC-00004  4569  5570  6755  1240  1484   902
16502 ERCC-00009   811   869  1123   909  1073   537
16503 ERCC-00012     0     0     0     0     0     0
16504 ERCC-00013     3     1     2     1     5     1
> 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 column name format Sample_Rep is required for the columns of any input count table.

These input count tables also must be unnormalized count data, because the default differential expression testing of RNA-Seq experiments in the erccdashboard is done with the QuasiSeq package, which requires the use of integer count data. During the erccdashboad analysis the input count data will be normalized using library size factors. These factors can be provided by the user as the input argument repNormFactor, a vector of total reads. The example total reads vectors we provide here were derived from the FASTQ files associated with each column in the count tables. Alternatively, if repNormFactor is undefined (or NULL) for an RNA-Seq analysis then a vector of total mapped reads (column sums of a count table) will be calculated and used for normalization.

1.2 Quick analysis: runDashboard

To run the default analysis function runDashboard on the MET-CTL rat toxicogenomics RNA-Seq experiment, the following input arguments are required:

```
> datType = "count" # count for RNA-Seq data, array for microarray data
> expTable = COH.RatTox.ILM.MET.CTL.countTable # the expression measure table
> repNormFactor = COH.RatTox.ILM.MET.CTL.totalReads # the library size factors
> filenameRoot = "COH.ILM" # user defined filename prefix for results files
> sample1Name = "MET" # name for sample 1 in the experiment
> sample2Name = "CTL" # name for sample 2 in the experiment
> erccmix = "RatioPair" # name of ERCC mixture design, "RatioPair" is default
> erccdilution = 1/100 # dilution factor used for Ambion spike-in mixtures
> spikeVol = 1 # volume (in microliters) of diluted spike-in mixture added to
>              #   total RNA mass
> totalRNAmass = 0.500 # mass (in micrograms) of total RNA
> choseFDR = 0.1 # user defined false discovery rate (FDR), default is 0.05
```

For any experiment 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.

The dilution factor of the pure Ambion ERCC mixes prior to spiking into total RNA samples is erccdilution. 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).

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), so this is the default choseFDR value. For the rat data sets a more liberal FDR was used, choseFDR = 0.1.

The function runDashboard.R is provided for convenient default erccdashboard analysis. Execution of the runDashboard function calls the default functions for erccdashboard analysis and reports parameters and progress to the R console. The functions called within runDashboard.R are also available to the user (details provided in Section 4). All data and analysis results are stored in the list object expDat. For convenience the main diagnostic figures are saved to a pdf file and the expDat object is saved to an .RData object named using the filenameRoot provided by the user.

```
> expDat <- runDashboard(datType = "count",
                         expTable = COH.RatTox.ILM.MET.CTL.countTable,
                         repNormFactor = COH.RatTox.ILM.MET.CTL.totalReads,
                         filenameRoot = "COH.ILM",sample1Name = "MET",
                         sample2Name = "CTL",erccmix = "RatioPair",
                         erccdilution = 1/100, spikeVol = 1,
                         totalRNAmass = 0.500, choseFDR = 0.1)
Initializing the expDat list structure...
choseFDR = 0.1
Filename root is: COH.ILM.MET.CTL

Transcripts were removed with a mean count < 1 or more than 2
replicates with 0 counts.
Original data contained  16590 transcripts.
After filtering  11570 transcripts remain for  analysis.
A total of 29 out of 92
ERCC controls were filtered from the data set
The excluded ERCCs are:
ERCC-00012 ERCC-00014 ERCC-00016 ERCC-00017 ERCC-00024
ERCC-00041 ERCC-00048 ERCC-00057 ERCC-00061 ERCC-00073
ERCC-00075 ERCC-00081 ERCC-00083 ERCC-00086 ERCC-00097
ERCC-00098 ERCC-00104 ERCC-00117 ERCC-00120 ERCC-00123
ERCC-00126 ERCC-00134 ERCC-00137 ERCC-00138 ERCC-00142
ERCC-00147 ERCC-00150 ERCC-00156 ERCC-00164

Library sizes:
41.4235 46.01615 44.32028 38.40036 47.51148 33.9101
Check for sample mRNA fraction differences(r_m)...

log.offset
3.723848 3.828992 3.791442 3.648067 3.860971 3.523713

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

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


Saving dynRangePlot to expDat

Starting differential expression tests

Using Total Reads
41423502 46016148 44320280 38400362 47511484 33910098
Disp = 0.0627 , BCV = 0.2504
Disp = 0.06263 , BCV = 0.2503
[1] "Analyzing Gene # 2"
[1] "Analyzing Gene # 10"
[1] "Analyzing Gene # 100"
[1] "Analyzing Gene # 500"
[1] "Analyzing Gene # 1000"
[1] "Analyzing Gene # 2500"
[1] "Analyzing Gene # 5000"
[1] "Analyzing Gene # 10000"
[1] "Analyzing Gene # 2"
[1] "Analyzing Gene # 10"
[1] "Analyzing Gene # 100"
[1] "Analyzing Gene # 500"
[1] "Analyzing Gene # 1000"
[1] "Analyzing Gene # 2500"
[1] "Analyzing Gene # 5000"
[1] "Analyzing Gene # 10000"
[1] "Note: 'test.mat' not provided. Comparing each model \nfrom 'design.list' to first model in 'desig
[1] "Spline scaling factor: 0.965600661690359"
[1] "Spline scaling factor: 0.962950223299725"
[1] "Analyzing Gene # 2"
[1] "Analyzing Gene # 10"
```

```
[1] "Analyzing Gene # 2"
[1] "Analyzing Gene # 10"
[1] "Note: 'test.mat' not provided. Comparing each model \nfrom 'design.list' to first model in 'desig
[1] "Spline scaling factor: 0.962950223299725"
[1] "Finished DE testing"
[1] "Spline scaling factor: 0.962950223299725"

Finished examining dispersions

Threshold P-value
0.006679732

Generating ROC curve and AUC statistics...

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

Estimating ERCC LODR
........................................
  Ratio LODR Estimate 90% CI Lower Bound 90% CI Upper Bound
1   4:1            25                 19                 31
3 1:1.5           Inf               <NA>               <NA>
4   1:2           240                120                340

Estimating Sim LODR
........................................
  Ratio LODR Estimate 90% CI Lower Bound 90% CI Upper Bound
1   4:1            35                 23                 45
3 1:1.5           Inf               <NA>               <NA>
4   1:2           Inf               <NA>               <NA>

 ERCC LODR estimates are available
   Fold Ratio Count Log2Count_normalized
1 4.000   4:1    25           -0.7460655
2 1.000   1:1    NA                   NA
3 0.667 1:1.5   Inf                  Inf
4 0.500   1:2   240            2.5169689

LODR estimates are available to code ratio-abundance plot

Saving main dashboard plots to pdf file...
Saving expDat list to .RData file...
Analysis completed.
```

1.3 Results of dashboard analysis

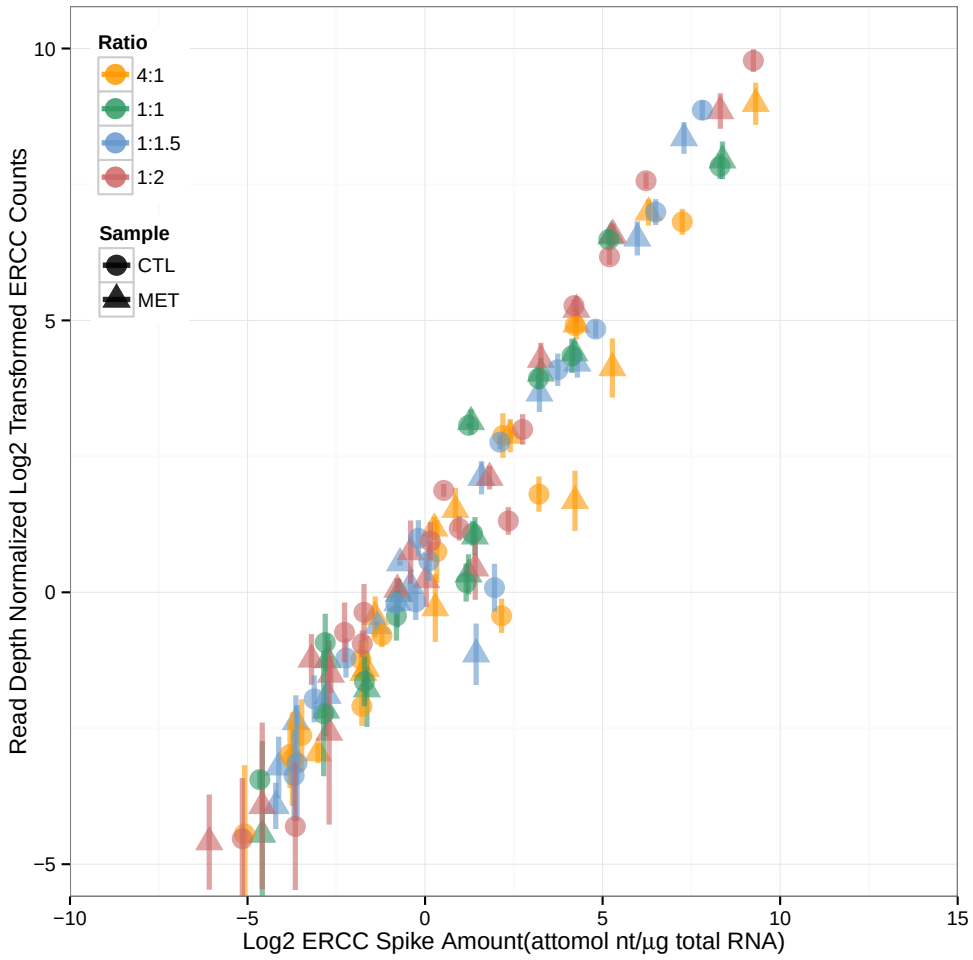
Look at the structure of expDat. The summary function will give a top level view of the list structure. The str function will give more detail. It is a good idea to set the max.level argument in the str function, because

by the end of the analysis the expDat structure is quite large.

```
> summary(expDat)
              Length Class      Mode
sampleInfo    11     -none-     list
plotInfo       8     -none-     list
erccInfo       4     -none-     list
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
mnLibeFactor   1     -none-     numeric
spikeFraction  1     -none-     numeric
idColsAdj      6     data.frame list
Results       14     -none-     list
Figures       13     -none-     list
```

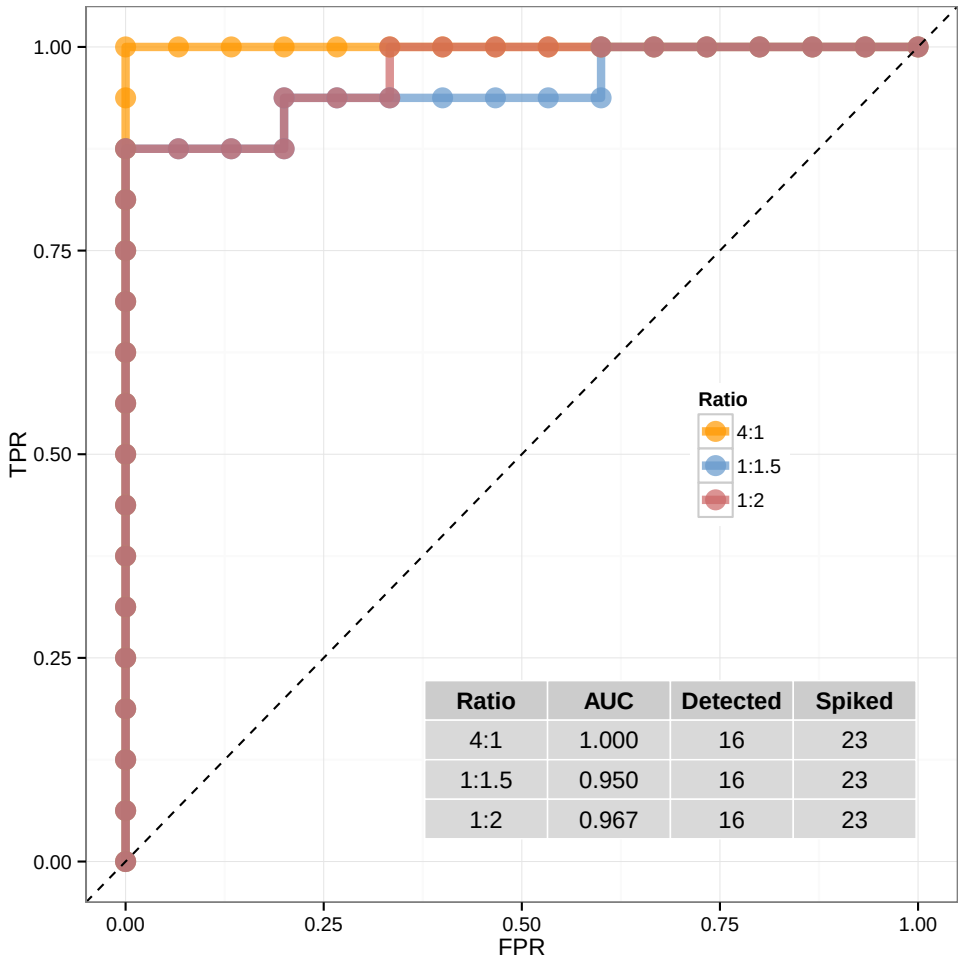
The figures from the analysis are stored in `expDat$Figures`. The four main diagnostic figures that are saved to pdf are the dynRangePlot, rocPlot, lodrERCCPlot, and maPlot.

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



For this particular experiment the relationship between abundance and signal for the ERCC controls show that the measurement results span a 2^{15} dynamic range. These ERCC mixtures were designed to span a 2^{20} dynamic range, but there was insufficient evidence to reliably quantify ERCC transcripts at low abundances.

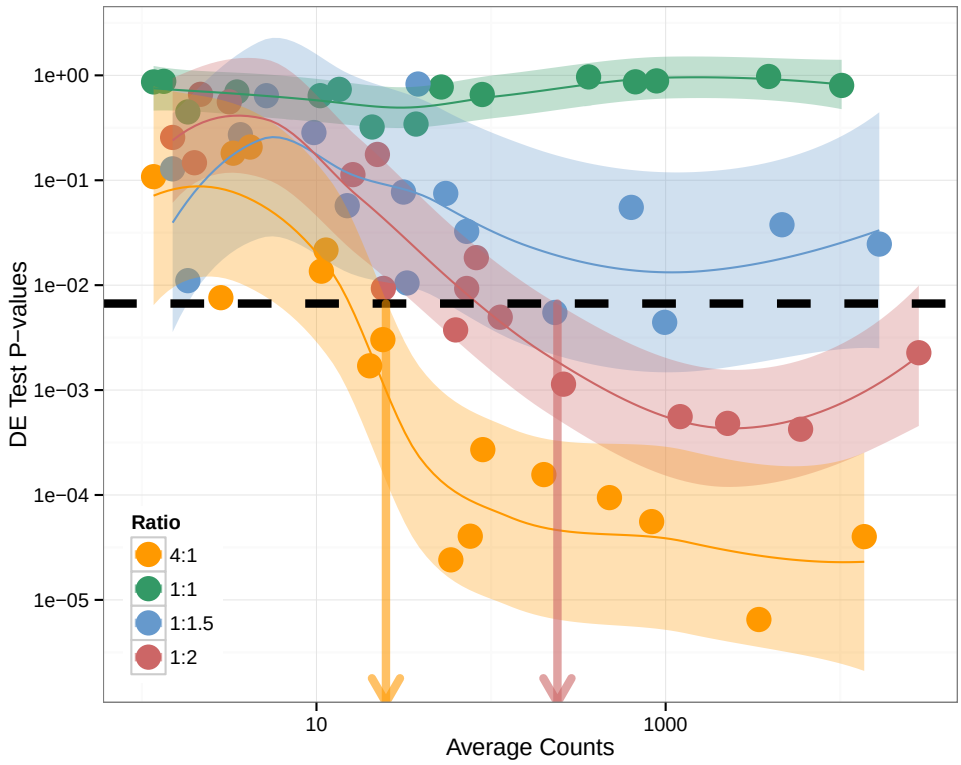
```
> expDat$Figures$rocPlot
```



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

The receiver operator characteristic (ROC) curve and the Area Under the Curve (AUC) statistic provide evidence of the diagnostic power for detecting differential expression in this rat toxicogenomics experiment. As expected with increased fold change, diagnostic power increases. The AUC summary statistic for different experiments can be used to compare diagnostic performance.

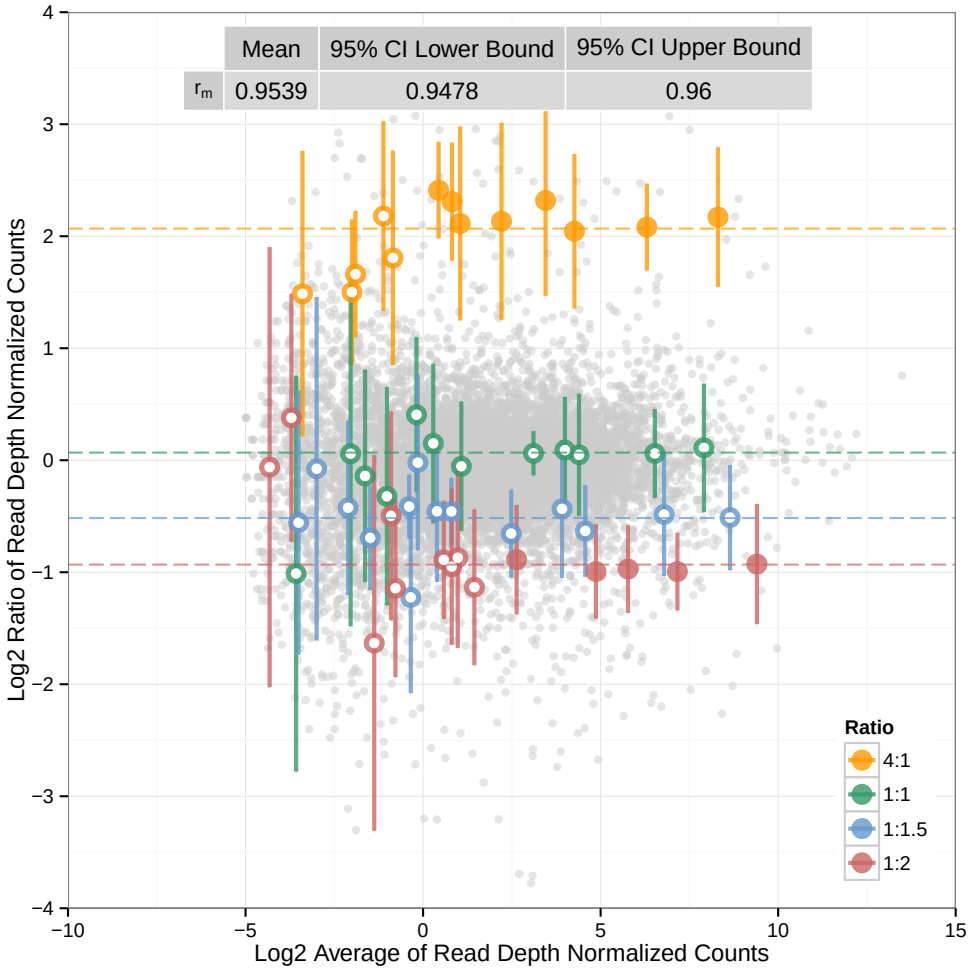
```
> expDat$Figures$lodrERCCPlot
```



Ratio	LODR Estimate	90% CI Lower Bound	90% CI Upper Bound
4:1	25	19	31
1:1.5	Inf	NA	NA
1:2	240	120	340

By modeling the relationship between average signal and p-values we can obtain Limit of Detection of Ratios (LODR) estimates for each differential fold change (or Ratio,indicated by color) and a threshold p-value, p.thresh, indicated by the dotted black line. LODR values can be compared between experiments to evaluate the ability to detect differences between samples as a function of transcript abundance.

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



An MA plot (Ratio of Signals vs Average Signals) shows the ratio measurements of transcripts in the pair of samples as a function of abundance. The ERCC control ratios measurements are coded to indicate which controls are above a given LODR (solid circles) or below the LODR (open circles). This plot also shows the v ariability in ratio measurements as a function of dynamic range and the bias in control ratio measurements (r_m), which is influenced by the mRNA fraction difference between the pair of samples.

2 SEQC Reference RNA (UHRR and HBRR) Examples

All UHRR and HBRR experiments for the SEQC interlaboratory and interplatform analysis were performed with aliquots from single large samples of UHRR and HBRR. The count data and microarray data provided here are from experiments using the same original UHRR and HBRR samples.

2.1 RNA-Seq UHRR vs. HBRR experiment analysis

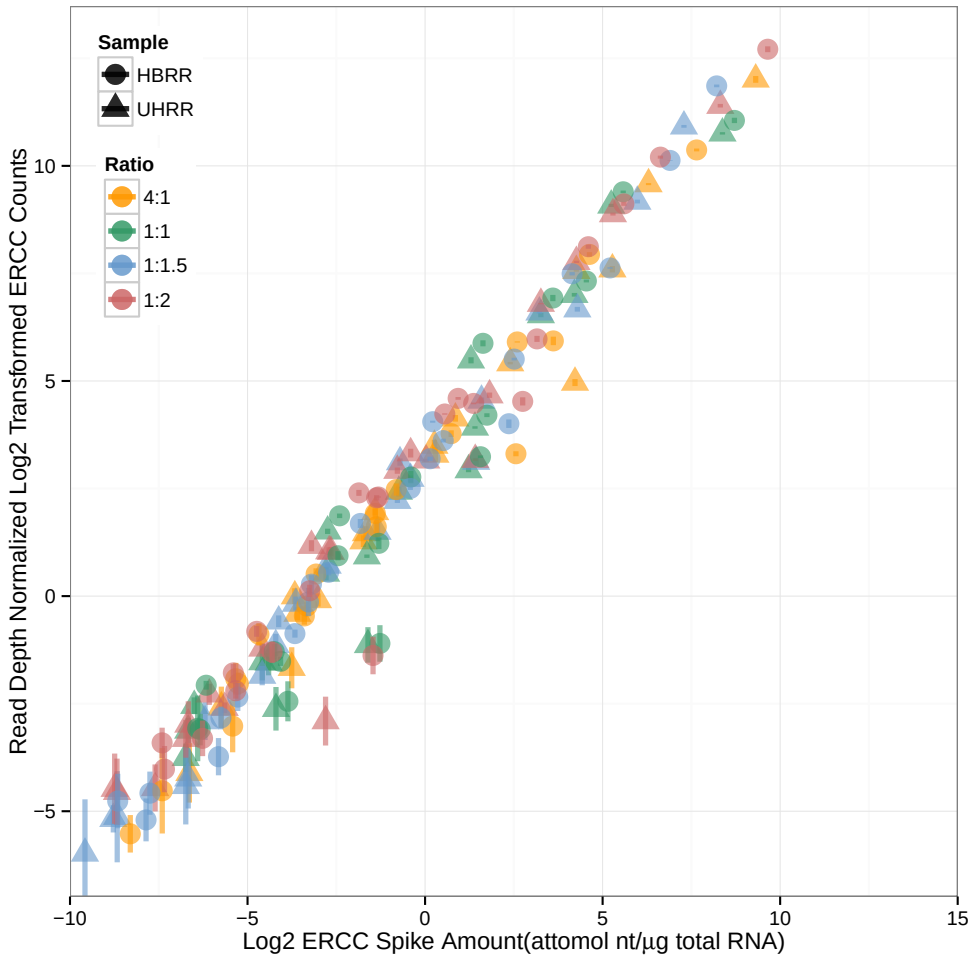
```
> expDat <- runDashboard(datType="count", expTable=Lab5.ILM.UHRR.HBRR.countTable,
                         repNormFactor=Lab5.ILM.UHRR.HBRR.totalReads,
                         filenameRoot="Lab5", sample1Name = "UHRR",
                         sample2Name = "HBRR", erccmix = "RatioPair",
                         erccdilution = 1, spikeVol = 50,
                         totalRNAmass = 2.5*10^(3), choseFDR = 0.01)
Initializing the expDat list structure...
choseFDR = 0.01
Filename root is: Lab5.UHRR.HBRR

Transcripts were removed with a mean count < 1 or more than 2
replicates with 0 counts.
Original data contained  43919 transcripts.
After filtering  39844 transcripts remain for  analysis.
A total of 2 out of 92
ERCC controls were filtered from the data set
The excluded ERCCs are:
ERCC-00057 ERCC-00083

Library sizes:
138.7869 256.0065 199.4683 431.9338 247.9856 219.3833 251.2658 257.5082
Check for sample mRNA fraction differences(r_m)...

log.offset
4.93294 5.545203 5.295655 6.068272 5.513371 5.39082 5.526511 5.551052

Number of ERCC Controls Used in r_m estimate
90

Outlier ERCCs for GLM r_m Estimate:
ERCC-00012 ERCC-00137 ERCC-00085 ERCC-00054 ERCC-00019

GLM log(r_m) estimate:
0.2332881

GLM log(r_m) estimate standard deviation:
0.002939789

GLM r_m estimate:
1.262745

GLM r_m upper limit
1.263703
```

```
GLM r_m lower limit
1.261788

Number of ERCCs in Mix 1 dyn range:  90

Number of ERCCs in Mix 2 dyn range:  90
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-00016 ERCC-00017 ERCC-00048 ERCC-00061 ERCC-00075
ERCC-00104 ERCC-00117 ERCC-00142 ERCC-00156 ERCC-00097
ERCC-00123 ERCC-00138


Saving dynRangePlot to expDat

Starting differential expression tests

Using Total Reads
138786892 256006510 199468322 431933806 247985592 219383270 251265814 257508210
Disp = 0.0014 , BCV = 0.0374
Disp = 0.00137 , BCV = 0.0369
[1] "Analyzing Gene # 2"
[1] "Analyzing Gene # 10"
[1] "Analyzing Gene # 100"
[1] "Analyzing Gene # 500"
[1] "Analyzing Gene # 1000"
[1] "Analyzing Gene # 2500"
[1] "Analyzing Gene # 5000"
[1] "Analyzing Gene # 10000"
[1] "Analyzing Gene # 15000"
[1] "Analyzing Gene # 20000"
[1] "Analyzing Gene # 25000"
[1] "Analyzing Gene # 30000"
[1] "Analyzing Gene # 35000"
[1] "Analyzing Gene # 40000"
[1] "Analyzing Gene # 2"
[1] "Analyzing Gene # 10"
[1] "Analyzing Gene # 100"
[1] "Analyzing Gene # 500"
[1] "Analyzing Gene # 1000"
[1] "Analyzing Gene # 2500"
[1] "Analyzing Gene # 5000"
[1] "Analyzing Gene # 10000"
[1] "Analyzing Gene # 15000"
[1] "Analyzing Gene # 20000"
[1] "Analyzing Gene # 25000"
[1] "Analyzing Gene # 30000"
[1] "Analyzing Gene # 35000"
```

```
[1] "Analyzing Gene # 40000"
[1] "Note: 'test.mat' not provided. Comparing each model \nfrom 'design.list' to first model in 'desig
[1] "Spline scaling factor: 1.1276452544631"
[1] "Spline scaling factor: 1.12762782588029"
[1] "Analyzing Gene # 2"
[1] "Analyzing Gene # 10"
[1] "Analyzing Gene # 2"
[1] "Analyzing Gene # 10"
[1] "Note: 'test.mat' not provided. Comparing each model \nfrom 'design.list' to first model in 'desig
[1] "Spline scaling factor: 1.12762782588029"
[1] "Finished DE testing"
[1] "Spline scaling factor: 1.12762782588029"

Finished examining dispersions

Threshold P-value
0.1498868
Threshold P-value is high for the chosen FDR of  0.01
The sample comparison indicates a large amount of
 differential expression in the measured transcript
 populations

Generating ROC curve and AUC statistics...

Area Under the Curve (AUC) Results:
  Ratio   AUC Detected Spiked
1   4:1 0.968       22     23
2 1:1.5 0.834       22     23
3   1:2 0.902       23     23

Estimating ERCC LODR
........................................
  Ratio LODR Estimate 90% CI Lower Bound 90% CI Upper Bound
1   4:1            17                5.9                 21
3 1:1.5           100                 57                120
4   1:2            43                 26                 49

Estimating Sim LODR
........................................
  Ratio LODR Estimate 90% CI Lower Bound 90% CI Upper Bound
1   4:1           5.3                3.4                6.9
3 1:1.5            86                 57                110
4   1:2            31                 22                 37

 ERCC LODR estimates are available
   Fold Ratio Count Log2Count_normalized
1 4.000   4:1    17            -3.880007
2 1.000   1:1    NA                   NA
3 0.667 1:1.5   100            -1.323614
4 0.500   1:2    43            -2.541205
```

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

Saving main dashboard plots to pdf file...
Saving expDat list to .RData file...
Analysis completed.
```

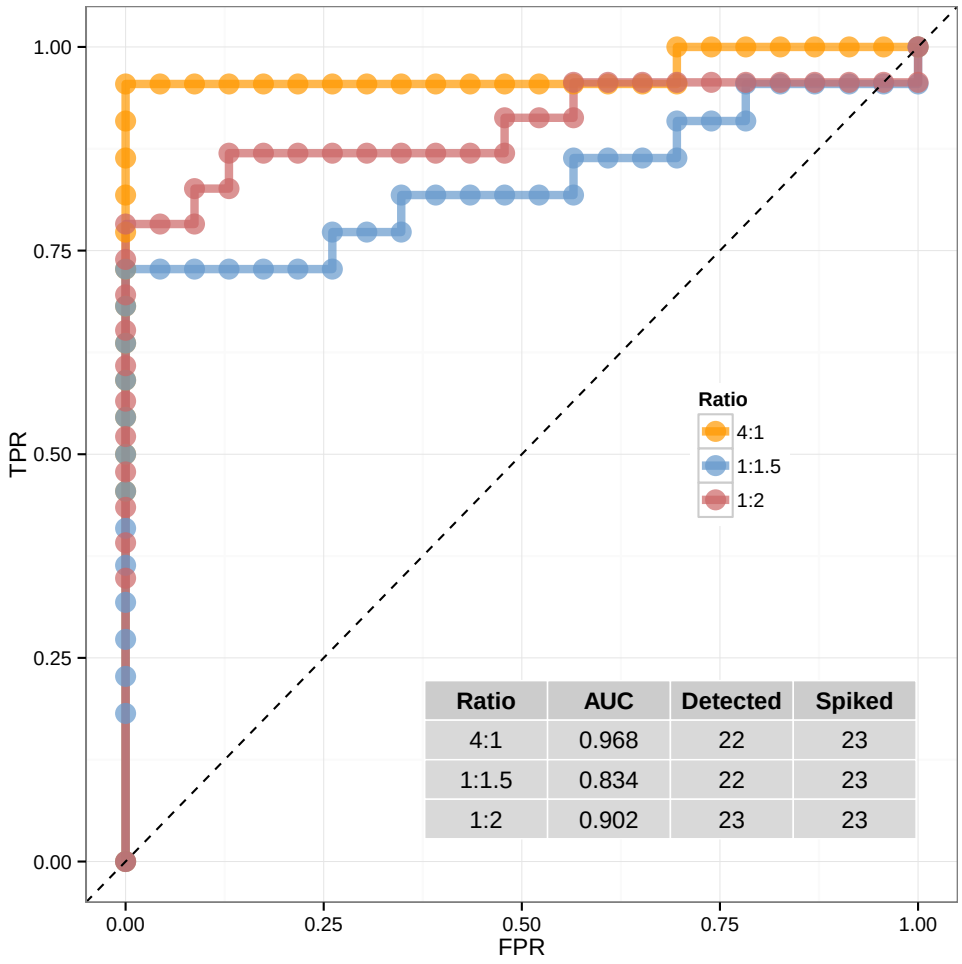
```
> expDat$Figures$dynRangePlot
```



Compared with the rat toxicogenomics experiment the UHRR/HBRR experiment from Lab 5 captures the full dynamic range of 2^{20} in the ERCC mixture design. This difference can be attributed to an much greater sequencing depth in the UHRR/HBRR experiments at each laboratory (full fluidic capacity of an instrument) compared to the MET-CTL rat toxicogenomics experiment. This difference is apparent in the total reads vectors for the two different experiments:

```
> COH.RatTox.ILM.MET.CTL.totalReads
[1] 41423502 46016148 44320280 38400362 47511484 33910098
> Lab5.ILM.UHRR.HBRR.totalReads
[1] 138786892 256006510 199468322 431933806 247985592
[6] 219383270 251265814 257508210
```

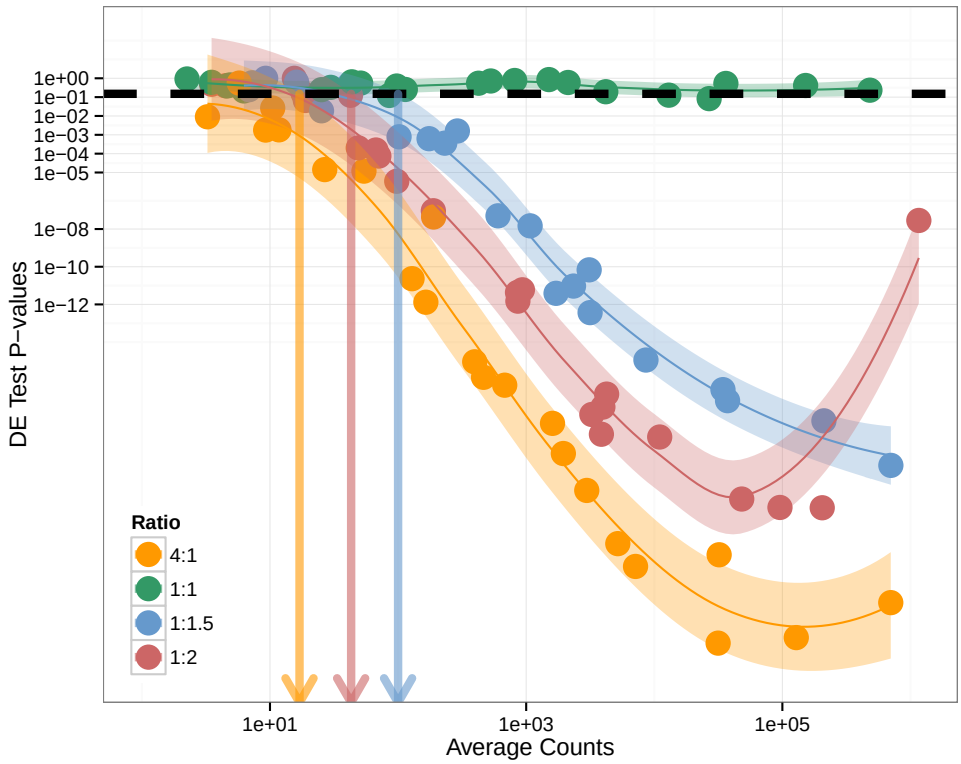
```
> expDat$Figures$rocPlot
```



Ratio	AUC	Detected	Spiked
4:1	0.968	22	23
1:1.5	0.834	22	23
1:2	0.902	23	23

Given the sequencing depth of the Lab 5 UHRR/HBRR experiment it is unsurprising that the ROC curve results show that almost all of the spiked ERCC transcripts were detected and tested for differential expression. At first glance, one may attempt to compare these ROC curve and AUC results to the rat toxicogenomics results, but this is ill-advised because the samples under comparison are different. Not only is the sequencing depth different for the two experiments, but the UHRR/HBRR sample transcripts also have high levels of differential expression, whereas the rat MET/CTL samples have much lower levels of differential expression. This difference in differential expression is shown in the spread and density of the points representing the endogenous transcripts in the MA plots for each experiment.

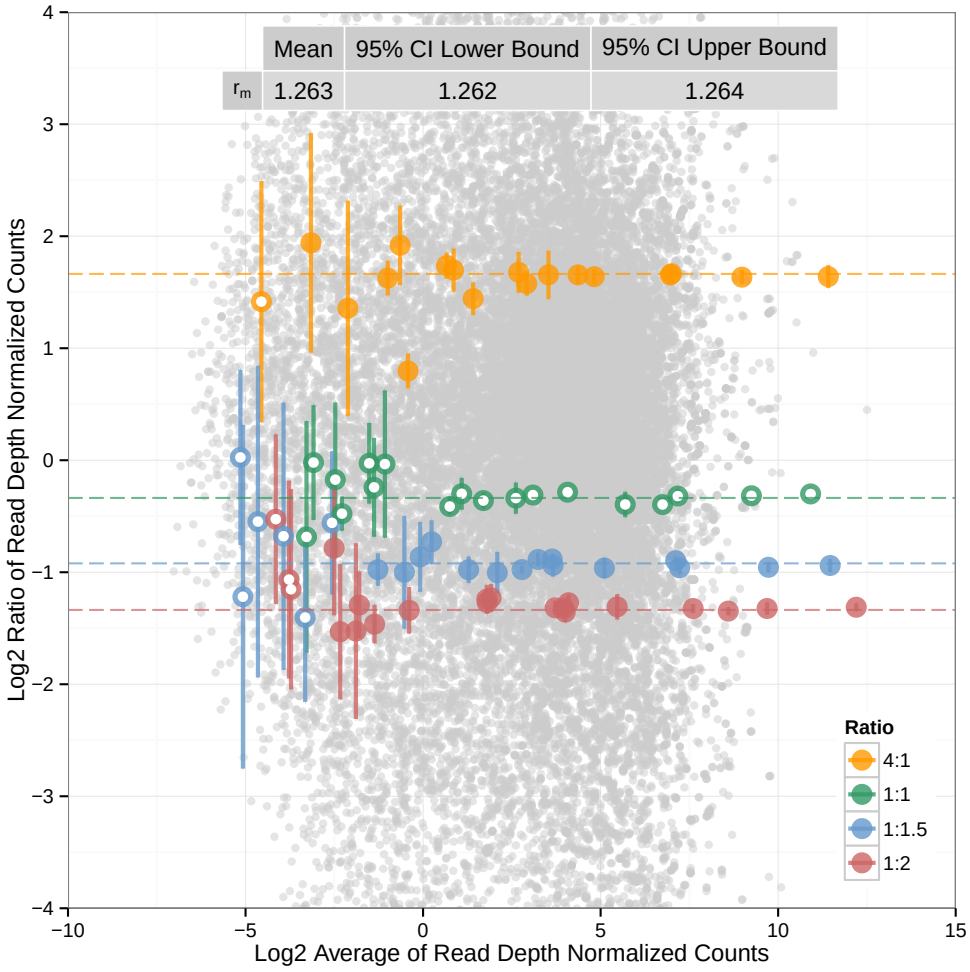
```
> expDat$Figures$lodrERCCPlot
```



Ratio	LODR Estimate	90% CI Lower Bound	90% CI Upper Bound
4:1	17	5.9	21
1:1.5	100	57	120
1:2	43	26	49

Much lower LODR estimates are found in the Lab 5 UHRR/HBRR experiment compared to the rat experiment. One can also see that the p-values in this analysis are much smaller than those in the rat experiment due to the large amount of differentially expressed transcripts in the UHRR/HBRR comparison.

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



Variability in the ratio measurements in this experiment is much lower than in the rat experiment. This is due to the nature of the replication in the two experiments. In the rat experiment the 3 replicates for each condition represent biological replicates, while in the UHRR/HBRR experiment the 4 replicates represent library preparation replicates of the same source samples and each library replicate count column was a sum of multiple fluidic replicates of the same library replicate. The bias in control ratios in this experiment is much greater than what was observed in the rat experiment. This bias is likely due to the known mRNA fraction difference between the UHRR and HBRR RNA samples.

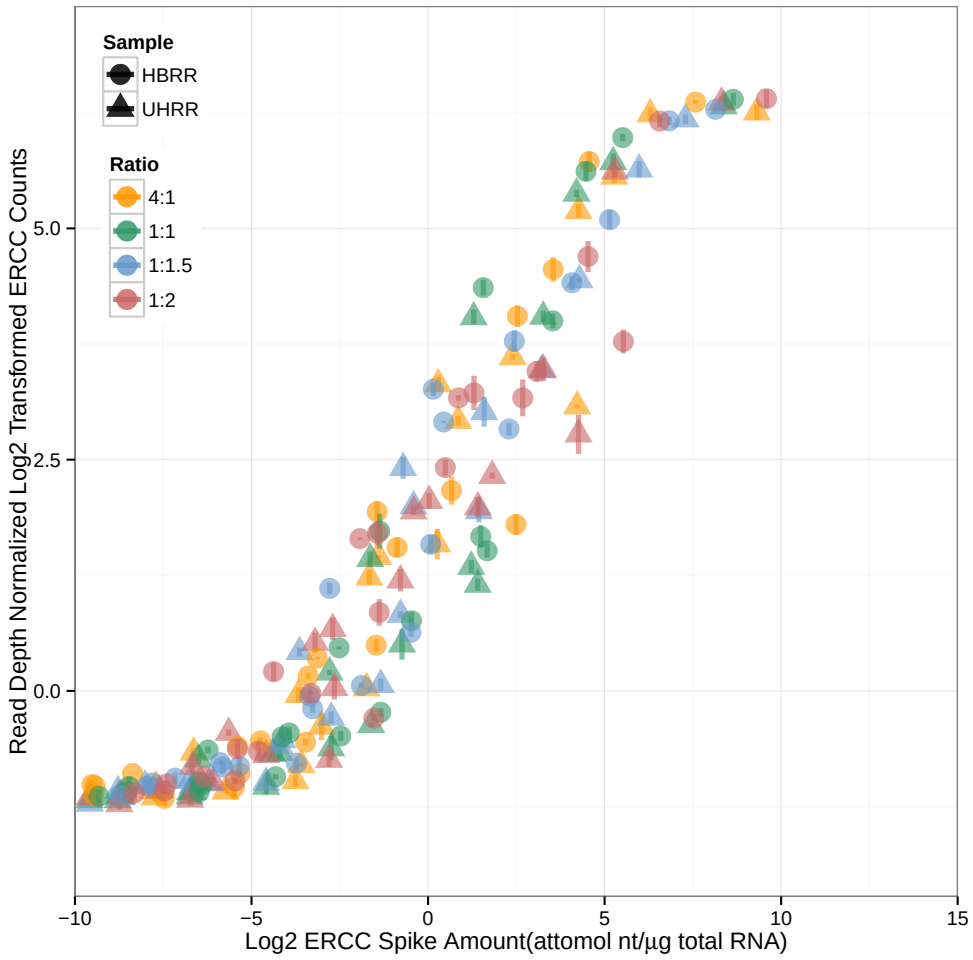
2.2 Microarray UHRR vs. HBRR experiment analysis

Unnormalized fluorescent signals are expected for microarray data, the data will be log2 transformed for DE testing with limma. The repNormFactor variable should be NULL for microarray analysis, and repNormFactor is set to NULL if it is missing in the list of arguments for the runDashboard function. Each array replicate is normalized by the 75th quantile fluoresence intensity value of that array.

```
> expDat <- runDashboard(datType="array", expTable=COH.array.UHRR.HBRR,
                         repNormFactor=NULL,
                         filenameRoot = "COH.Array",
                         sample1Name = "UHRR", sample2Name="HBRR",
                         erccmix = "RatioPair", erccdilution = 1,
                         spikeVol = 50, totalRNAmass = 2.5*10^(3), choseFDR=0.01)
Initializing the expDat list structure...
choseFDR = 0.01
repNormFactor is NULL
Filename root is: COH.Array.UHRR.HBRR

Using 75th quantile intensity to normalize each array

Library sizes:
209.4 224.65 212.25 187.3 204.9 201.2
Check for sample mRNA fraction differences(r_m)...

log.offset
5.344246 5.414544 5.357765 5.232712 5.322522 5.304299

Number of ERCC Controls Used in r_m estimate
92

Outlier ERCCs for GLM r_m Estimate:
ERCC-00083 ERCC-00057 ERCC-00048 ERCC-00017 ERCC-00098
ERCC-00061 ERCC-00156 ERCC-00123 ERCC-00097 ERCC-00012
ERCC-00024 ERCC-00086 ERCC-00081 ERCC-00016 ERCC-00134
ERCC-00041 ERCC-00147 ERCC-00164 ERCC-00120 ERCC-00168
ERCC-00137 ERCC-00040 ERCC-00013 ERCC-00077 ERCC-00033
ERCC-00154 ERCC-00028 ERCC-00058 ERCC-00069 ERCC-00039
ERCC-00085 ERCC-00143 ERCC-00054 ERCC-00160 ERCC-00170
ERCC-00144 ERCC-00157 ERCC-00014 ERCC-00019 ERCC-00059
ERCC-00163 ERCC-00099 ERCC-00062 ERCC-00095 ERCC-00131
ERCC-00092 ERCC-00042 ERCC-00116 ERCC-00108 ERCC-00136
ERCC-00145 ERCC-00004 ERCC-00046 ERCC-00113 ERCC-00074
ERCC-00096 ERCC-00130 ERCC-00002

GLM log(r_m) estimate:
0.1867726

GLM log(r_m) estimate standard deviation:
0.005072296

GLM r_m estimate:
1.205353
```

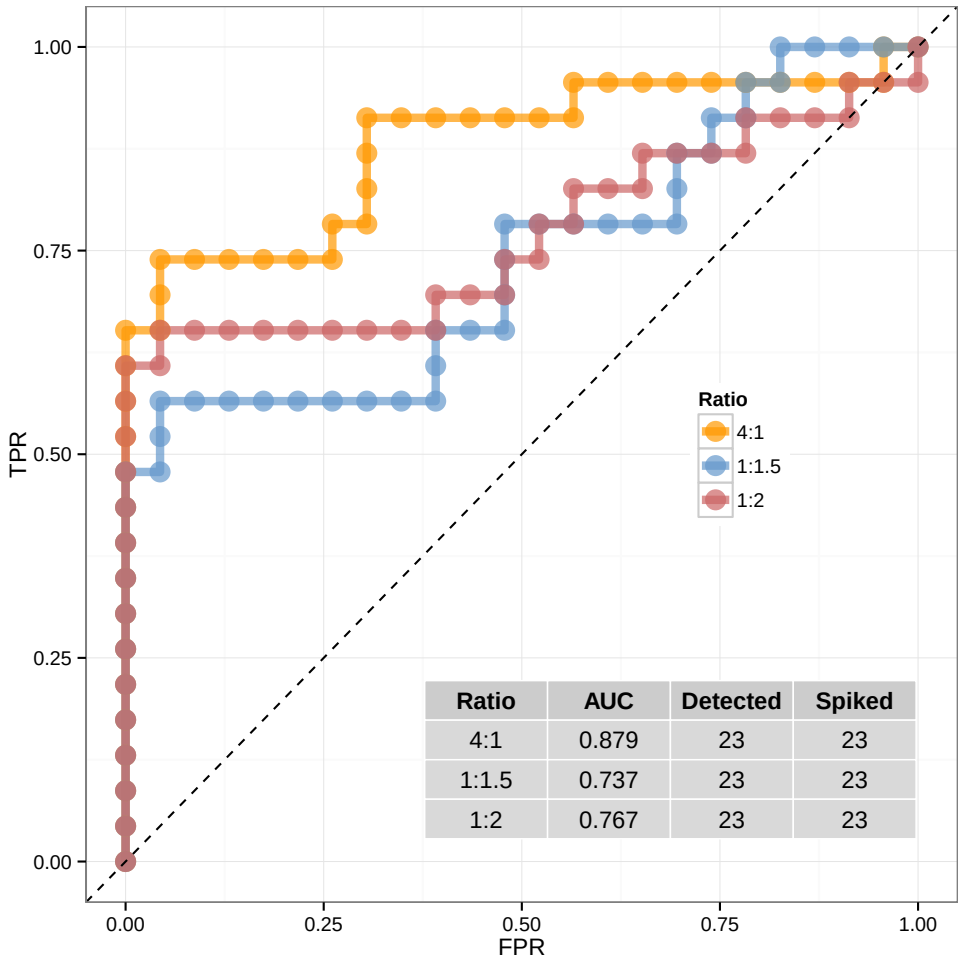
```
GLM r_m upper limit
1.206914

GLM r_m lower limit
1.203794

Number of ERCCs in Mix 1 dyn range:  92

Number of ERCCs in Mix 2 dyn range:  92

datType is array, no linear model fit for ERCC specific effects

rangeResidPlot is empty

Saving dynRangePlot to expDat

Finished DE testing

Generating ROC curve and AUC statistics...

Area Under the Curve (AUC) Results:
  Ratio   AUC Detected Spiked
1   4:1 0.879       23     23
2 1:1.5 0.737       23     23
3   1:2 0.767       23     23

Estimating ERCC LODR
.........................................
  Ratio LODR Estimate 90% CI Lower Bound 90% CI Upper Bound
1   4:1           130                120                140
3 1:1.5           220                180                250
4   1:2           170                150                190

COH.Array.UHRR.HBRR Sim Pvals.csv is missing, because simulated DE data was not generated
and tested for differential expression. Continuing with analysis...

 ERCC LODR estimates are available
   Fold Ratio Count Log2Count_normalized
1 4.000   4:1   130          -0.66844501
2 1.000   1:1    NA                   NA
3 0.667 1:1.5   220           0.09054689
4 0.500   1:2   170          -0.28142189

LODR estimates are available to code ratio-abundance plot

Saving main dashboard plots to pdf file...
Saving expDat list to .RData file...
Analysis completed.
```

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



As expected for a hybridization measurement technology all ERCC transcripts were quantified in this microarray experiment, but although the designed dynamic range is fully captured in this experiment the signal compression of high abundance transcripts and the noise floor for low abundance transcripts are apparent.

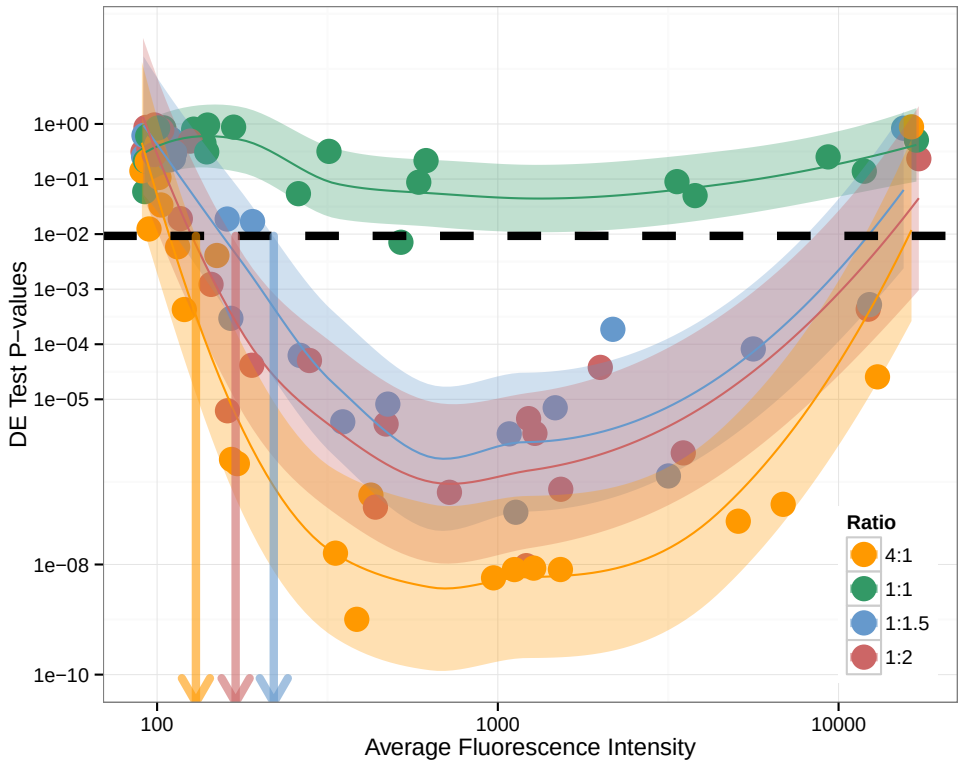
```
> expDat$Figures$rocPlot
```



Ratio	AUC	Detected	Spiked
4:1	0.879	23	23
1:1.5	0.737	23	23
1:2	0.767	23	23

Comparison of these ROC curves and AUC statistics to the results in the RNA-Seq UHRR/HBRR experiment from Lab 5 clearly show improvements in diagnostic power for the Lab 5 data set compared to this microarray data set. Although the relative cost of running microarrays compared with using the full fluidic capacity of a sequencing instrument should be considered in this comparison and at lower sequencing depths diagnostic performance between the two technologies would likely be more comparable.

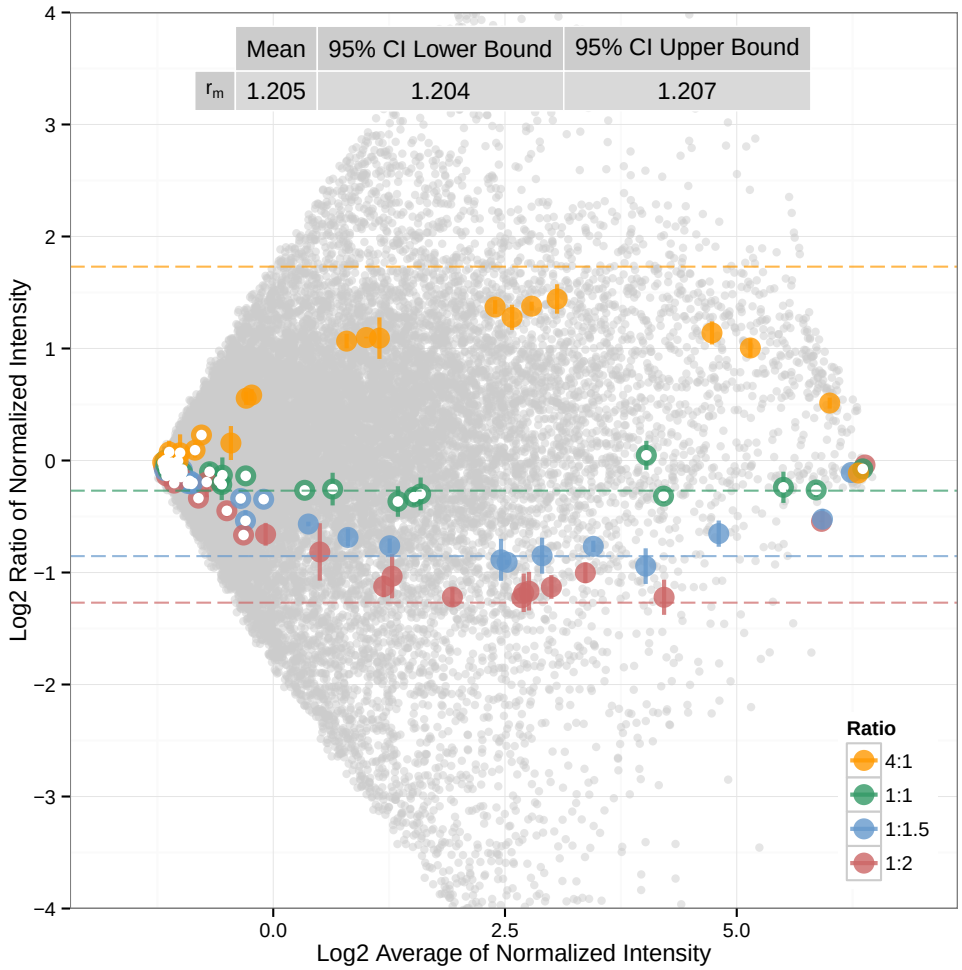
```
> expDat$Figures$lodrERCCPlot
```



Ratio	LODR Estimate	90% CI Lower Bound	90% CI Upper Bound
4:1	130	120	140
1:1.5	220	180	250
1:2	170	150	190

LODR estimates for microarray data are derived from the intersection with the threshold p-value at the lowest signal. The same FDR was chosed for the RNA-Seq and microarray measurements of the UHRR/HBRR samples, it is interesting to note that different threshold p-values were derived for each type of experiment using the same FDR = 0.01, p.thresh = 0.1512 for RNA-Seq at Lab 5 and p.thresh = 0.0093 for the microarray experiment.

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



The variability and bias in the control ratio measurements in the microarray data measurement of UHRR/HBRR are similar to the observations in the RNA-Seq UHRR/HBRR experiment.

3 Comparison of Performance Between Experiments

The performance metrics provided here derived from measurements of ERCC ratios in gene expression experiments (AUC, LODR, r_m, and the standard deviations of the ERCC ratio measurements) can be used to assess performance between experiments within the same laboratory, or between different laboratories or technology platforms.

4 Analysis Details: Advanced Use of Functions in runDashboard

The analysis functions contained in the convenience wrapper function `runDashboard` can also be used directly by the user. Comments are provided above each analysis step included in `runDashboard` to describe the purpose and ordering constraints. View the runDashboard script to see comments describing the analysis functions and the ordering constraints:

```
> runDashboard
function(datType=NULL, expTable=NULL, repNormFactor=NULL,
                         filenameRoot = NULL,
                         sample1Name = NULL,sample2Name = NULL,
                         erccmix = "RatioPair", erccdilution = 1,
                         spikeVol = 1, totalRNAmass = 1,choseFDR = 0.05,
                         userMixFile=NULL){

  # Initialize expDat structure
  # Required for all subsequent functions
  expDat <- initDat(datType=datType, expTable=expTable,
                    repNormFactor=repNormFactor, filenameRoot=filenameRoot,
                    sample1Name=sample1Name, sample2Name=sample2Name,
                    erccmix=erccmix, erccdilution=erccdilution,
                    spikeVol=spikeVol, totalRNAmass=totalRNAmass,
                    choseFDR=choseFDR,userMixFile=userMixFile)

  # Estimate the difference in mRNA fraction of total RNA for the two samples
  # Required for all subsequent functions
  expDat <- est_r_m(expDat)

  # Evaluate the dynamic range of the experiment (Signal-Abundance plot)
  # Not required for subsequent functions
  expDat <- dynRangePlot(expDat)

  # Evaluate pairwise ERCC control ratios within Mix1 and Mix2
  # Not required for subsequent functions
  if(erccmix == "RatioPair"){
    expDat <- withinMixRatios(expDat)
  }


  # Test for differential expression between samples
  # Required for all subsequent functions
  expDat <- geneExprTest(expDat)

  # Generate ROC curves and AUC statistics
  # Not Required for subsequent functions
  expDat <- erccROC(expDat)

  # Estimate LODR for ERCC controls
  # Required for subsequent functions
  expDat = estLODR(expDat,kind = "ERCC", prob=0.9)

  # Estimate LODR using Simulated data from endogenous transcripts
  # Not required for subsequent functions
  expDat = estLODR(expDat,kind = "Sim", prob=0.9)

  # Generate MA plot (Ratio vs. Average Signal) with ERCC controls below LODR
  #   annotated also flags possible False Negatives on DE gene list based on LODR
  #   threshold from DE gene list
```

```
  # Not required for subsequent functions
  expDat <- annotLODR(expDat)


  ### Saving plots and results
  # Convenience function to save 4 main figures to PDF
  saveERCCPlots(expDat)

  # Save expDat to a RData file for later use
  cat("\nSaving expDat list to .RData file...")
  nam <- paste(expDat$sampleInfo$filenameRoot, "expDat",sep = ".")
  assign(nam,expDat)

  to.save <- ls()

  save(list = to.save[grepl(pattern = nam,x=to.save)],
       file=paste0(expDat$sampleInfo$filenameRoot,".RData"))

  # End analysis and return expDat to global environ. / workspace
  cat("\nAnalysis completed.")
  return(expDat)

}
<environment: namespace:erccdashboard>
```

4.1 Flexibility in Differential Expression Testing

The `geneExprTest` function wraps the QuasiSeq differential expression testing package for datType = "count" or uses the limma package for differential expression testing if datType = "array". The function uses the DE testing p-value results and choseFDR parameter to select a threshold p-value for LODR estimation. In the case of count data 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 columns with names "Feature", "pvals", and "qvals" must be in the file.

4.2 Options for LODR Estimation

The default behavior of runDashboard is to use the `estLODR` function to obtain an LODR estimate using empirical data from the ERCCs and a model-based simulation using the endogenous genes in the sample. The type of LODR estimation is selected using the argument kind in the `estLODR` function. The other parameter that may be adjusted is the probability for the LODR estimate, in the default analysis prob = 0.9 is selected.

4.3 Options for Printing Plots to File

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.

To print 4 plots from manuscript to a single page pdf file use plotsPerPg = "manuscript" in the `saveERCCPlots` function or to create a multiple page pdf of all plots produced use plotsPerPg = "single".

4.4 Analysis of 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 control 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 (`initDat`, `est_r_m`, and `dynRangePlot`. 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.

5 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 3.0.2 with the following session information.

```
> sessionInfo()
R version 3.0.2 (2013-09-25)
Platform: x86_64-apple-darwin10.8.0 (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] splines   grid      stats     graphics  grDevices
[6] utils     datasets  methods   base

other attached packages:
 [1] ROCR_1.0-5          gplots_2.12.1
 [3] edgeR_3.4.2         limma_3.18.12
 [5] QuasiSeq_1.0-3      MASS_7.3-29
 [7] reshape2_1.2.2      erccdashboard_0.9.5
 [9] gridExtra_0.9.1     ggplot2_0.9.3.1

loaded via a namespace (and not attached):
 [1] bitops_1.0-6       caTools_1.16
 [3] colorspace_1.2-4   dichromat_2.0-0
 [5] digest_0.6.4       gdata_2.13.2
 [7] gtable_0.1.2       gtools_3.3.0
 [9] KernSmooth_2.23-10 labeling_0.2
[11] lattice_0.20-24    locfit_1.5-9.1
[13] Matrix_1.1-2       mgcv_1.7-28
[15] munsell_0.4.2      nlme_3.1-113
[17] plyr_1.8           proto_0.3-10
[19] qvalue_1.36.0      RColorBrewer_1.0-5
```

```
[21] scales_0.2.3       stringr_0.6.2
[23] tcltk_3.0.2        tools_3.0.2
```