EDASeq: Exploratory Data Analysis and Normalization for RNA-Seq

Davide Risso

Modified: October 7, 2011. Compiled: March 30, 2012

1 Introduction

In this document, we show how to conduct Exploratory Data Analysis (EDA) and normalization for a typical RNA-Seq experiment using the package *EDASeq*.

One can think of EDA for RNA-Seq as a two-step process: "read-level" EDA helps in discovering lanes with low sequencing depths, quality issues, and unusual nucleotide frequencies, while "gene-level" EDA can capture mislabeled lanes, issues with distributional assumptions (e.g., over-dispersion), and GC-content bias.

The package also implements both "within-lane" and "between-lane" normalization procedures, to account, respectively, for within-lane gene-specific (and possibly lane-specific) effects on read counts (e.g., related to gene length or GC-content) and for between-lane distributional differences in read counts (e.g., sequencing depths).

To illustrate the functionality of the *EDASeq* package, we make use of the *Saccharomyces cerevisiae* RNA-Seq data from Lee et al. [1]. Briefly, a wild-type strain and three mutant strains were sequenced using the Solexa 1G Genome Analyzer. For each strain, there are two technical replicate lanes from the same library preparation. The reads were aligned using Bowtie [2] with unique mapping and allowing up to two mismatches.

The *leeBamViews* package provides a subset of the aligned reads in BAM format. In particular, only the reads mapped between bases 800,000 and 900,000 of chromosome XIII are considered. We use these reads to illustrate read-level EDA.

The *yeastRNASeq* package contains gene-level read counts for four lanes: the two replicates of the wild-type strain ("wt") and the two replicates of one of the mutant strains ("mut"). We use these data to illustrate gene-level EDA.

```
> require(EDASeq)
> require(yeastRNASeq)
> require(leeBamViews)
```

2 Reading in unaligned and aligned read data

Unaligned reads. Unaligned (unmapped) reads stored in FASTQ format may be managed via the class *FastqFileList* imported from *ShortRead*. Information related to the libraries sequenced in each lane can be stored in the `elementMetadata` slot of the *FastqFileList* object.

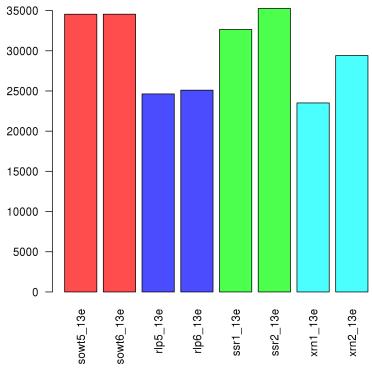
```
> files <- list.files(file.path(system.file(package = "yeastRNASeq"),
+                     "reads"), pattern = "fastq", full.names = TRUE)
> names(files) <- gsub("\\.fastq.*", "", basename(files))
> met <- DataFrame(conditions=c(rep("mut",2),rep("wt",2)),
+                  row.names=names(files))
> fastq <- FastqFileList(files)
> elementMetadata(fastq) <- met
> fastq

FastqFileList of length 4
names(4): mut_1_f mut_2_f wt_1_f wt_2_f
```

Aligned reads. The package can deal with aligned (mapped) reads in BAM format, using the class *BamFileList* from *Rsamtools*. Again, the `elementMetadata` slot can be used to store lane-level sample information.

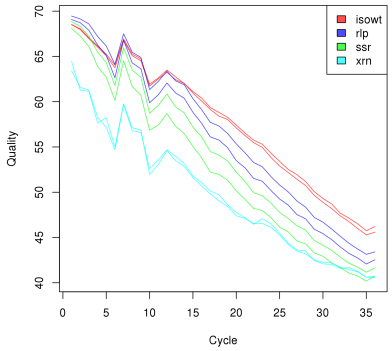
```
> files <- list.files(file.path(system.file(package = "leeBamViews"),
+                     "bam"), pattern = "bam$", full.names = TRUE)
> names(files) <- gsub("\\.bam", "", basename(files))
> gt <- gsub(".*/", "", files)
> gt <- gsub("_.*", "", gt)
> lane <- gsub(".*(.)$", "\\1", gt)
> geno <- gsub(".$", "", gt)
> pd <- DataFrame(geno=geno, lane=lane, row.names=paste(geno,lane,sep="."))
> bfs <- BamFileList(files)
> elementMetadata(bfs) <- pd
> bfs

BamFileList of length 8
names(8): isowt5_13e isowt6_13e ... xrn1_13e xrn2_13e
```

3 Read-level EDA

Numbers of unaligned and aligned reads. One important check for quality control is to look at the total number of reads produced in each lane, the number of reads mapped to a reference genome, and the percentage of mapped reads. A low total number of reads might be a symptom of low quality of the input RNA, while a low mapping percentage might indicate poor quality of the reads (low complexity) or problems with the reference genome.



(a) Number of mapped reads



(b) Mean per-base quality of mapped reads

Figure 1: Per-lane number of mapped reads and quality scores.

```
> colors <- c(rep(rgb(1,0,0,alpha=0.7),2),rep(rgb(0,0,1,alpha=0.7),2),
+             rep(rgb(0,1,0,alpha=0.7),2),rep(rgb(0,1,1,alpha=0.7),2))
> barplot(bfs,las=2,col=colors)
```

Figure 1(a), produced using the `barplot` method for the *BamFileList* class, displays the number of mapped reads for the subset of the yeast dataset included in the package *leeBamViews*. Unfortunately, *leeBamViews* does not provide unaligned reads, but barplots of the total number of reads can be obtained using the `barplot` method for the *FastqFileList* class. Analogously, one can plot the percentage of mapped reads with the `plot` method with signature *BamFileList*, *FastqFileList*. See the manual pages for details.

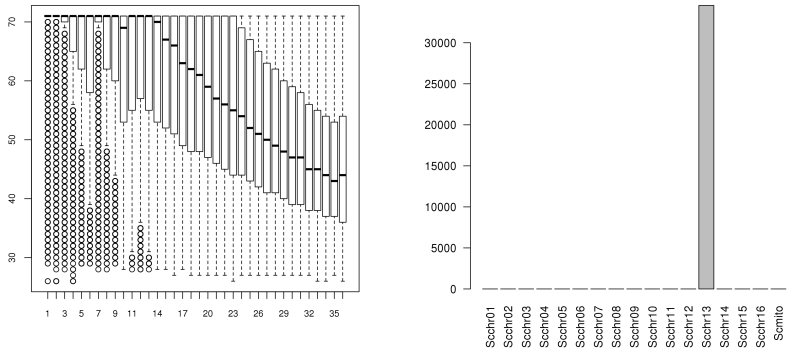
Read quality scores. As an additional quality check, one can plot the mean per-base (i.e., per-cycle) quality of the unmapped or mapped reads in every lane (Figure 1(b)).

```
> plotQuality(bfs,col=colors,lty=1)
> legend("topright",unique(elementMetadata(bfs)[,1]),
+        fill=unique(colors))
```

Individual lane summaries. If one is interested in looking more thoroughly at one lane, it is possible to display the per-base distribution of quality scores for each lane (Figure 2(a)) and the number of mapped reads stratified by chromosome (Figure 2(b)) or strand. As expected, all the reads are mapped to chromosome XIII.

```
> plotQuality(bfs[[1]],cex.axis=.8)
```

```
> barplot(bfs[[1]],las=2)
```



(a) Per-base quality of mapped reads

(b) Number of mapped reads per-chromosome

Figure 2: Quality scores and number of mapped reads for lane "isowt5_13e".

Read nucleotide distributions. A potential source of bias is related to the sequence composition of the reads. The function `plotNtFrequency` plots the per-base nucleotide frequencies for all the reads in a given lane (Figure 3).

```
> plotNtFrequency(bfs[[1]])
```

4 Gene-level EDA

Examining statistics and quality metrics at a read level can help in discovering problematic libraries or systematic biases in one or more lanes. Nevertheless, some biases can be difficult to detect at this scale and gene-level EDA is equally important.

Classes and methods for gene-level counts. There are several Bioconductor packages for aggregating reads over genes (or other genomic regions, such as, transcripts and exons) given a particular genome annotation, e.g., *IRanges*, *ShortRead*, *Genominator*, *Rsubread*. See their respective vignettes for details.

Here, we consider this step done and load the object `geneLevelData` from *yeastRNASeq*, which provides gene-level counts for 2 wild-type and 2 mutant lanes from the yeast dataset of Lee et al. [1] (see the *Genominator* vignette for an example on the same dataset).

```
> data(geneLevelData)
> head(geneLevelData)
```

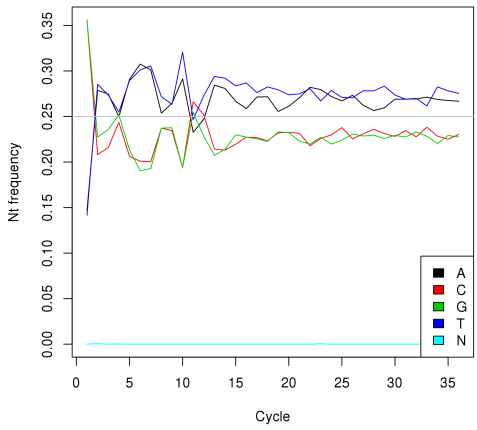


Figure 3: Per-base nucleotide frequencies of mapped reads for lane "isowt5_13e".

```
        mut_1 mut_2 wt_1 wt_2
YHR055C     0     0    0    0
YPR161C    38    39   35   34
YOL138C    31    33   40   26
YDR395W    55    52   47   47
YGR129W    29    26    5    5
YPR165W   189   180  151  180
```

Since it is useful to explore biases related to length and GC-content, the *EDASeq* package provides, for illustration purposes, length and GC-content for *S. cerevisiae* genes (based on SGD annotation [3]).

```
> data(yeastGC)
> head(yeastGC)

  YAL001C   YAL002W   YAL003W   YAL004W   YAL005C   YAL007C
0.3712317 0.3717647 0.4460548 0.4490741 0.4406428 0.3703704

> data(yeastLength)
> head(yeastLength)

YAL001C YAL002W YAL003W YAL004W YAL005C YAL007C
   3483    3825     621     648    1929     648
```

First, we filter the non-expressed genes, i.e., we consider only the genes with an average read count greater than 10 across the four lanes and for which we have length and GC-content information.

```
> filter <- apply(geneLevelData,1,function(x) mean(x)>10)
> table(filter)

filter
FALSE  TRUE
 1988  5077

> common <- intersect(names(yeastGC),rownames(geneLevelData[filter,]))
> length(common)

[1] 4994
```

This leaves us with 4994 genes.

The *EDASeq* package provides the *SeqExpressionSet* class to store gene counts, (lane-level) information on the sequenced libraries, and (gene-level) feature information. We use the data frame `met` created in Section 2 for the lane-level data. As for the feature data, we use gene length and GC-content.

```
> feature <- data.frame(gc=yeastGC,length=yeastLength)
> data <- newSeqExpressionSet(exprs=as.matrix(geneLevelData[common,]),
+                             featureData=feature[common,],
+                             phenoData=data.frame(
+                               conditions=c(rep("mut",2),rep("wt",2)),
+                               row.names=colnames(geneLevelData)))
> data

SeqExpressionSet (storageMode: lockedEnvironment)
assayData: 4994 features, 4 samples
  element names: exprs, offset
protocolData: none
phenoData
  sampleNames: mut_1 mut_2 wt_1 wt_2
  varLabels: conditions
  varMetadata: labelDescription
featureData
  featureNames: YAL001C YAL002W ... YPR201W (4994
    total)
  fvarLabels: gc length
  fvarMetadata: labelDescription
experimentData: use 'experimentData(object)'
Annotation:
```

Note that the row names of `exprs` and `featureData` must be the same; likewise for the row names of `phenoData` and the column names of `exprs`. As in the *ExpressionSet* class, the expression values can be accessed with `exprs`, the lane information with `pData`, and the feature information with `fData`.

```
> head(exprs(data))
```

```
        mut_1 mut_2 wt_1 wt_2
YAL001C    80    83   27   40
YAL002W    33    38   53   66
YAL003W  1887  1912  270  270
YAL004W    90   110  276  295
YAL005C   325   316  874  935
YAL007C    27    30   19   24
```

```
> pData(data)
```

```
      conditions
mut_1        mut
mut_2        mut
wt_1          wt
wt_2          wt
```

```
> head(fData(data))
```

```
               gc length
YAL001C 0.3712317   3483
YAL002W 0.3717647   3825
YAL003W 0.4460548    621
YAL004W 0.4490741    648
YAL005C 0.4406428   1929
YAL007C 0.3703704    648
```

The *SeqExpressionSet* class has an additional slot called `offset` (matrix of the same dimension as `exprs`), which may be used to store a normalization offset to be supplied to a model for read counts in differential expression analysis (see Section 5 and the vignette for *edgeR* for details on the role offsets). If not specified, the offset is initialized as a matrix of zeros.

```
> head(offst(data))
```

```
        mut_1 mut_2 wt_1 wt_2
YAL001C     0     0    0    0
YAL002W     0     0    0    0
YAL003W     0     0    0    0
YAL004W     0     0    0    0
YAL005C     0     0    0    0
YAL007C     0     0    0    0
```

Between-lane distribution of gene-level counts. One of the main considerations when dealing with gene-level counts is the difference in count distributions between lanes. The `boxplot` method provides an easy way to produce boxplots of the logarithms of the gene counts in each lane (Figure 4).

```
> boxplot(data,col=colors[1:4])
```

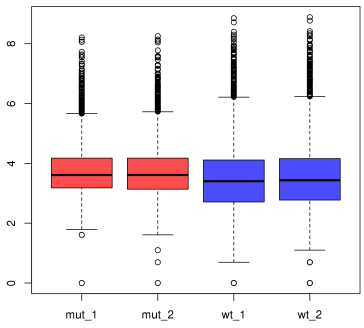


Figure 4: Between-lane distribution of gene-level counts (log).

The `MDPlot` method produces a mean-difference plot (MD-plot) of read counts for two lanes (Figure 5).

```
> MDPlot(data,c(1,3))
```

Over-dispersion. Although the Poisson distribution is a natural and simple way to model count data, it has the limitation of assuming equality of the mean and variance. For this reason, the negative binomial distribution has been proposed as an alternative when the data show over-dispersion. The function `meanVarPlot` can be used to check whether the count data are over-dispersed (for the Poisson distribution, one would expect the points in Figures 6(a) and 6(b) to be evenly scattered around the black line).

```
> meanVarPlot(data[,1:2],log=T)
```

```
> meanVarPlot(data,log=T)
```

Note that the mean-variance relationship should be examined within replicate lanes only (i.e., conditional on variables expected to contribute to differential expression). For the yeast dataset, it is not surprising to see no evidence of over-dispersion for the two mutant technical replicate lanes (Figure 6(a)); likewise for the two wild-type lanes. However, one expects over-dispersion in the presence of biological variability, as seen in Figure 6(b) when considering at once all four mutant and wild-type lanes [6, 4, 5].

Gene-specific effects on read counts. Several authors have reported selection biases related to sequence features such as gene length, GC-content, and mappability [4, 9, 7, 8].

In Figure 7, obtained using `biasPlot`, one can see the dependence of gene-level counts on GC-content. The same plot could be created for gene length or mappability instead of GC-content.

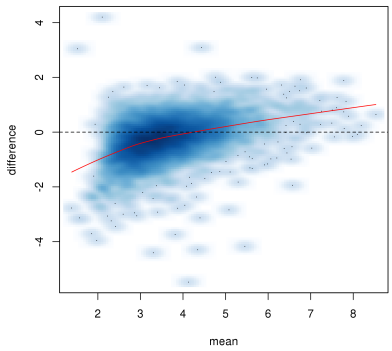
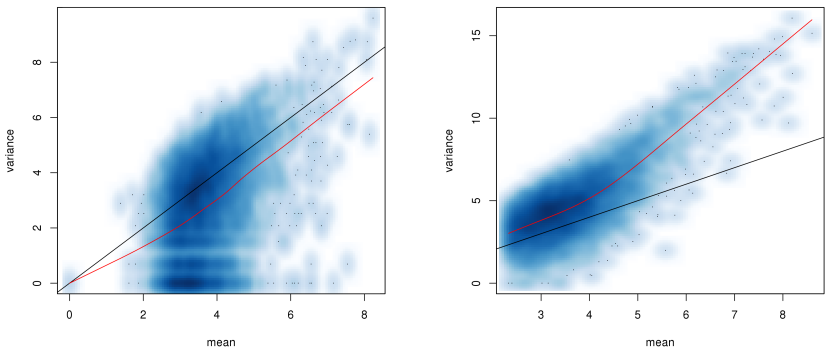


Figure 5: Mean-difference plot of the gene-level counts (log) of lanes "mut_1" and "wt_1".



(a) Mutant lanes

(b) All four lanes

Figure 6: Mean-variance relationship for the two mutant lanes and all four lanes: the black line corresponds to the Poisson distribution (variance equal to the mean), while the red curve is a lowess fit.

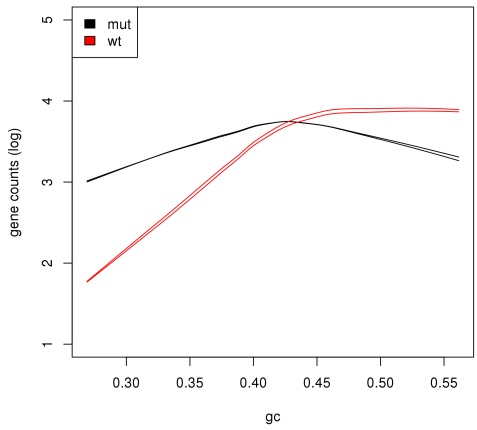


Figure 7: Lowess regression of the gene-level counts (log) on GC-content for each lane, color-coded by experimental condition.

```
> biasPlot(data,"gc",log=T,ylim=c(1,5))
```

To show that GC-content dependence can bias differential expression analysis, one can produce stratified boxplots of the log-fold-change of read counts from two lanes using the `biasBoxplot` method (Figure 8). Again, the same type of plots can be created for gene length or mappability.

```
> lfc <- log(exprs(data)[,3]+0.5)-log(exprs(data)[,1]+0.5)
> biasBoxplot(lfc,fData(data)$gc)
```

5 Normalization

Following Risso et al. [8], we consider two main types of effects on gene-level counts: (1) within-lane gene-specific (and possibly lane-specific) effects, e.g., related to gene length or GC-content, and (2) effects related to between-lane distributional differences, e.g., sequencing depth. Accordingly, `withinLaneNormalization` and `betweenLaneNormalization` adjust for the first and second type of effects, respectively. We recommend to normalize for within-lane effects prior to between-lane normalization.

We implemented four within-lane normalization methods, namely: loess robust local regression of read counts (log) on a gene feature such as GC-content (`loess`), global-scaling between feature strata using the median (`median`), global-scaling between feature strata using the upper-quartile (`upper`), and full-quantile normalization between feature strata (`full`). For a discussion of these methods in context of GC-content normalization see Risso et al. [8].

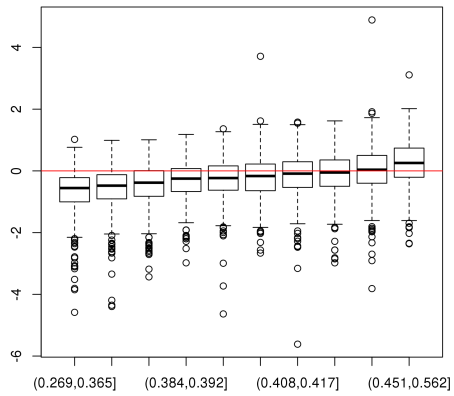


Figure 8: Boxplots of the log-fold-change between "mut_1" and "wt_1" lanes stratified by GC-content.

```
> dataWithin <- withinLaneNormalization(data,"gc",which="full")
> dataNorm <- betweenLaneNormalization(dataWithin,which="full")
```

Regarding between-lane normalization, the package implements three of the methods introduced in Bullard et al. [4]: global-scaling using the median (`median`), global-scaling using the upper-quartile (`upper`), and full-quantile normalization (`full`).

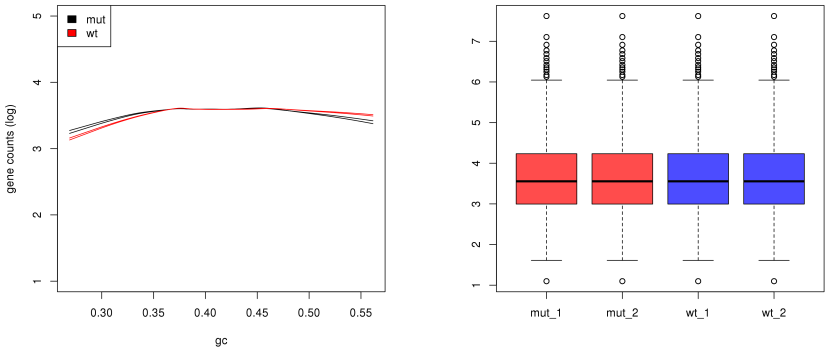
Figure 9 shows how after full-quantile within- and between-lane normalization,the GC-content bias is reduced and the distribution of the counts is the same in each lane.

```
> biasPlot(dataNorm,"gc",log=T,ylim=c(1,5))
```

```
> boxplot(dataNorm,col=colors)
```

Offset. Some authors have argued that it is better to leave the count data unchanged to preserve their sampling properties and instead use an offset for normalization purposes in the statistical model for read counts [6, 9, 5]. This can be achieved easily using the argument `offset` in both normalization functions.

```
> dataOffset <- withinLaneNormalization(data,"gc",
+                                       which="full",offset=TRUE)
> dataOffset <- betweenLaneNormalization(dataOffset,
+                                        which="full",offset=TRUE)
```



(a) GC-content (b) Count distribution

Figure 9: Full-quantile within- and between-lane normalization. (a) Lowess regression of normalized gene-level counts (log) on GC-content for each lane. (b) Between-lane distribution of normalized gene-level counts (log).

6 Differential expression analysis

One of the main applications of RNA-Seq is differential expression analysis. The normalized counts (or the original counts and the offset) obtained using the *EDASeq* package can be supplied to packages such as *edgeR* [5] or *DESeq* [6] to find differentially expressed genes. This section should be considered only as an illustration of the compatibility of the results of *EDASeq* with two of the most widely used packages for differential expression; our aim is not to compare differential expression strategies (e.g., normalized counts vs. offset).

6.1 edgeR

We can perform a differential expression analysis with *edgeR* based on the original counts by passing an offset to the generalized linear model. Here, we estimate a common dispersion parameter for all genes.

```
> library(edgeR)
> design <- model.matrix(~conditions,data=pData(data))
> disp <- estimateGLMCommonDisp(exprs(dataOffset),
+                               design, offst(dataOffset))
> fit <- glmFit(exprs(dataOffset),design,disp,offst(dataOffset))
> lrt <- glmLRT(exprs(dataOffset),fit,coef=2)
> head(lrt$table)

             logFC   logCPM          LR        PValue
YAL001C -1.4703669 25.05815  46.2021114  1.066620e-11
```

```
YAL002W -0.1016666 24.97245   0.1915494  6.616301e-01
YAL003W -2.6345550 30.97185 624.1390687 9.408699e-138
YAL004W  2.0406310 28.40104 213.9816983  1.858703e-48
YAL005C  2.1836077 30.42233 423.9484651  3.373072e-94
YAL007C -1.2308101 24.14651  16.6665308  4.456028e-05
```

6.2 DESeq

We can perform a differential expression analysis with *DESeq* based on the normalized counts by using the coerce method from the *SeqExpressionSet* class to the *CountDataSet* class of *DESeq*. When working with data that have been normalized for both within- and between-lane effects, we force the size factors to be one, since differences in lane sequencing depths have already been accounted for in our between-lane normalization. One could also consider only within-lane normalization and account for differences in sequencing depth by estimating the size factors using *DESeq*.

```
> library(DESeq)
> counts <- as(dataNorm,"CountDataSet")
> sizeFactors(counts) <- rep(1,4)
> counts <- estimateDispersions(counts)
> res <- nbinomTest(counts, "wt", "mut" )
> head(res)
```

7 SessionInfo

```
> toLatex(sessionInfo())
```

- R version 2.15.0 (2012-03-30), `x86_64-unknown-linux-gnu`
- Locale: `LC_CTYPE=en_US.UTF-8`, `LC_NUMERIC=C`, `LC_TIME=en_US.UTF-8`, `LC_COLLATE=C`, `LC_MONETARY=en_US.UTF-8`, `LC_MESSAGES=en_US.UTF-8`, `LC_PAPER=C`, `LC_NAME=C`, `LC_ADDRESS=C`, `LC_TELEPHONE=C`, `LC_MEASUREMENT=en_US.UTF-8`, `LC_IDENTIFICATION=C`
- Base packages: base, datasets, grDevices, graphics, methods, stats, utils
- Other packages: BSgenome 1.24.0, Biobase 2.16.0, BiocGenerics 0.2.0, Biostrings 2.24.0, EDASeq 1.2.0, GenomicRanges 1.8.0, IRanges 1.14.0, R.methodsS3 1.2.2, R.oo 1.9.3, RColorBrewer 1.0-5, Rsamtools 1.8.0, ShortRead 1.14.0, aroma.light 1.24.0, edgeR 2.6.0, lattice 0.20-6, latticeExtra 0.6-19, leeBamViews 0.99.17, limma 3.12.0, yeastRNASeq 0.0.5
- Loaded via a namespace (and not attached): AnnotationDbi 1.18.0, DBI 0.2-5, DESeq 1.8.0, KernSmooth 2.23-7, RSQLite 0.11.1, annotate 1.34.0, bitops 1.0-4.1, genefilter 1.38.0, geneplotter 1.34.0,

grid 2.15.0, hwriter 1.3, splines 2.15.0, stats4 2.15.0, survival 2.36-12, tools 2.15.0, xtable 1.7-0, zlibbioc 1.2.0

References

[1] A. Lee, K.D. Hansen, J. Bullard, S. Dudoit, and G. Sherlock. Novel low abundance and transient RNAs in yeast revealed by tiling microarrays and ultra high-throughput sequencing are not conserved across closely related yeast species. *PLoS Genet*, 4(12):e1000299, 2008.

[2] B. Langmead, C. Trapnell, M. Pop, and S.L. Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*, 10(3):R25, 2009.

[3] Saccharomyces Genome Database. http://www.yeastgenome.org, r64.

[4] J.H. Bullard, E. Purdom, K.D. Hansen, and S. Dudoit. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC bioinformatics*, 11(1):94, 2010.

[5] M.D. Robinson, D.J. McCarthy, and G.K. Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139, 2010.

[6] S. Anders and W. Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11(10):R106, 2010.

[7] A. Oshlack and M.J. Wakefield. Transcript length bias in RNA-seq data confounds systems biology. *Biology Direct*, 4(1):14, 2009.

[8] D. Risso, K. Schwartz, G. Sherlock, and S. Dudoit. GC-Content Normalization for RNA-Seq Data. Technical report #291, University of California, Berkeley, Division of Biostatistics, 2011. http://www.bepress.com/ucbbiostat/paper291/.

[9] K.D. Hansen, R.A. Irizarry, and Z. Wu. Removing technical variability in RNA-Seq data using conditional quantile normalization. Technical report #227, Johns Hopkins University, Dept. of Biostatistics Working Papers, 2011. http://www.bepress.com/jhubiostat/paper227/.