Regularized logarithm for sample clustering

Michael Love

October 14, 2014

1 Regularized logarithm of the Hammer dataset

To demonstrate the use of the regularized logarithm, we chose to use the Hammer *et al.* RNA-Seq dataset which was compiled by the authors of the ReCount project at http://bowtie-bio.sourceforge.net/recount/. We include the *eSet* object in the /data directory so that the vignettes can be built without a network connection. Below we contrast the rlog transformation with the log2 of normalized counts plus a pseudocount of 1.

```
library("DESeq2")
library("DESeq2paper")
library("Biobase")

## Welcome to Bioconductor
##
##     Vignettes contain introductory material; view with
##     ’browseVignettes()’.  To cite Bioconductor, see
##     ’citation("Biobase")’, and for packages ’citation("pkgname")’.

library("vsn")
```

```
# download Hammer count matrix from ReCount project site
# http://bowtie-bio.sourceforge.net/recount/ExpressionSets/hammer_eset.RData
data("hammer_eset")
e <- hammer.eset
pData(e)$Time <- as.character(pData(e)$Time)
pData(e)["SRX020105", "Time"] <- "2 months"
pData(e)$Time <- factor(pData(e)$Time)
dds <- DESeqDataSetFromMatrix(exprs(e), pData(e), ~1)
levels(colData(dds)$protocol) <- c("CTRL", "SNL")
levels(colData(dds)$Time) <- c("2mn", "2wk")
lab <- factor(with(colData(dds), paste(protocol, Time, sep = ":")))

dds <- dds[rowSums(counts(dds)) > 0, ]
dds <- estimateSizeFactors(dds)

log2m <- log2(counts(dds, normalized = TRUE) + 1)

rld <- rlogTransformation(dds)
rlogm <- assay(rld)
```

2 Plot

```r
library("RColorBrewer")
line <- 0.5
adj <- -0.3
cex <- 1.5
plotHclustColors <- function(matrix, labels, hang = 0.1, ...) {
    colnames(matrix) <- labels
    d <- dist(t(matrix))
    hc <- hclust(d)
    labelColors <- brewer.pal(nlevels(labels), "Paired")
    colLab <- function(n) {
        if (is.leaf(n)) {
            a <- attributes(n)
            labCol <- labelColors[which(levels(lab) == a$label)]
            attr(n, "nodePar") <- c(a$nodePar, lab.col = labCol, pch = NA)
        }
        n
    }
    clusDendro <- dendrapply(as.dendrogram(hc, hang = hang), colLab)
    plot(clusDendro, ...)
}

par(mfrow = c(2, 2), mar = c(4.5, 4.5, 2, 2))
meanSdPlot(log2m, main = expression(log[2]), ylim = c(0, 3))
```

```
## KernSmooth 2.23 loaded
## Copyright M. P. Wand 1997-2009
```

```r
mtext("A", side = 3, line = line, adj = adj, cex = cex)
meanSdPlot(rlogm, main = "rlog", ylim = c(0, 3))
mtext("B", side = 3, line = line, adj = adj, cex = cex)
plotHclustColors(log2m, lab, main = expression(log[2]), ylab = "height")
mtext("C", side = 3, line = line, adj = adj, cex = cex)
plotHclustColors(rlogm, lab, main = "rlog", ylab = "height")
mtext("D", side = 3, line = line, adj = adj, cex = cex)
```

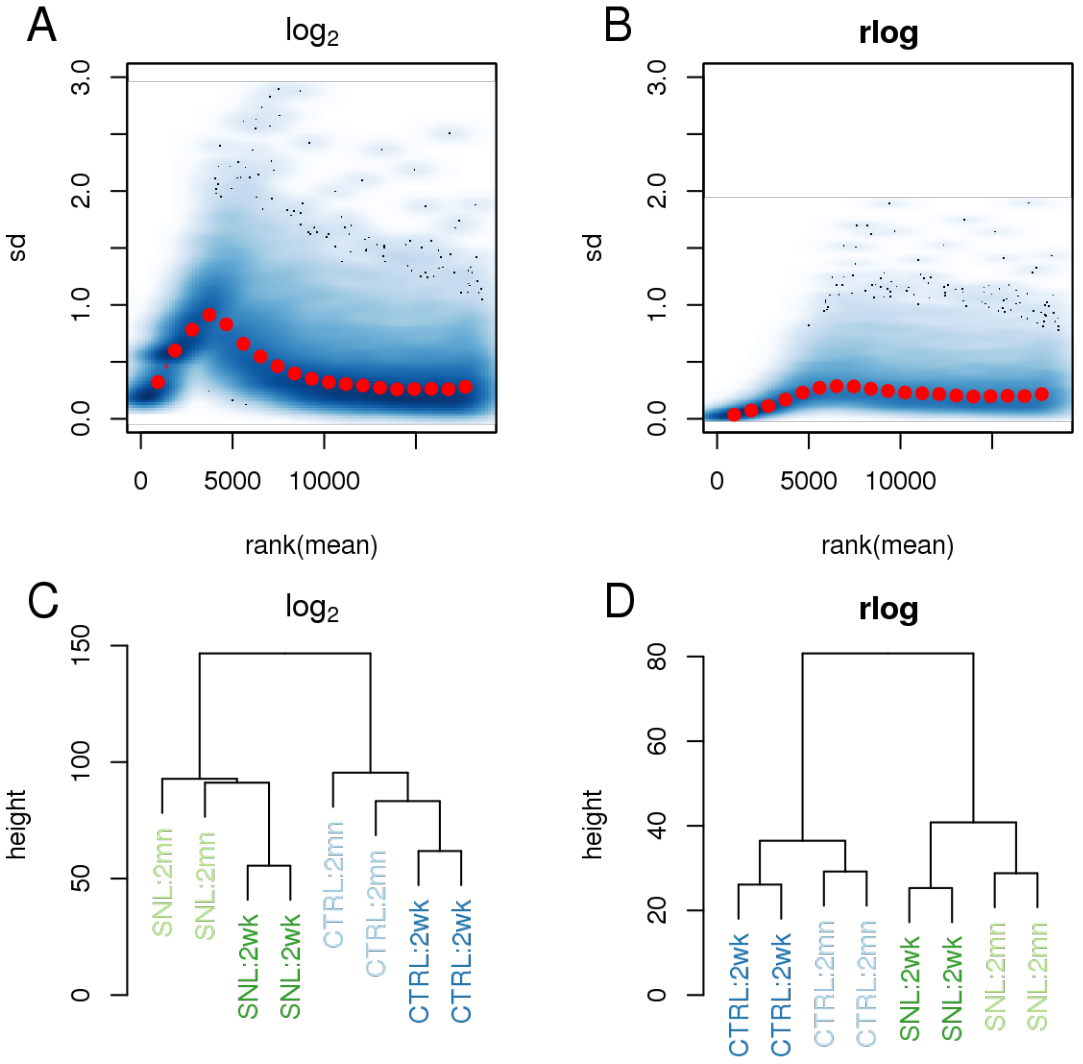

Figure 1: Comparison of the log2 of normalized counts plus a pseudocount and the rlog transformation. Plots (A) and (B) show the standard deviation across samples for every gene, plotted over the average expression strength. Plots (C) and (D) show the hierarchical clustering using Euclidean distance and complete linkage.

3 Session information

- R version 3.1.0 (2014-04-10), `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=en_US.UTF-8`, `LC_NAME=C`, `LC_ADDRESS=C`, `LC_TELEPHONE=C`, `LC_MEASUREMENT=en_US.UTF-8`, `LC_IDENTIFICATION=C`
- Base packages: base, datasets, grDevices, graphics, grid, methods, parallel, stats, utils
- Other packages: Biobase 2.24.0, BiocGenerics 0.10.0, DESeq2 1.4.0, DESeq2paper 1.3, GenomeInfoDb 1.0.0, GenomicRanges 1.16.0, IRanges 1.21.45, LSD 2.5, MASS 7.3-31, RColorBrewer 1.0-5, Rcpp 0.11.1, RcppArmadillo 0.4.200.0, colorRamps 2.3, ellipse 0.3-8, ggplot2 0.9.3.1, gridExtra 0.9.1, gtools 3.3.1, hexbin 1.27.0, knitr 1.5, schoolmath 0.4, vsn 3.32.0, xtable 1.7-3
- Loaded via a namespace (and not attached): AnnotationDbi 1.26.0, BiocInstaller 1.14.2, DBI 0.2-7, KernSmooth 2.23-12, RSQLite 0.11.4, XML 3.98-1.1, XVector 0.4.0, affy 1.42.2, affyio 1.32.0, annotate 1.42.0, codetools 0.2-8, colorspace 1.2-4, dichromat 2.0-0, digest 0.6.4, evaluate 0.5.5, formatR 0.10, genefilter 1.46.0, geneplotter 1.42.0, gtable 0.1.2, highr 0.3, labeling 0.2, lattice 0.20-29, limma 3.20.1, locfit 1.5-9.1, munsell 0.4.2, plyr 1.8.1, preprocessCore 1.26.1, proto 0.3-10, reshape2 1.4, scales 0.2.3, splines 3.1.0, stats4 3.1.0, stringr 0.6.2, survival 2.37-7, tools 3.1.0, zlibbioc 1.10.0