Estimation of mean dispersion pairs from Pickrell et al dataset

Michael Love

October 14, 2014

The following vignette displays the relationship of dispersion values over the mean of normalized counts from the Pickrell et al dataset. This joint distribution was then subsequently used in a number of simulations of Negative Binomial distributed counts. In addition, this vignette demonstrates the simulation of a set of Negative Binomial counts, using the mean values from the genes of the Pickrell dataset but with a single dispersion value, chosen as the asymptotic dispersion for genes in the Pickrell dataset. This simulation had the same number of genes and number of samples as the Pickrell dataset, yet it resulted in a distribution of dispersion estimates which was flat across the range of the mean of counts.

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

```
data("pickrell_sumexp")
ddspickrell <- DESeqDataSet(pickrell, ~1)
ddspickrell <- estimateSizeFactors(ddspickrell)
ddspickrell <- estimateDispersionsGeneEst(ddspickrell)
ddspickrell <- estimateDispersionsFit(ddspickrell)
meanDispPairs <- mcols(ddspickrell)[which(mcols(ddspickrell)$dispGeneEst > 1e-06),
    c("baseMean", "dispGeneEst")]
names(meanDispPairs) <- c("mean", "disp")
# save(meanDispPairs, file='../data/meanDispPairs.RData')
```

```
asymptDisp <- attr(dispersionFunction(ddspickrell), "coefficients")["asymptDisp"]
asymptDisp

## asymptDisp
##     0.1614

# the fit gives roughly the same asymptotic dispersion as the average
# dispersion for genes with average expression > 100
with(mcols(ddspickrell), mean(dispGeneEst[baseMean > 100]))

## [1] 0.1746

rm <- rowMeans(counts(ddspickrell, normalized = TRUE))
dim(ddspickrell)

## [1] 56299    69

m <- ncol(pickrell)
n <- nrow(pickrell)
nbdata <- matrix(rnbinom(n * m, mu = rm, size = 1/asymptDisp), ncol = m)
ddsSim <- DESeqDataSetFromMatrix(nbdata, DataFrame(row.names = seq_len(m)),
    ~1)
ddsSim <- estimateSizeFactors(ddsSim)
ddsSim <- estimateDispersionsGeneEst(ddsSim)
```

```r
par(mfrow = c(1, 2))
line <- 0.4
adj <- -0.2
cex <- 1.5
with(mcols(ddspickrell), plot(baseMean, dispGeneEst, cex = 0.1, log = "xy",
    col = rgb(0, 0, 0, 0.2), ylim = c(0.01, 10), xlim = c(0.1, 1e+05), xaxt = "n",
    xlab = "mean of normalized counts", ylab = "genewise dispersion estimate"))

## Warning:  12736 x values <= 0 omitted from logarithmic plot

axis(1, at = c(1, 100, 10000))
with(mcols(ddspickrell)[which(mcols(ddspickrell)$dispGeneEst < 0.01), ], points(baseMean,
    rep(0.01, length(baseMean)), cex = 0.1, col = rgb(0, 0, 0, 0.2)))
abline(v = 1/(asymptDisp * quantile(sizeFactors(ddspickrell), 1:3/4)))
mtext("A", side = 3, line = line, adj = adj, cex = cex)
with(mcols(ddsSim), plot(baseMean, dispGeneEst, cex = 0.1, log = "xy", col = rgb(0,
    0, 0, 0.2), ylim = c(0.01, 10), xlim = c(0.1, 1e+05), xaxt = "n", xlab = "mean of normalized counts",
    ylab = "genewise dispersion estimate"))

## Warning:  14906 x values <= 0 omitted from logarithmic plot

axis(1, at = c(1, 100, 10000))
with(mcols(ddsSim)[which(mcols(ddsSim)$dispGeneEst < 0.01), ], points(baseMean,
    rep(0.01, length(baseMean)), cex = 0.1, col = rgb(0, 0, 0, 0.2)))
abline(v = 1/(asymptDisp * quantile(sizeFactors(ddsSim), 1:3/4)))
mtext("B", side = 3, line = line, adj = adj, cex = cex)
```

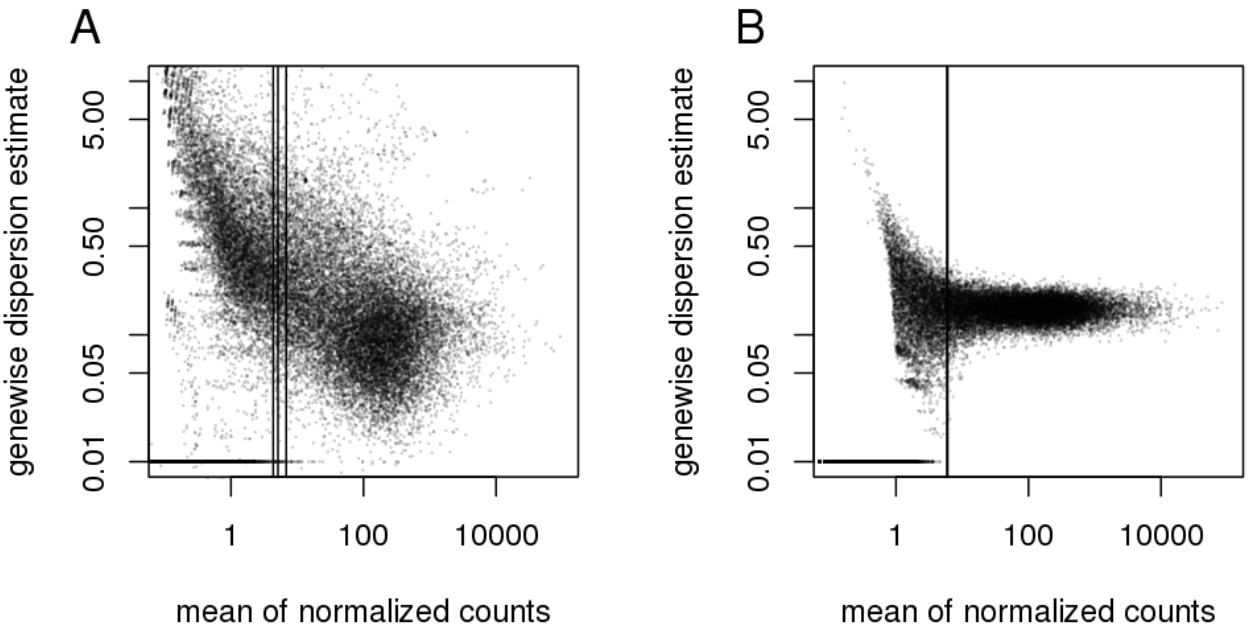

Figure 1: Demonstration through simulation that the dependence of dispersions on the mean seen in the Pickrell dispersion plot is not an artifact of estimation bias. (A) The gene-wise estimates of dispersion for the 69 samples of the Pickrell et al dataset. (B) The gene-wise estimates of dispersion for a simulated Negative Binomial dataset, using a fixed dispersion of 0.16, equal to the asymptotic gene-wise dispersion estimate seen in the original dataset (A), and with the same means and the same number of genes and samples as the original dataset. Genes with dispersion estimates below the plotting range are depicted at the bottom of the frame. For genes with mean counts greater than around 5, the gene-wise dispersion estimates do not exhibit a dependence on the mean count for the simulated data in panel B. Vertical lines indicate the reciprocal of dispersion on the scale of the samples with size factors in the 1st, 2nd and 3rd quartile.

1 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: 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, xtable 1.7-3
- Loaded via a namespace (and not attached): AnnotationDbi 1.26.0, Biobase 2.24.0, DBI 0.2-7, RSQLite 0.11.4, XML 3.98-1.1, XVector 0.4.0, 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, locfit 1.5-9.1, munsell 0.4.2, plyr 1.8.1, 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