Assessing dispersion estimation and shrinkage

Michael Love

October 14, 2014

1 Estimates of dispersion in DESeq2

We constructed a simulated dataset with 10 samples divided into 2 groups, with no true difference between the means of the two groups. The means and dispersions of the Negative Binomial simulated data were drawn from the estimates from the Pickrell et al dataset.

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

```
makeSimScript <- system.file("script/makeSim.R", package = "DESeq2paper", mustWork = TRUE)
source(makeSimScript)
data("meanDispPairs")
n <- 4000
m <- 10
condition <- factor(rep(c("A", "B"), each = m/2))
x <- model.matrix(~condition)
beta <- rep(0, n)
sim <- makeSim(n, m, x, beta, meanDispPairs)
mat <- sim$mat
dds <- DESeqDataSetFromMatrix(mat, DataFrame(condition), ~condition)
dds <- DESeq(dds)

## estimating size factors
## estimating dispersions
## gene-wise dispersion estimates
## mean-dispersion relationship
## final dispersion estimates
## fitting model and testing

data <- log10(cbind(genewise = mcols(dds)$dispGeneEst, fit = mcols(dds)$dispFit,
    maximum = pmax(mcols(dds)$dispGeneEst, mcols(dds)$dispFit), MAP = dispersions(dds),
    true = sim$disp))
data <- data[data[, "genewise"] > -7, ]
data <- data[rowSums(is.na(data)) == 0, ]
```

```
library("LSD")

## Loading required package:  MASS
## Loading required package:  gtools
## Loading required package:  RColorBrewer
## Loading required package:  colorRamps
## Loading required package:  schoolmath
## Loading required package:  ellipse

heatpairs(data, xlim = c(-2, 2), ylim = c(-2, 2), main = "")
```

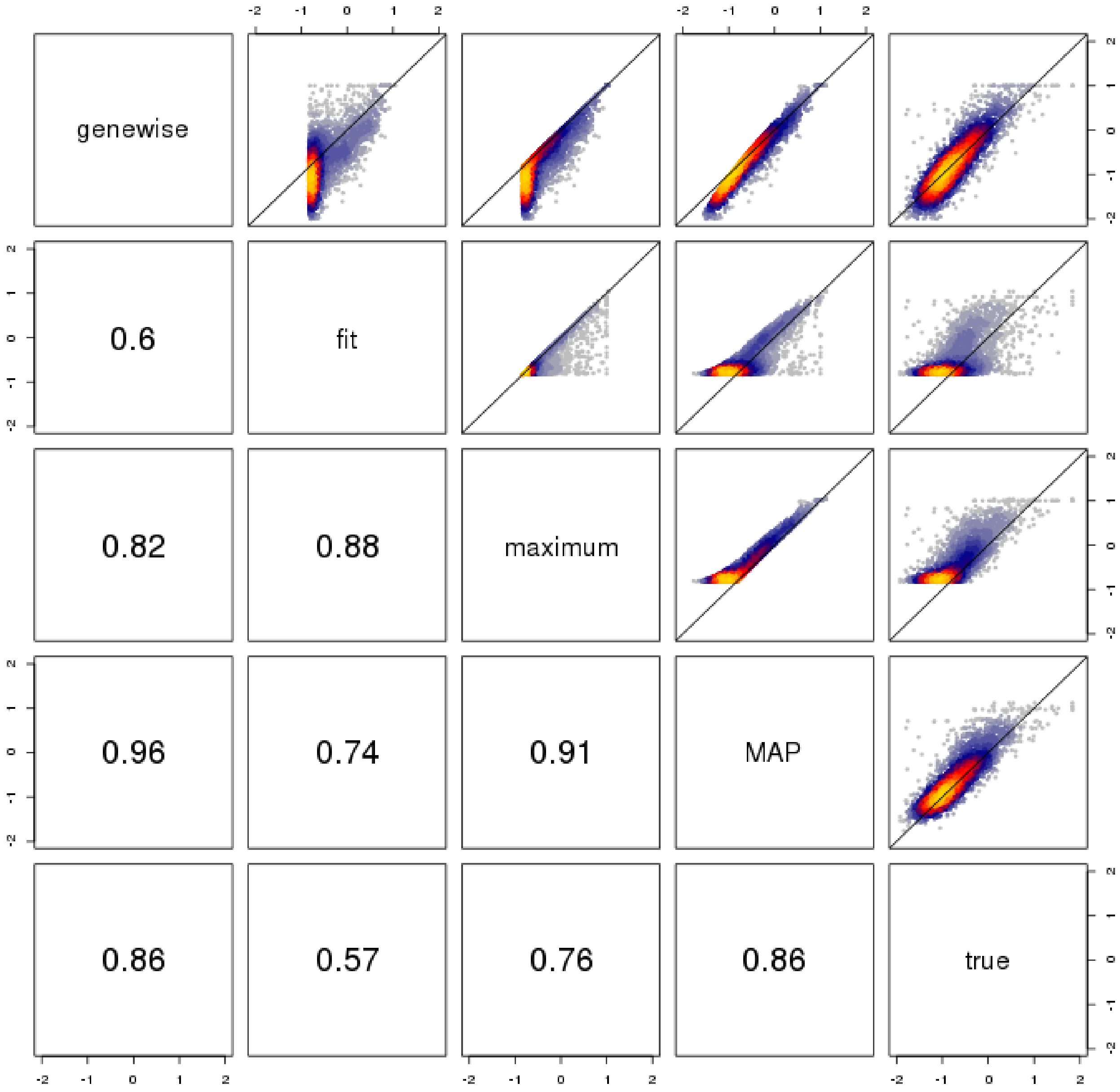

Figure 1: Scatterplot of various estimates of dispersion using DESeq2, against the true dispersion in the logarithmic scale (base 10) from simulated counts. The blue, red, and yellow colors indicate regions of increasing density of points. Counts for 4000 genes and for 10 samples were simulated for two groups with no true difference in means. The Negative Binomial counts had mean and dispersion drawn from the joint distribution of the mean and gene-wise dispersion estimates from the Pickrell et al dataset. The estimates shown are genewise, the CR-adjusted maximum likelihood estimate; fit the value from the fitted curve; maximum, the maximum of the two previous values (the estimate used in the older version of DESeq); and MAP, the maximum a posteriori estimate used in DESeq2. The correlations shown in the bottom panels do not include the very low gene-wise estimates of dispersion which can result in potential false positives. The MAP, shrunken estimates used in DESeq2 were closer to the diagonal, while the maximum estimate was typically above the true value of dispersion, which can lead to overly-conservative inference of differential expression.

2 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, 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, gtools 3.3.1, schoolmath 0.4
- 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, digest 0.6.4, evaluate 0.5.5, formatR 0.10, genefilter 1.46.0, geneplotter 1.42.0, grid 3.1.0, highr 0.3, knitr 1.5, lattice 0.20-29, locfit 1.5-9.1, splines 3.1.0, stats4 3.1.0, stringr 0.6.2, survival 2.37-7, tools 3.1.0, xtable 1.7-3