Design

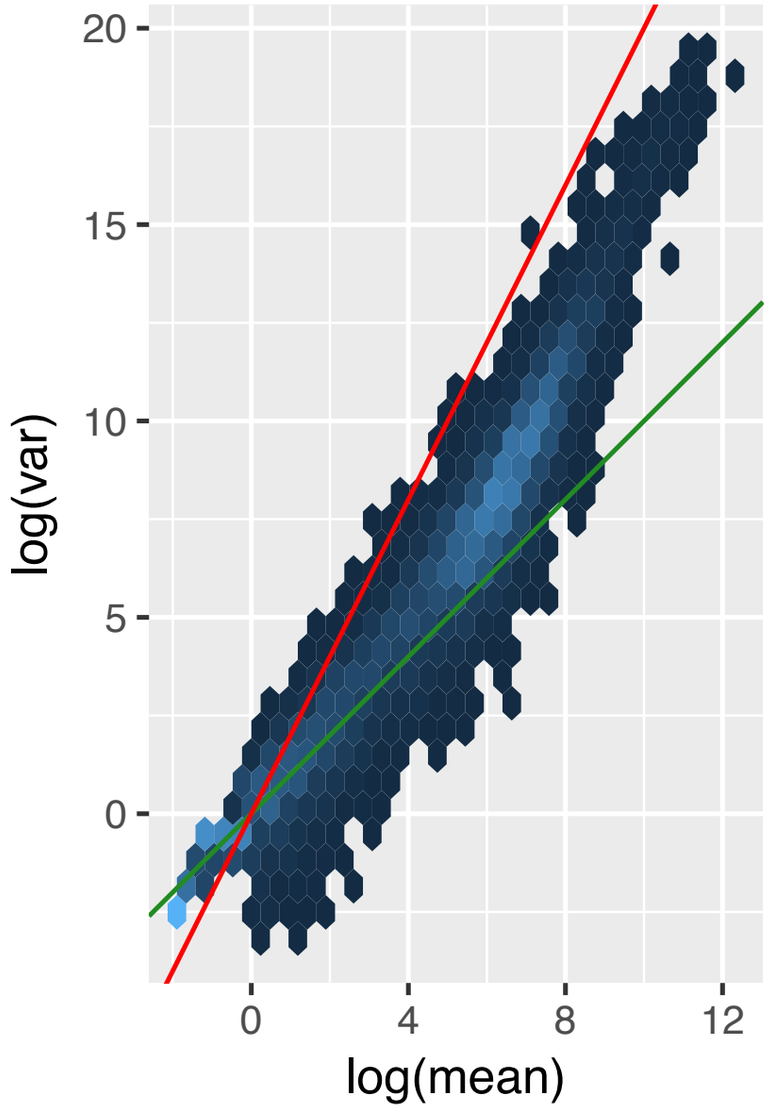
Susan Holmes
Wolfgang Huber

Overview

Variance-stabilizing transformations

Hypothesis screening

Variance-mean relationship in the pasilla data



Variance and mean are computed for each row (gene), across the columns (samples)

$$v = c \cdot m^k$$

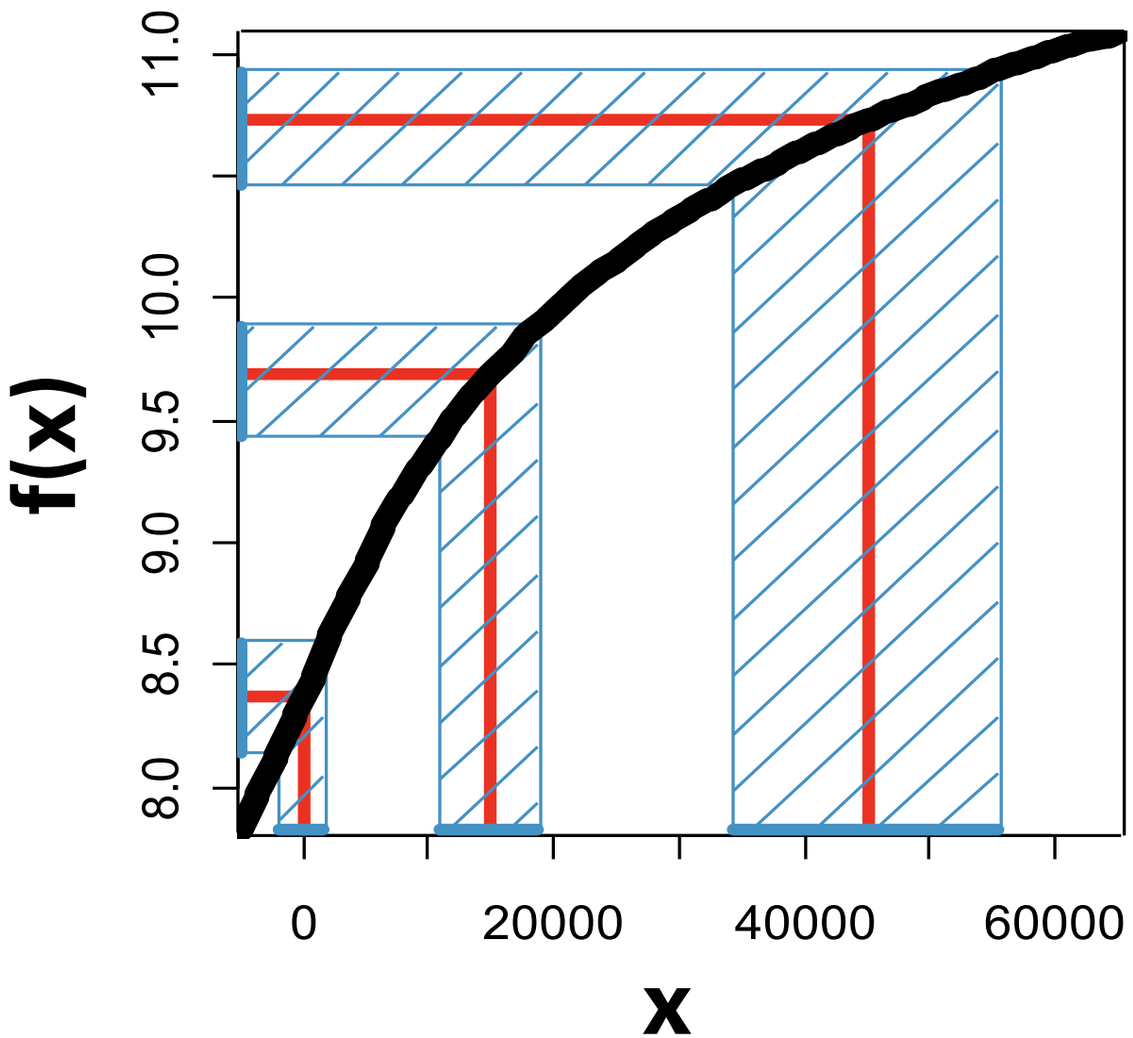
$$\Updownarrow$$

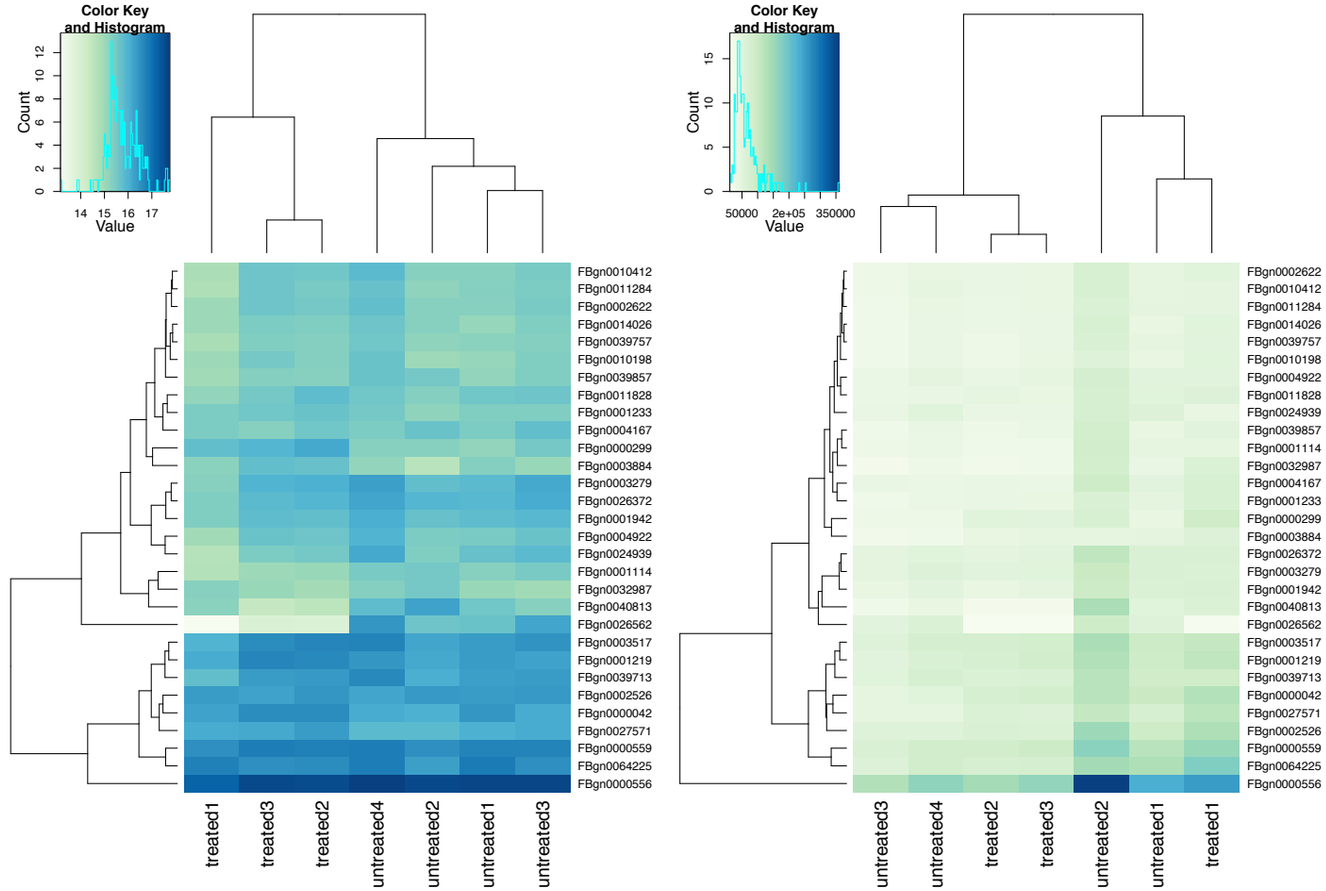
$$\log(v) = k \cdot \log(m) + \log(c)$$

Figure 8.4: Variance versus mean for the (size factor adjusted) `counts` data. The axes are logarithmic. Also shown are lines through the origin with slopes 1 (green) and 2 (red).

Variance-stabilizing transformation

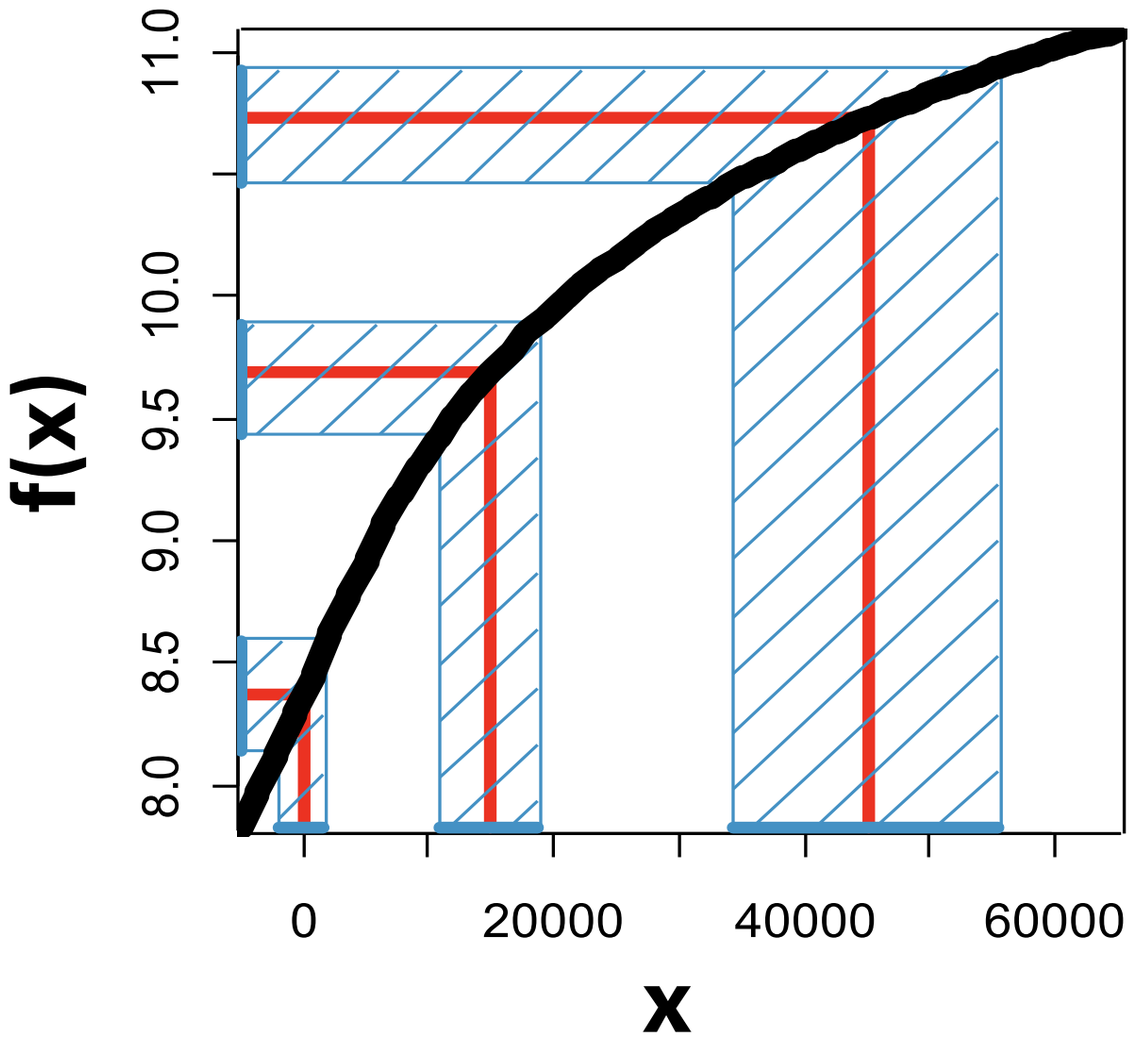
The estimated variance-mean dependence allows deriving a logarithm-like transformation that removes it

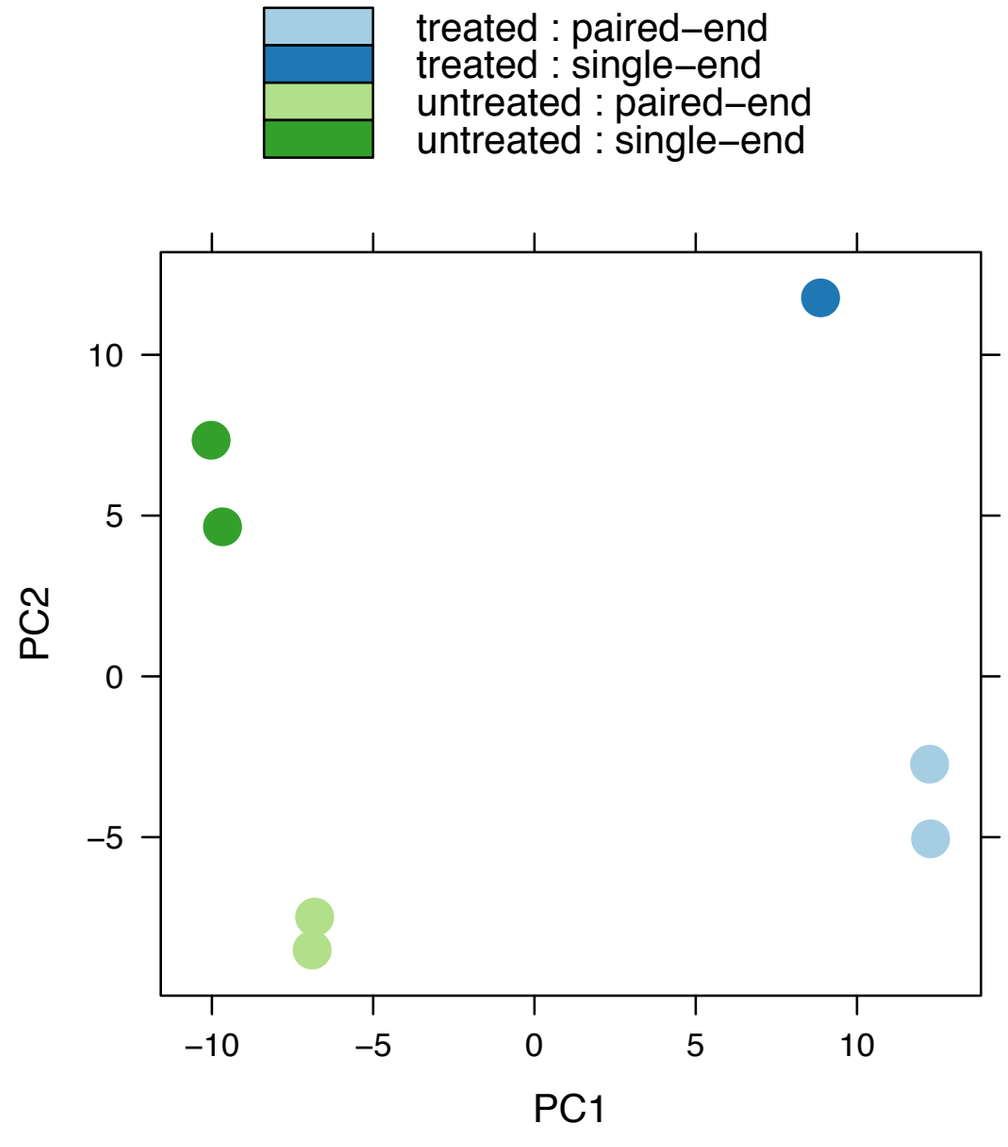




Variance-stabilizing transformation

The estimated variance-mean dependence allows deriving a logarithm-like transformation that removes it





Variance Stabilizing Transformation

$$f(x) = \int_x \frac{\mathrm{d}u}{\sqrt{v(u)}}$$

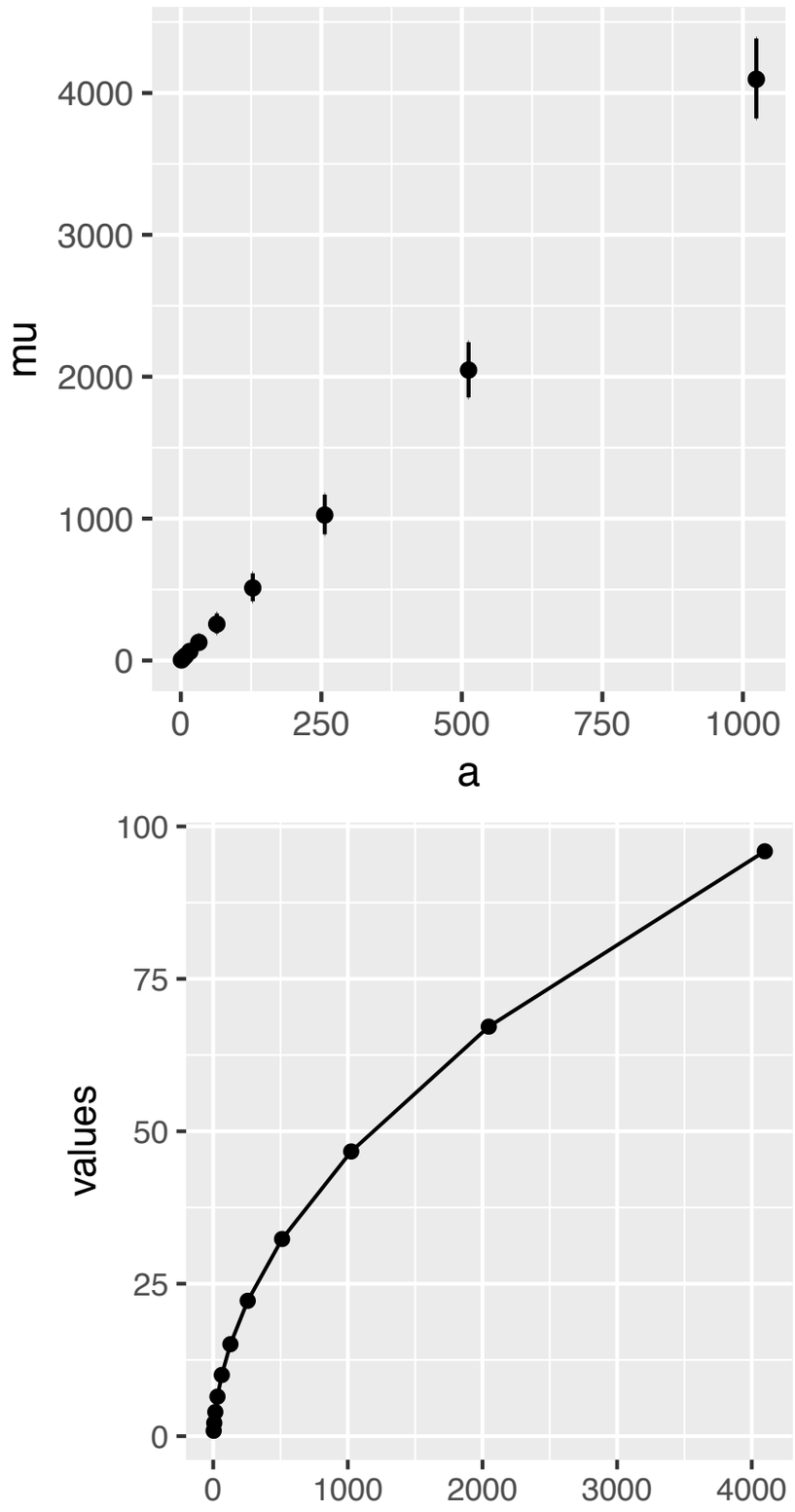
For Gamma-Poisson distributed data:

$$f_{a,b}(x) = \frac{1}{\log(2)} \log\left(\frac{1 + 2ax + b + 2\sqrt{ax(1 + ax + b)}}{4a}\right)$$

Demo

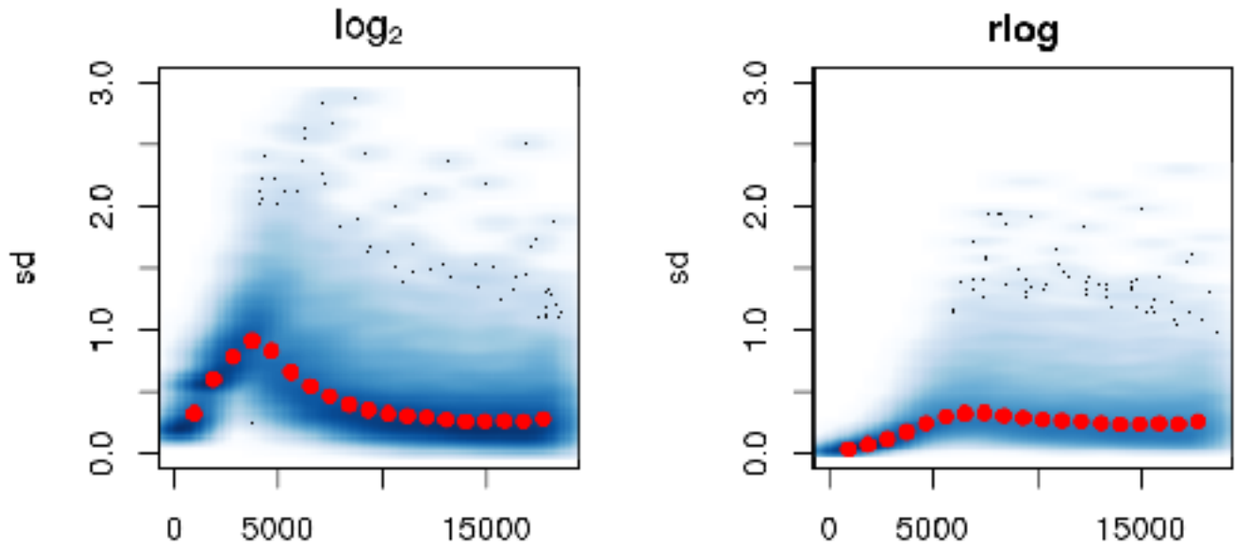
Deriving a variance-stabilizing transformation from empirical variances

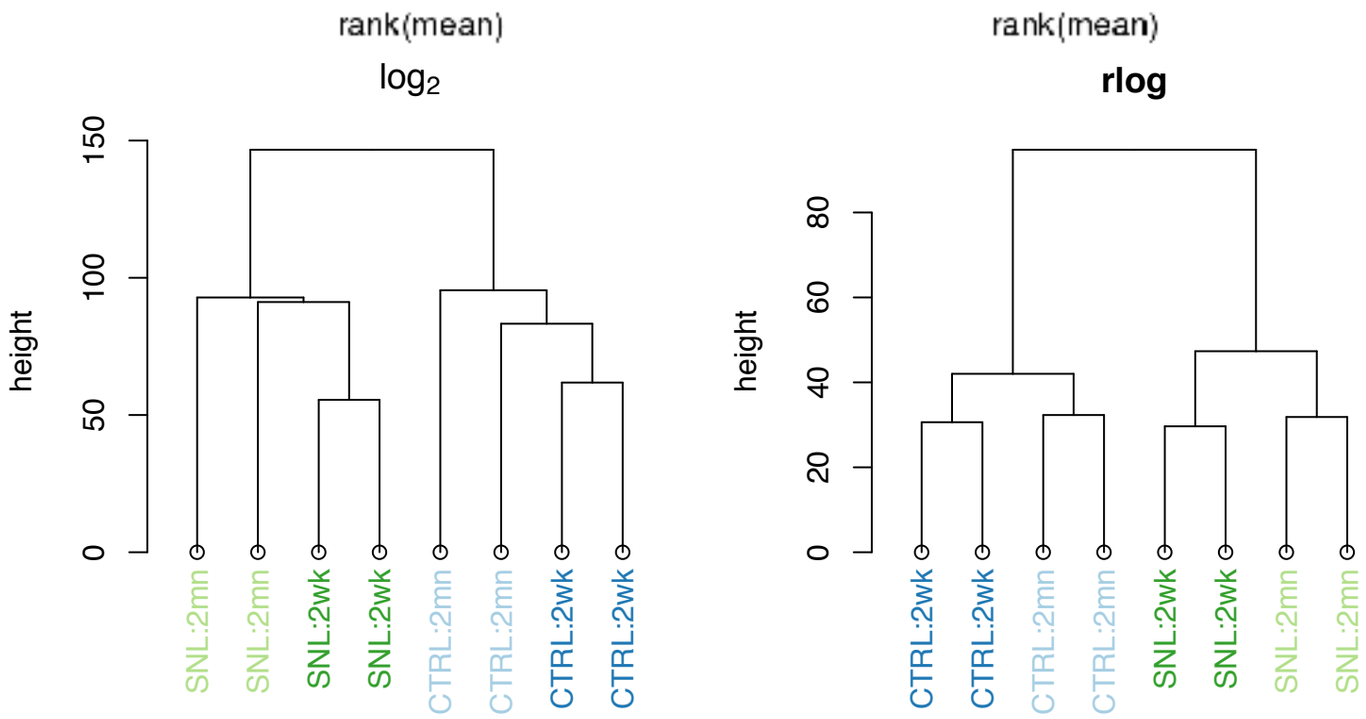
```
nba = 2^seq(0, 10, by = 1)
simnb = lapply(nba, function(a) {
 u = rnbinom(1e4, a, 0.2)
 tibble(mu = mean(u), sd = sd(u),
            lower = quantile(u, 0.025),
            upper = quantile(u, 0.975),
            a = a)
 } ) %>% bind_rows

head(as.data.frame(simnb), 2)

##        mu       sd lower upper a
## 1 3.9129 4.402028     0    16 1
## 2 8.0493 6.297113     0    24 2

ggplot(simnb, aes(x = a, y = mu, ymin = lower, ymax = upper)) +
 geom_point() + geom_errorbar()
```




```
slopes= 1/simnb$sd
datacurve = data.frame(mns=simnb$mu, values = cumsum(slopes * simnb$mu))
ggplot(datacurve,aes(x=mns, y=values)) +
geom_point()+geom_line() +xlab("")
```

Regularized log-transformation: Visualization, Clustering, PCA

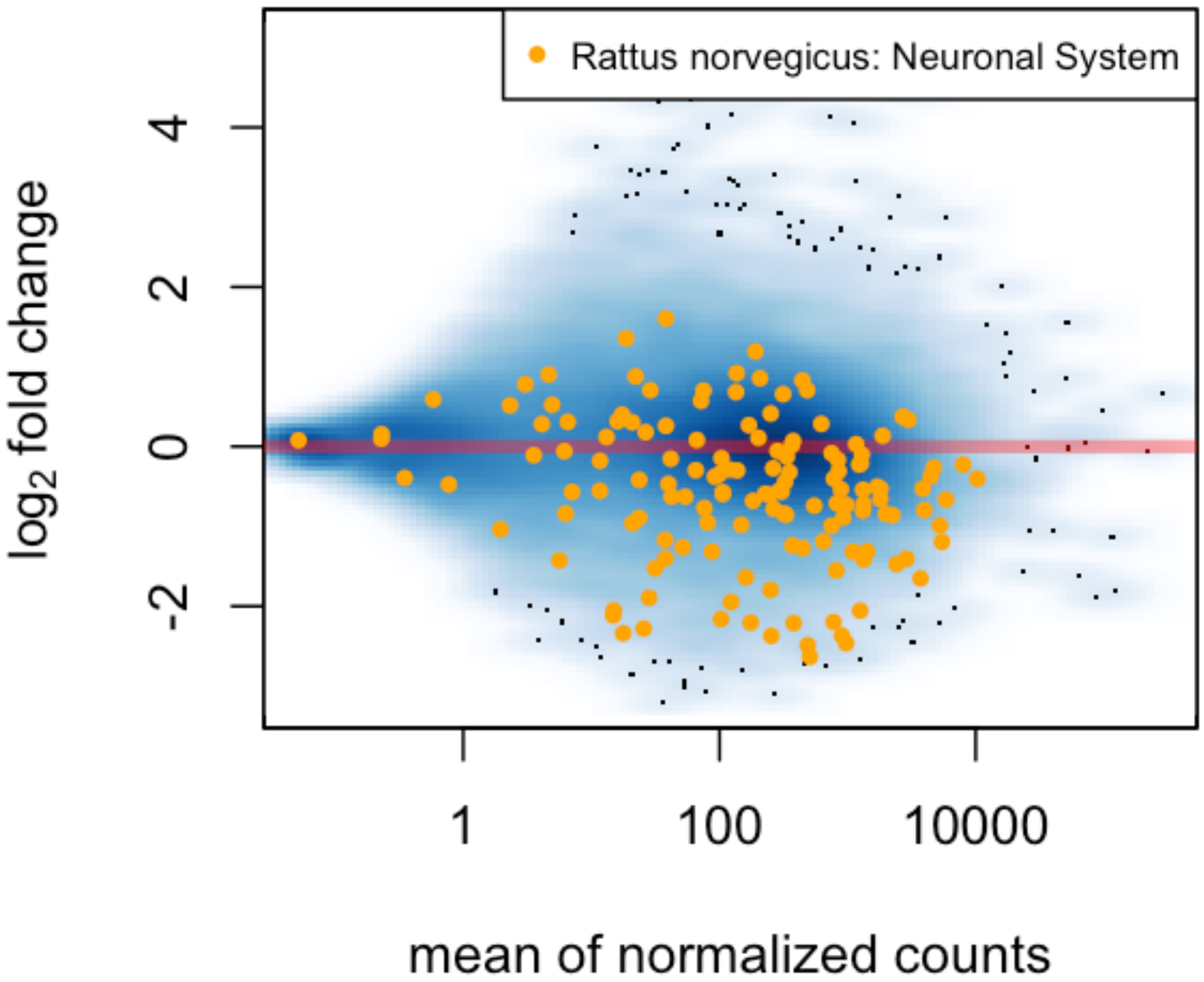




"rlog":
Shrunken log fold changes for every sample:
reduces effect of shot noise on inter-sample distances

RNA from the dorsal root ganglion of rats that underwent spinal nerve ligation and controls, 2 weeks & 2 months after the ligation. Hammer, …, Beutler AS, Genome Research 2010.

GSEA with shrunken log fold changes



Reactome gene set
one-sample t-statistic

Neuronal System
144 genes
avg LFC: -0.55
adjusted p-value: 10^{-8}

RNA from the dorsal root ganglion of rats that underwent spinal nerve ligation and controls, 2 weeks & 2 months after the ligation. Hammer, …, Beutler AS, Genome Research 2010.

Considerations on hypothesis screening

(a.k.a. ‘multiple testing)

FDR is a set quantity. Subsequent subsetting invalides it.
An FDR of 10% for a result list DOES NOT mean local fdr for each component gene is <= 10%

Tests against point-like null hypotheses can be too powerful.
Consider banded nulls.

If you get astronomically small p-values, something is wrong.

Banded hypothesis testing:
integrate testing with fold-change cutoff

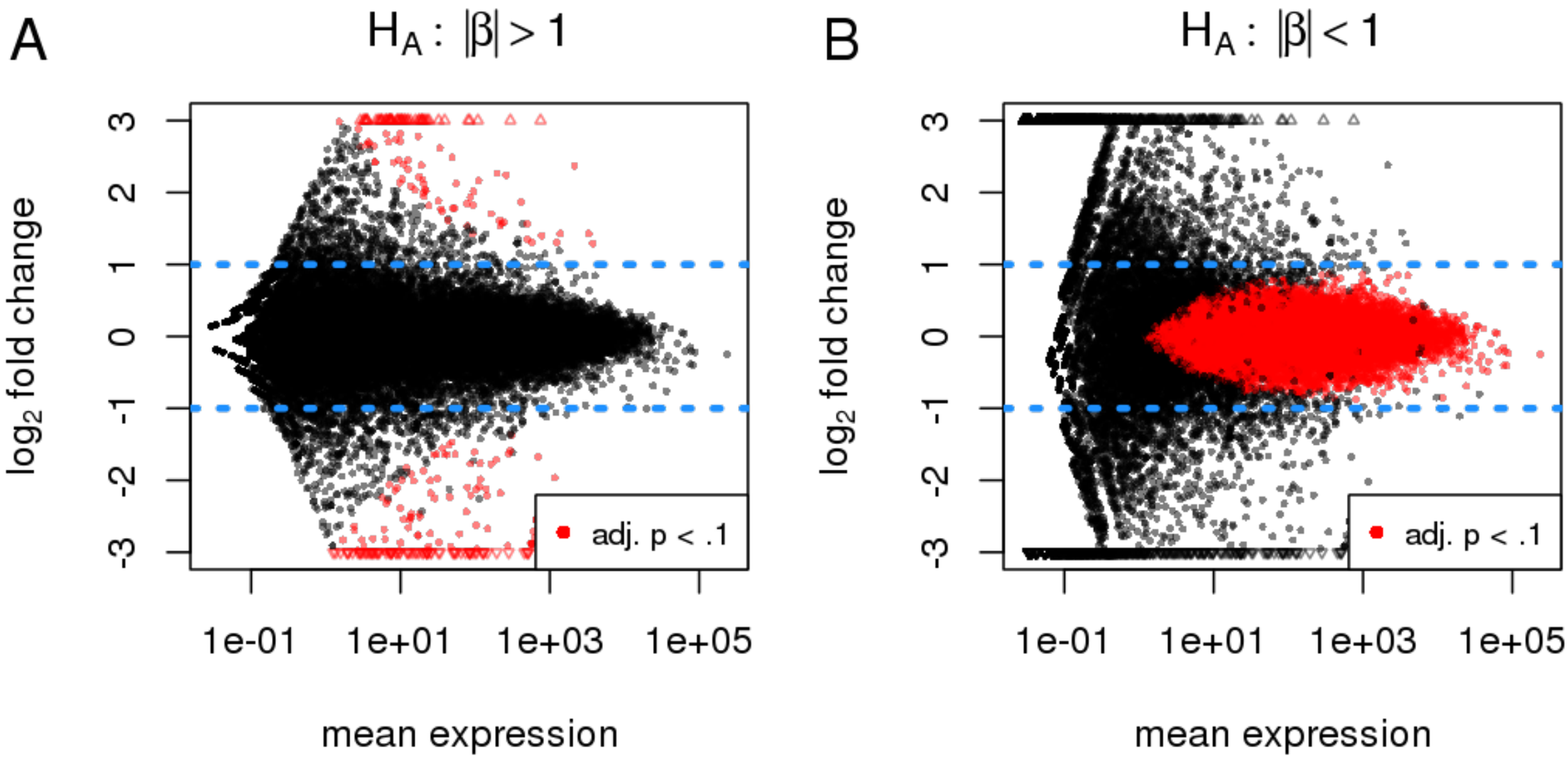


Figure 4 Hypothesis testing involving non-zero thresholds. Shown are MA-plots for a 10 vs 11 comparison using the Bottomly *et al.* [15] dataset, with highlighted points indicating low adjusted p-values. The alternate hypotheses are that logarithmic (base 2) fold changes are (A) greater than 1 in absolute value or (B) less than 1 in absolute value.