UNIVERSITÄT LEIPZIG

Normalization and differential expression II

Katharina Hößel

Statistical Analysis of RNA-Seq Data

May 29th, 2012

Overview

- **Differential expression analysis for sequence count data** (Anders, Huber 2010)
- **Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments** (Bullard, Purdom, Hansen, Dudoit 2010)

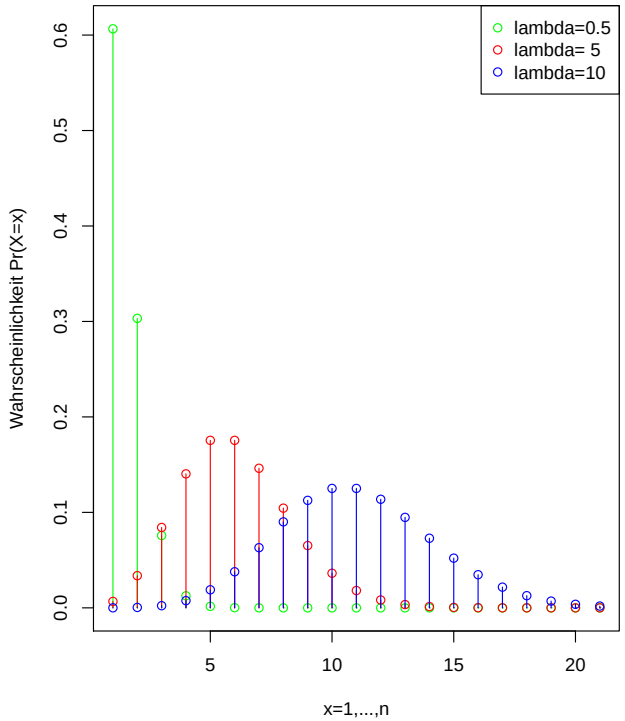
Background

- RNA-sequencing: reads are mapped to a class (=gene)
- the number of reads in a class is called 'read count'
- *read count* is linearly related to the abundance of the target transcript
- interest: comparing counts between different biological conditions

$\rightarrow$ statistical testing

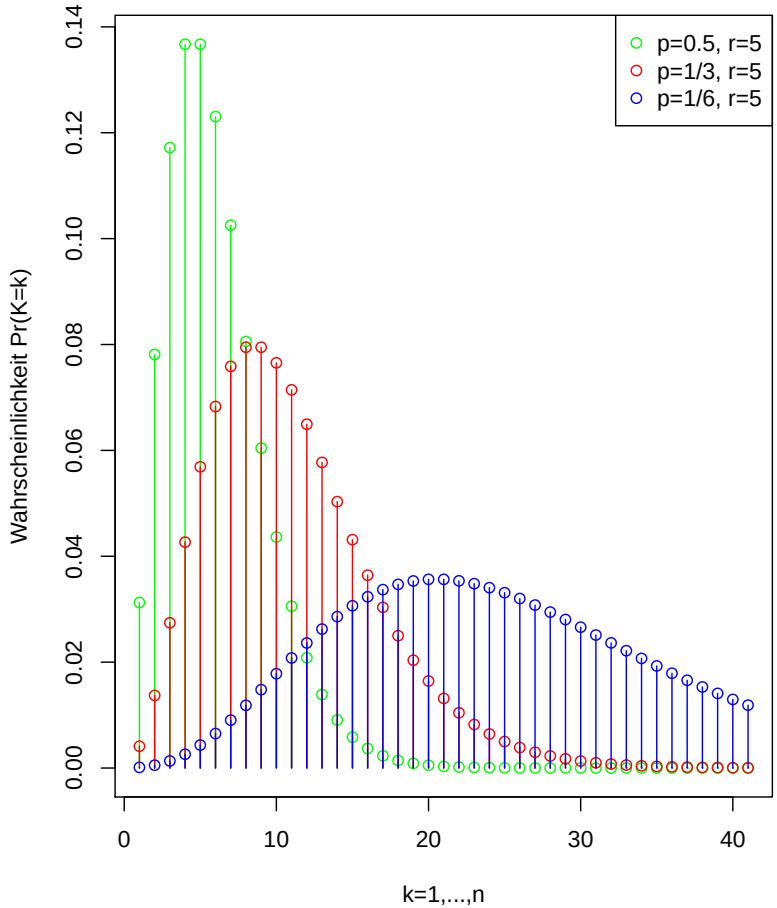
DESeq - Statistics

- read counts can be approximated by a Poisson distribution



- Poisson leads to overdispersion problem

$\rightarrow$ use of negative binomial distribution



Comparison: Poisson vs. NB

	Poisson distribution	negative binomial distribution
parameters	λ	r, p
distr.function	$\Pr(X = x) = \frac{\lambda^x}{x!} e^{-\lambda}$	$\Pr(K = k) = \binom{k+r-1}{r-1} p^r (1-p)^k$
expectation	$\mathrm{E}(X) = \lambda$	$\mathrm{E}(K) = \frac{r(1-p)}{p}$
variance	$\mathrm{var}(X) = \lambda$	$\mathrm{var}(K) = \frac{r(1-p)}{p^2}$

DESeq - Model I

distribution

$$K_{ij} \sim \mathrm{NB}(\mu_{ij}, \sigma_{ij}^2), \tag{1}$$

i – genes, j – samples, K – read counts

expectation value

$$\mu_{ij} = q_{i,\rho(j)} \cdot s_j \tag{2}$$

$q_{i,\rho(j)}$ – expected read count (per gene and condition)

s_j – scaling factor across genes and groups (depends on sampling depth resp. coverage of sample j)
$\rightarrow$ normalization and adjusting for coverage

DESeq - Model II

variance

$$\sigma_{ij}^2 = \underbrace{\mu_{ij}}_{\textit{shot noise}} + \underbrace{\underbrace{s_j^2}_{\text{size factor}} \cdot \underbrace{v_{i,\rho(j)}}_{\text{raw variance parameter}}}_{\text{raw variance}} \tag{3}$$

$v_{i,\rho(j)}$ – per-gene raw variance parameter is assumed to be a smooth function of $q_{i,\rho}$:

$$v_{i,\rho(j)} = v_\rho(q_{i,\rho(j)}) \tag{4}$$

$\rightarrow$ allows pooling of data from genes with similar expression strength

DESeq - Parameter reduction

example:

- $n = 10.000$ genes
- $m = 20$ samples
- $G = 2$ groups à 10 samples each

number of parameters for model fit is reduced in two steps:

1. mean
2. variance

parameters needed for ...

	mean	variance	total
naive NB	$n \cdot m = 200.000$	$n \cdot m = 200.000$	400.000
after step 1	$n \cdot G + m = 20.020$	$n \cdot m = 200.000$	220.020
after step 2	$n \cdot G + m = 20.020$	$n \cdot G = 20.000$	**40.020**

DESeq - Fitting I

size factors

$$\hat{s}_j = \text{median}_i \frac{k_{ij}}{(\prod_{v=1}^{m} k_{iv})^{\frac{1}{m}}} \tag{5}$$

empirical expectation values (common scale)

$$\hat{q}_{i\rho} = \frac{1}{m_\rho} \sum_{j:\rho(j)=\rho} \frac{k_{ij}}{\hat{s}_j} \tag{6}$$

DESeq - Fitting II

sample variances (common scale)

$$w_{i\rho} = \frac{1}{m_\rho - 1} \sum_{j:\rho(j)=\rho} \left(\frac{k_{ij}}{\hat{s}_j} - \hat{q}_{i\rho} \right)^2 \tag{7}$$

they define

$$z_{i\rho} = \frac{\hat{q}_{i\rho}}{m_\rho} \sum_{j:\rho(j)=\rho} \frac{1}{\hat{s}_j} \tag{8}$$

$w_{i\rho} - z_{i\rho}$ is an unbiased estimator of $v_{i\rho}$.
local regression

$$\Rightarrow \hat{v}_\rho(\hat{q}_{i\rho}) = w_\rho(\hat{q}_{i\rho}) - z_{i\rho} \tag{9}$$

DESeq - Testing I

We have two biological conditions, A and B.
null hypothesis: counts for A and B are identical

$$q_{iA} = q_{iB}$$

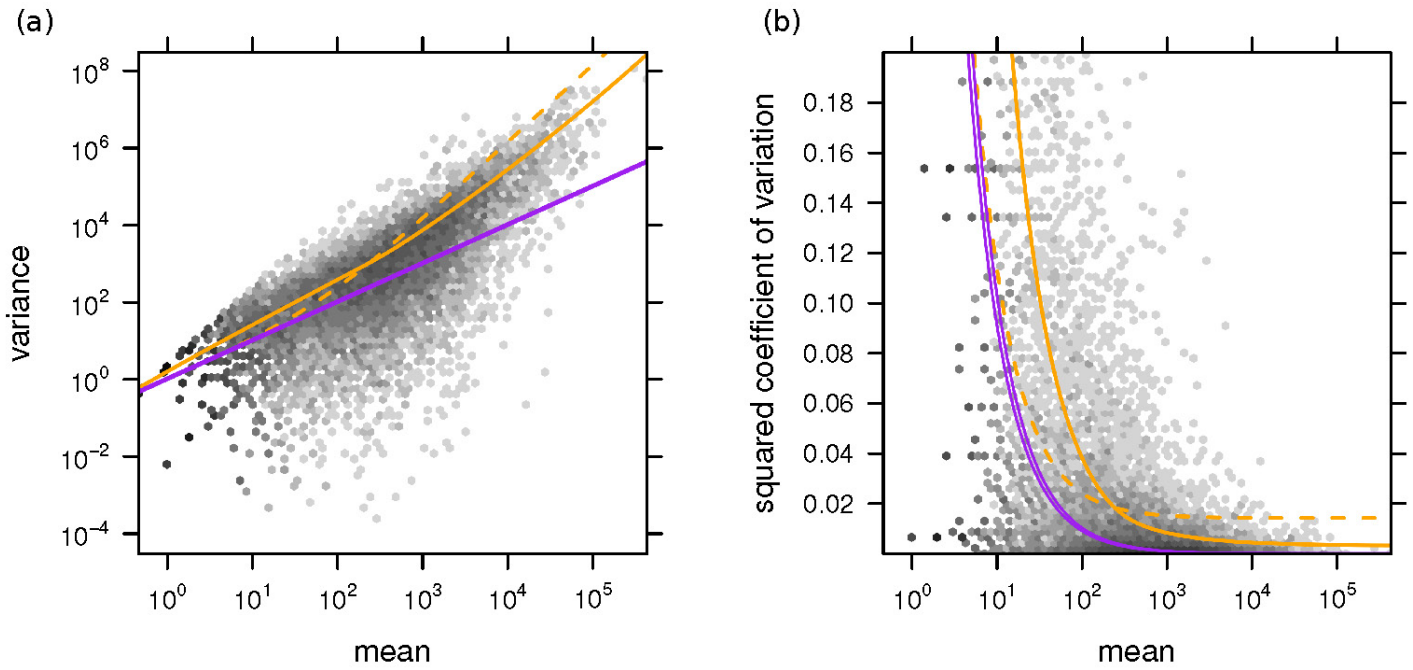
test statistic: counting reads for each condition: K_{iA}, K_{iB}
sum: $K_{iS} = K_{iA} + K_{iB}$

$$p(a, b) = \Pr(K_{iA} = a) \Pr(K_{iB} = b)$$

performing `nbinomTest` as fisher's exact test on negative binomial data
p **value**

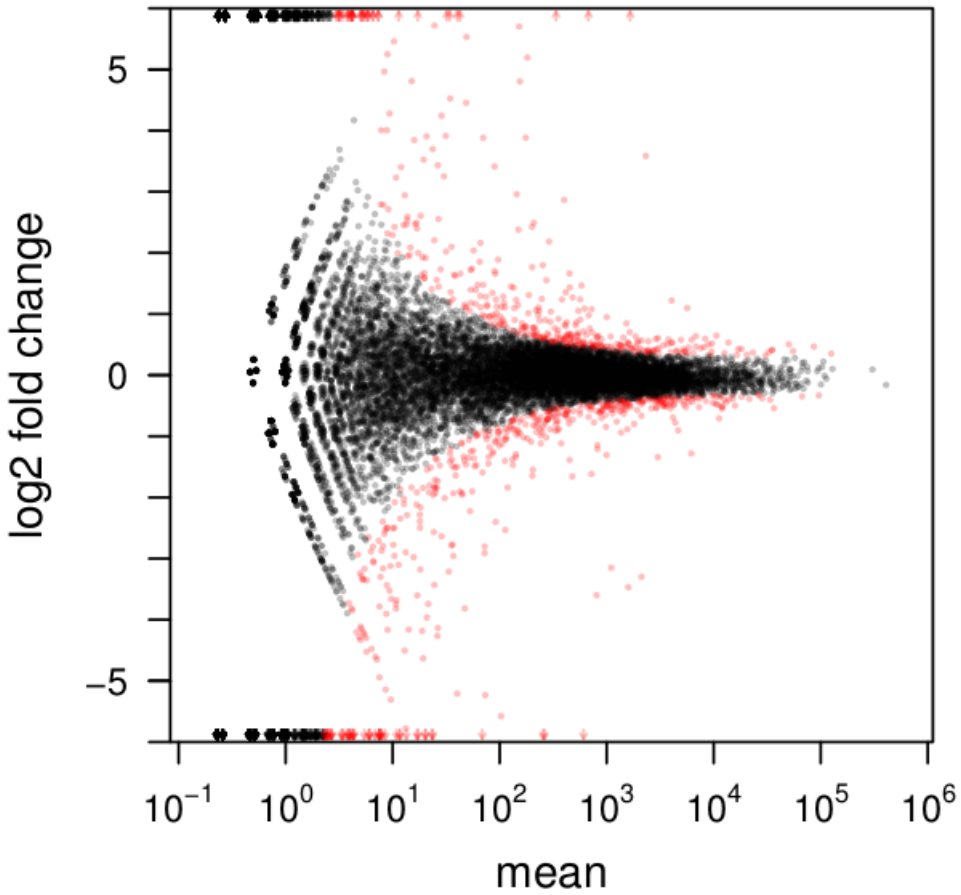
$$p_i = \frac{\sum_{\substack{a+b=k_{iS} \\ p(a,b) \leq p(k_{iA}, k_{iB})}} p(a, b)}{\sum_{a+b=k_{iS}} p(a, b)} \tag{10}$$

DESeq - Applications I (Fly embryos)



orange variance estimate by *DESeq* (fit $w(q)$)
dotted orange variance estimate by *edgeR*
purple variance via Poisson distribution

DESeq - Applications II



Testing for differential expression between conditions A and B: Scatter plot of log2 ratio (fold change) versus mean.

DESeq - Conclusions

- using parametric methods (e.g., tests)
- sharing information between genes
- Poisson distribution is adequate for modelling read counts within technical replicates (small dispersion)

$\rightarrow$ using NB for *biological* replicates

DESeq - R/Bioconductor package

- available via Bioconductor
- current version 1.9.7 by 2012/05/25 (example computations in paper were done in 1.1.12)
- huge changelog: bugfixes, addition/removal/renaming of functions, adding/removing/extending functionality, new methods etc.
  - handling of variance
  - variance stabilization
  - testing procedure
  - diagnose plots

$\rightarrow$ this software is evolving!

Overview

- **Differential expression analysis for sequence count data** (Anders, Huber 2010)
- **Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments** (Bullard, Purdom, Hansen, Dudoit 2010)

Evaluation of statistical methods . . . - Motivation

- Microarrays vs. RNA-Seq
- different statistical tests
- different approaches of normalization
- calibration
- assess biases based on seq. technology
  - length biases
  - flow cell effects
  - library preparation effects

Evaluation - Methods

- 2 biological samples: brain vs. universal human reference (UHR)
- performing Microarray, RNA-Seq analysis and qRT-PCR on $\sim$ 1000 genes
- compare expression values obtained from Microarray and RNA-Seq experiments using qRT-PCR as benchmark
- nested RNA-Seq setup

Evaluation - Normalization

global vs. quantile-based methods

1. total lane counts
   (RNA-Seq standard)
2. per-lane counts for "housekeeping gene" POLR2A
   (borrowed from qRT-PCR)
3. per-lane quantile for genes with reads in at least 1 lane
   (borrowed from Microarrays)

Evaluation - Differential Expression

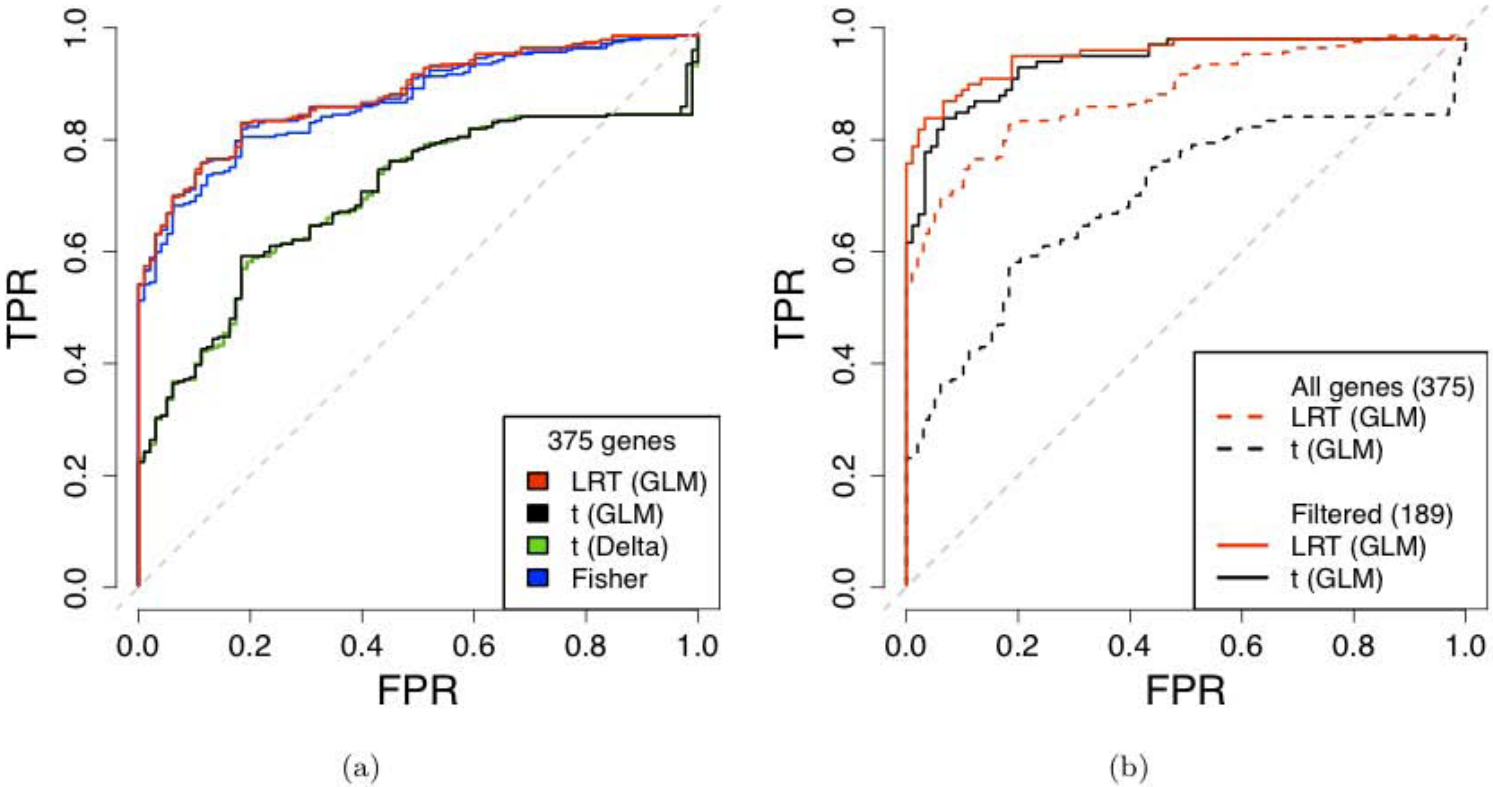
generalized linear model (GLM)

$$\log(E(X_{ij}|d_i)) = \underbrace{\log d_i}_{\text{offset}} + \underbrace{\lambda_{a(i,j)}}_{\text{expression level}} + \underbrace{\theta_{ij}}_{\text{technical effects}}$$

tests

- fisher's exact test
- likelihood ratio test (GLM based)
- t-test (GLM based + delta)

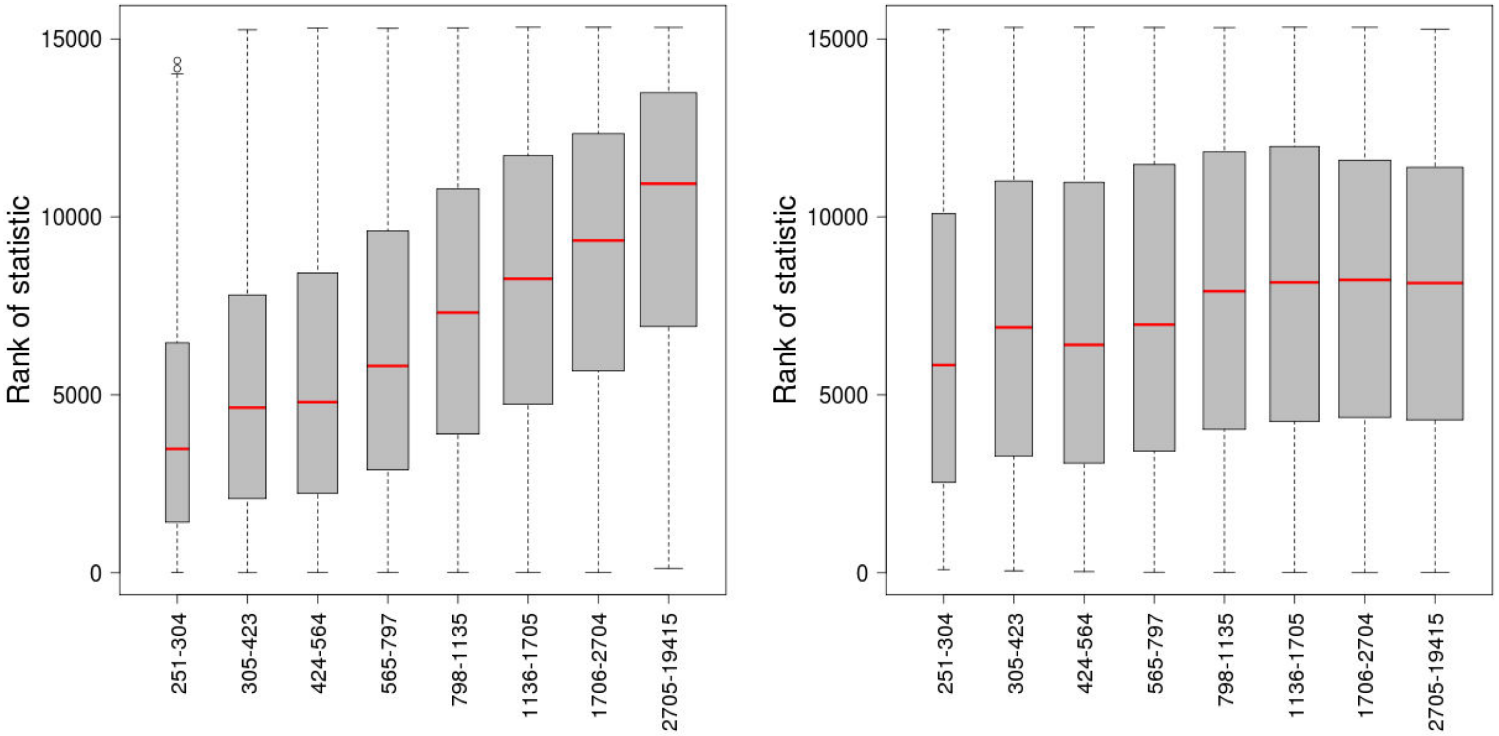
Evaluation results - ROC curves



a) no filtering

b) removing all genes with < 20 reads in either condition

Evaluation results - influence of gene length



(a) t-statistics

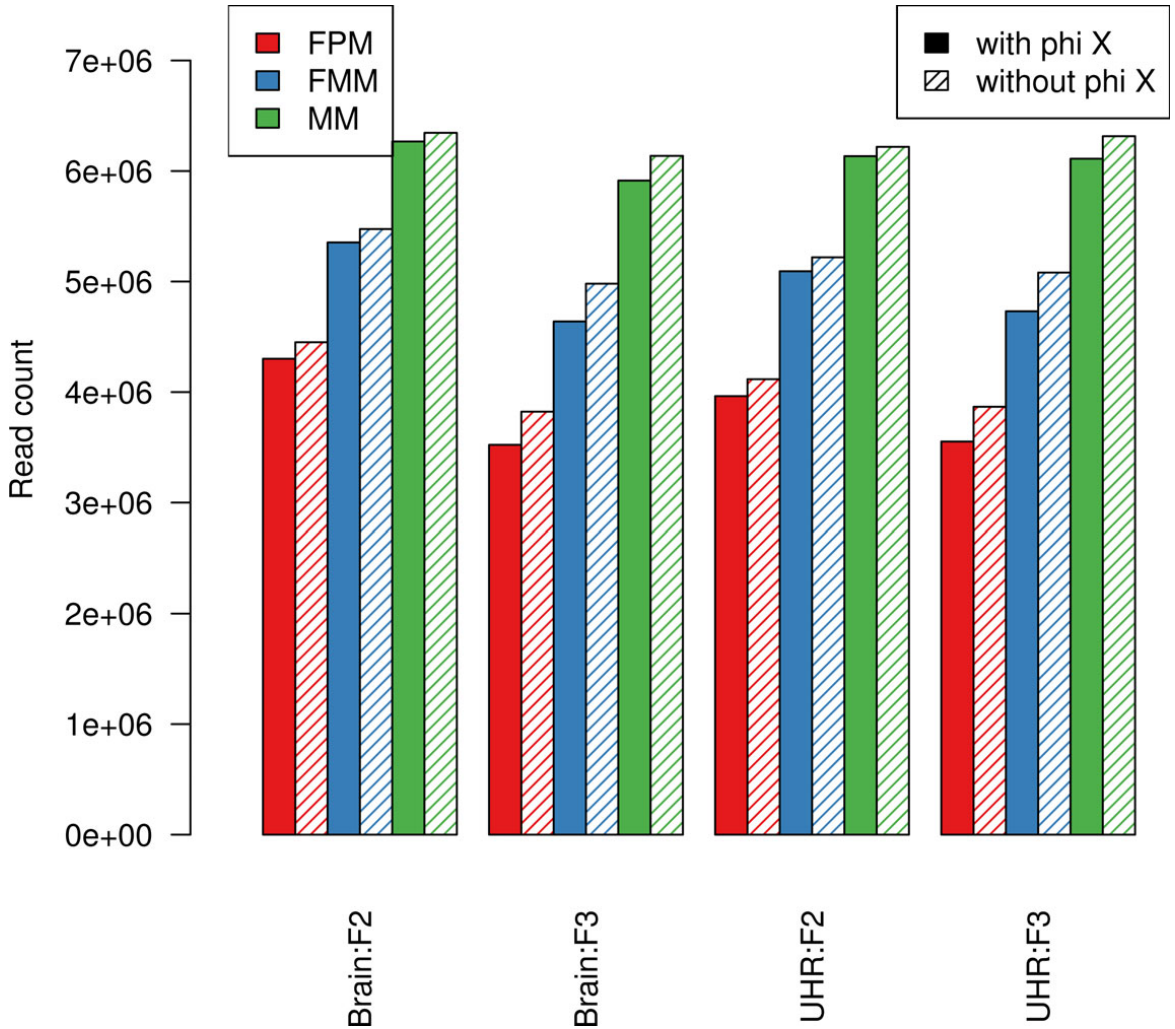
(b) Length-weighted t-statistics

ranks of DE statistics vs. gene lengths

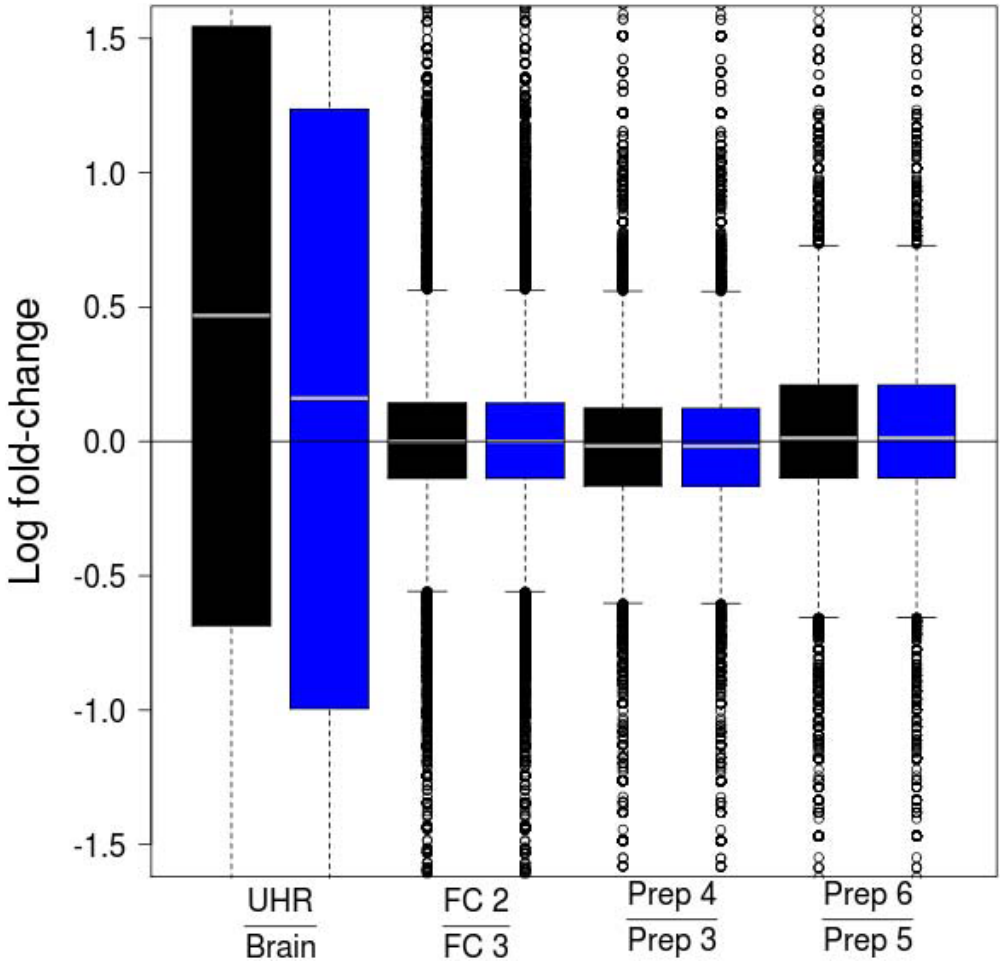
a) no weighting

b) weighting by $\frac{1}{\sqrt{\text{length}}}$

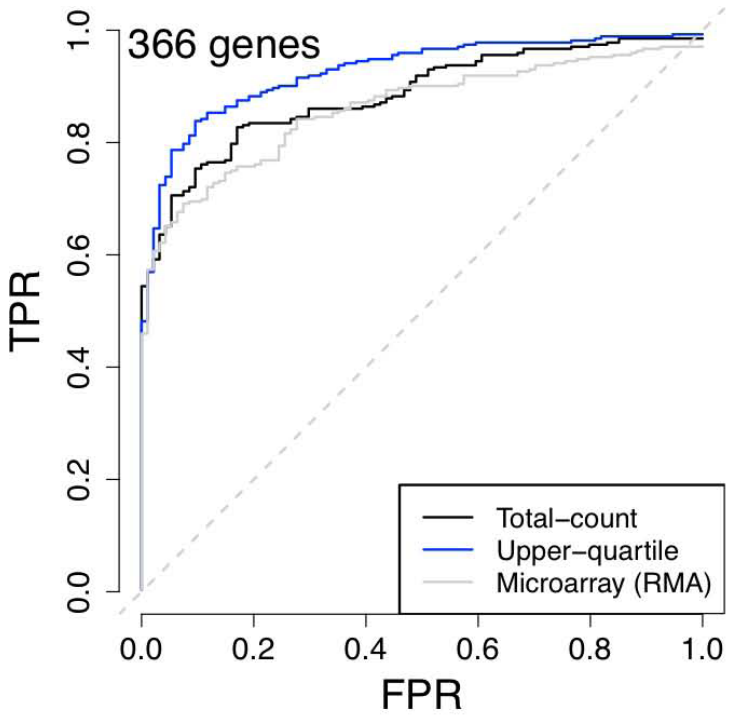
Evaluation results - calibration method



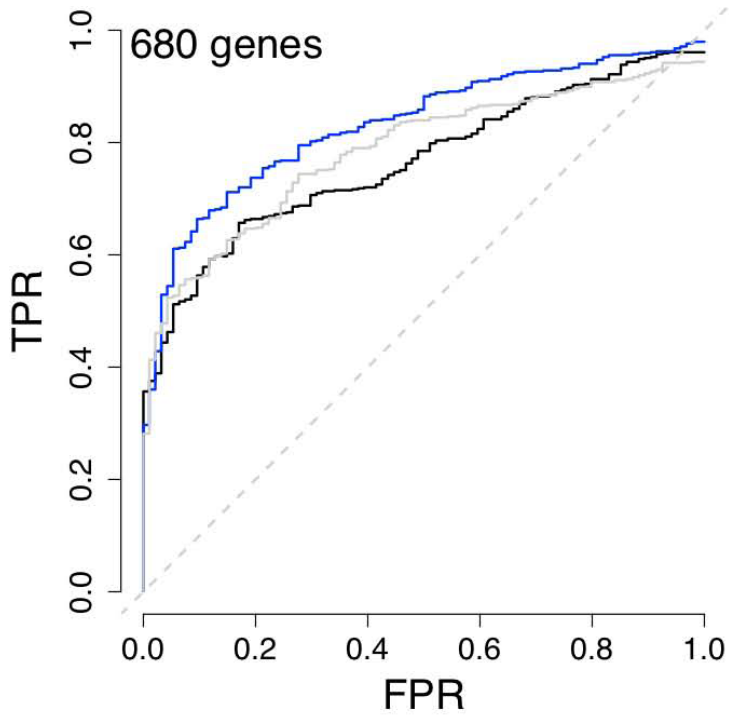
Evaluation results - biological and technical effects



Evaluation results - ROC curves RNA-Seq vs. Microarrays



(a) qRT-PCR positives: LR> 2



(b) qRT-PCR positives: LR> 0.5

Evaluation - summary

- LRT + fisher's test provide best results (t-tests fail if read count = 0)
- weighting by length
- phi-X calibration not neccessary
- larger variation between biological samples than between flow cells/library preparations
- sensitivity varies more between normalization procedures than between test statistics (!)

Thank you for your attention.

List of references

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

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