RNAseq:
Normalization and differential expression I

Jens Gietzelt

22.05.2012

- Robinson, Oshlack. *A scaling normalization method for differential expression analysis of RNA-seq data.* Genome Biology. 2010
- Hardcastle, Kelly. *baySeq: Empirical Bayesian methods for identifying differential expression in sequence count data.* BMC Bioinformatics. 2010

Outline of the presentation

1. Introduction

2. Pairwise calibration (EdgeR)

3. Differential expression

Introduction

normalization:

- comparison of expression levels between genes within a sample (same scale)
- however technical effects introduce a bias in the comparison between samples
- ⇒ normalization is crucial before performing differential expression
- calibration method EdgeR takes advantage of within-sample comparability

differential expression:

- appropriate distribution for count data
- incorporate calibration parameters

Framework

$Y_{g,k}$ ... observed count for gene g in library k

$N_k = \sum_{g=1}^{G} Y_{g,k}$ ... total number of reads for library k

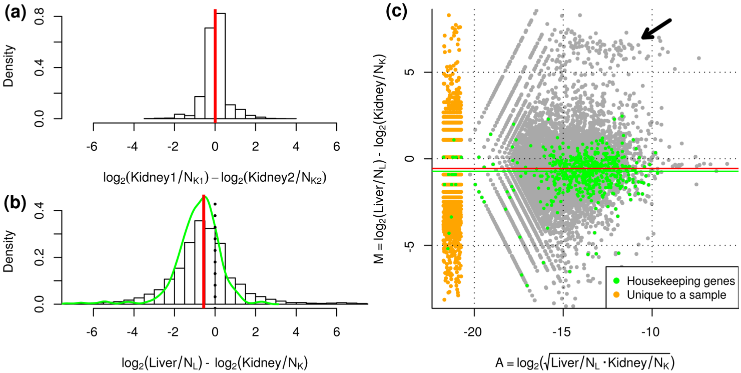
$\eta_{g,k}$ ... number of transcripts of gene g in library k
L_g ... length of gene g

$S_k = \sum_{g=1}^{G} \eta_{g,k} L_g$ ... total RNA output of sample k

$$E\left(Y_{g,k}\right) = \frac{\eta_{g,k} L_g}{S_k} N_k$$

- counts are a linear function of the number of transcripts
- library size calibration $\left(Y_{g,k}/N_k\right)$ is appropriate for the comparison of replicates
- comparison of biologically different samples may be biased by varying RNA composition

kidney vs. liver dataset



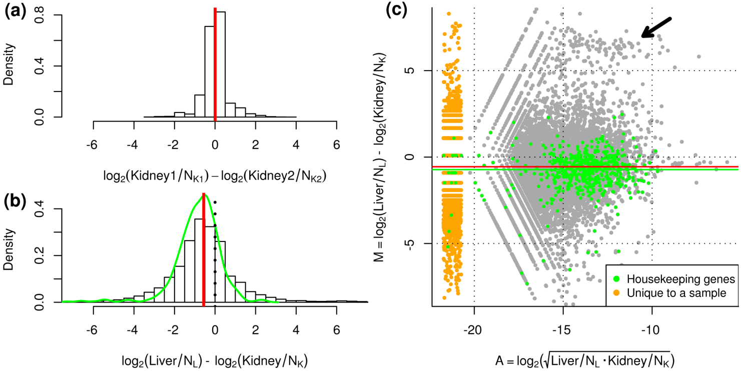
Trimmed mean of log-foldchange

- RNA production S_k of one sample cannot be determined directly
- estimation of relative differences of RNA production $f_k = S_k / S_r$ of a pair of samples (k, r)
- assumption: most genes are not differentially expressed
- $\Rightarrow$ compute robust mean over log-foldchanges:
  - double filtering over both mean and difference of log-values
  - calculate a weighted mean over the log-foldchanges
  - $\Rightarrow$ resacle factors $f_k = TMM_{(k,r)}$, where r is reference sample

$$\log_2\left(TMM_{(k,r)}\right) = \frac{\sum_{g \in G^*} w_{g,(k,r)} \left(\log_2\left(Y_{g,k}/N_k\right) - \log_2\left(Y_{g,r}/N_r\right)\right)}{\sum_{g \in G^*} w_{g,(k,r)}}$$

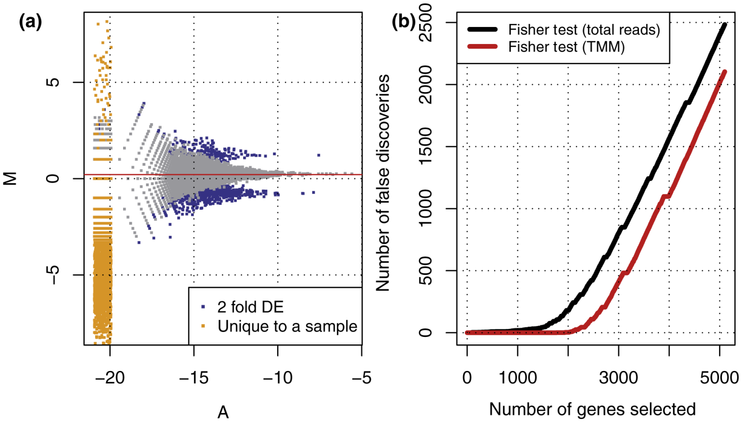
$$w_{g,(k,r)} = \left(\frac{1}{Y_{g,k}} - \frac{1}{N_k} + \frac{1}{Y_{g,r}} - \frac{1}{N_r}\right)^{-1}$$

kidney vs. liver dataset



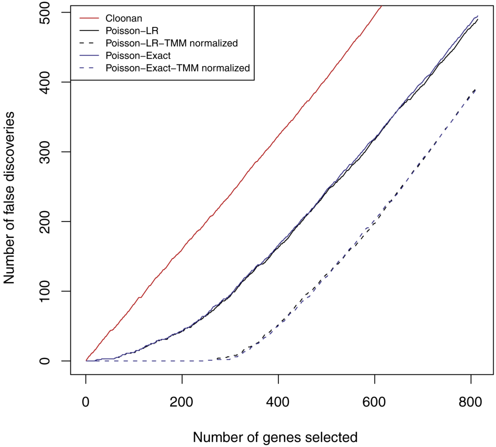
Simulation: pair of samples

simulated data sampled from poisson distribution



Simulation: replicates

Cloonan: log-transformation and quantile normalization



Differential expression

methods in use:

- DegSeq (normal distr.)
- EdgeR (negative binomial)
- DEseq (negative binomial, multiple groups)
- baySeq (negative binomial, multiple groups)
- Myrna (permutation based)

EdgeR

technical replicates: poisson distr.
biologically different samples: negative binomial distr.

$$Y \sim NB\left(p,\ m\right)$$

Y ... number of successes in a sequence of Bernoulli trials with probability p before r failures occur

alternative parametrization:
$q_{g,e}$ ... proportion of sequenced RNA of gene g for experimental group e

$$Y_{g,k,e} \sim NB\left(q_{g,e} N_k f_k,\ \phi_g\right)$$

$$E\left(Y_{g,k,e}\right) = \mu_{g,k,e} = q_{g,e} N_k f_k,\ \ Var\left(Y_{g,k,e}\right) = \mu_{g,k,e} + \mu_{g,k,e}^2 \phi_g$$

- test if $q_{g,1}$ is significantly different from $q_{g,2}$
- dispersons ϕ_g are moderated towards a common disperson

baySeq I

empirical Bayes approach to detect differential expression

$D_g \;=\; \{Y_{g,k},\; N_k, f_k\}_{k=1,\dots,K}$
M ... user specified model
θ_M ... vector of parameters of model M

$$P\left(M|D_g\right) = \frac{P\left(D_g|M\right)P\left(M\right)}{P\left(D_g\right)}$$

calculate marginal likelihood:

$$P\left(D_g|M\right) = \int P\left(D_g|\theta_M, M\right)P\left(\theta_M|M\right)d\theta_M$$

baySeq II

$$P\left(D_g|M\right) = \int P\left(D_g|\theta_M, M\right) P\left(\theta_M|M\right) d\theta_M$$

- e.g. Poisson-Gamma conjugacy, however no such conjugacy with negative binomial data
- $\Rightarrow$ define an empirical distribution on θ_M and estimate the marginal likelihood numerically

prior $P\left(M\right)$ is estimated by iteration:

$$P\left(M\right) = p_g, \quad p_g^* = P\left(M|D_g\right)$$

baySeq:

- applicable to complex experimental designs
- computationally intensive