Technologie w skali genomowej 2/ Algorytmiczne i statystyczne aspekty sekwencjonowania DNA

Expression analysis for RNA-seq data

Ewa Szczurek

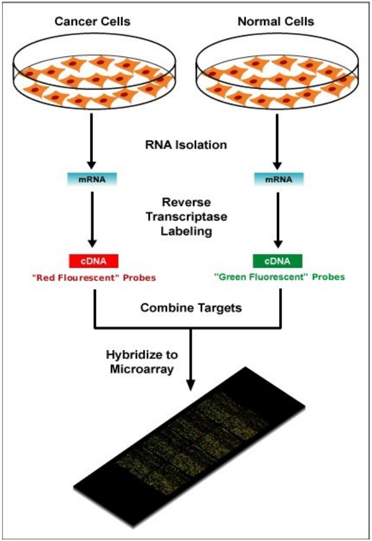
Instytut Informatyki
Uniwersytet Warszawski

The problem

We need to assess relative levels of transcript abundances in multiple samples This requires:

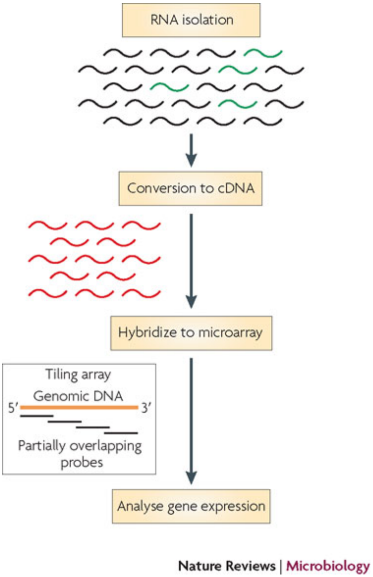
- Sample collection (gene arrays, tiling arrays or RNASeq)
- Signal normalization (bringing the measured signals to comparable values)
- Assessment of differential expression significance

Microarrays



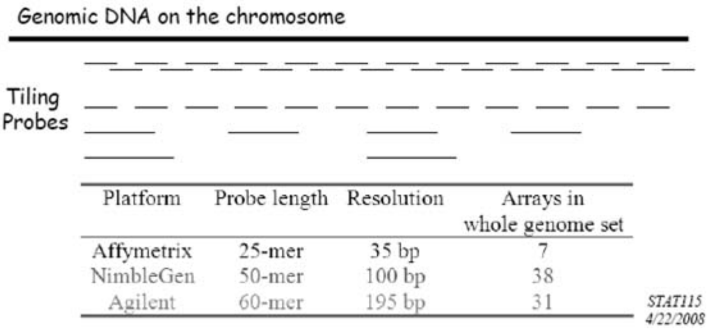
- designed to look at gene expression
- use a few probes for each known or predicted gene
- prehistory

Tiling arrays



- subtype of microarray chips.
- differ in the nature of the probes
- short fragments, designed to cover the entire genome or contiguous regions of the genome
- depending on probe lengths and spacing, different degrees of resolution can be achieved

Tiling arrays



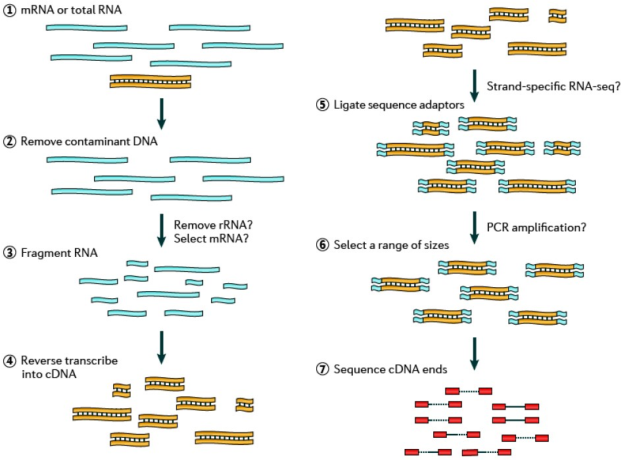
Platform	Probe length	Resolution	Arrays in whole genome set
Affymetrix	25-mer	35 bp	7
NimbleGen	50-mer	100 bp	38
Agilent	60-mer	195 bp	31

STAT115
4/22/2008

For Affymetrix tiling arrays:

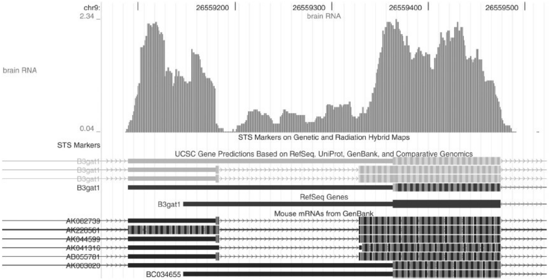
- contain 25-nt probes tiled every 35 bp of DNA sequence.
- whole genome arrays (2.0R) are comprised of 7 chips covering entire human or mouse genome that is masked of repeat sequences.

RNA-seq data preparation



J. A. Martin and Z. Wang *Next-generation transcriptome assembly*. Nature Reviews 2011.

Read count matrix



- A $n \times m$ count matrix N, where N_{gs} is the number of reads assigned to gene g in sample s
- Produced from alignment data (eg using HTSeq, or Picard)
- Not a direct measure of gene expression!
- Rather, $N_{gs} \propto l_g \mu_{gs}$, where l_g is the gene g length, and μ_{gs} is the expected expression.

Normalization

Definition (Normalization)

Normalization is a process designed to identify and correct technical biases removing the least possible biological signal. This step is technology and platform-dependant.

Nomenclature

sample material with a specific source, e.g. culture or tissue.

replicates several independent samples with the same material type and origin

condition environment in which samples are prepared (e.g. added chemicals). There can be several samples per condition

flow cell a glass slide where the sequencing takes place

line one of eight independent sequencing areas that a flow cell is made up from

library contains cDNAs representative of the RNA molecules that are extracted from a given sample, pre-processed and deposited on lanes in order to be sequenced

library size the number of mapped short reads obtained from sequencing of the library.

Here, one sample ⇔ one line ⇔ one library ⇔ one condition.

Dillies et al. *A comprehensive evaluation of normalization ...*. Brief. in Bioinformatics 2012.

Two issues calling for normalization

1. Bias in sample size
2. Bias in over-represented genes - genes whose counts dominate the sample size

Normalization by scaling factor

Total count (TC)

$$N_{gs} \times \frac{\text{mean}(\sum_i N_{ij}) \text{ across samples } j}{\sum_j N_{js}} \tag{1}$$

Upper Quartile (UQ)

$$N_{gs} \times \frac{\text{mean upper quartile of } N_{ij} \neq 0 \text{ across samples } j}{\text{mean upper quartile of } N_{js}} \tag{2}$$

Median (Med)

$$N_{gs} \times \frac{\text{med}(N_{ij} \neq 0) \text{ across samples } j}{\text{med}(N_{js} \neq 0)} \tag{3}$$

Normalization by scaling factor – cont.

Hypothesis: most genes are not DE, should have similar read counts across samples.

DESeq

$$N_{gs} \times \text{med}\left(\frac{\text{geometric mean}(N_{ij}) \text{ across samples } j}{N_{is}}\right), \tag{4}$$

across genes i.

Trimmed Mean of M-values (TMM) Similar to DESeq but uses means and removes outliers.

RPKM/FPKM normalization

Reads/Fragments Per

- Kilobase of transcript sequence
- Million base pairs sequenced

RPKM/FPKM normalization

Reads/Fragments Per

- Kilobase of transcript sequence
- Million base pairs sequenced

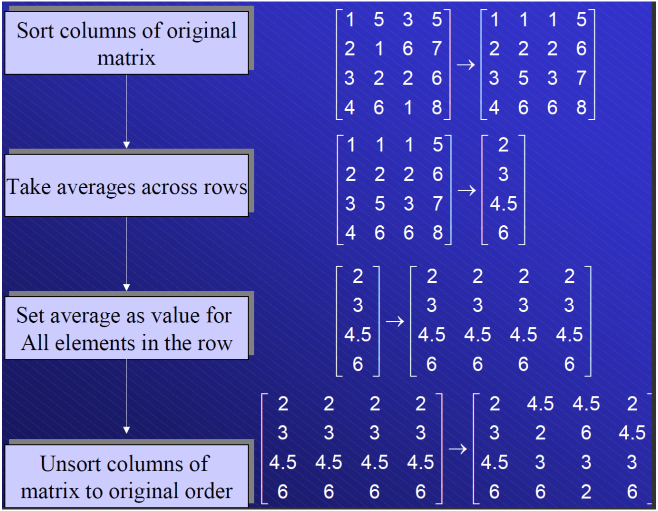
+ correction for gene/transcript length
+ correction for sequencing depth

– no correction for difference in expression distribution between samples

– relation between read number and variation is lost

Quantile normalization (Q)

A technique for making two distributions identical in statistical properties.



Sort columns of original matrix

$$\begin{bmatrix}1&5&3&5\\2&1&6&7\\3&2&2&6\\4&6&1&8\end{bmatrix} \to \begin{bmatrix}1&1&1&5\\2&2&2&6\\3&5&3&7\\4&6&6&8\end{bmatrix}$$

Take averages across rows

$$\begin{bmatrix}1&1&1&5\\2&2&2&6\\3&5&3&7\\4&6&6&8\end{bmatrix} \to \begin{bmatrix}2\\3\\4.5\\6\end{bmatrix}$$

Set average as value for All elements in the row

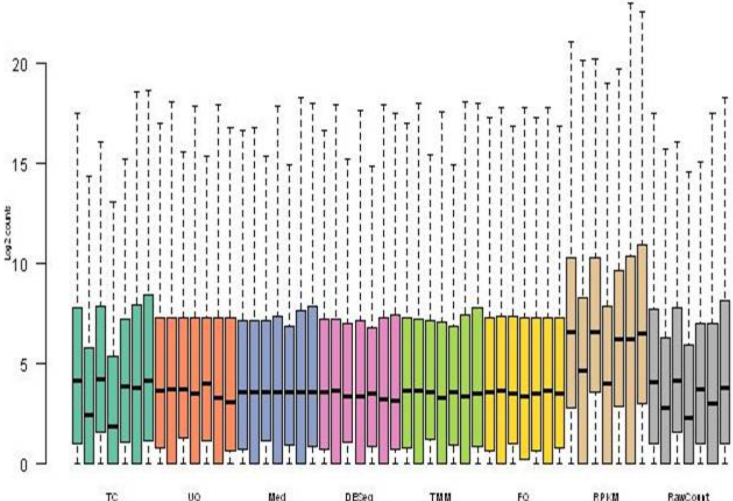
$$\begin{bmatrix}2\\3\\4.5\\6\end{bmatrix} \to \begin{bmatrix}2&2&2&2\\3&3&3&3\\4.5&4.5&4.5&4.5\\6&6&6&6\end{bmatrix}$$

Unsort columns of matrix to original order

$$\begin{bmatrix}2&2&2&2\\3&3&3&3\\4.5&4.5&4.5&4.5\\6&6&6&6\end{bmatrix} \to \begin{bmatrix}2&4.5&4.5&2\\3&2&6&4.5\\4.5&3&3&3\\6&6&2&6\end{bmatrix}$$

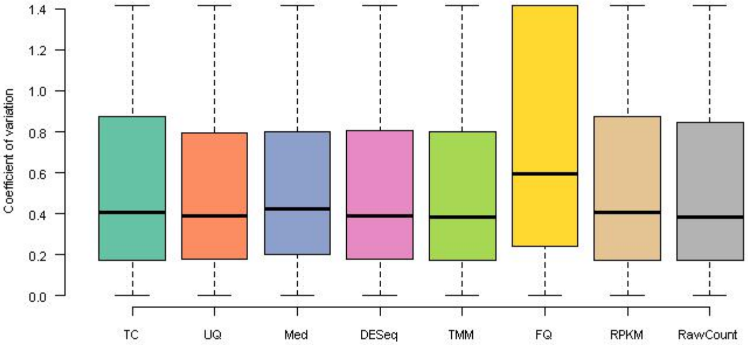
Comparison of normalization methods on real data - normalized data distribution

When large differences in library size, TC and RPKM do not improve over the raw counts.



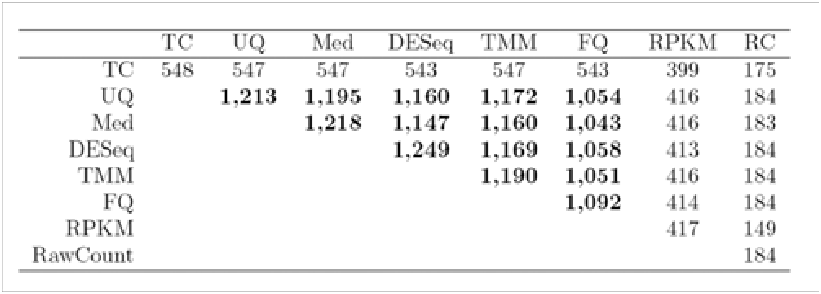
Comparison of normalization methods on real data - within-condition variability

Example: *Mus musculus*, condition D dataset



M.-A. Gillies et al. *A comprehensive evaluation of normalization . . .* Brief. in Bioinformatics 2012.

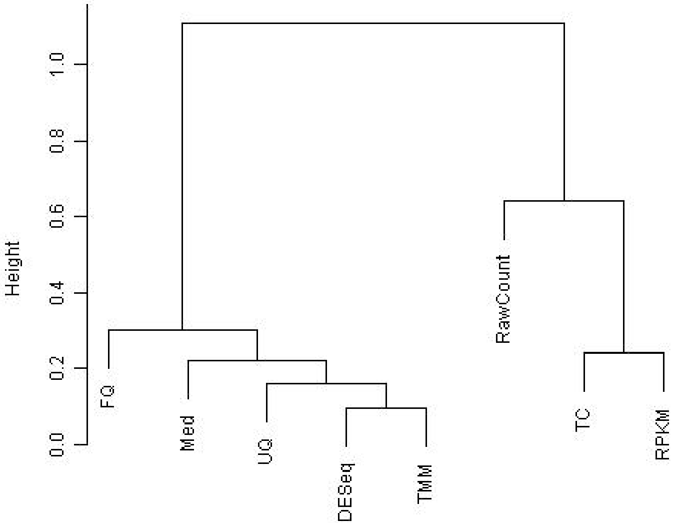
Comparison of normalization methods by DE gene number



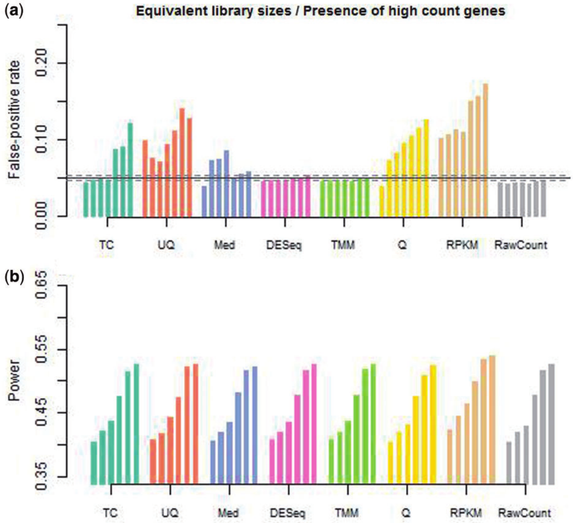
	TC	UQ	Med	DESeq	TMM	FQ	RPKM	RC
TC	548	547	547	543	547	543	399	175
UQ		**1,213**	**1,195**	**1,160**	**1,172**	**1,054**	416	184
Med			**1,218**	**1,147**	**1,160**	**1,043**	416	183
DESeq				**1,249**	**1,169**	**1,058**	413	184
TMM					**1,190**	**1,051**	416	184
FQ						**1,092**	414	184
RPKM							417	149
RawCount								184

M.-A. Gillies et al. *A comprehensive evaluation of normalization . . .* Brief. in Bioinformatics 2012.

Consensus dendogram



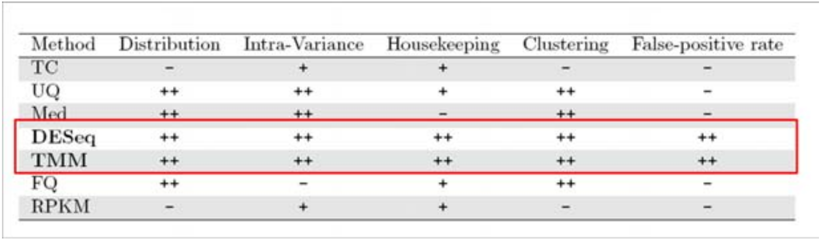
Comparison of normalization methods on simulated data - error rate and power



M.-A. Gillies et al. *A comprehensive evaluation of normalization . . .* . Brief. in Bioinformatics 2012.

So the winner is...?

- in most cases, the methods give similar results
- the differences appear in data characteristics



Method	Distribution	Intra-Variance	Housekeeping	Clustering	False-positive rate
TC	-	+	+	-	-
UQ	++	++	+	++	-
Med	++	++	-	++	-
DESeq	++	++	++	++	++
TMM	++	++	++	++	++
FQ	++	-	+	++	-
RPKM	-	+	+	-	-

Interpretation

RawCount Often fewer differential expressed genes (e.g. *A. fumigatus*: no DE gene)

TC, RPKM

- Sensitive to the presence of predominant genes
- Less effective stabilization of distributions
- Ineffective (similar to RawCount)

Q

- Can increase between group variance
- Is based on a (too) strong assumption (similar distributions)

Med High variability of housekeeping genes

TC, RPKM, Q, Med, UQ Adjustment of distributions, implies a similarity between RNA repertoires expressed

Concusions on normalization

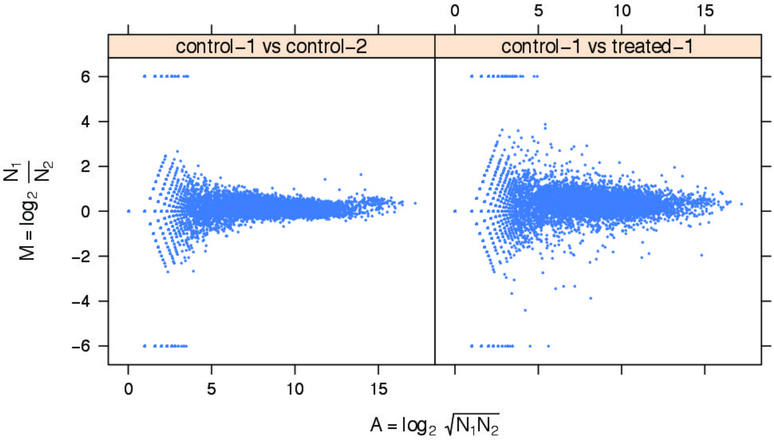
- RNA-seq data are affected by technical biaises (total number of mapped reads per lane, gene length, composition bias)
- Normalization is needed and has a great impact on the DE genes
- Detection of differential expression in RNA-seq data is inherently biased (more power to detect DE of longer genes)
- Do not normalise by gene length in a context of differential analysis.
- TMM and DESeq : performant and robust methods in a DE analysis context on the gene scale.

Differential analysis

Aim : To detect differentially expressed genes between two conditions

- Discrete quantitative data
- Few replicates
- Overdispersion problem

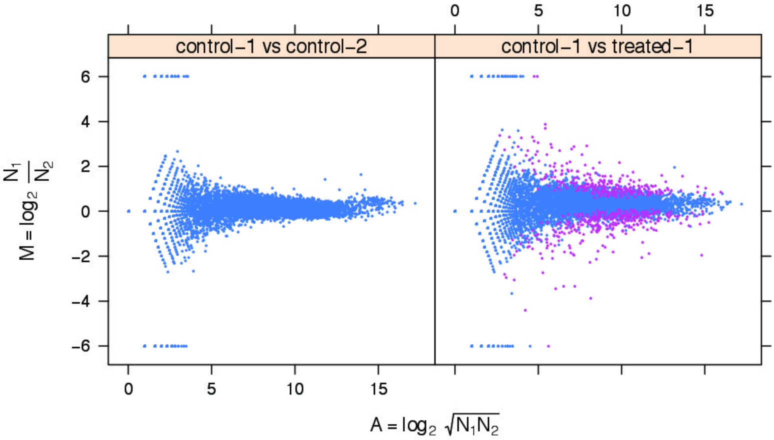
Differential analysis



two biological replicates

treatment vs control

Differential analysis



two biological replicates

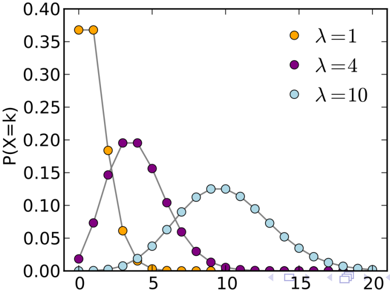
treatment vs control

Poisson distribution

For X Poisson-distributed

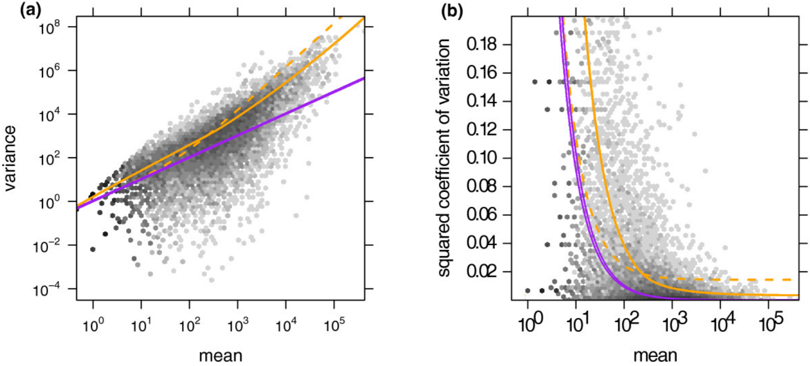
$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

- expresses the probability of a given number of events occurring in a fixed interval of time or space
- assumes these events occur with a known average rate and independently of the time since the last event.
- variance equal to the mean (λ)



Overdispersion

- Poisson distribution was proposed to model read count data
- No need to estimate the variance. This is convenient
- E.g., Wang *et al* (2010), Bloom *et al* (2009), Kasowski *et al* (2010), ...
- however, when models are fit, the observed variance is higher than the variance of theoretical models ⇒ overdispersion ⇒ type-I errors (false DE discoveries).



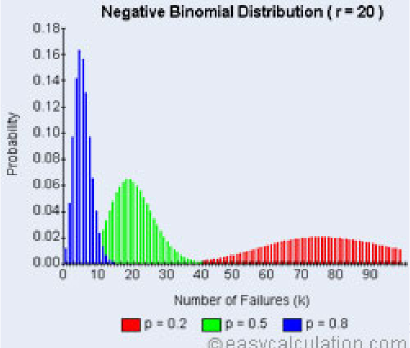
Negative binomial distribution

- Suppose a sequence of independent Bernoulli trials.
- The probability of success is p and of failure is $(1-p)$.
- We observe this sequence until r failures occurr.

Then for the random number of successes we have seen,

$$P(X=k)=\binom{k+r-1}{k}(1-p)^r p^k,$$

with mean $\mu=\frac{pr}{1-p}$ and variance $\sigma^2=\frac{pr}{(1-p)^2}$.



Negative binomial distribution re-parametrized

Let

$$\alpha = \frac{1}{r},$$

and mean as before

$$\mu = \frac{pr}{1-p}.$$

Then the mean for NB is μ and variance $\mu + \alpha\mu^2$.

edgeR

- Model count data with NB distributions
- The number of replicates in read count data is often too small to reliably estimate mean μ and variance σ^2 parameters reliably
- Assume mean and variance are related by $\sigma^2 = \mu + \alpha\mu^2$, with a single proportionality constant α, estimated for each gene.

Robinson and Smyth, 2010

DEseq

$$N_{gs} = NB(\mu_{gs}, \sigma^2_{gs})$$

Three assumptions:

1. $\mu_{gs} = q_{g,\rho(s)} f_s$, where
   - $\rho(s)$ denotes the experimental condition of sample s
   - $q_{g,\rho(s)}$ is proportional to the expectation value of the true (but unknown) concentration of reads from gene g under condition $\rho(s)$
   - f_s is the DEseq normalization factor.
2. $\sigma^2_{gs} = \mu_{gs} + v_{g,\rho(s)} f_s^2$. Here μ_{gs} is the technical, Poisson-distributed variance (shot noise), and $v_{g,\rho(s)} f_s^2$ refers to *raw variance*.
3. $v_{g,\rho}$ is a smooth function of $q_{s,\rho}$: $v_{g,\rho(s)} = v_\rho(q_{g,\rho(s)})$. This allows to pool the data from genes with similar expression strength for the purpose of variance estimation.

Anders and Huber, 2010

DEseq model fitting

Assume n genes and m samples, and k experimental conditions. We have the following parameters estimated:

1. m size factors f_s (expected values of all counts from sample s proportional to f_s)
2. kn expression strength parameters $q_{g,\rho}$, for each condition ρ and gene g, (expected values of counts for gene g in condition ρ are proportional to $q_{g,\rho}$):

$$\hat{q}_{g,\rho} = \frac{1}{m_\rho} \sum_{s:\rho(s)=\rho} \frac{N_{g,s}}{f_s},$$

   i.e., the averaged normalized counts from samples for condition ρ, with $m_\rho =$ the number of samples for condition ρ.
3. k smooth functions $v_\rho : R^+ \Rightarrow R^+$. For each condition ρ, v_ρ models the dependence of the raw variance $v_{g,\rho}$ on the expected mean $q_{g,\rho}$.

Anders and Huber, 2010

DEseq model fitting

For estimation of raw variance $v_{g,\rho}$:

1. Calculate normalized condition variance estimates

$$w_{g,\rho} = \frac{1}{m_\rho - 1} \sum_{s:\rho(s)=\rho} \left(\frac{N_{g,s}}{f_s} - \hat{q}_{g,\rho} \right)^2$$

2. define

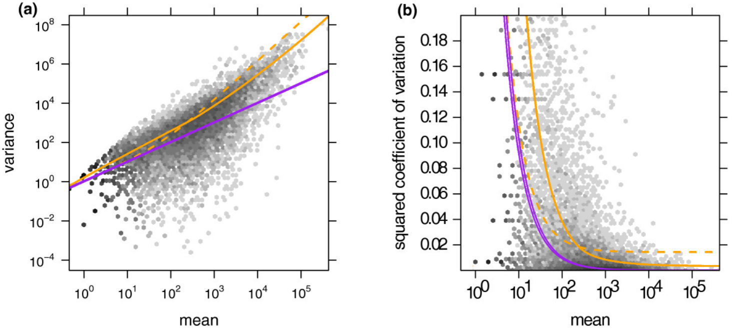
$$z_{g,\rho} = \frac{\hat{q}_{g,\rho}}{m_\rho} \sum_{s:\rho(s)=\rho} \frac{1}{f_s}$$

3. Theorem: $w_{g,\rho} - z_{g,\rho}$ is an unbiased estimator of raw variance.
4. For small m_ρ not useful. Instead regress on $(\hat{q}_{g,\rho}, w_{g,\rho})$ to obtain a smooth function $w_\rho(q)$ and estimate raw variance with

$$\hat{v}_\rho(\hat{q}_{g,\rho}) = w_\rho(\hat{q}_{g,\rho}) - z_{g,\rho}.$$

Anders and Huber, 2010

DEseq model fitting



Orange line regression w_ρ of y-axis: condition variance estimator $w_{g,\rho}$, on x-axis: means estimator $\hat{q}_{g,\rho}$.

Dashed orange line edgeR variance estimator

Violet line Poisson variance (=mean)

Anders and Huber, 2010

DEseq DE calling

- For gene g and two conditions, 1 and 2, we want to evaluate whether gene g is differentially expressed between 1 and 2
- Hypothesis testing: H_0: the means $q_{g,1} = q_{g,2}$ equal.