DESeq2

Laurent Modolo

March 16 2018

Table of Contents

1 mRNA level measurements

2 Counts distributions

3 DESeq2 model

Table of Contents

1 mRNA level measurements

2 Counts distributions

3 DESeq2 model

mRNA level measurements

Microarrays:

- fluorescence level
- continuous measurement

$$X_{ij} \sim \mathcal{N}(\mu, \sigma)$$

for genes i in condition j, μ is the average of the signal and σ it's dispersion.

RNASeq:

- number of reads
- discreet measurement

$$X_{ij} \sim \mathcal{P}(\mu)$$

μ is the number of reads transcribed from the genes i in condition j by unit of time.

Table of Contents

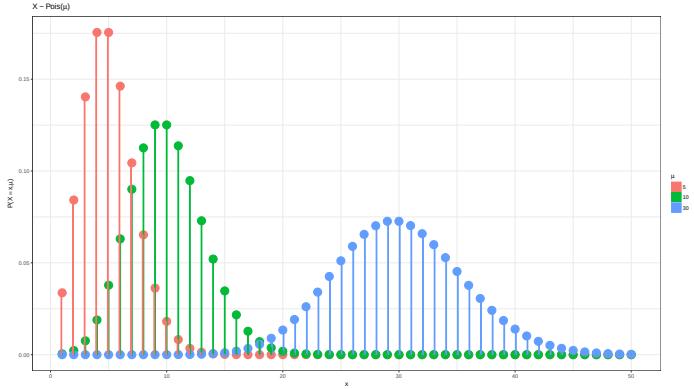
1 mRNA level measurements

2 Counts distributions

3 DESeq2 model

Counts distributions

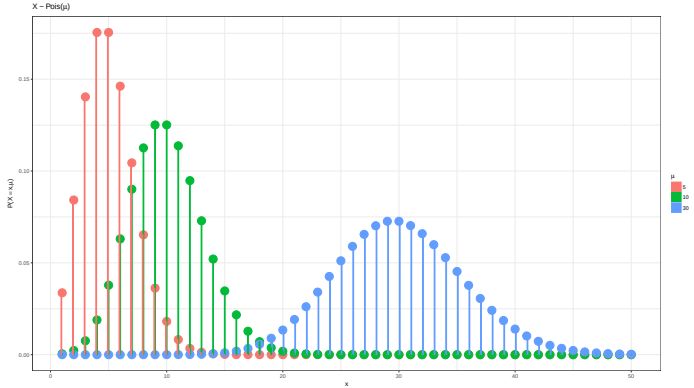
$P(X = x)$ for $\mathcal{P}(\mu)$



μ the rate of reads production is equal to the variability in the number of reads.

Counts distributions

$P(X = x)$ for $\mathcal{P}(\mu)$

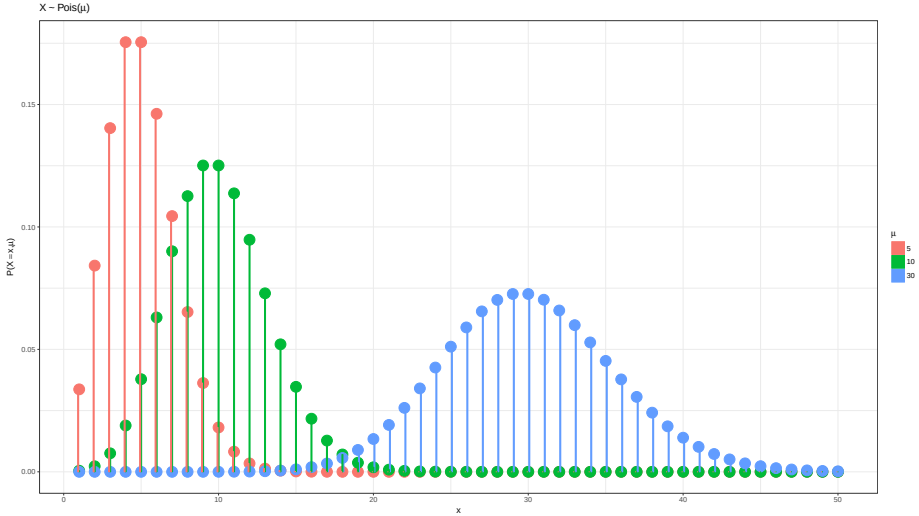


μ the rate of reads production is equal to the variability in the number of reads.

We often have more variability! (broader distributions)

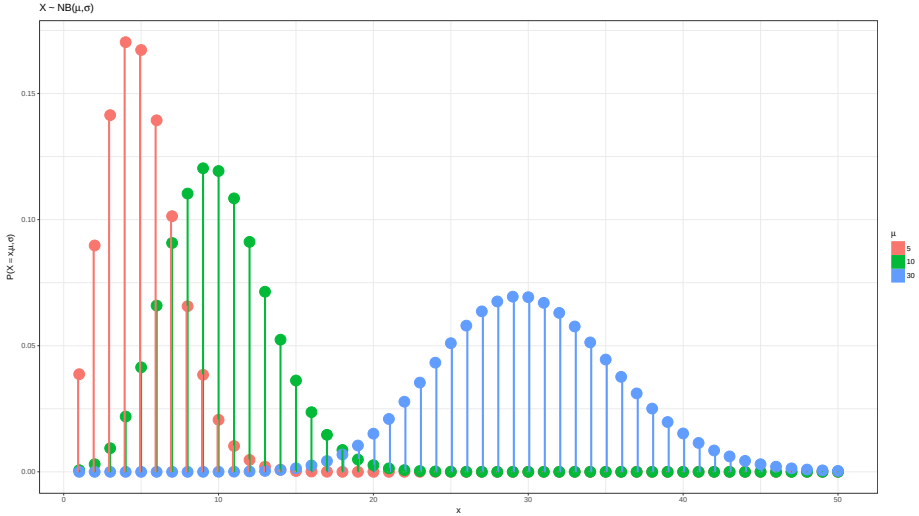
Counts distributions

$P(X = x)$ for $\mathcal{NB}(\mu, \sigma)$



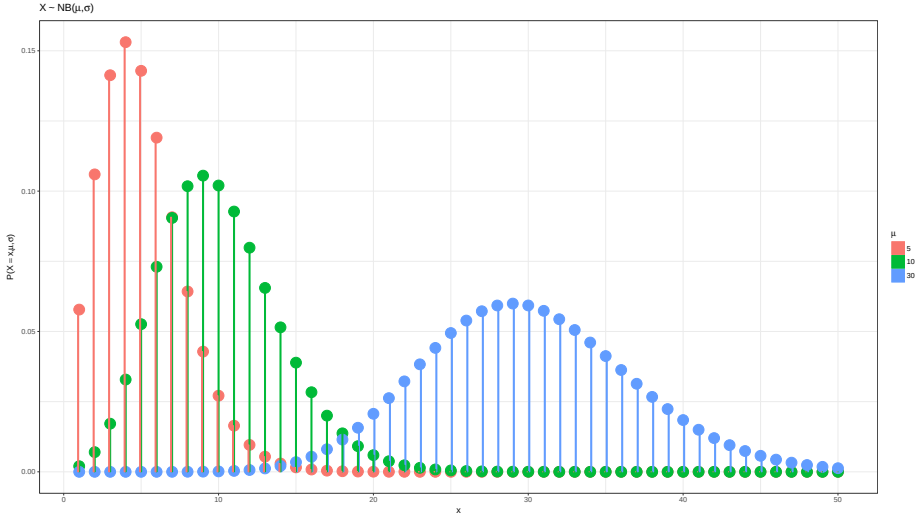
Counts distributions

$P(X = x)$ for $\mathcal{NB}(\mu, \sigma)$



Counts distributions

$P(X = x)$ for $\mathcal{NB}(\mu, \sigma)$



Counts distributions

$P(X = x)$ for $\mathcal{NB}(\mu, \sigma)$

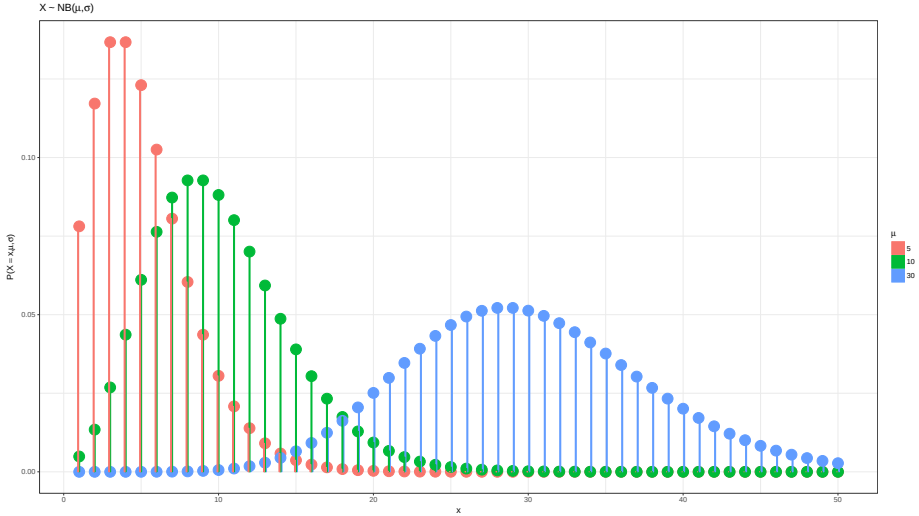


Table of Contents

1 mRNA level measurements

2 Counts distributions

3 DESeq2 model

DESeq2 model: size factors

The observed counts are modeled as following:

$$K_{ij} \sim \mathcal{NB}(\mu_{ij}, \alpha_i)$$

for genes $i \in \{1, \ldots, n\}$ in condition $j \in \{1, \ldots, m\}$, with:

DESeq2 model: size factors

The observed counts are modeled as following:

$$K_{ij} \sim \mathcal{NB}(\mu_{ij}, \alpha_i)$$

for genes $i \in \{1, \ldots, n\}$ in condition $j \in \{1, \ldots, m\}$, with:

$$\mu_{ij} = s_j q_{ij}$$

with s_j the size factor of replicate j and q_{ij} proportional to the number of cDNA fragments.

DESeq2 model: dispersion

The size factors are computed as following:

$$s_j = \underset{i}{\text{median}} \frac{K_{ij}}{K_i^R}$$

with

$$K_i^R = \left(\prod_{j=1}^{m} K_{ij} \right)^{\frac{1}{m}}$$

DESeq2 model

The observed counts are modeled as following:

$$K_{ij} \sim \mathcal{NB}(\mu_{ij}, \alpha_i)$$

$$Var(K_{ij}) = \mu_{ij} + \alpha_i \mu_{ij}^2$$

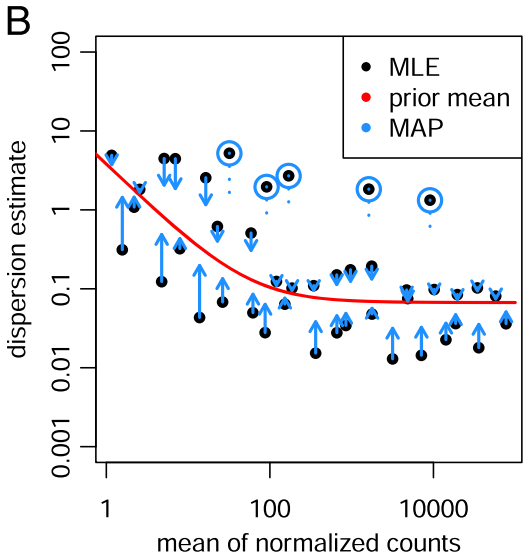
$$\log \alpha_i \sim \mathcal{N}(\ \log \alpha_{tr}(\bar{\mu}_i), \sigma_d^2)$$

with:

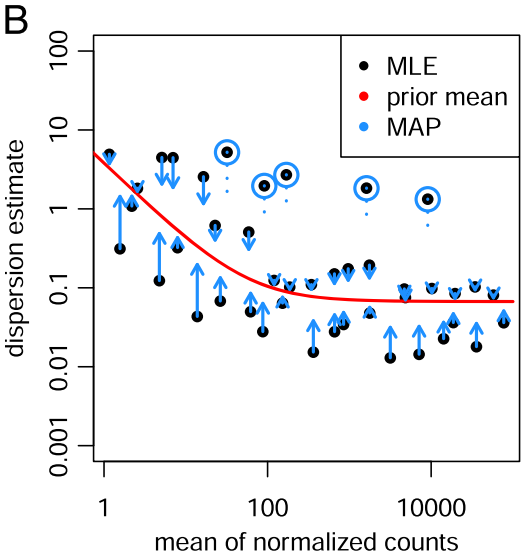
$$\bar{\mu}_i = \frac{1}{m} \sum_j \frac{K_{ij}}{s_{ij}}$$

and:

$$\alpha_{tr}(\bar{\mu}) = \frac{a_1}{\bar{\mu}} + \alpha_0$$

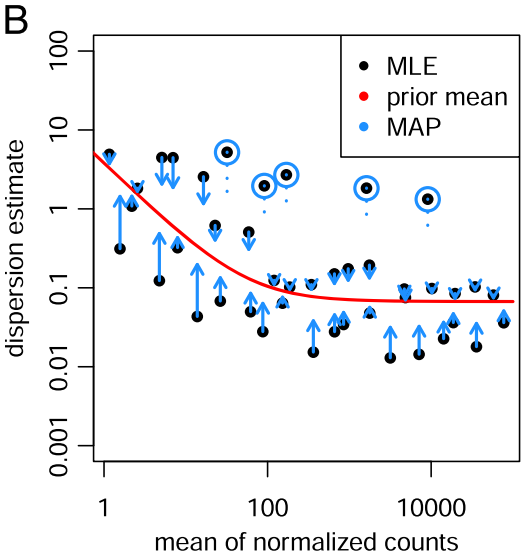


DESeq2 model: dispersion



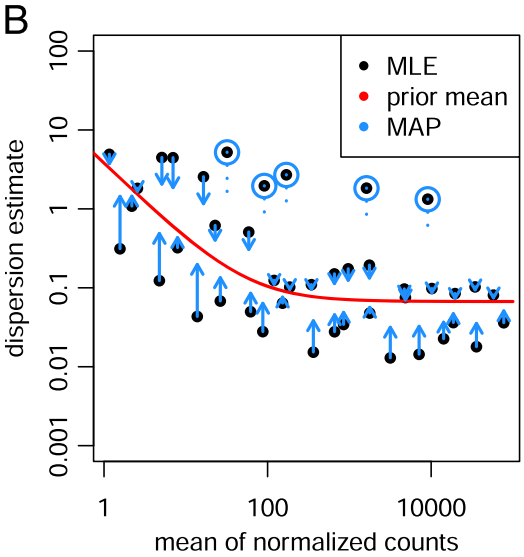
1. MLE computation of each genes α_i^{gw}
   - black dots
   - extremely noisy (low number of replicates)
   - would compromise the accuracy of the analysis if used direclty
2. fit a smooth curve for the dispersion trend $\alpha_{tr}(\bar{\mu})$
3. compute α_i by MAP using $\alpha_{tr}(\bar{\mu})$
4. keep α_i^{gw} for genes more than 2 residual standard deviations above the curve.

DESeq2 model: dispersion



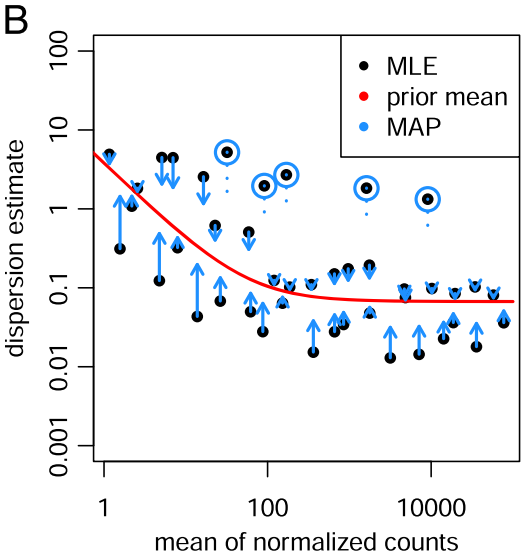
1. MLE computation of each genes α_i^{gw}
2. fit a smooth curve for the dispersion trend $\alpha_{tr}(\bar{\mu})$
   - red line
   - share information across genes
   - high dependence between α and μ for low counts
   - asymptotic dispersion of α_0
3. compute α_i by MAP using $\alpha_{tr}(\bar{\mu})$
4. keep α_i^{gw} for genes more than 2 residual standard deviations above the curve.

DESeq2 model: dispersion



1. MLE computation of each genes α_i^{gw}
2. fit a smooth curve for the dispersion trend $\alpha_{tr}(\bar{\mu})$
3. compute α_i by MAP using $\alpha_{tr}(\bar{\mu})$
   - blue arrow
   - shrink α_i^{gw} toward $\alpha_{tr}(\bar{\mu})$
   - shrinkage decreases with the distance to $\alpha_{tr}(\bar{\mu})$
   - shrinkage decreases with the degree of freedom
4. keep α_i^{gw} for genes more than 2 residual standard deviations above the curve.

DESeq2 model: dispersion



1. MLE computation of each genes α_i^{gw}
2. fit a smooth curve for the dispersion trend $\alpha_{tr}(\bar{\mu})$
3. compute α_i by MAP using $\alpha_{tr}(\bar{\mu})$
4. keep α_i^{gw} for genes more than 2 residual standard deviations above the curve.
   - blue circles
   - decreases false positives

DESeq2 model: hypothesis testing

The observed counts are modeled as following:

$$K_{ij} \sim \mathcal{NB}(\mu_{ij}, \alpha_i)$$

for genes $i \in \{1, \ldots, n\}$ in condition $j \in \{1, \ldots, m\}$, with:

$$\mu_{ij} = s_j q_{ij}$$

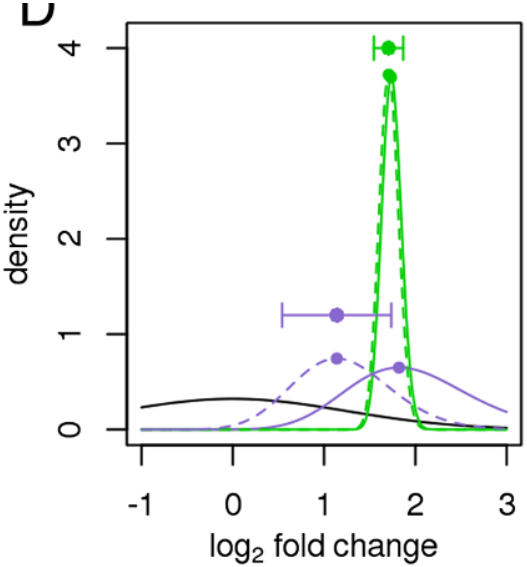
with s_j the size factor of replicate j and q_{ij} proportional to the number of cDNA fragments.

Each gene can be analysed with the following GLM:

$$\log q_{ij} = \sum_r x_{jr} \beta_{ir}$$

with x_r a factor (treated or control) and β_r the corresponding coefficient.
The use of linear models, however, provides the flexibility to also analyze more complex designs

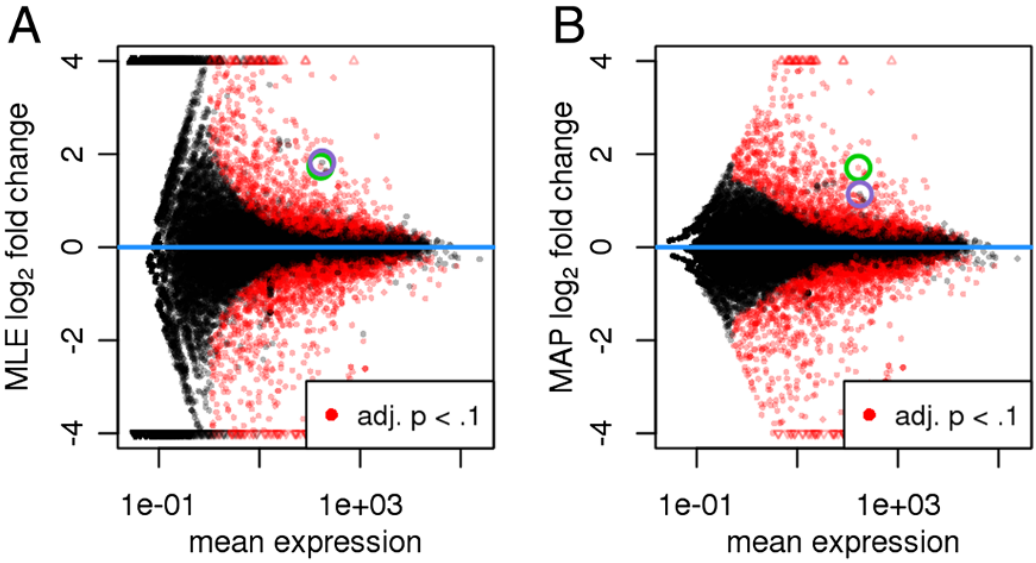
DESeq2 model: LFCs



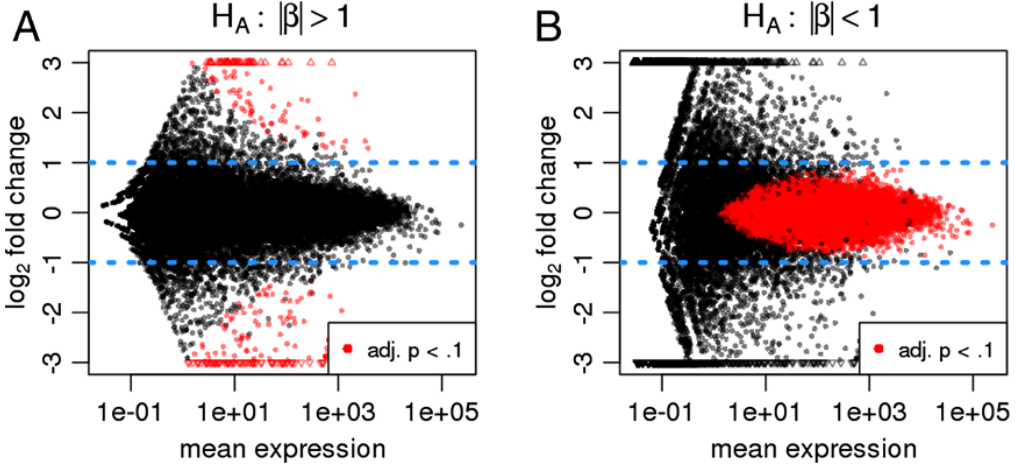
The same MAP approach is used to compute the LFCs

1. MLE estimage of the LFCs
2. fit a zero centred Gaussian overs the LFCs
3. compute final LFCs by MAP
   - shrinkage is stronger for genes with low information
   - low counts
   - high dispersion

DESeq2 model



DESeq2 model: hypothesis testing



Thank you

DOI10.1186/s13059-014-0550-8