Normalising RNA-seq data

Ernest Turro

University of Cambridge

31 Oct 2012

Aims of normalisation

Normalisation aims to ensure our expression estimates are:

- **comparable across features** (genes, isoforms, etc)
- **comparable across libraries** (different samples)
- **on a human-friendly scale** (interpretable magnitude)

Aims of normalisation

Normalisation aims to ensure our expression estimates are:

- **comparable across features** (genes, isoforms, etc)
- **comparable across libraries** (different samples)
- **on a human-friendly scale** (interpretable magnitude)

Necessary for valid inference about DE

- between transcripts within samples
- between samples belonging to different biological conditions

Basic Poisson model

Number of reads from gene g in library i can be captured by a Poisson model (Marioni et al. 2008):

$$\begin{aligned} r_{ig} &\sim Poisson(k_{ig}\mu_{ig}), \\ \implies \mathbb{E}(r_{ig}) &= k_{ig}\mu_{ig} \end{aligned}$$

where μ_{ig} is the concentration of RNA in the library and k_{ig} is a normalisation constant.

$$\boxed{\hat{\mu}_{ig} = \frac{r_{ig}}{k_{ig}}}$$

RPKM normalisation

Normalisation is all about deciding how to set k_{ig} such that the estimates of μ_{ig} are comparable between genes and across libraries.

$$\hat{\mu}_{ig} = \frac{r_{ig}}{k_{ig}}$$

RPKM normalisation

Normalisation is all about deciding how to set k_{ig} such that the estimates of μ_{ig} are comparable between genes and across libraries.

$$\hat{\mu}_{ig} = \frac{r_{ig}}{k_{ig}}$$

The number of reads r_{ig} is roughly proportional to

- the length of the gene, l_g
- the total number of reads in the library, N_i

Thus it is natural to include them in the normalisation constant.

RPKM normalisation

Normalisation is all about deciding how to set k_{ig} such that the estimates of μ_{ig} are comparable between genes and across libraries.

$$\hat{\mu}_{ig} = \frac{r_{ig}}{k_{ig}}$$

The number of reads r_{ig} is roughly proportional to

- the length of the gene, l_g
- the total number of reads in the library, N_i

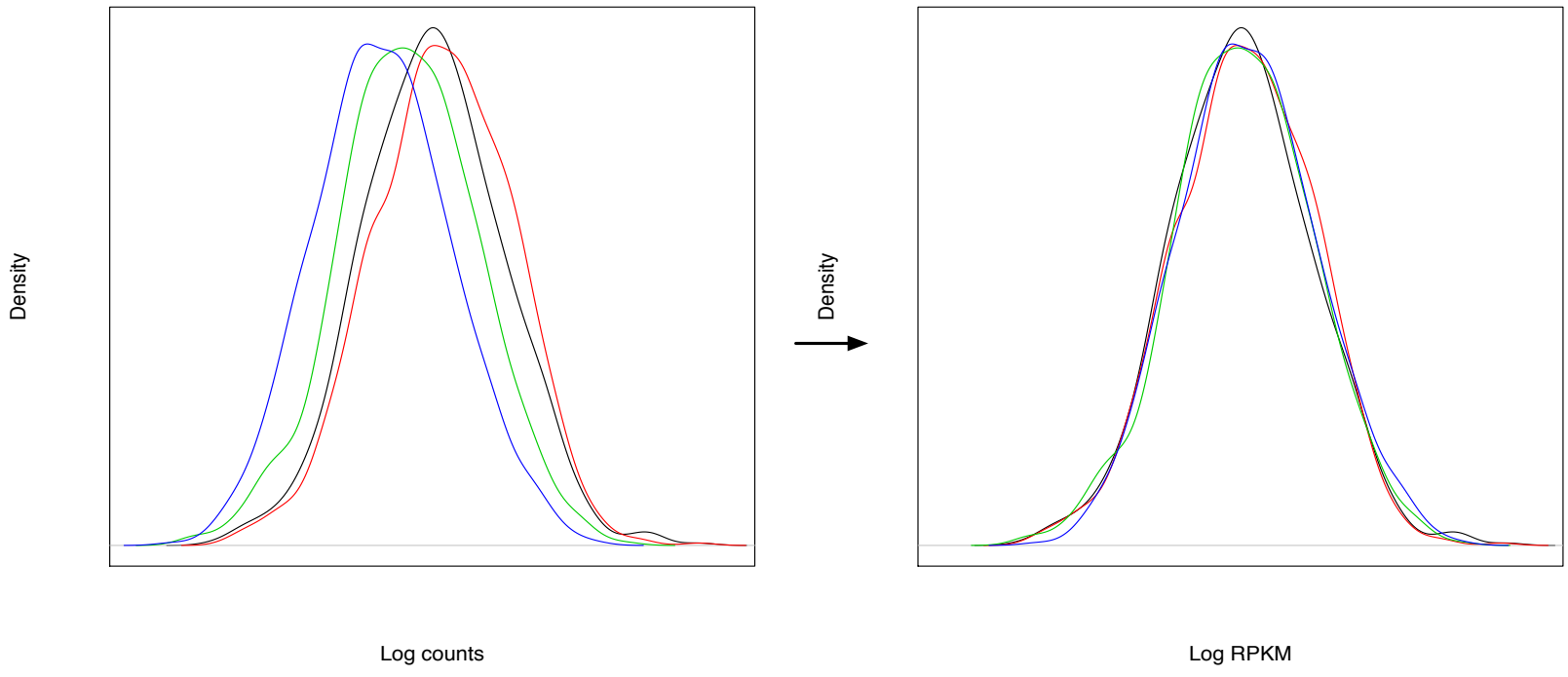
Thus it is natural to include them in the normalisation constant.

If $k_{ig} = 10^{-9} N_i l_g$, the units of $\hat{\mu}_{ig}$ are Reads Per Kilobase per Million mapped reads (RPKM) (Mortazavi et al. 2008).

This is the most elementary form of normalisation.

RPKM normalisation

- RPKM works well for technical and some biological replicates
- $\mu_{ig} \simeq \mu_{jg}$ for all libraries i and j
- RPKM units obtained by scaling of counts by N_i^{-1}



Sample to sample normalisation

- Between different biological samples, homogeneity assumption does not hold
- Why is this a problem?

Sample to sample normalisation

- Between different biological samples, homogeneity assumption does not hold
- Why is this a problem?

Number of reads is limited

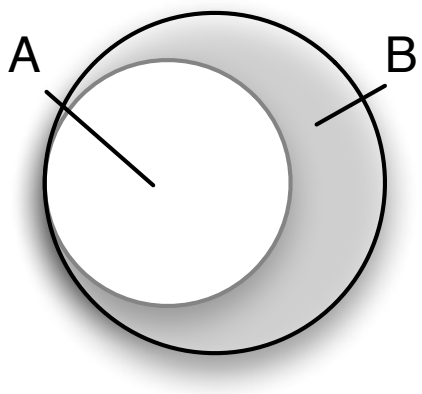
E.g. counts from very highly expressed genes leave less real estate available for counts from lowly expressed genes

Sample to sample normalisation

- Between different biological samples, homogeneity assumption does not hold
- Why is this a problem?

Number of reads is limited

E.g. counts from very highly expressed genes leave less real estate available for counts from lowly expressed genes



- Suppose you have two RNA populations A and B sequenced at same depth
- A and B are identical except half of genes in B are unexpressed in A
- Only $\sim$ half of reads from B come from shared gene set
- Estimates for shared genes differ by factor of ~ 2!

Robinson and Oslack 2010

Trimmed Mean of Ms (TMM) normalisation

- RPKM normalisation implicitly assumes total RNA output $\sum_g \mu_{ig} l_g$ (unknown) is the same for all libraries
- Poisson model is an approximation of Binomial model: $r_{ig} \sim Binomial\left(N_i, \frac{\mu_{ig} l_g}{\sum_j \mu_{ij} l_j}\right)$, $\mathbb{E}(r_{ig}) = N_i \frac{\mu_{ig} l_g}{\sum_j \mu_{ij} l_j}$

Trimmed Mean of Ms (TMM) normalisation

- RPKM normalisation implicitly assumes total RNA output $\sum_g \mu_{ig} l_g$ (unknown) is the same for all libraries
- Poisson model is an approximation of Binomial model: $r_{ig} \sim Binomial\left(N_i, \frac{\mu_{ig} l_g}{\sum_j \mu_{ij} l_j}\right)$, $\mathbb{E}(r_{ig}) = N_i \frac{\mu_{ig} l_g}{\sum_j \mu_{ij} l_j}$
- Better assumption: the output between samples for a *core set only* of genes G is similar: $\sum_{g \in G} \mu_{ig} l_g = \sum_{g \in G} \mu_{jg} l_g$

TMM normalisation

The naive MLE is proportional to the normalised counts:

$$\hat{\mu}_{jg} = \frac{r_{jg}}{k_{jg}} = \frac{1}{10^{-9} l_g} \frac{r_{jg}}{N_j}$$

If $\sum_{g \in G} \hat{\mu}_{ig} l_g \neq \sum_{g \in G} \hat{\mu}_{jg} l_g$, the MLEs need to be adjusted.

TMM normalisation

The naive MLE is proportional to the normalised counts:

$$\hat{\mu}_{jg} = \frac{r_{jg}}{k_{jg}} = \frac{1}{10^{-9} l_g} \frac{r_{jg}}{N_j}$$

If $\sum_{g \in G} \hat{\mu}_{ig} l_g \neq \sum_{g \in G} \hat{\mu}_{jg} l_g$, the MLEs need to be adjusted.

Calculate scaling factor for sample j relative to reference sample i:

$$\sum_{g \in G} \frac{r_{ig}}{N_i} \simeq S^{(i,j)} \sum_{g \in G} \frac{r_{jg}}{N_j}.$$

Adjust the MLEs for sample j for *all* genes:

$$\hat{\mu}_{jg} = \frac{r_{jg}}{k_{jg}} = \frac{r_{jg}}{10^{-9} N_j l_g} \cdot S^{(i,j)}.$$

Robinson and Oslack 2010

TMM normalisation

How to choose the subset G used to calculate $S^{(i,j)}$?

TMM normalisation

How to choose the subset G used to calculate $S^{(i,j)}$?

- For pair of libraries (i, j) determine log fold change of normalised counts

$$M_g^{(i,j)} = \log \frac{r_{ig}}{N_i} - \log \frac{r_{jg}}{N_j}.$$

- and the mean of the log normalised counts

$$A_g^{(i,j)} = \frac{1}{2}\left[\log \frac{r_{ig}}{N_i} + \log \frac{r_{jg}}{N_j}\right].$$

TMM normalisation

How to choose the subset G used to calculate $S^{(i,j)}$?

- For pair of libraries (i, j) determine log fold change of normalised counts

$$M_g^{(i,j)} = \log \frac{r_{ig}}{N_i} - \log \frac{r_{jg}}{N_j}.$$

- and the mean of the log normalised counts

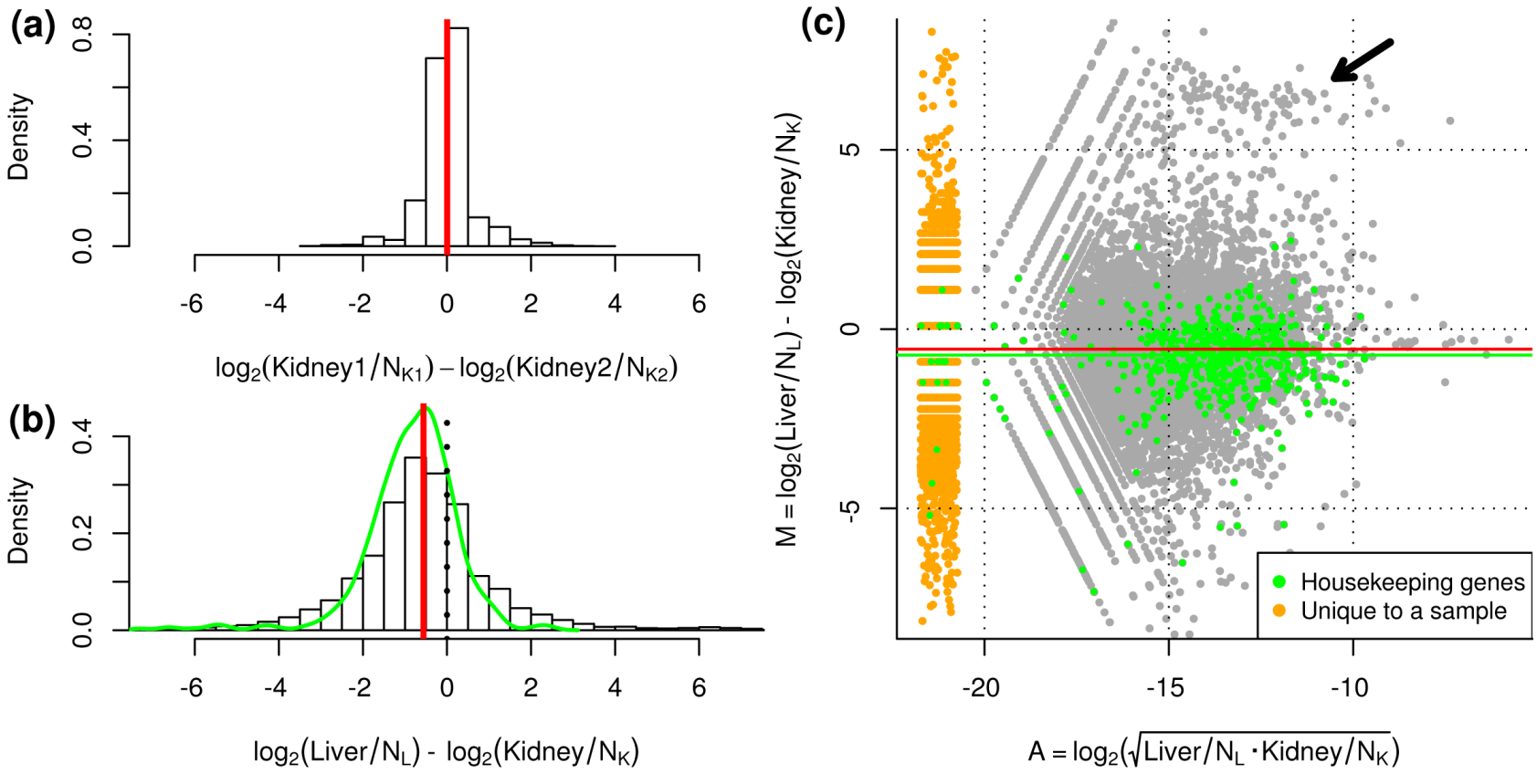
$$A_g^{(i,j)} = \frac{1}{2}\left[\log \frac{r_{ig}}{N_i} + \log \frac{r_{jg}}{N_j}\right].$$

- Set G to genes remaining after trimming upper and lower $x\%$ of the $\{A_g\}$ and $\{M_g\}$. I.e. genes in G have unexceptional values of $A_g^{(i,j)}$ and $M_g^{(i,j)}$

Robinson and Oslack 2010

TMM normalisation (with edgeR)

- Compute summary of $\{M_g^{(i,j)}\}$ for genes in G (weighted mean)
- Let $S^{(i,j)}$ be the exponential of this summary
- Adjust $\hat{\mu}_{jg}$ by a factor of $S^{(i,j)}$ for all genes g



Robinson and Oslack 2010

Median log deviation normalisation (with DESeq)

An alternative normalisation provided in DESeq package

- For each gene g in sample i, calculate deviation of $\log r_{ig}$ from the mean $\log r_{ig}$ over all libraries: $d_{ig} = \log r_{ig} - \frac{1}{I}\sum_i \log r_{ig}$.
- Calculate median over all genes: $\log S^{(i)} = \text{median}_i(d_{ig})$
- Adjust $\hat{\mu}_{ig}$ by a factor of $S^{(i)}$ for all genes g

Median log deviation normalisation (with DESeq)

An alternative normalisation provided in DESeq package

- For each gene g in sample i, calculate deviation of $\log r_{ig}$ from the mean $\log r_{ig}$ over all libraries: $d_{ig} = \log r_{ig} - \frac{1}{I}\sum_i \log r_{ig}$.
- Calculate median over all genes: $\log S^{(i)} = \text{median}_i(d_{ig})$
- Adjust $\hat{\mu}_{ig}$ by a factor of $S^{(i)}$ for all genes g

edgeR and DESeq are both robust across genes (weighted mean of core set vs. median of all genes)

Anders and Huber 2010

Normalisation between genes

- So far we have looked at library-level scaling to make the expression of a given gene comparable across libraries
- In other words, we have been seeking to account for factors affecting all genes in a library similarly

Normalisation between genes

- So far we have looked at library-level scaling to make the expression of a given gene comparable across libraries
- In other words, we have been seeking to account for factors affecting all genes in a library similarly
- Are there factors affecting different genes differently?
- Recall normalisation equation:

$$\boxed{\hat{\mu}_{ig} = \frac{r_{ig}}{k_{ig}}}$$

Normalisation between genes

- So far we have looked at library-level scaling to make the expression of a given gene comparable across libraries
- In other words, we have been seeking to account for factors affecting all genes in a library similarly
- Are there factors affecting different genes differently?
- Recall normalisation equation:

$$\boxed{\hat{\mu}_{ig} = \frac{r_{ig}}{k_{ig}}}$$

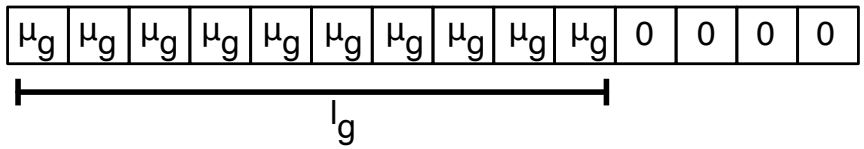
Consider the decomposition of $k_{ig} = kk_i k_g$

- k: global scaling to get more convenient units. E.g. 10^{-9}.
- k_i: library-specific normalisation factors. E.g. $\tilde{N}_i = N_i / S^{(i)}$
- k_g: gene-specific normalisation factors. E.g. l_g

Normalisation between genes

Where does the l_g factor come from anyway?

Underlying assumption: constant Poisson rate across bases.

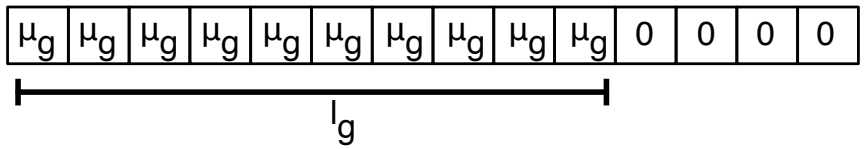


$$r_{igp} \sim Pois(kk_i\mu_g)$$

Normalisation between genes

Where does the l_g factor come from anyway?

Underlying assumption: constant Poisson rate across bases.



$$r_{igp} \sim Pois(kk_i \mu_g)$$

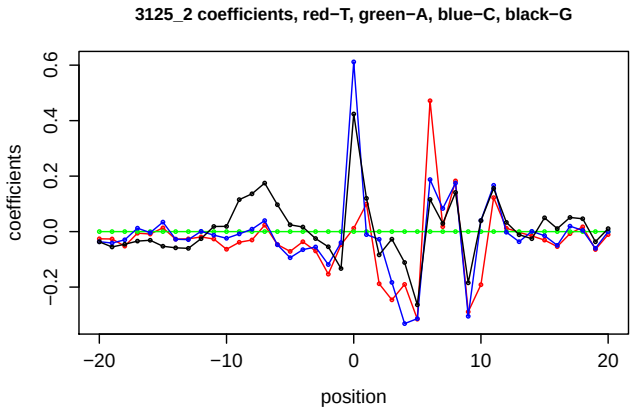
$$r_{ig} = \sum_{p=1}^{l_g} r_{igp}$$

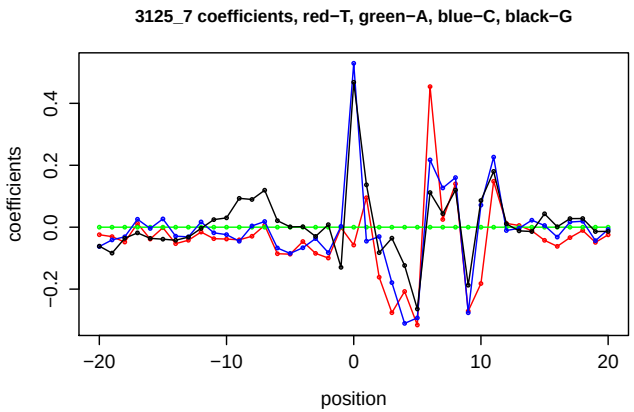
$$r_{ig} \sim Pois(kk_i \sum_{p=1}^{l_g} \mu_g)$$

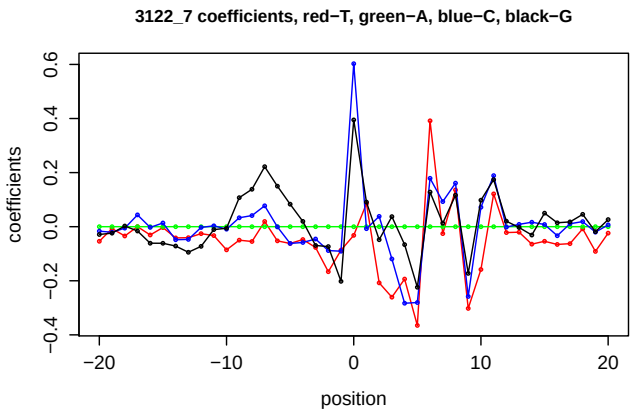
$$\sim Pois(kk_i l_g \mu_g)$$

$$\sim Pois(10^{-9} \tilde{N}_i l_g \mu_{ig})$$

Normalisation between genes

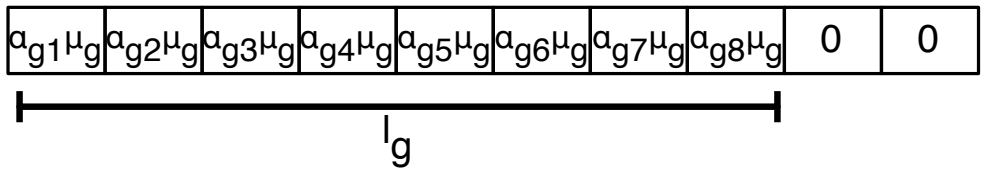






There are in fact local sequence-specific biases (Li et al. 2010, Hansen et al. 2010) (non-random amplification?).

This suggests a variable-rate model with weights α_{gp}:

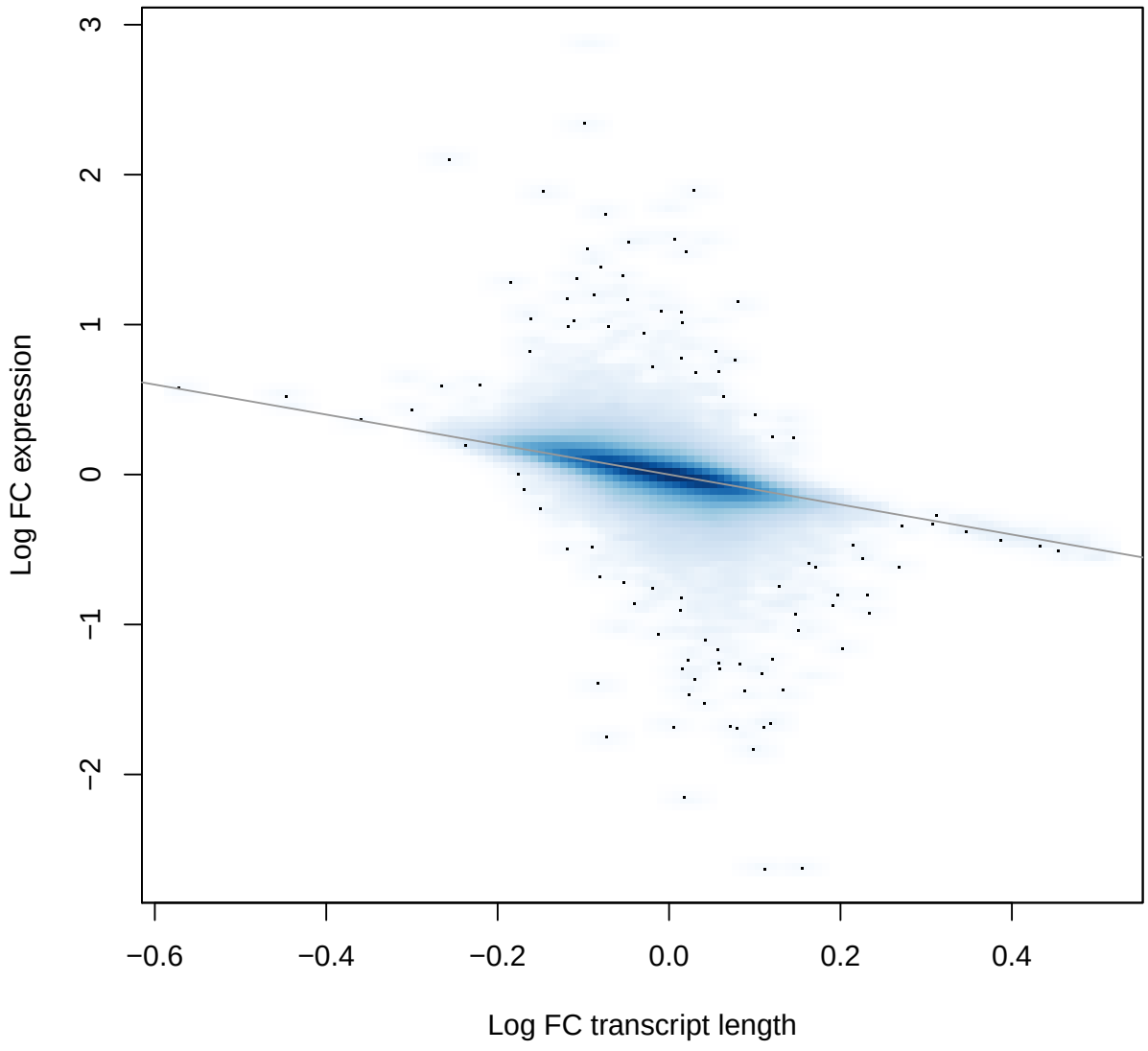


$$r_{ig} \sim Pois(kk_i \sum_{p=1}^{l_g} \alpha_{gp}\mu_{ig})$$

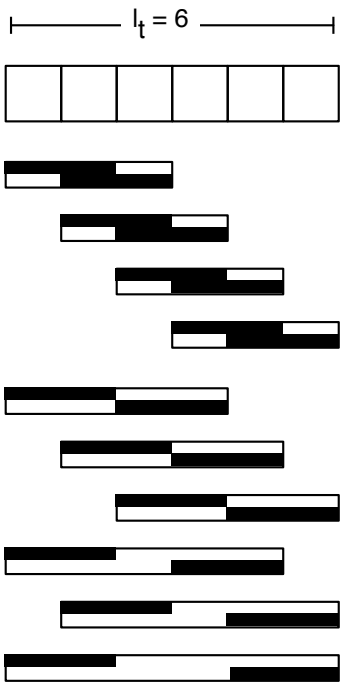
$$\sim Pois(kk_i \tilde{l}_g \mu_{ig})$$

$$\sim Pois(10^{-9}\tilde{N}_i \tilde{l}_g \mu_{ig})$$

Accounting for sequencing biases with mseq



Normalisation between genes (adjust for insert size distro)



$l_r = 2$

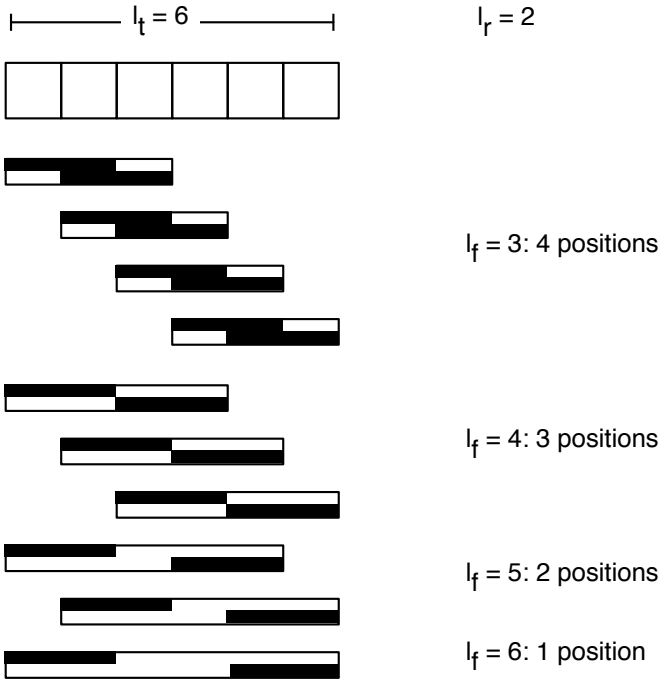
$l_f = 3$: 4 positions

$l_f = 4$: 3 positions

$l_f = 5$: 2 positions

$l_f = 6$: 1 position

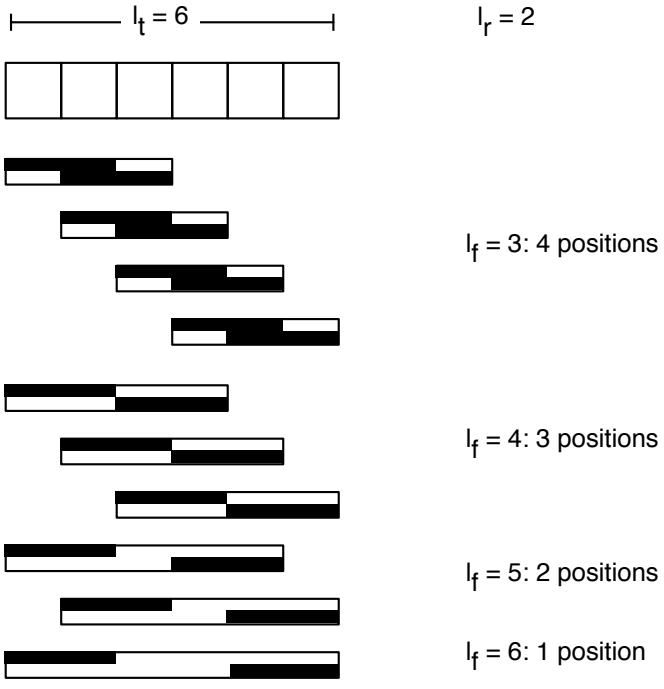
Normalisation between genes (adjust for insert size distro)



$$\tilde{l}_t = \sum_{l_f=l_r+1}^{l_t} p(l_f|l_t)(l_t - l_f + 1)$$

(assuming each position equally likely)

Normalisation between genes (adjust for insert size distro)



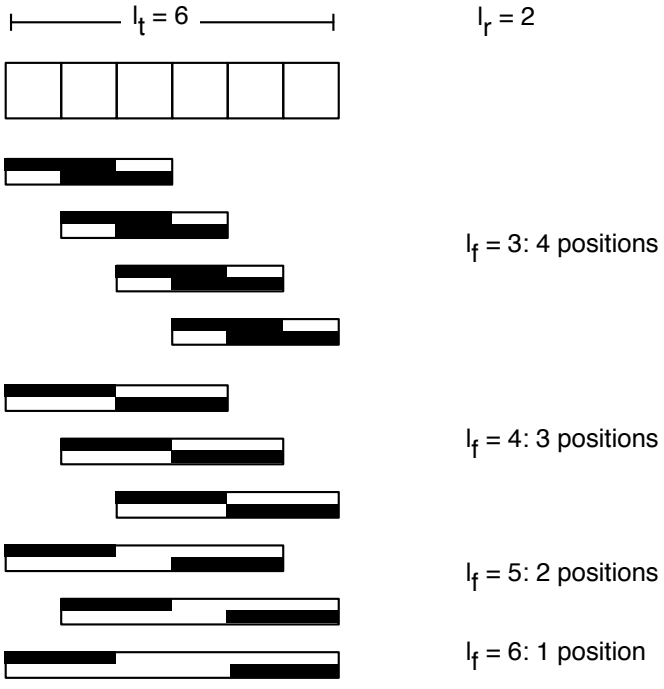
$$\tilde{l}_t = \sum_{l_f=l_r+1}^{l_t} p(l_f|l_t)(l_t - l_f + 1)$$

(assuming each position equally likely)

$$\tilde{l}_t = \sum_{l_f=l_r+1}^{l_t} p(l_f|l_t) \sum_{p=1}^{l_t-l_f+1} \alpha(p,t,l_f)$$

(weight $\alpha(p,t,l_f)$ to fragments of length l_f at position p of transcript t)

Normalisation between genes (adjust for insert size distro)



$$\tilde{l}_t = \sum_{l_f=l_r+1}^{l_t} p(l_f|l_t)(l_t - l_f + 1)$$

(assuming each position equally likely)

$$\tilde{l}_t = \sum_{l_f=l_r+1}^{l_t} p(l_f|l_t) \sum_{p=1}^{l_t-l_f+1} \alpha(p,t,l_f)$$

(weight $\alpha(p,t,l_f)$ to fragments of length l_f at position p of transcript t)

If pre-selection fragments roughly uniform up to l_t within main support of insert size distribution, then $p(l_f|l_t) \simeq p(l_f)$

Glaus et al 2012, to appear

Differential expression

We have obtained library and gene specific normalisation factors to make counts/concentration estimates as comparable as possible.

This allows us to:

- obtain reasonably unbiased log fold changes between two groups of samples
- obtain p-values under the null hypothesis of no differential expression

Differential expression

We have obtained library and gene specific normalisation factors to make counts/concentration estimates as comparable as possible.

This allows us to:

- obtain reasonably unbiased log fold changes between two groups of samples
- obtain p-values under the null hypothesis of no differential expression

Recall hypothesis testing (e.g. limma for microarrays):

- define a function of the data, T (the *test statistic*)
- derive distribution of T under the null (e.g. no DE)
- define critical regions of T
- compute observed value t from actual data
- reject null if t is in a critical region

Concluding remarks

- Variation in total RNA output per sample leads to biases in expression estimates (limited real estate)
- Variation in sequence composition of genes leads to biases (non-random hexamer priming)
- Fragment size selection leads to positional biases
- Normalisation seeks to correct for these biases

Concluding remarks

- Variation in total RNA output per sample leads to biases in expression estimates (limited real estate)
- Variation in sequence composition of genes leads to biases (non-random hexamer priming)
- Fragment size selection leads to positional biases
- Normalisation seeks to correct for these biases
- Only then can we reliably begin to draw inferences about differential expression