Transcript expression estimation, normalisation, differential expression

Ernest Turro

University of Cambridge

14 Sep 2016

Gene expression

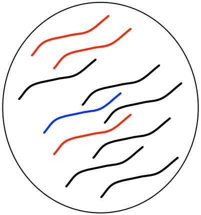
An important aim in genomics is the characterisation of RNA samples. Specifically:

1. What is the sequence of each distinct RNA in a sample?
2. What is the concentration of each RNA in a sample?

Gene expression

An important aim in genomics is the characterisation of RNA samples. Specifically:

1. What is the sequence of each distinct RNA in a sample?
2. What is the concentration of each RNA in a sample?



6/10

3/10

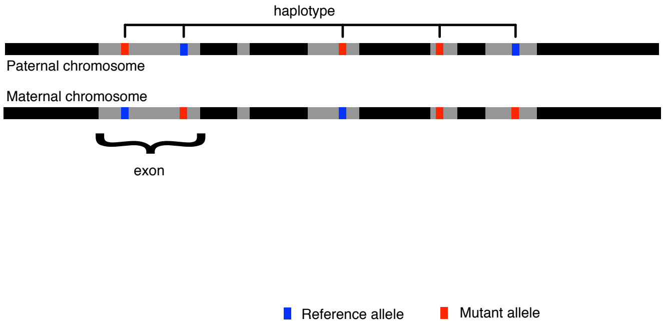
1/10

NB: in general only relative proportions available

Gene expression

Different kinds of RNAs (tRNAs, rRNAs, mRNAs, other ncRNAs...).

Messenger RNAs of particular interest as they code for proteins.

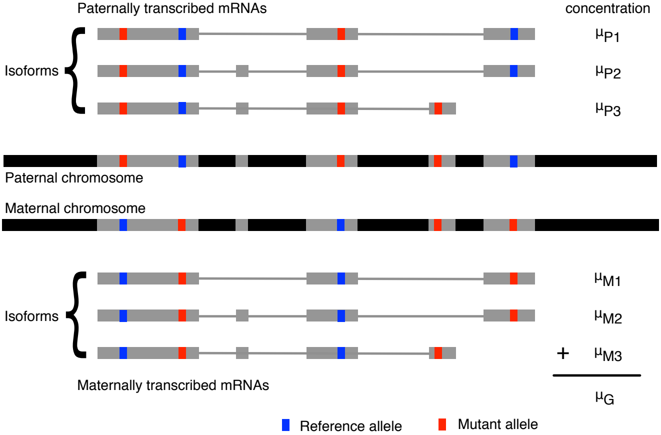


Gene expression

Different kinds of RNAs (tRNAs, rRNAs, mRNAs, other ncRNAs...).

Messenger RNAs of particular interest as they code for proteins.

1. Alternative isoforms have distinct sequences
2. Two versions of each isoform sequence in diploid organisms



RNA-seq read counts

To infer concentrations, we need to identify

- the set of transcript sequences in the sample
- the set of reads (potentially) emanating from each transcript

RNA-seq read counts

To infer concentrations, we need to identify

- the set of transcript sequences in the sample
- the set of reads (potentially) emanating from each transcript

Approaches:

- Select transcript sequences from a database (e.g. Ensembl) and align reads to them
- Align reads to genome and infer transcript sequences
- Assemble reads into contigs

RNA-seq read counts

To infer concentrations, we need to identify

- the set of transcript sequences in the sample
- the set of reads (potentially) emanating from each transcript

Approaches:

- Select transcript sequences from a database (e.g. Ensembl) and align reads to them
- Align reads to genome and infer transcript sequences
- Assemble reads into contigs

In any case, we get a **mapping of reads to features of interest** (e.g. genes, isoforms, haplotype-specific isoforms).

RNA-seq read counts

To infer concentrations, we need to identify

- the set of transcript sequences in the sample
- the set of reads (potentially) emanating from each transcript

Approaches:

- Select transcript sequences from a database (e.g. Ensembl) and align reads to them
- Align reads to genome and infer transcript sequences
- Assemble reads into contigs

In any case, we get a **mapping of reads to features of interest** (e.g. genes, isoforms, haplotype-specific isoforms).

How do we model the alignments?

The Poisson distribution

If independent events occur at a known given rate, then the number of such events follows a **Poisson distribution**.

The Poisson distribution

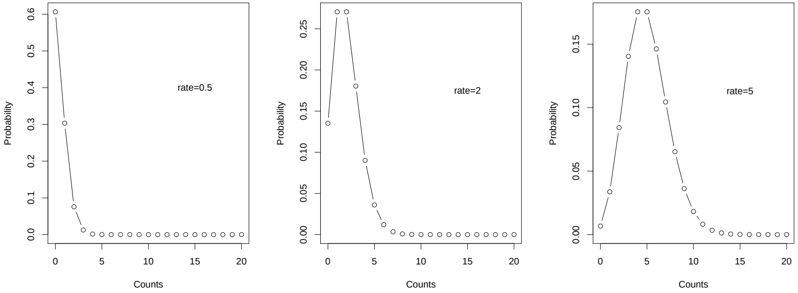
If independent events occur at a known given rate, then the number of such events follows a **Poisson distribution**.

Examples:

- Number of cars crossing a milestone every hour
- Number of raindrops falling on a rooftop every minute

Single rate parameter λ (pets rate = cats rate + dogs rate).

Mean = variance = rate.



Basic Poisson model for expression quantification

Number of reads aligning to a transcript increases with

- Total number of reads
- Length of transcript
- Abundance of transcript

Basic Poisson model for expression quantification

Number of reads aligning to a transcript increases with

- Total number of reads
- Length of transcript
- Abundance of transcript

Number of reads from gene *g* captured by Poisson model (Marioni et al. 2008):

$$r_g \sim Poisson(b\mu_g l_g),$$

- μ_g: concentration of RNAs from gene *g*
- l_g: effective length of the gene
- *b*: normalisation constant (e.g. total no. of reads)

Basic Poisson model for expression quantification

Basic model is useful but:

- “gene length” ambiguous — fragments from several isoforms with different lengths are sequenced
- reads counts not always observed due to sequence sharing (e.g. paralogous families)

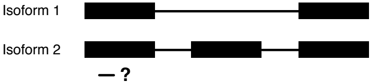
Basic Poisson model for expression quantification

Basic model is useful but:

- “gene length” ambiguous — fragments from several isoforms with different lengths are sequenced
- reads counts not always observed due to sequence sharing (e.g. paralogous families)

Can we estimate expression for each isoform?

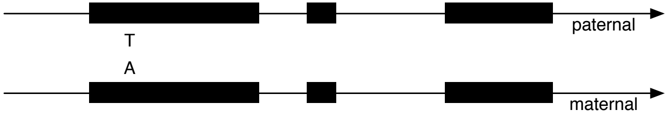
- Isoform read counts in general not observed:



- We need a **read count model for isoforms**

Basic Poisson model for expression quantification

Recall that sequencing allows us to distinguish alleles at heterozygous positions.



Can we use RNA-seq to detect allelic imbalance?

We need a **read count model for alleles**

The binomial distribution

A Bernoulli trial is an experiment in which “success” occurs with probability p and “failure” occurs with probability $1 - p$.

The number of successes given n Bernoulli trials follows a **binomial distribution** with parameters n and p. $\mathbb{E}(X) = np$.

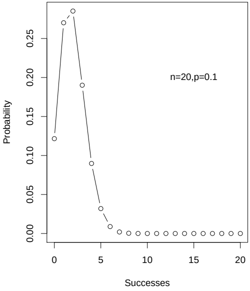
The binomial distribution

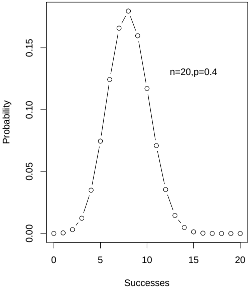
A Bernoulli trial is an experiment in which “success” occurs with probability *p* and “failure” occurs with probability 1 – *p*.

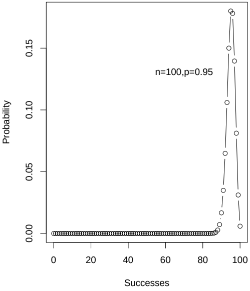
The number of successes given *n* Bernoulli trials follows a **binomial distribution** with parameters *n* and *p*. $\mathbb{E}(X) = np$.

Examples:

- Number of heads after *n* coin tosses. $p \neq 0.5$ if not fair
- Number of times you win the lottery (tiny *p* (but £££))







The multinomial distribution

A Bernoulli trial is an experiment in which “success” occurs with probability p and “failure” occurs with probability $1 - p$.

The number of successes given n Bernoulli trials follows a **binomial distribution** with parameters n and p.

Examples:

- Number of heads after n coin tosses. $p \neq 0.5$ if it is unfair
- Number of times you hit the bullseye out of n shots

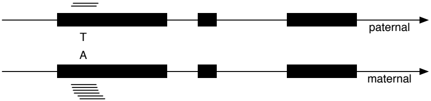
If there are > 2 categories, the per-category counts follow a **multinomial distribution** with parameters n and $(p_1, p_2, \ldots)$.

Example:

- Number of 1s, 2s, 3s, 4s, 5s, 6s if you roll a die n times. If $\{p_i\} \neq \frac{1}{6}$ then the die is not fair.

Basic Binomial model for allelic imbalance

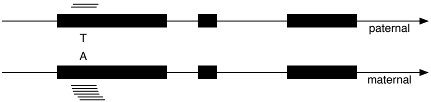
- Reads permit discrimination between two copies of an isoform



- Binomial test: $\sum_{r=0}^{r_0} P(r|p = 0.5, n = r_0 + r_1) < \alpha$? (Degner et al. 2009)

Basic Binomial model for allelic imbalance

- Reads permit discrimination between two copies of an isoform



- Binomial test: $\sum_{r=0}^{r_0} P(r|p = 0.5, n = r_0 + r_1) < \alpha$? (Degner et al. 2009). E.g. suppose $r_0 = 2; r_1 = 6$:

$$P(r = 0|p = 0.5, n = 8) = 0.00390625$$

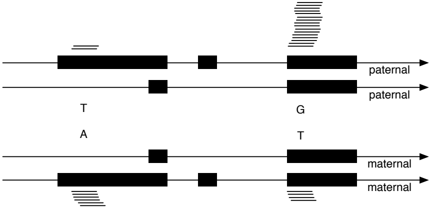
$$P(r = 1|p = 0.5, n = 8) = 0.03125$$

$$P(r = 2|p = 0.5, n = 8) = 0.109375$$

$$\sum_{r=0}^{2} P(r|p = 0.5, n = 8) = 0.1445312 \text{ (not significant)}$$

Basic Binomial model for allelic imbalance

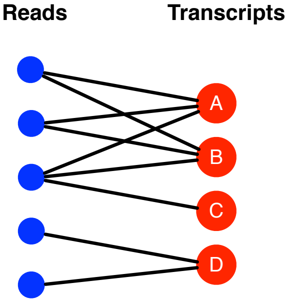
- What if there are multiple SNPs and isoforms?



- Binomial test not appropriate
- We need a **read count model for haplotype-specific isoforms**

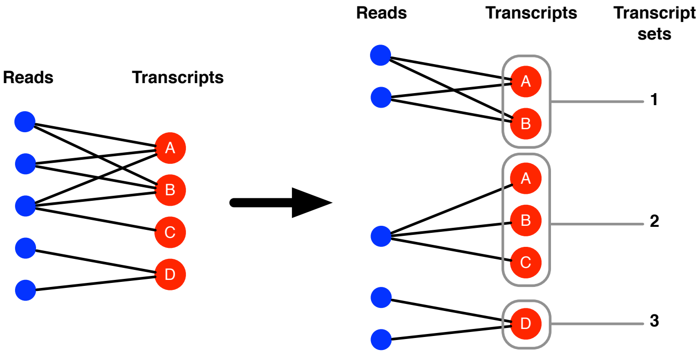
Multi-mapping reads

- Align reads back to reference transcript sequences with Bowtie (Langmead et al. 2009), allowing multiple alignments per read
- Multi-mapping structure between reads and transcripts



Multi-mapping reads

- Obtain transcript sets, such that each read maps to only 1 set
- Transcripts may belong to more than one set
- Read counts per set can be observed
- Transcripts can be isoforms sharing exons or from multiple genes



Poisson model for transcript set reads counts

Model reads per transcript set instead of per gene (Turro et al. 2011).

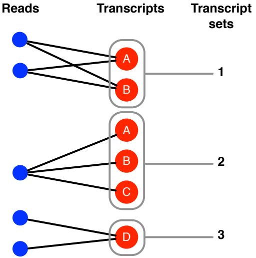
$$\text{Define } M_{it} = \begin{cases} 1 & \text{if transcript } t \text{ in set } i, \\ 0 & \text{otherwise.} \end{cases}$$

Now model for reads counts is:

$$k_i \sim Poisson(bs_i \sum_t M_{it}\mu_t),$$

where s_i is the effective length shared by transcripts in set i.

Latent variables for read counts

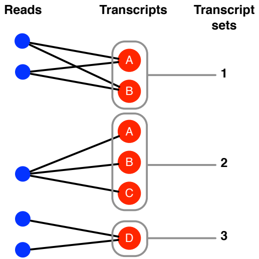


$$M = \begin{array}{c} \overbrace{\begin{array}{cccc} A & B & C & D \end{array}}^{\text{Transcripts}} \\ \left(\begin{array}{cccc} 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{array}\right) \end{array} \begin{array}{c} 1 \\ 2 \\ 3 \end{array} \left.\vphantom{\begin{array}{c}1\\2\\3\end{array}}\right\} \text{Transcript sets}$$

$$\mathbf{k} = \overbrace{\begin{pmatrix} 2 \\ 1 \\ 2 \end{pmatrix}}^{\text{Observed set counts}} \qquad X = \begin{pmatrix} X_{11} & X_{12} & 0 & 0 \\ X_{21} & X_{22} & X_{23} & 0 \\ 0 & 0 & 0 & X_{34} \end{pmatrix} \begin{array}{c} 1 \\ 2 \\ 3 \end{array} \left.\vphantom{\begin{array}{c}1\\2\\3\end{array}}\right\} \text{Transcript sets}$$

$$\text{Unobserved transcript counts} \left\{ \mathbf{r} = \begin{pmatrix} r_1 & r_2 & r_3 & r_4 \end{pmatrix} \right.$$

Latent variables for read counts



$$
\begin{array}{cc}
 & \overbrace{\begin{array}{cccc}\text{A} & \text{B} & \text{C} & \text{D}\end{array}}^{\text{Transcripts}} \\
M = & \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \left.\begin{array}{c}1\\2\\3\end{array}\right\} \text{Transcript sets}
\end{array}
$$

$$
\overbrace{\mathbf{k} = \begin{pmatrix} 2 \\ 1 \\ 2 \end{pmatrix}}^{\text{Observed set counts}} \qquad X = \begin{pmatrix} X_{11} & X_{12} & 0 & 0 \\ X_{21} & X_{22} & X_{23} & 0 \\ 0 & 0 & 0 & X_{34} \end{pmatrix} \left.\begin{array}{c}1\\2\\3\end{array}\right\} \text{Transcript sets}
$$

$$
\text{Unobserved transcript counts} \left\{ \mathbf{r} = \begin{pmatrix} r_1 & r_2 & r_3 & r_4 \end{pmatrix} \right.
$$

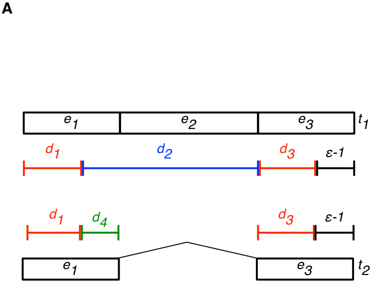
$$
X_{it} \sim Poisson(bs_i M_{it}\mu_t),
$$

$$
k_i \sim Poisson(bs_i \sum_t M_{it}\mu_t),
$$

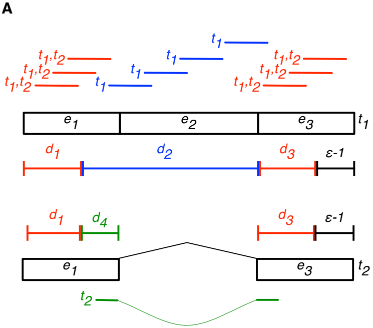
$$
r_t \sim Poisson(b\mu_t \sum_i M_{it} s_i) = Poisson(bl_t \mu_t),
$$

$$
\{X_{i1}, \ldots, X_{in}\} | \{\mu_1, \ldots, \mu_n\}, k_i \sim Mult(k_i, \frac{M_{i1}\mu_1}{\sum_t M_{it}\mu_t}, \ldots, \frac{M_{in}\mu_n}{\sum_t M_{it}\mu_t}).
$$

Concrete example

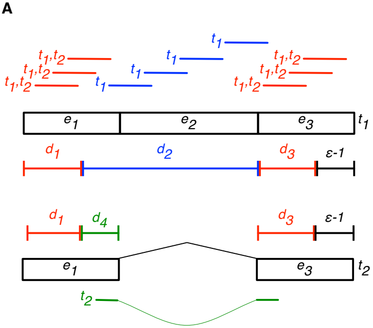


Concrete example



$$M = \begin{array}{c} \begin{array}{cc} t_1 & t_2 \end{array} \\ \begin{pmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \end{array} \quad \mathbf{k} = \begin{pmatrix} 6 \\ 4 \\ 1 \end{pmatrix}$$

Concrete example

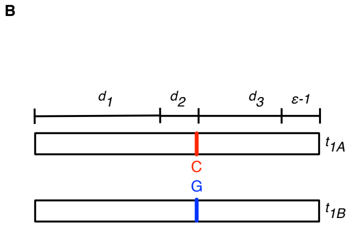


$$M = \begin{array}{c} \begin{array}{cc} t_1 & t_2 \end{array} \\ \begin{pmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \end{array} \quad \mathbf{k} = \begin{pmatrix} 6 \\ 4 \\ 1 \end{pmatrix}$$

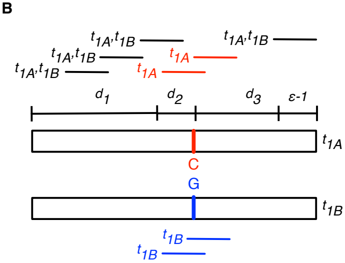
$$\mathbf{s} = \begin{pmatrix} d_1 + d_3 \\ d_2 \\ d_4 \end{pmatrix} = \begin{pmatrix} e_1 + e_3 - 2(\epsilon - 1) \\ e_2 + \epsilon - 1 \\ \epsilon - 1 \end{pmatrix}$$

$$l_1 = s_1 + s_2 = e_1 + e_2 + e_3 - (\epsilon - 1)$$
$$l_2 = s_1 + s_3 = e_1 + e_3 - (\epsilon - 1)$$

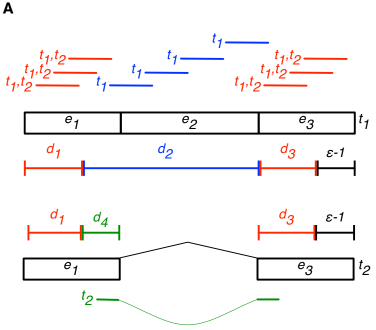
Heterozygotes and haplo-isoforms



Heterozygotes and haplo-isoforms



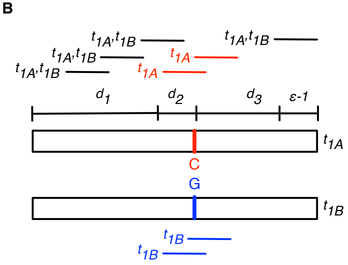
Same model structure for isoforms and haplo-isoforms



$$M = \begin{array}{c} \begin{array}{cc} t_1 & t_2 \end{array} \\ \begin{pmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \end{array} \quad \mathbf{k} = \begin{pmatrix} 6 \\ 4 \\ 1 \end{pmatrix}$$

$$\mathbf{s} = \begin{pmatrix} d_1 + d_3 \\ d_2 \\ d_4 \end{pmatrix} = \begin{pmatrix} e_1 + e_3 - 2(\epsilon - 1) \\ e_2 + \epsilon - 1 \\ \epsilon - 1 \end{pmatrix}$$

$$l_1 = s_1 + s_2 = e_1 + e_2 + e_3 - (\epsilon - 1)$$
$$l_2 = s_1 + s_3 = e_1 + e_3 - (\epsilon - 1)$$



$$M = \begin{array}{c} \begin{array}{cc} t_{1A} & t_{1B} \end{array} \\ \begin{pmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \end{array} \quad \mathbf{k} = \begin{pmatrix} 4 \\ 2 \\ 2 \end{pmatrix}$$

Heterozygotes can be treated like alternative exons!

Remarks on expression estimation

- Poisson distribution captures the unavoidable variance due to counting independent events
- The mapping of a read or read pair to a feature can be ambiguous
- Deconvolution methods help quantify expression of different isoforms and even haplotype-specific isoforms
- This really sets RNA-seq and microarrays apart!

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)

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):

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

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 procedure for setting 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 procedure for setting 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 procedure for setting 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

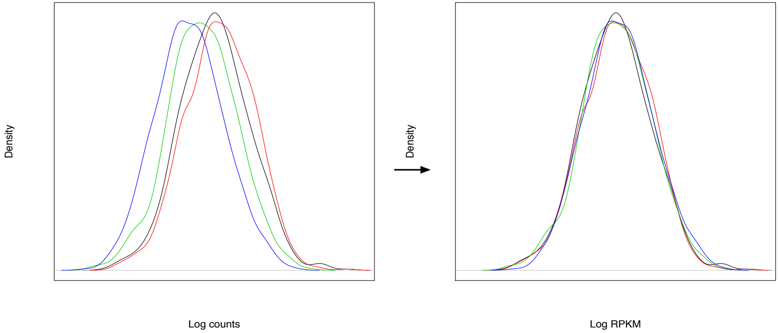
Thus it is natural to include them in the normalisation constant.

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

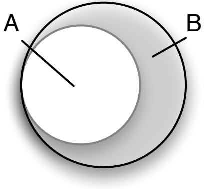
E.g. counts from very highly expressed genes leave less real estate available for counts from lowly expressed genes

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 ~ half of reads from B come from shared gene set
- Estimates for shared genes differ by factor of ~ 2!

Robinson and Oslack 2010

Poisson approximation to Binomial

- Total RNA output, $\sum_g \mu_{ig} l_g$, inversely affects read counts r_{ig} (for fixed μ_{ig})
- RPKM normalisation assumes implicitly that total RNA output (unknown) is the same for all libraries:

Poisson approximation to Binomial

- Total RNA output, $\sum_g \mu_{ig} l_g$, inversely affects read counts r_{ig} (for fixed μ_{ig})
- RPKM normalisation assumes implicitly that total RNA output (unknown) is the same for all libraries:

$$
\begin{aligned}
r_{ig} &\sim \mathit{Binomial}\left(N_i, \frac{\mu_{ig} l_g}{\sum_g \mu_{ig} l_g}\right) \\
&\sim \mathit{Poisson}\left(N_i \frac{\mu_{ig} l_g}{\sum_g \mu_{ig} l_g}\right) \text{ as } N \to \infty \\
\implies \mathbb{E}(r_{ig}) &= N_i \frac{\mu_{ig} l_g}{\sum_g \mu_{ig} l_g}
\end{aligned}
$$

- RPKM assumption: $\forall i, \sum_g \mu_{ig} l_g = 10^9$ (so $\hat{\mu}_{ig} = \frac{r_{ig}}{10^{-9} N_i l_g}$)

Poisson approximation to Binomial

- Total RNA output, $\sum_g \mu_{ig} l_g$, inversely affects read counts r_{ig} (for fixed μ_{ig})
- RPKM normalisation assumes implicitly that total RNA output (unknown) is the same for all libraries:

$$
\begin{aligned}
r_{ig} &\sim \mathit{Binomial}\left(N_i, \frac{\mu_{ig} l_g}{\sum_g \mu_{ig} l_g}\right) \\
&\sim \mathit{Poisson}\left(N_i \frac{\mu_{ig} l_g}{\sum_g \mu_{ig} l_g}\right) \text{ as } N \to \infty \\
\implies \mathbb{E}(r_{ig}) &= N_i \frac{\mu_{ig} l_g}{\sum_g \mu_{ig} l_g}
\end{aligned}
$$

- RPKM assumption: $\forall i, \sum_g \mu_{ig} l_g = 10^9$ (so $\hat{\mu}_{ig} = \frac{r_{ig}}{10^{-9} N_i l_g}$)
- Better assumption: 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 for *all* genes 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 for *all* genes 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

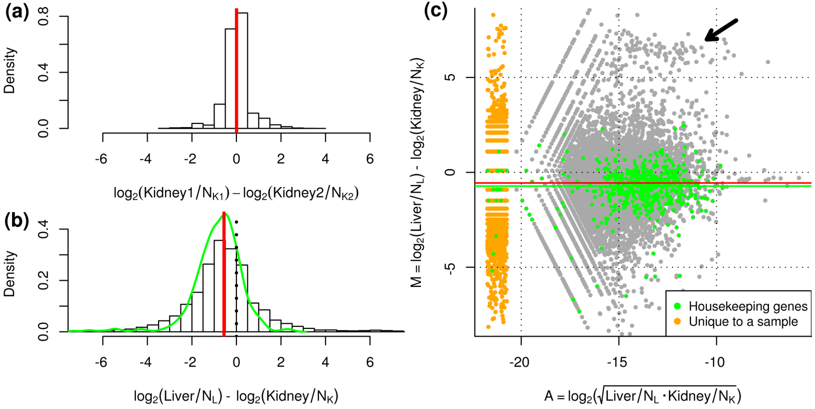
$$A_g^{(i,j)} = \frac{1}{2}\left[\log \frac{r_{ig}}{N_i} + \log \frac{r_{jg}}{N_j}\right].$$

- 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)

Call $\tilde{N}_i = \frac{N_i}{S_i}$ the “adjusted library size”.

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:

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

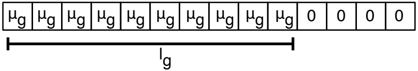
Consider the decomposition of $k_{ig} = k k_i k_g$

- 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.

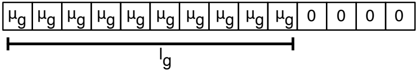

μ_g μ_g μ_g μ_g μ_g μ_g μ_g μ_g μ_g μ_g 0 0 0 0
l_g

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

Normalisation between genes

Where does the l_g factor come from anyway?

Underlying assumption: constant Poisson rate across bases.

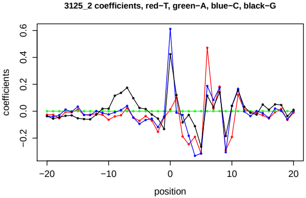


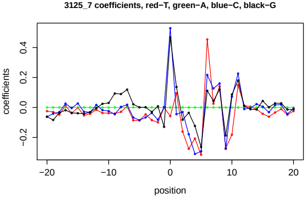
$$r_{igp} \sim Pois(kk_i\mu_g)$$

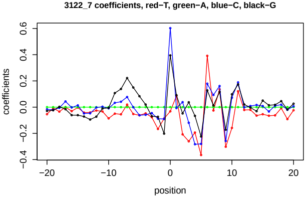
$$r_{ig} = \sum_{p=1}^{l_g} r_{igp}$$

$$\begin{aligned} r_{ig} &\sim Pois(kk_i \sum_{p=1}^{l_g} \mu_g) \\ &\sim Pois(kk_i l_g \mu_g) \\ &\sim Pois(10^{-9}\tilde{N}_i l_g \mu_{ig}) \end{aligned}$$

Normalisation between genes

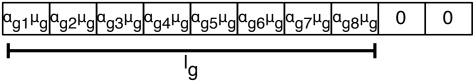

3125_2 coefficients, red−T, green−A, blue−C, black−G


3125_7 coefficients, red−T, green−A, blue−C, black−G


3122_7 coefficients, red−T, green−A, blue−C, black−G

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}:

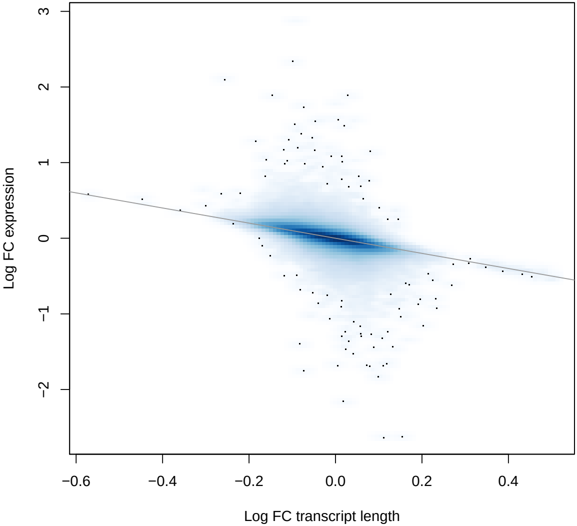

$\alpha_{g1}\mu_g$	$\alpha_{g2}\mu_g$	$\alpha_{g3}\mu_g$	$\alpha_{g4}\mu_g$	$\alpha_{g5}\mu_g$	$\alpha_{g6}\mu_g$	$\alpha_{g7}\mu_g$	$\alpha_{g8}\mu_g$	0	0

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

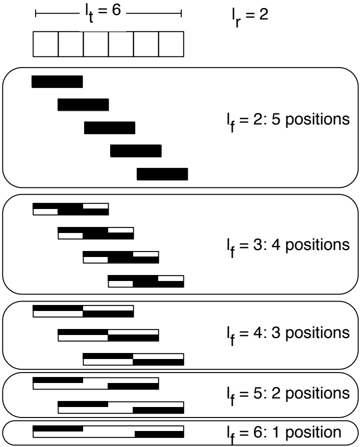
$$\sim Pois(kk_i \tilde{l}_g \mu_{ig})$$

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

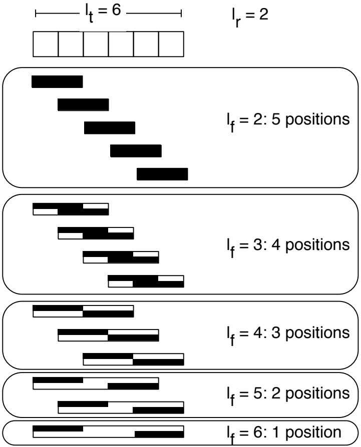
Accounting for sequencing biases with mseq



Normalisation between genes (adjust for insert size distro)



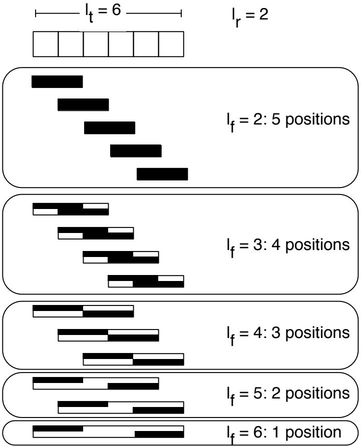
Normalisation between genes (adjust for insert size distro)



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

(assuming each position equally likely)

Normalisation between genes (adjust for insert size distro)



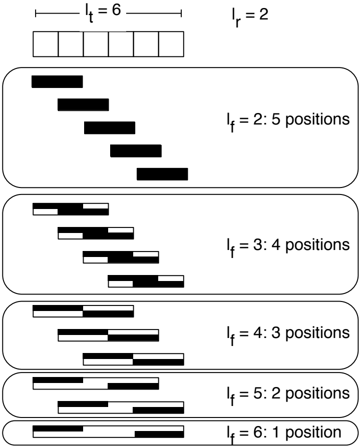
$$\tilde{l}_t = \sum_{l_f = l_r}^{l_t} p(l_f | l_t)(l_t - l_f + 1)$$

(assuming each position equally likely)

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

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

Normalisation between genes (adjust for insert size distro)



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

(assuming each position equally likely)

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

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

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

Glaus et al 2012

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:

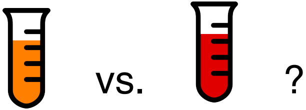
- obtain reasonably unbiased log fold changes between two groups of samples
- obtain *p*-values under the null hypothesis of no differential expression

Recall hypothesis testing:

- 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

Differential expression

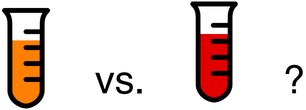
Option 1. $H_0 : \mu_{1g} = \mu_{2g}$



vs. ?

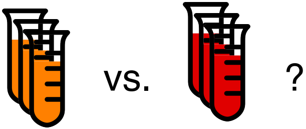
Differential expression

Option 1. $H_0 : \mu_{1g} = \mu_{2g}$



vs. ?

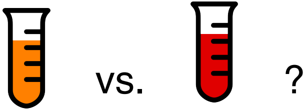
Option 2. $H_0 : \mu_{1g} = \mu_{2g} = \mu_{3g} = \mu_g^{(A)} = \mu_{4g} = \mu_{5g} = \mu_{6g} = \mu_g^{(B)}$



vs. ?

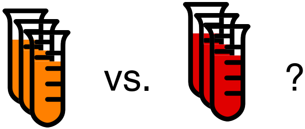
Differential expression

Option 1. $H_0 : \mu_{1g} = \mu_{2g}$



vs. ?

Option 2. $H_0 : \mu_{1g} = \mu_{2g} = \mu_{3g} = \mu_g^{(A)} = \mu_{4g} = \mu_{5g} = \mu_{6g} = \mu_g^{(B)}$

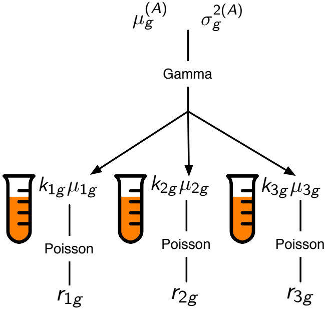


vs. ?

Both options are inadequate!

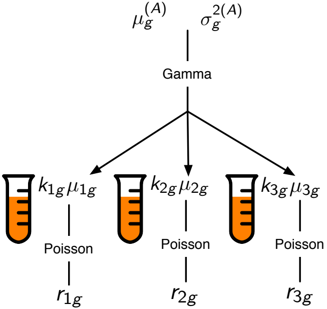
μ_{ig} is the RNA concentration parameter for library i, which *varies* across biological replicates.

Negative binomial distribution



If the rate parameter of the Poisson distribution is not fixed, but varies according to a Gamma distribution, then the counts follow a **negative binomial distribution**.

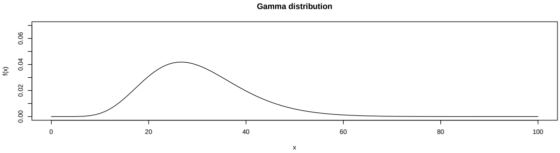
Negative binomial distribution



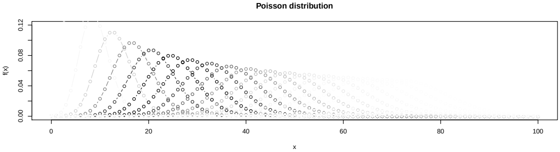
If the rate parameter of the Poisson distribution is not fixed, but varies according to a Gamma distribution, then the counts follow a **negative binomial distribution**.

Unlike the Poisson, the variance is greater than the mean.

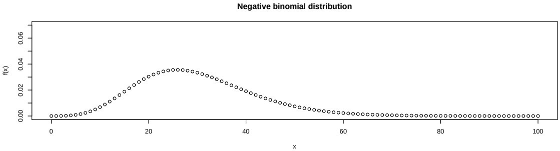
Negative binomial distribution



Biological variance



Poisson noise



Read count variance

$$\text{Mean(NB)} = \text{mean(Gamma)} = \mathbb{E}(\text{mean(Poisson)})$$

$$\text{Var(NB)} = \text{var(Gamma)} + \mathbb{E}(\text{var(Poisson)})$$

Negative binomial distribution

Number of reads from gene *g* in library *i* of condition *c* can be captured by a negative binomial model:

$$r_{cig} = NB(k_{ig}\mu_{cg}, s_{cg})$$

where μ_{cg} and s_{cg} are, respectively, the mean and dispersion for reads from gene *g* in condition *c*.

The variance has two components:

$$\sigma^2_{cg} = k_{ig}\mu_{cg} + k^2_{ig}\mu^2_{cg}s_{cg}$$

Poisson noise Overdispersion

Negative binomial distribution

Number of reads from gene *g* in library *i* of condition *c* can be captured by a negative binomial model:

$$r_{cig} = NB(k_{ig}\mu_{cg}, s_{cg})$$

where μ_{cg} and s_{cg} are, respectively, the mean and dispersion for reads from gene *g* in condition *c*.

The variance has two components:

$$\sigma^2_{cg} = k_{ig}\mu_{cg} + k^2_{ig}\mu^2_{cg}s_{cg}$$

Poisson noise Overdispersion

- Notice there are now **two parameters to estimate**
- How do we obtain precise estimates of the dispersion if we have a small number of libraries per condition?

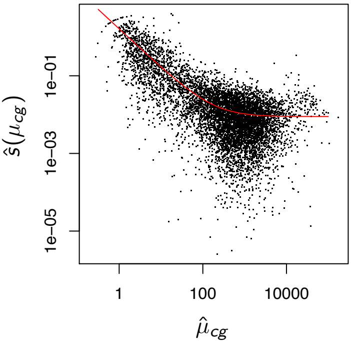
DESeq

How do we estimate the variance robustly?

Assumption: **dispersion is a smooth function of the mean**.

$$\sigma_{cg}^2 = k_{ig}\mu_{cg} + k_{ig}^2\mu_{cg}^2 s(\mu_{cg})$$

Poisson noise Overdispersion



Use fitted values (or values above the line) instead of raw estimates.

This is a form of pooling (sharing of information across genes) to stabilise the estimates.

Anders and Huber 2010

Differential expression with DESeq

Back to hypothesis testing...

$$r_{cig} = NB(k_{ig}\mu_{cg}, s_{cg})$$

$H_0 : \mu_{1g} = \mu_{2g}$.

Differential expression with DESeq

Back to hypothesis testing...

$$r_{cig} = NB(k_{ig}\mu_{cg}, s_{cg})$$

$H_0 : \mu_{1g} = \mu_{2g}$.

Perform a negative binomial exact test.

How extreme is the partitioning of counts between the two conditions under the null?

Differential expression with DESeq

Let the observed condition-specific counts be $q^*_{cg} = \sum_i r_{cig}$.

The probability of the data under the null is $P^* = P(q^*_{1g}, q^*_{2g} | \hat{\mu}_g, \hat{\sigma}^2_g)$.

Differential expression with DESeq

Let the observed condition-specific counts be $q^*_{cg} = \sum_i r_{cig}$.

The probability of the data under the null is
$P^* = P(q^*_{1g}, q^*_{2g} | \hat{\mu}_g, \hat{\sigma}^2_g)$.

Obtain gene-wise exact p-values:

$$p_g = \frac{\sum\limits_{q_{1g}, q_{2g} : P(q_{1g}, q_{2g} | \hat{\mu}_g, \hat{\sigma}^2_g) < P^* \wedge q_{1g} + q_{2g} = q^*_{1g} + q^*_{2g}} P(q_{1g}, q_{2g} | \hat{\mu}_g, \hat{\sigma}^2_g)}{\sum\limits_{q_{1g}, q_{2g} : q_{1g} + q_{2g} = q^*_{1g} + q^*_{2g}} P(q_{1g}, q_{2g} | \hat{\mu}_g, \hat{\sigma}^2_g)},$$

where $\hat{\mu}_g$ and $\hat{\sigma}^2_g$ are estimates for the mean and variance under the null.

Anders and Huber 2010

Differential isoform expression

- At the gene level, counts are often observed (however beware of isoform switching)
- At other levels (isoforms, haplo-isoforms) counts almost always have to be estimated (e.g. with MMSEQ) because reads map to multiple overlapping transcripts
- Count-based methods such as DESeq can be used to obtain differential isoform expression results by using estimated counts instead of observed counts
- A more powerful approach is to take into account posterior uncertainty in expression estimates (MMDIFF; Turro et al 2014)

Closing remarks

- Variation in total RNA output per sample leads to biases in expression estimates (limited real estate)

Closing 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)
- Normalisation seeks to correct for these biases

Closing 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)
- Normalisation seeks to correct for these biases
- Biological and Poisson variability can be modelled with a negative binomial distribution

Closing 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)
- Normalisation seeks to correct for these biases
- Biological and Poisson variability can be modelled with a negative binomial distribution
- Variance of negative binomial hard to estimate gene-by-gene (best to share information acrosss genes)

Closing 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)
- Normalisation seeks to correct for these biases
- Biological and Poisson variability can be modelled with a negative binomial distribution
- Variance of negative binomial hard to estimate gene-by-gene (best to share information acrosss genes)
- Negative binomial exact test produces *p*-values under the null of no differential expression

Further reading (transcriptome-based analysis)

Turro E, Su S-Y, Gonçalves Â , Coin LJM, Richardson S, Lewin A. **Haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads**. *Genome Biology*, 2011 Feb; 12:R13.

Turro E, Astle WJ, Tavaré S. **Flexible analysis of RNA-seq data using mixed effects models**. *Bioinformatics*, 2014 Jan; 30(2):180-188.

MMSEQ, MMDIFF

- Haplotype-specific, isoform, gene-level expression estimation
- Flexible model comparison (polytomous model selection)

https://github.com/eturro/mmseq