Stories from the Supplement

Lior Pachter
Department of Mathematics and Molecular & Cell Biology
UC Berkeley

November 1, 2013
Genome Informatics, CSHL
The Cufflinks supplements


A supplementary arithmetic progression?

• C. Trapnell, B.A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M.J. van Baren, S.L. Salzberg, B.J. Wold and L. Pachter,
  **Supplementary Material: 42 pages.**

• C. Trapnell, D.G. Hendrickson, M. Sauvageau, L. Goff, J.L. Rinn and L. Pachter,
  **Supplementary Material: 70 pages.**

  **Supplementary Material: 98 pages?**
The nature methods manuscript checklist

ONLINE METHODS

☐ Maximum 16,000 characters including spaces. Provide additional information in a separate Supplementary Note.

☐ If the Online Methods section contains more than 10 equations, move the equation-heavy portions to a separate Supplementary Note.

☐ Do not refer to previous publications regarding specific methodologies unless the technical description is comprehensive in the reference cited. If necessary, provide the protocol used.

☐ For all products specified by a brand name, provide manufacturer name in parentheses (but NO locations).

☐ Divide Methods into short subheads (for example, “Subjects.” or “In situ hybridization.”).

☐ Only one level of subheadings can be included; lists cannot be included.
The nature methods manuscript checklist

**Online Methods**
- Maximum 16,000 characters including spaces. Provide additional information in a separate Supplementary Note.
- If the Online Methods section contains more than 10 equations, move the equation-heavy portions to a separate Supplementary Note.
- Do not refer to previous publications regarding specific methodologies unless the technical description is comprehensive in the reference cited. If necessary, provide the protocol used.
- For all products specified by a brand name, provide manufacturer name in parentheses (but NO locations).
- Divide Methods into short subheads (for example, “Subjects.” or “In situ hybridization.”).
- Only one level of subheadings can be included; lists cannot be included.

**Emperor Joseph II:** My dear young man, don't take it too hard. Your work is ingenious. It's quality work. And there are simply too many notes, that's all. Just cut a few and it will be perfect.

**Mozart:** Which few did you have in mind, Majesty?
ONLINE METHODS

Probabilistic model. Our approach to fragment assignment is based on a probabilistic graphical model for sequencing experiments (Supplementary Fig. 11). In this framework, experiments produce multiple random fragments (according to the random variable $P$) that consist of pairs of sequences (reads) from a set of target sequences. $P$ depends on random variables describing fragment length ($L$), target sequences of origin ($T$) and starting positions within target sequences ($P$). The probability distributions for the random variables are based on parameters for target abundances, fragment length probabilities, sequence biases affecting the start and end locations of the fragment and probabilities for sequencing errors in reads. The generative model stipulates that the joint probability of obtaining a fragment $f = f_l$ sequenced from position $p$ in target $t$ is given by

$$P(f = f_l, p, T = t, L = L) = \lambda_f p(p) p(T = t) p(L = L | p)$$

where the parameters of the model are the conditional probabilities $p(p)$. $p(T = t)$ and $p(L = L | p)$ are based on truncated geometrically distributed priors ($0.73$) on the insertion/deletion (indel) length.

The model described here is similar to the Cufflinks model$^{1,5}$, although it incorporates a target order selectivity for fragment-length selection in the generative model and includes the modeling of errors and indels. The online EM algorithm for maximum-likelihood estimation.

The online EM algorithm is an iterative algorithm that consists of computing vectors $\gamma = [\gamma(T), \gamma(p)]$ where $\gamma(T) = 1, \ldots, T$. If the fragments are ordered as $f_1, \ldots, f_n$, then each $\gamma(p)$ represents an estimate of the parameter $\gamma(p)$ at the end of processing fragments $1, \ldots, p$. The update procedure is given by

$$\gamma(p)_{\text{new}} = \gamma(p)_{\text{old}} + (1 - \gamma(p)_{\text{old}}) \gamma(p)_{\text{old}}$$

where $\gamma(T) = 1$, for some constant $0 < c < 1$ and

$$\gamma(p)_{\text{old}} = \gamma(p)_{\text{old}} + (1 - \gamma(p)_{\text{old}}) \gamma(p)_{\text{old}}$$

The online EM algorithm is shown to converge under the appropriate conditions.

Counts. We distinguish between two forms of useful output in an RNA-seq experiment. The relative abundances of targets (encoded in the parameter $p$ discussed above) are of primary interest. However, also of interest are the posterior distributions of counts. The latter describe the number of fragments that are connected to each target. The effective count distribution for a transcript is defined as the posterior count distribution assuming the experiment had no sequence bias and all fragments had length 1. Effective counts are useful because they can be directly compared across experiments after they are normalized for sequencing depth. They can be calculated from estimated counts and the effective lengths ($S$), by rescaling the effective length of each fragment.

In Xpress, fragment-count distributions are modeled by shifted beta-binomial distributions as follows. For each target sequence $t$, the number of unique (Uniq) and total (Tot) fragments mapping to the target sequence are computed. Xpress then uses a beta-binomial distribution with parameters $\text{Tot}$, Uniq to approximate the posterior distribution by fitting moments. Specifically, for each target, the mean of the beta-binomial distribution is estimated as the mean of the posterior assignment probabilities $\gamma$. This is specified as a prior distribution that approximates the distribution for the fragment assignments.

To address instability of assignment in targets with few mapped reads, a flag (referred to as the solvability flag) is assigned to each target sequence. Initially all target sequences are unsolvable: if a fragment maps ambiguously to a set of target sequences and all but one is solvable, then the remaining one becomes unsolvable. Unsolvable targets are assigned the uniform distribution, whereas solvable targets are assigned a beta-binomial with support $[0, 1]$.

For the above procedure to work, it is necessary that the estimated counts always lie between Uniq and Tot. Moreover, the structure of (2) can be used to provide bounds on the scaled maximum-likelihood solution for the abundances described below.

Implementation of Xpress. Xpress takes input alignments in SAM or BAM format and makes use of only the required fields. It can therefore be applied to alignments made with any tool that outputs SAM format. The output of Xpress consists of three files containing the estimated abundances and fragment counts for each target sequence, parameter estimates and the variance-covariance matrix for the posterior count distributions. A modified SAM file containing the posterior probability of each alignment can be written to standard output. The output file from the posterior distribution, can also be optionally output.

Fragments are processed in a multithreaded pipeline. One thread processes each input stream into individual alignment objects. A second thread takes a set of fragment read alignments, computes the likelihood of each and updates the model parameters on the basis of the posterior probabilities of fragment origin. A third
1. Fragment assignment is fundamental for *Seq assays

The "RESCUE" method. Figure 3, Mortazavi et al. 2008 (and supplement)
What is a *Seq assay?

**Sequence**

**Desired measurement**

Creativity

Chemistry & Physics

Solve inverse problem

Mathematics & Computer Science

Computational Biology

Analyze

**Biological**

Molecular Biology

Statistics

Biology

\[
P(f = (p, t, l)) \approx \frac{\lambda_l \cdot \frac{\tau}{I(t)} \cdot w_{p|t,l} \cdot \phi_{f|p,t,l}}{\sum_{(q,r,m) \in A(f)} \lambda_m \cdot \frac{\tau}{I(r)} \cdot w_{q|r,m} \cdot \phi_{f|q,r,m}}
\]
First *Seq assay: ChIP-Seq

Desired measurement → **reduce to sequencing** → Sequence → Solve inverse problem → Analyze

Protein-DNA binding → Cross link, shear, IP → Sequence → Infer binding → Binding affinity analysis

Johnson et al., Science, 2007
Most popular *Seq assay: RNA-Seq

Desired measurement → reduce to sequencing → Sequence → Solve inverse problem → Analyze

RNA abundance → cDNA library preparation → Sequence → Estimate abundances → Differential analysis

Mortazavi et al., Nature Methods, 2008
More than 50 different assays have been published
Detecting WIMPs (weakly interacting massive particles)

Druker et al., arXiv, 2012


Full list at http://liorpachter.wordpress.com/seq/
Analysis of the sequencing output

- Desired measurement
- reduce to sequencing
- Sequence
- Solve inverse problem
- Mathematics & Computer Science
- Analyze
- Molecular Biology
- Creativity
- Chemistry & Physics
- Statistics
- Biology
- Computational Biology
- Creativity
- Chemistry & Physics
- Statistics
- Biology
- Computational Biology
Analysis of the sequencing output

Solve inverse problem

Fragment assignment

Density deconvolution

Cole Trapnell

Adam Roberts
Analysis of the sequencing output

For every read, find the target sequence and position in that target that it originated from.

After normalizing for non-uniform generation of reads from target sequences, for every position in every target sequence estimate the number of reads that originated there.
Analysis of the sequencing output

- Solve inverse problem
- Fragment assignment
- Density deconvolution

Reads

Target sequences
Analysis of the sequencing output

1. Solve inverse problem
2. Fragment assignment
3. Density deconvolution

Reads → Target sequences

Diagram showing the process of analyzing sequencing output with stages involving reads, target sequences, and a chick hatching from an egg.
The fragment assignment problem

target sequences
The fragment assignment problem

target sequences

?
The fragment assignment problem
The fragment assignment problem

Data:
q
q
q

target sequences
The fragment assignment problem

Data: \( q \quad q \quad q \)

Target sequences: \( l_1, l_2, l_3 \)
The fragment assignment problem

Data: $q$ $q$

Target sequences:

$q$ $q$ $q$

$q + q + q$
The fragment assignment problem

target sequences

read
The fragment assignment problem

target sequences

?
The fragment assignment problem

Data: q q q q q q q q
The fragment assignment problem

Data: $q_1$ $q_2$ $q_3$ $q_4$ $q_5$ $q_6$ $q_7$ $q_8$ $q_9$ $q_{10}$
A simple model

• The model parameters are the relative target abundances $\rho$.

• Equivalently, but more convenient is to parameterize the model with $\alpha$ as follows:

$$
\alpha_t = \frac{\rho_t l_t}{\sum_r \rho_r l_r}
$$

$$
\rho_t = \frac{\alpha_t l_t}{\sum_r \alpha_r l_r}
$$

probabilities of selecting fragments from each of the targets

relative target abundances
A likelihood function for fragment assignment

\[ L(\alpha_1, \ldots, \alpha_n) = \prod_{A \in 2^{[n]}} \left( \sum_{i \in A} \frac{\alpha_i}{\tilde{l}_i} \right)^{q_A} \]

- The product is over all subsets \( A \) of the set \( \{1,2,\ldots,n\} \) of targets.
- The notation \( q_A \) refers to the number of reads mapping to the targets in subset \( A \).
- The likelihood function is concave. (Proposition 1.4, Algebraic Statistics for Computational Biology).
An EM solution

Assuming equal length fragments and transcript, uniform coverage, and no sequencing errors.

\[ L(\rho|\mathcal{F}) = \prod_{f \in \mathcal{F}} \sum_{t \in \mathcal{T}} y_{f,t} \rho_t \]

\[ \sum_{t \in \mathcal{T}} \rho_t = 1 \]

\[ y_{f,t} = \begin{cases} 
1, & f \text{ compatible with } t \\
0, & \text{otherwise} 
\end{cases} \]

Xing et al 2006
A more realistic likelihood function


\[
L(\lambda, \tau, \pi, \phi | \mathcal{F}) = \prod_{f \in \mathcal{F}} \sum_{l=1}^{M_l} \sum_{t \in \mathcal{T}} \sum_{p=1}^{l(t)-l+1} \lambda_l \cdot \tau_t \cdot \pi_{p|t,l} \cdot \phi_{f|p,t,l}
\]

\[
\propto \prod_{f \in \mathcal{F}} \sum_{l=1}^{M_l} \sum_{t \in \mathcal{T}} \sum_{p=1}^{l(t)-l+1} \lambda_l \cdot \rho_t \cdot w_{p|t,l} \cdot \phi_{f|p,t,l}
\]

\[
\approx \prod_{f \in \mathcal{F}} \sum_{(p,t,l) \in \mathcal{A}(f)} \lambda_l \cdot \frac{\tau_t}{\bar{l}(t)} \cdot w_{p|t,l} \cdot \phi_{f|p,t,l}
\]

\[
\implies P(f = (p,t,l)) \approx \frac{\lambda_l \cdot \frac{\tau_t}{\bar{l}(t)} \cdot w_{p|t,l} \cdot \phi_{f|p,t,l}}{\sum_{(q,r,m) \in \mathcal{A}(f)} \lambda_m \cdot \frac{\tau_r}{\bar{l}(r)} \cdot w_{q|r,m} \cdot \phi_{f|q,r,m}}
\]
ONLINE METHODS

Probabilistic model. Our approach to fragment assignment is based on a probabilistic graphical model for sequencing experiments (Supplementary Fig. 11). In this framework, experiments produce multiple random fragments (according to the random variable \( F \)) that consist of pairs of sequencing reads (targets) from a set of target sequences. \( F \) depends on hidden random variables describing fragment length \( L \), target sequences of \( T \) and starting positions within target sequences \( P \). The probability distributions for the random variables are based on parameters for target abundances, fragment-length probabilities, sequence biases affecting the start and end locations of the fragment and probabilities for sequencing errors in reads. The generative model stipulates that the joint probability of obtaining a fragment \( f \) of length \( l \) sequenced from position \( p \) in target \( t \) is given by

\[
P(f = p, t, l) = \lambda f \cdot \rho_{p, t, l} 
\]

where the parameters of the model are the conditional probabilities \( \rho_{p, t, l} = P(F = f | p, t, l) = \lambda_{p, t, l} P(T = t | l) \). From this we obtain the following function

\[
L(
\begin{array}{c}
\alpha, \beta, \gamma
\end{array}
| f) = \sum_{t} \sum_{l} \sum_{p} \rho_{p, t, l} \cdot w_{p} \cdot \alpha \cdot \beta \cdot \gamma 
\]

The online EM algorithm for maximum-likelihood estimation. The online EM algorithm is an iterative algorithm that consists of computing vectors \( \alpha^{(t)} = \{\alpha_{i}^{(t)}\}_{i=1}^{N} \), where the \( i \)th fragment \( f_{i} \) is considered as an independent event. The algorithm iterates the following steps:

1. E-step: Evaluating \( \alpha_{i}^{(t)} \) and \( \beta_{i}^{(t)} \) for \( t \in T \).
2. M-step: Updating the parameters \( \rho_{p, t, l} \), \( \omega_{p} \), and \( \alpha_{i}^{(t)} \) and \( \beta_{i}^{(t)} \) for \( t \in T \).

The probabilities in equation (8) can be calculated using Bayes' rule from the conditional probabilities described in the section above.

**Theorem 1 (refs. 9,10).** The online EM algorithm is asymptotically equivalent to stochastic gradient ascent in the space of efficient statistics. Moreover, under the assumption that \( 0.5 < \epsilon \) is together with mild regularity assumptions \( \beta \), the algorithm converges to a local maximum of the likelihood function.

The weights \( w_{p} \) reflect sequence bias resulting in preferential selection of certain fragments \( \epsilon \) so that the weights \( w_{f} \) are normalized to \( w_{f} = 1 \) for each fragment.
ONLINE METHODS

Probabilistic model. Our approach to fragment assignment is based on a probabilistic graphical model for sequencing experiments (Supplementary Fig. 11). In this framework, experiments produce multiple random fragments (according to the random variable \( P \)) that consist of pairs of sets of reads (\( P \)) from a set of target sequences. \( P \) depends on hidden random variables describing fragment length (\( L \)), target sequences of origin (\( T \)) and starting positions within target sequences (\( T \)). The probability distributions for the random variables are based on parameters for target abundances, fragment-length probabilities, sequence biases affecting the start and end locations of the fragment and probabilities for sequencing errors in reads. The generative model stipulates that the joint probability of obtaining a fragment of length \( L \) sequenced from position \( p \) in target \( t \) is given by

\[
P(F = f, P = p, T = t, L = l) = \lambda f_p \pi(T = t | L = l) \lambda_p \pi(T = l | p)
\]

where the parameters of the model are the conditional probabilities \( \pi(T = t | L = l) \) and \( \pi(T = l | p) \).

From this we obtain the likelihood function

\[
L(\lambda, \pi, \sigma, \alpha) = \prod_{f \in F} \prod_{l \in L} \prod_{p \in P} \frac{1}{(e - 1) + 1} \left( \frac{1}{2} \right)^{e - 1}
\]

where \( F \) is the set of observed fragments, \( L \) is the set of target sequences, \( M \) is the maximum length of a fragment and \( H \) is the length of target sequence \( t \). The likelihood function is derived from the generative model, but there is a more convenient form that is useful computationally and that corresponds more directly to the main point of interest: the relative abundances of target sequences. If \( \rho_t \) denotes the relative abundance of target \( t \), and the probability of generating a fragment of arbitrary length from a target \( t \) is \( \lambda t \), then rewriting the likelihood function in terms of \( \rho_t \) yields

\[
L(\lambda, \pi, \sigma, \alpha) = \prod_{f \in F} \prod_{l \in L} \prod_{p \in P} \frac{1}{(e - 1) + 1} \left( \frac{1}{2} \right)^{e - 1}
\]

where the remaining weights \( w_p \), \( \lambda \), \( \sigma \), and \( \alpha \) are normalized to unity, and an effective length

\[
\tilde{L} = \sum_{l \in L} \lambda l \pi(l | p)
\]

The weights \( w_p \) reflect sequence bias resulting in preferential selection of certain fragments so that only at the end of reads. To confirm that such heuristic does not affect performance in practice, we provide a detailed analysis of the weight term, which finds all mappings within a specified Hamming distance, and compares the results to mappings from Bowtie (Supplementary Table 2). It is important to note that it is important that it incorporates a dataset of fragment lengths whenever possible, and that the model for assignment is independent of the technology used. The second difficulty can be addressed by a change of coordinates that greatly simplifies the calculation.

We replace \( \pi(T = l | p) \) with \( \omega(T = l), \) and instead of \( \sigma \), we compute

\[
a_t = \frac{\alpha_t}{\alpha_t + \epsilon}
\]

where

\[
\alpha_t = \sum_{l \in L} \lambda t \pi(l | t)
\]

is called the forgetting mass and depends on the forgetting factor \( \gamma \). It is convenient to use the form \( \gamma = 1/\alpha_t \), where \( 0 < \gamma < 1 \).

In that case, the recursion (10) reduces to the formula in Figure 1:

\[
\pi(T = l | p) = \frac{\pi(l | p) + \gamma \pi(T = l | p + 1)}{\gamma + 1}
\]

where \( \gamma = 1/\alpha_t \) for some constant 0 < \( \gamma < 1 \). This recursion is used to calculate \( \pi(T = l | p) \) at each position of the fragment.

The probabilities in equation (8) can be calculated using Bayes' rule from the conditional probabilities described in the section above.

Theorem 1 (refs. 9, 10). The online EM algorithm is asymptotically equivalent to stochastic gradient ascent in the space of efficient statistics. Moreover, under the assumption that \( 0 < \gamma < 1 \), it is together with mild regularity assumptions, the algorithm converges to a local maximum of the likelihood function.

For fixed auxiliary variables, the model (2) is convex, and it follows that the online algorithm (also called the stepwise EM algorithm) converges to the (unique) global maximum of the likelihood function.

Updating (7) requires O(1) operations at each step, making the algorithm attractive for large numbers of target sequences. The two main reasons for this. First, (computing \( \gamma \), requires, in principle, calculation of a normalization constant that is based on a sum taken over all positions in all targets. Second, the update in (7) requires changing \( \gamma \) for all \( t \). The first difficulty can be overcome by limiting the calculation to locations where fragments map using one of the known alignment programs, such as Bowtie. This is reasonable because the probabilities \( \pi(T = l | f) \) are approximately zero when a fragment does not align to a target because there is a relatively low probability of sequencing errors. Nevertheless, if only one strict fastAlign might lead to biased quantification results because they restrict mappings in all hco ways. For example, in Bowtie an exact matching seed is required, and three mismatches at most can be allowed.
No sample is an island

- Population surveys: LCLs of 462 individuals from 1000 genomes.

- Tissues: dozens of tissue samples from hundreds of dead people.

- Single-cell RNA-Seq experiments: $O(10^1)$ in 2012 — $O(10^2)$ in 2013 — $O(10^3)$ in 2014.

Regularized pooling for joint analysis of RNA-Seq
Regularized pooling for joint analysis of RNA-Seq
Regularized pooling for joint analysis of RNA-Seq

Nicolas Bray

Lee-Seung algorithm
The GEUVADIS data

• RNA-Seq on LCLs of 462 individuals from 1000 genomes.

• 30 million reads per sample.

• Recently published: T. Lappalainen et al., Nature (2013)

• We tested regularized pooling on the GEUVADIS data. Our implementation of the algorithm is based on eXpress (Roberts and Pacher, 2013).
Regularized pooling magnifies coverage

Nicolas Bray
2. Counting reads: wrong does not cancel out wrong

Trapnell et al., Nature Biotechnology 2012
The math

\[ \frac{1}{2} + \frac{2}{3} \neq \frac{3}{5} \]

but

\[ \frac{3}{4} + \frac{5}{6} = \frac{3}{5} \quad \frac{3}{8} \quad = \frac{3}{4} \]

(answer is: \( \frac{14}{19} = 0.7368421053\ldots \))
A thought experiment

<table>
<thead>
<tr>
<th>Exon-union model</th>
<th>Exon-intersection model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Isoform A</td>
<td></td>
</tr>
<tr>
<td>Isoform B</td>
<td></td>
</tr>
<tr>
<td>L</td>
<td>e</td>
</tr>
<tr>
<td>e</td>
<td>e</td>
</tr>
<tr>
<td>L - e</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Condition A</th>
<th>Condition B</th>
<th>log fold change (union count)</th>
<th>log fold change (intersect count)</th>
<th>log fold change (expression)</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Condition A" /></td>
<td><img src="image2" alt="Condition B" /></td>
<td>0</td>
<td>log (\frac{8}{7}) = 0.06</td>
<td>log (\frac{10}{L + \frac{4}{2L}}) = -0.20</td>
</tr>
<tr>
<td><img src="image3" alt="Condition A" /></td>
<td><img src="image4" alt="Condition B" /></td>
<td>log_2 (\frac{6}{8}) = -0.41</td>
<td>0</td>
<td>log_2 (\frac{6/L}{8/2L}) = 0.58</td>
</tr>
<tr>
<td><img src="image5" alt="Condition A" /></td>
<td><img src="image6" alt="Condition B" /></td>
<td>log_2 (\frac{5}{10}) = -1</td>
<td>log (\frac{4}{5}) = -0.1</td>
<td>0</td>
</tr>
</tbody>
</table>

Trapnell et al., Nature Biotechnology 2012
Not just a thought experiment

Trapnell et al., Nature Biotechnology 2012
DESeq fold-change estimation is inaccurate

Trapnell et al., Nature Biotechnology 2012
3. Units for RNA-Seq: CPM, RPKM, FPKM or TPM?

From SeqAnswers:

- The point with RPKM that I do not like, it is that I do not feel that it can handle different coverages.
> I totally agree with you. RPKM is biased for testing differential expression for longer genes.
Review: transcript abundance

- The model parameters are the relative target abundances $\rho$.

- Equivalently, but more convenient is to parameterize the model with $\alpha$ as follows:

\[
\rho_t = \frac{\alpha_t}{l_t} \sum_r \frac{\alpha_r}{l_r},
\]

\[
\alpha_t = \frac{\rho_t l_t}{\sum_r \rho_r l_r}.
\]
Special case: no ambiguously mapped reads

\[
\hat{\rho}_t = \frac{\hat{\alpha}_t}{\tilde{l}_t} = \frac{q_t \cdot \frac{1}{N \cdot \tilde{l}_t}}{\sum_{r \in T} \frac{\hat{\alpha}_r}{\tilde{l}_r}} \left( \frac{1}{\sum_{r \in T} \frac{q_r}{N \cdot \tilde{l}_r}} \right)
\]

\[
\propto \frac{q_t}{\left( \frac{\tilde{l}_t}{10^3} \right) \cdot \left( \frac{N}{10^6} \right)} \cdot \left( \frac{1}{\sum_{r \in T} \frac{q_r}{N \cdot \tilde{l}_r}} \right)
\]

\[
\propto \frac{q_t}{\left( \frac{\tilde{l}_t}{10^3} \right) \cdot \left( \frac{N}{10^6} \right)}
\]

RPKM

“read per kilobase per million mapped reads”
Special case: no ambiguously mapped reads

\[ \hat{\rho}_t = \frac{\sum_{r \in T} \frac{\hat{\alpha}_r}{\tilde{l}_r}}{\frac{q_t}{N} \cdot \frac{1}{\tilde{l}_t}} \cdot \left( \frac{1}{\sum_{r \in T} \frac{q_r}{N \tilde{l}_r}} \right) \]

\[ \propto \frac{q_t}{\left( \frac{\tilde{l}_t}{10^3} \right) \cdot \left( \frac{N}{10^6} \right)} \cdot \left( \frac{1}{\sum_{r \in T} \frac{q_r}{N \tilde{l}_r}} \right) \]

FPKM

“fragments per kilobase per million mapped reads”
The problem with FPKM

- Although abundances in FPKM are proportional to the relative abundances $\hat{\rho}_t$ the proportionality constant is *experiment specific*.

- Li and Dewey go back to the basics in the RSEM paper (BMC Bioinformatics, 2011). Instead of RPKM/FPKM, why not use a *universal* proportionality constant? Instead of $\hat{\rho}_t$, they propose **TPM**:

$$\hat{\rho}_t \times 10^6$$

- Please use TPM in your papers!
A final thought
A final thought

13,765 Cufflinks hits