FPKM, RPKM, TPM and KPKM (and other RNA-Seq Normalization Issues)

BYOB February 18, 2014

I may cite work by:

Rob Patro (programmer, Sailfish developer)

Carl Kingsford (PI)

Michael Kleyman

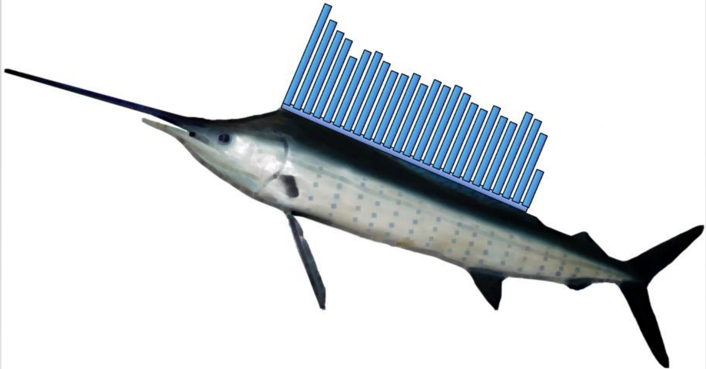
Julien Buchbinder (segments programmer)

Helen Salz (data provider)

ongen.us/SFish

Errors are my own

Steve Mount



Context:

Sailfish provides both RPKM and TPM?

Which should I use?

Outline:

- Definitions
- A toy example
- Considerations
  - meaning of RPKM vs. TPM
  - pros and cons of each

Conclusions:

The difference is small

RPKM is proportional in any one sample

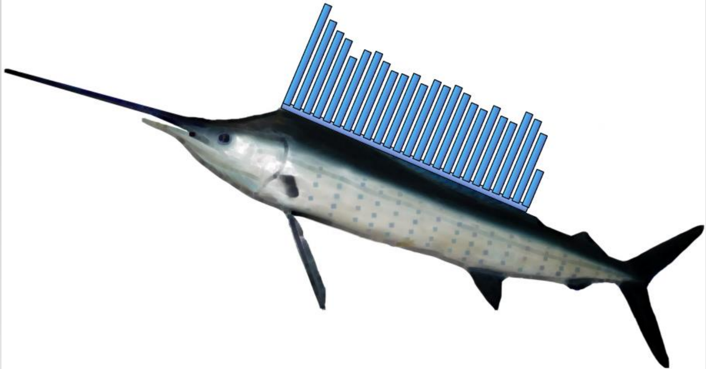
The ratio depends on the average length and on the length spread.

Recommendations:

- I like KPKM, but that’s not available yet.

What is Sailfish?

rapid k-mer-based RNAseq quantification of transcripts and gene segments.



Distinct goals of RNA-Seq

- Gene discovery (no genome, impossibly large genome). Trinity, etc.
- Isoform discovery
- Differential Expression Analysis
  - there is a transcriptome
  - How does expression of genes and isoforms vary? by
    - genotype
    - environmental change
    - pathogen
    - tissue or developmental stage

Distinct goals of RNA-Seq

- Gene discovery (no genome). Trinity, etc.

- Isoform discovery

- Differential Expression Analysis
  - there is a transcriptome
  - How does expression of genes and isoforms vary? by
    - genotype
    - environmental change
    - pathogen
    - tissue or developmental stage

Distinct goals of RNA-Seq

- Gene discovery (no genome). Trinity, etc.
- Isoform discovery
- Differential Expression Analysis
  - there is a transcriptome
  - How does expression of genes and isoforms vary? by
    - genotype
    - environmental change
    - pathogen
    - tissue or developmental stage

Transcript isoform discovery is no longer necessary once all isoforms have been described.

Maybe not today, maybe not tomorrow,
but soon, and for the rest of time.

Citation –
Humphrey Bogart as Rick in Casablanca

"If that plane leaves the ground and you're not on it,
you'll regret it.
Maybe not today, maybe not tomorrow,
but soon, and for the rest of your life."

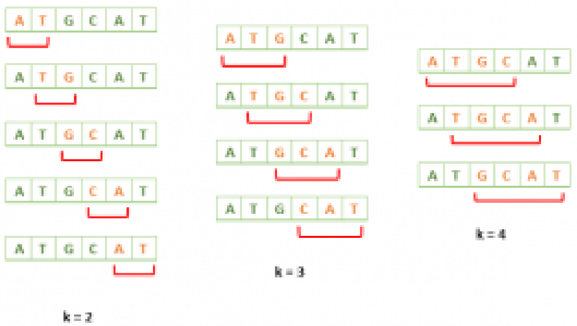
http://www.youtube.com/watch?v=xLQwphwP0ys

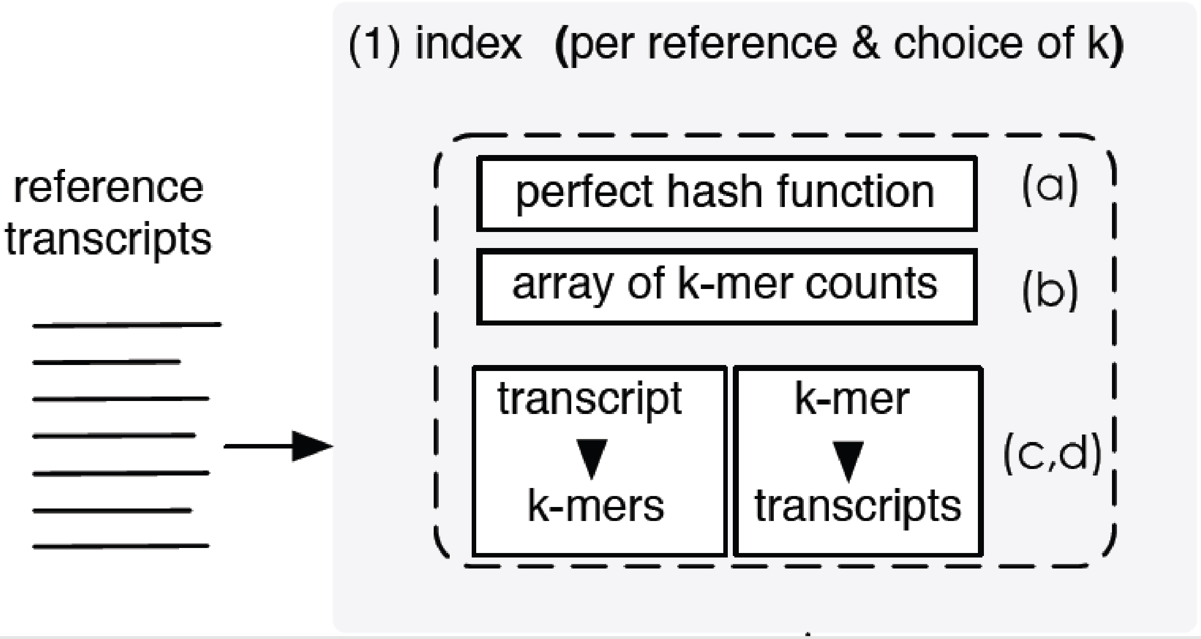
Distinct goals of RNA-Seq

- Gene discovery (no genome, impossibly large genome). Trinity, etc.
- Isoform discovery
- Differential Expression Analysis ←
  - there is a transcriptome
  - How does expression of genes and isoforms vary? by
    - genotype
    - environmental change
    - pathogen
    - tissue or developmental stage

Sailfish is for differential expression analysis.

Sailfish





reference transcripts

(1) index (per reference & choice of k)

perfect hash function (a)

array of k-mer counts (b)

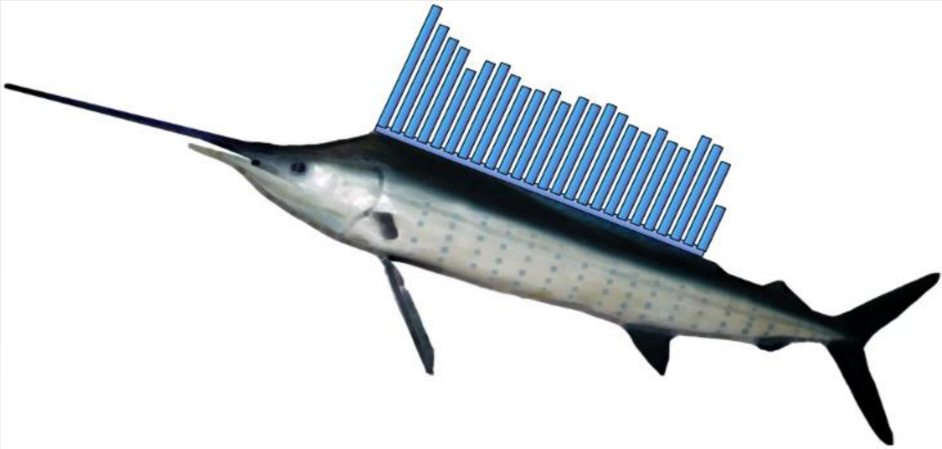
transcript ▼ k-mers

k-mer ▼ transcripts

(c,d)

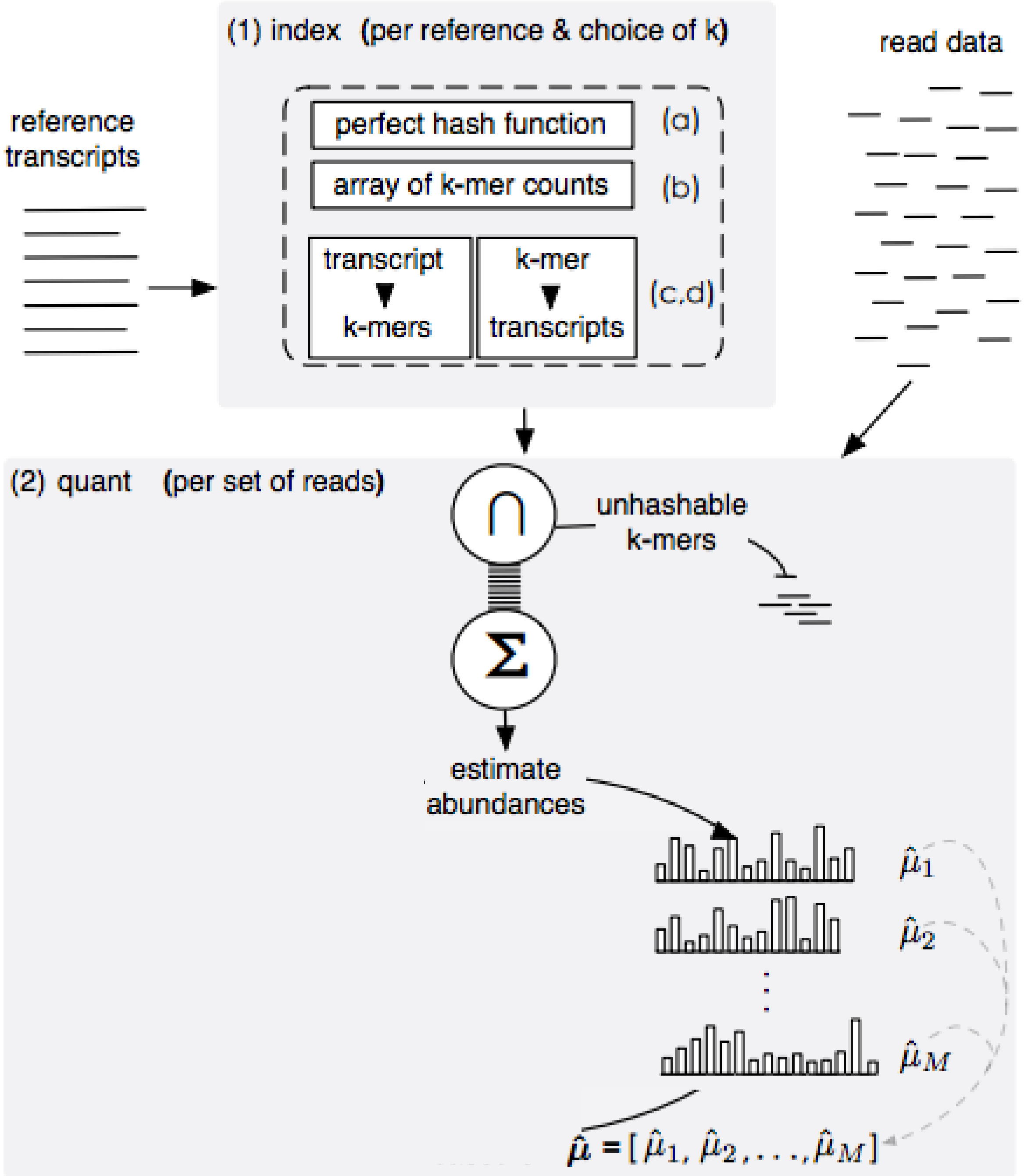
Sailfish generates an index of kmers that correlates transcripts to kmers and kmers to transcripts in a "perfect hash."

This index is built once per transcriptome.



Sailfish

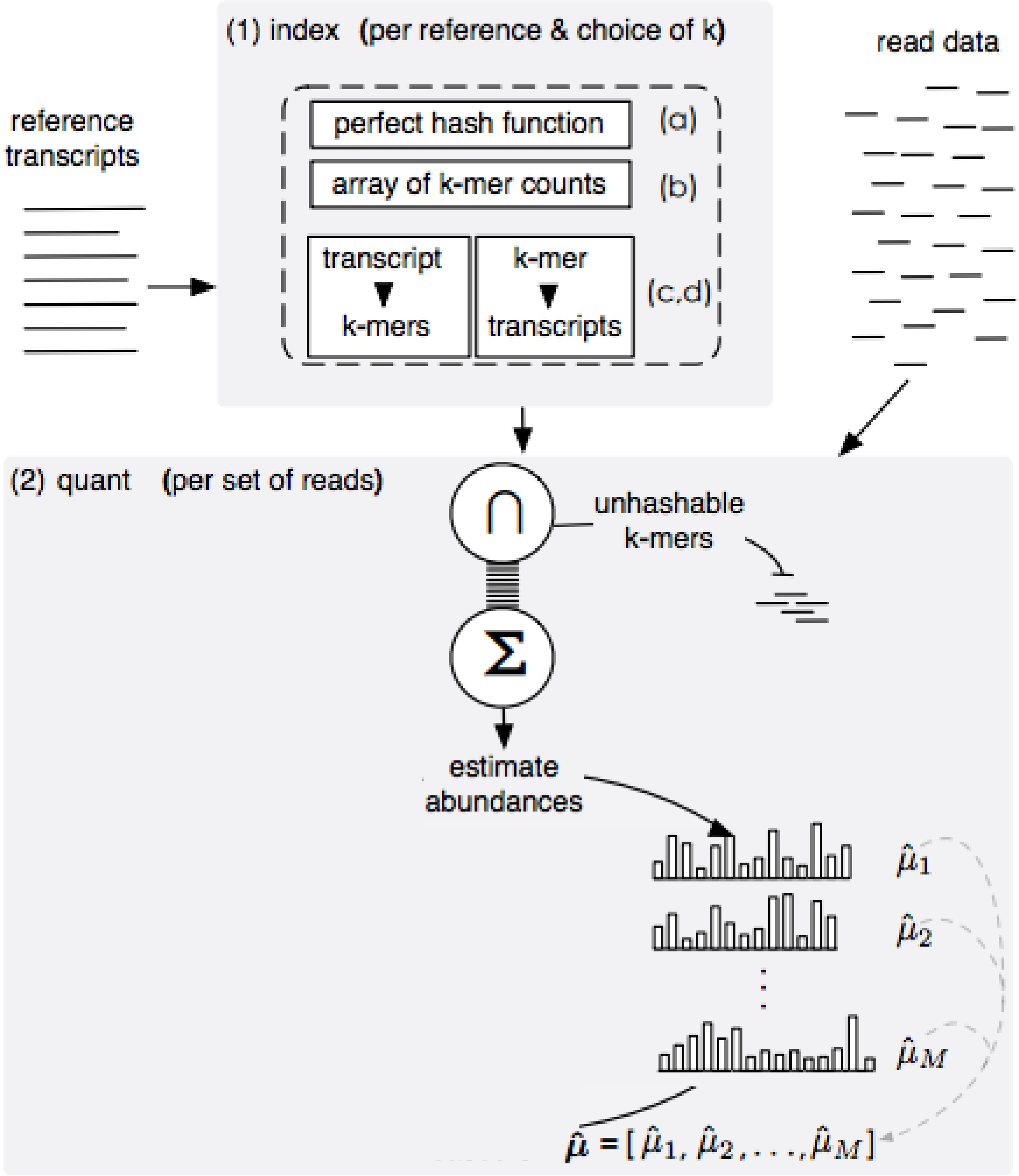
k-mers in reads are then tabulated.



Sailfish

k-mers in reads are then tabulated.

k-mer counts are then used to estimate transcript abundances.

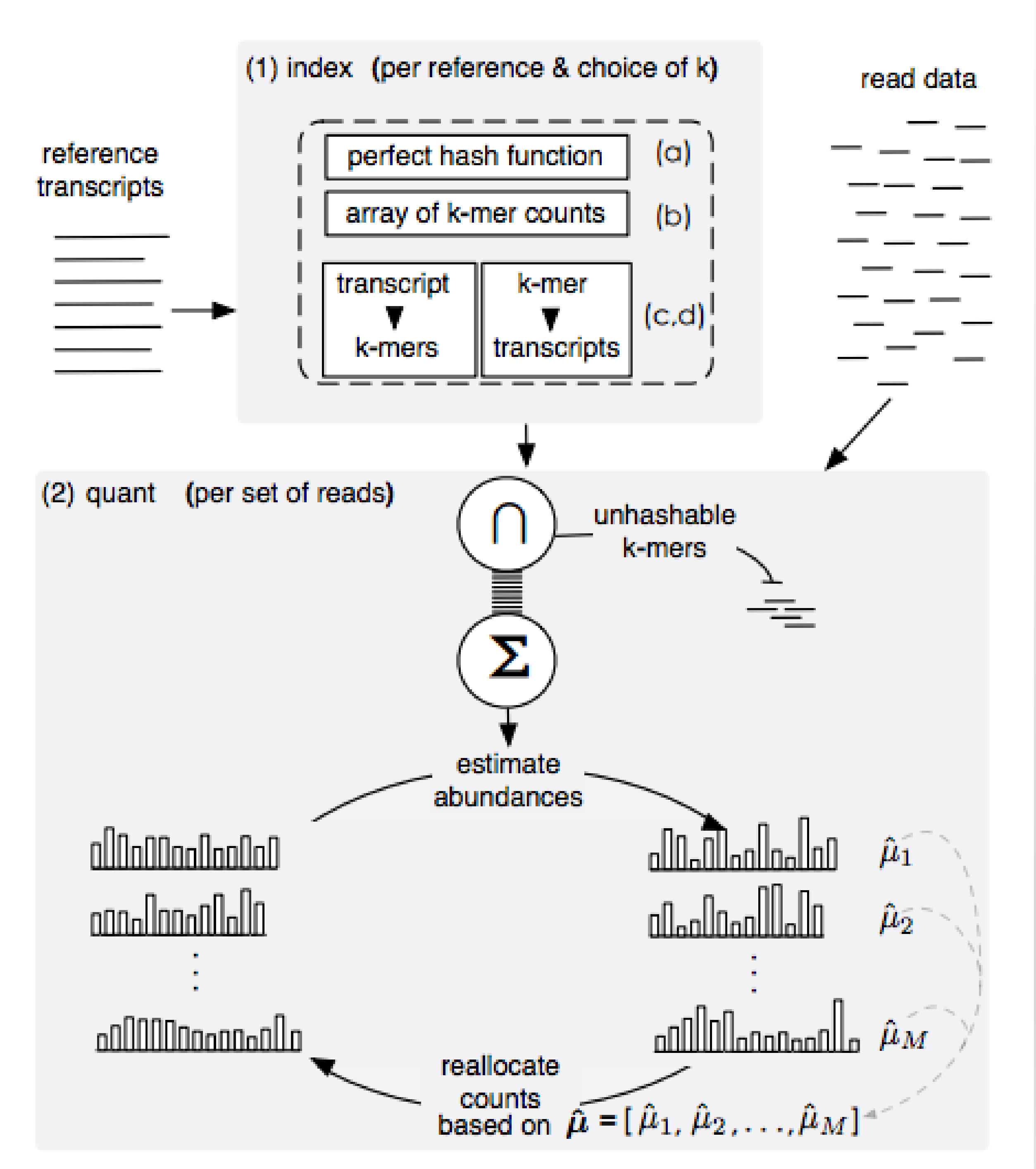


Sailfish

k-mers in reads are then tabulated.

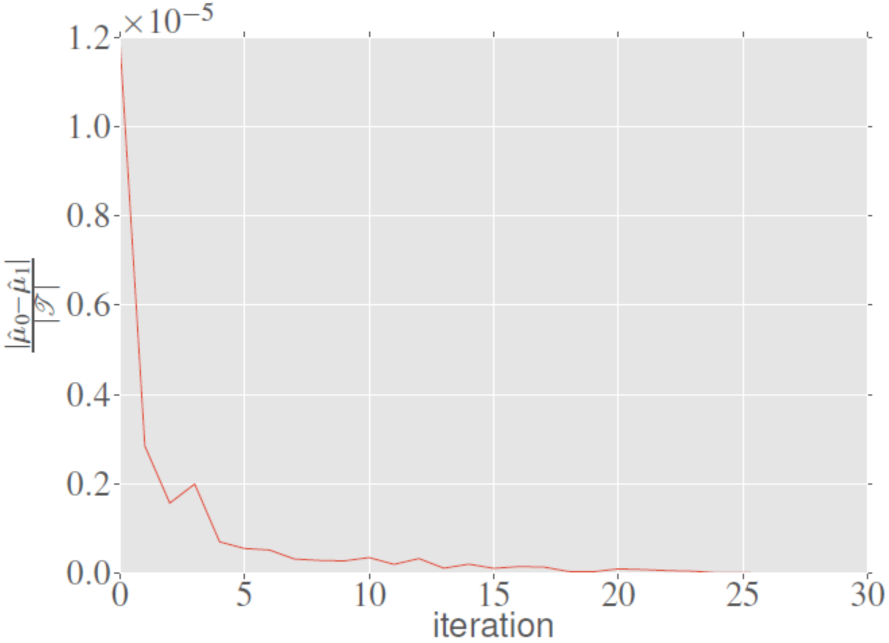
k-mer counts are then used to estimate transcript abundances.

k-mers are then reassigned using a maxmimum likelihood formula.

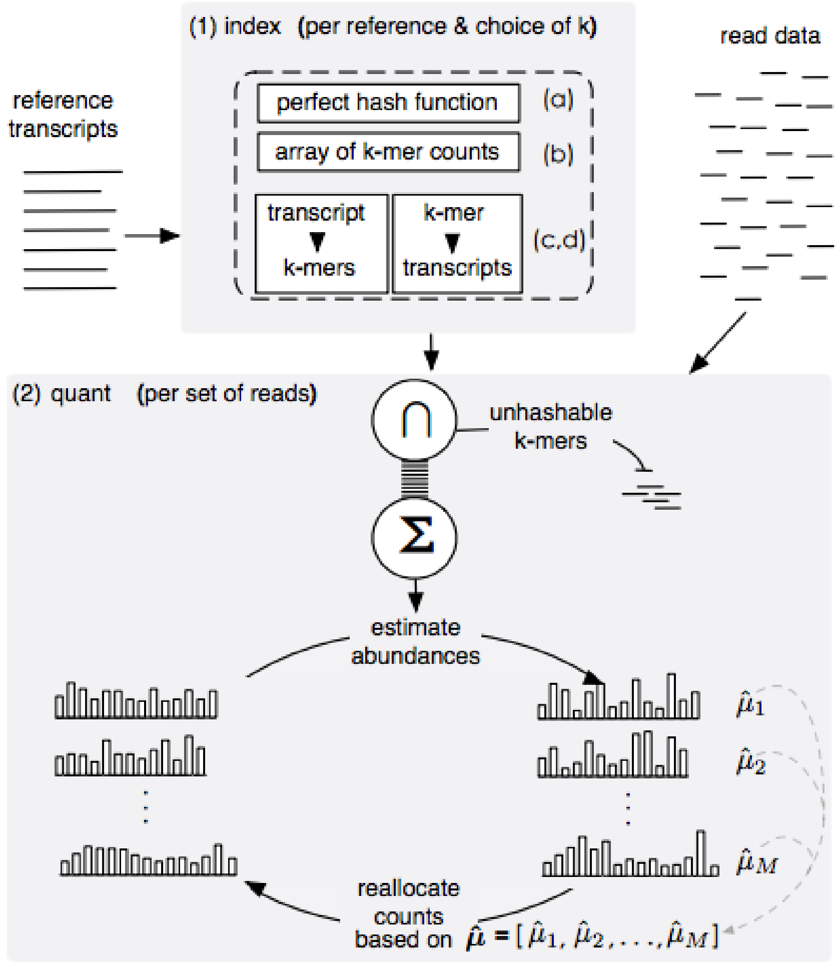


Sailfish

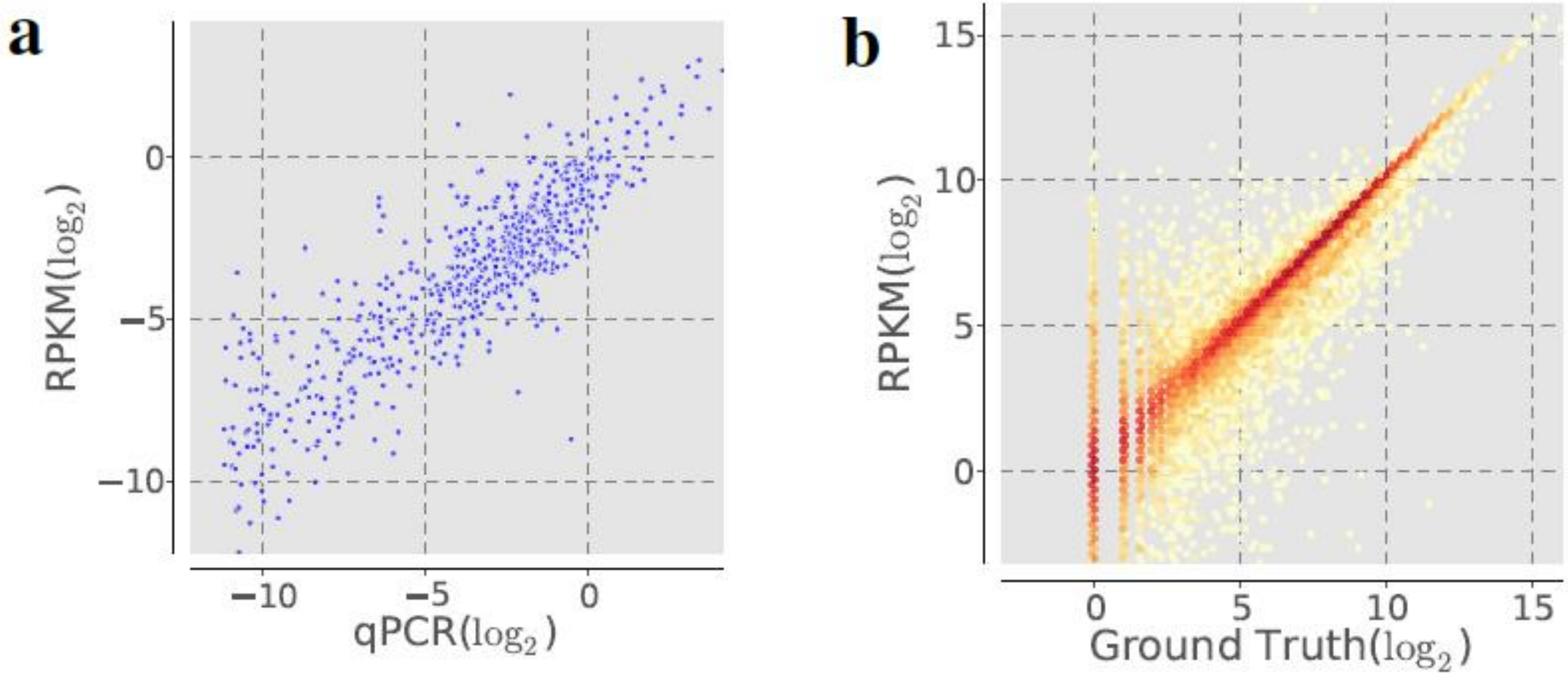
Convergence is rapid.

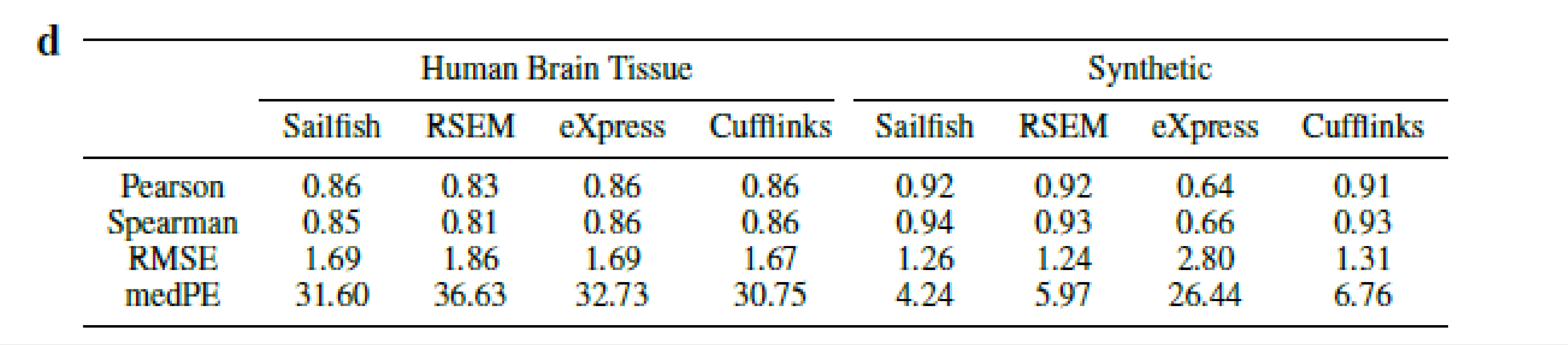


Supplementary Figure 5: The average difference between the relative abundance as estimated by two successive applications of the EM step (Algo. 2 lines 1–2) versus iterations of the SQUAREM algorithm (in the Universal Human Reference tissue experiment). We can see that the residual drops off quickly, and appears to have converged before 30 iterations of the SQUAREM procedure have been performed.



Sailfish is accurate!

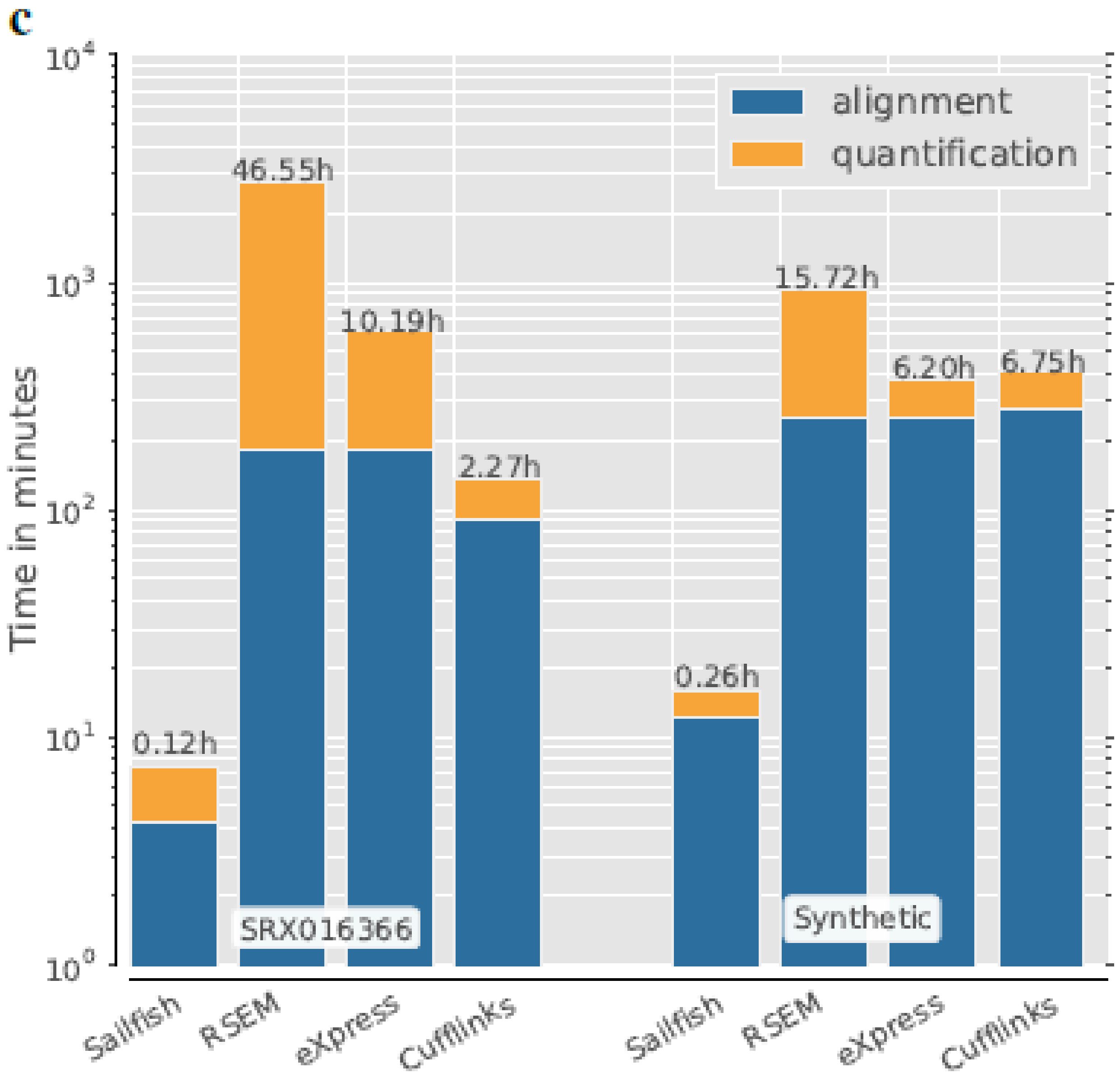


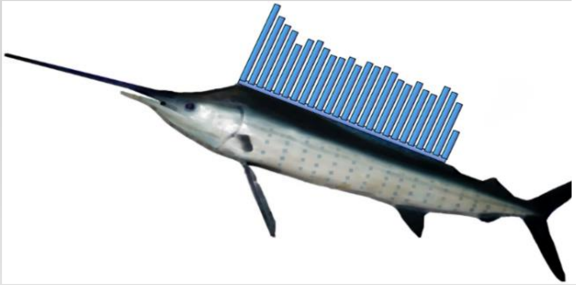

d

	Human Brain Tissue				Synthetic			
	Sailfish	RSEM	eXpress	Cufflinks	Sailfish	RSEM	eXpress	Cufflinks
Pearson	0.86	0.83	0.86	0.86	0.92	0.92	0.64	0.91
Spearman	0.85	0.81	0.86	0.86	0.94	0.93	0.66	0.93
RMSE	1.69	1.86	1.69	1.67	1.26	1.24	2.80	1.31
medPE	31.60	36.63	32.73	30.75	4.24	5.97	26.44	6.76

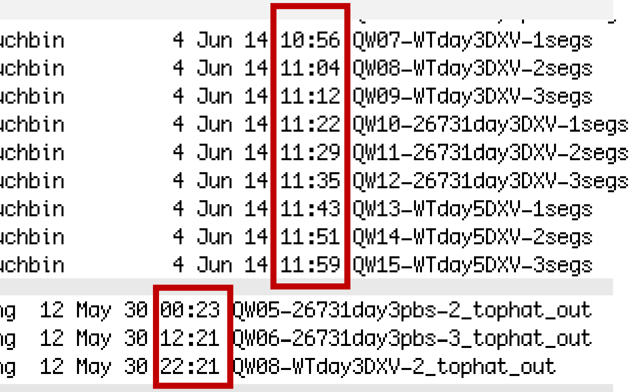
Sailfish is fast!

Note the log scale for time.





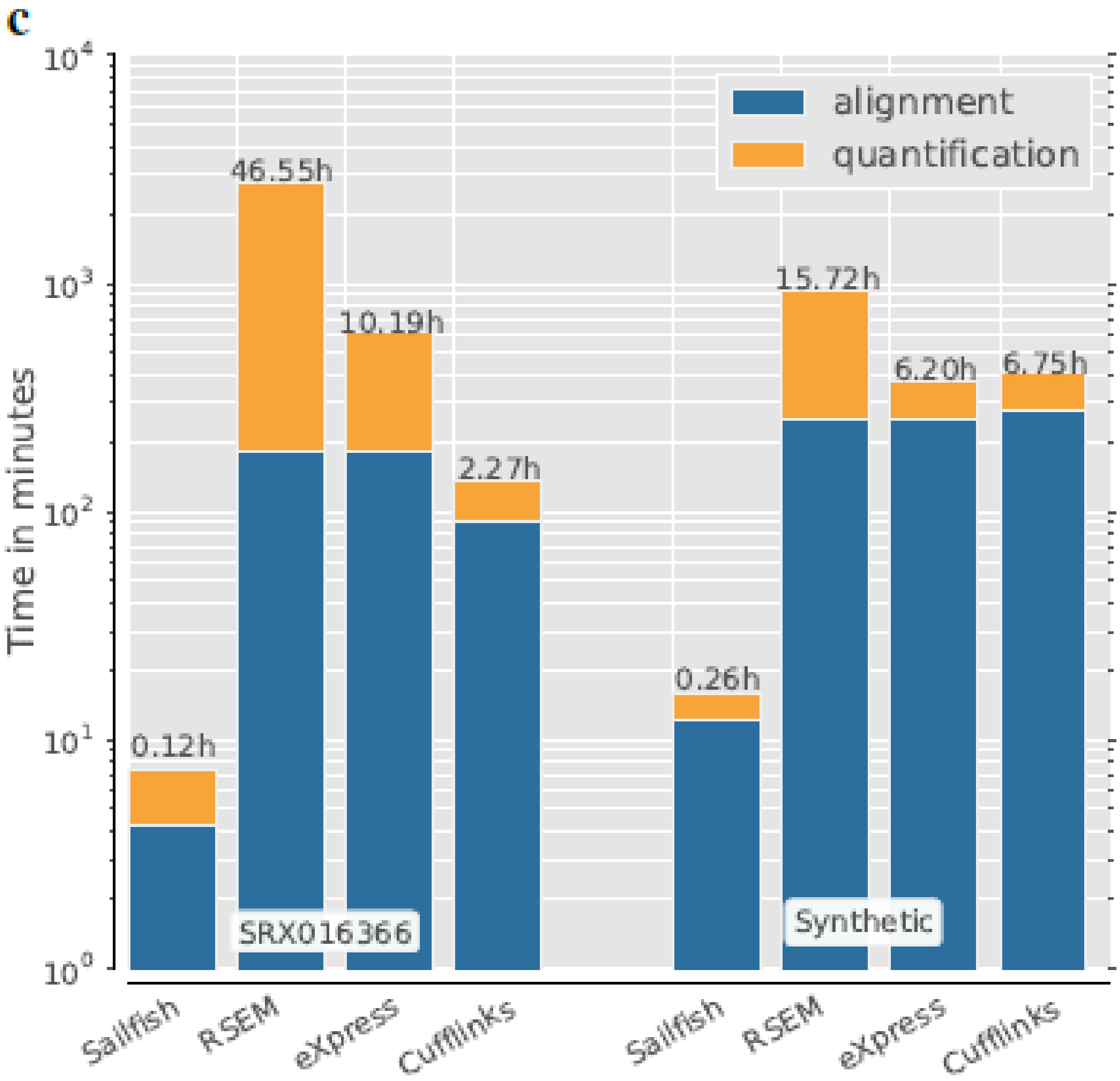
Sailfish is fast!

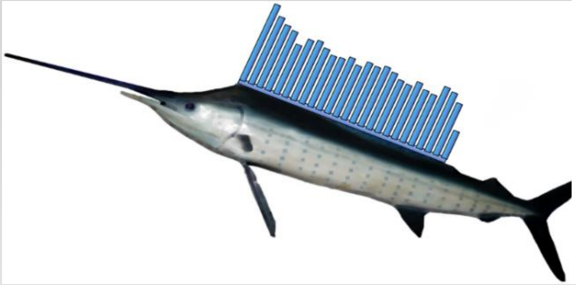


```
uchbin          4 Jun 14 10:56 QW07-WTday3DXV-1segs
uchbin          4 Jun 14 11:04 QW08-WTday3DXV-2segs
uchbin          4 Jun 14 11:12 QW09-WTday3DXV-3segs
uchbin          4 Jun 14 11:22 QW10-26731day3DXV-1segs
uchbin          4 Jun 14 11:29 QW11-26731day3DXV-2segs
uchbin          4 Jun 14 11:35 QW12-26731day3DXV-3segs
uchbin          4 Jun 14 11:43 QW13-WTday5DXV-1segs
uchbin          4 Jun 14 11:51 QW14-WTday5DXV-2segs
uchbin          4 Jun 14 11:59 QW15-WTday5DXV-3segs
```

```
ng  12 May 30 00:23 QW05-26731day3pbs-2_tophat_out
ng  12 May 30 12:21 QW06-26731day3pbs-3_tophat_out
ng  12 May 30 22:21 QW08-WTday3DXV-2_tophat_out
```

In practice, about 10 minutes per sample vs. about 10 hours

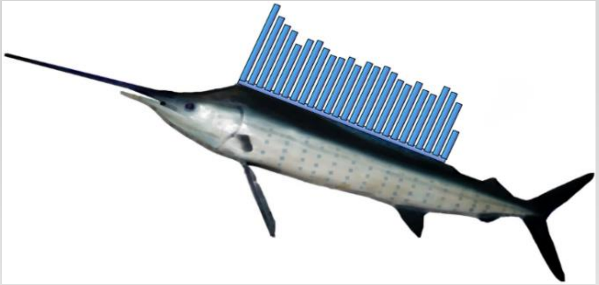




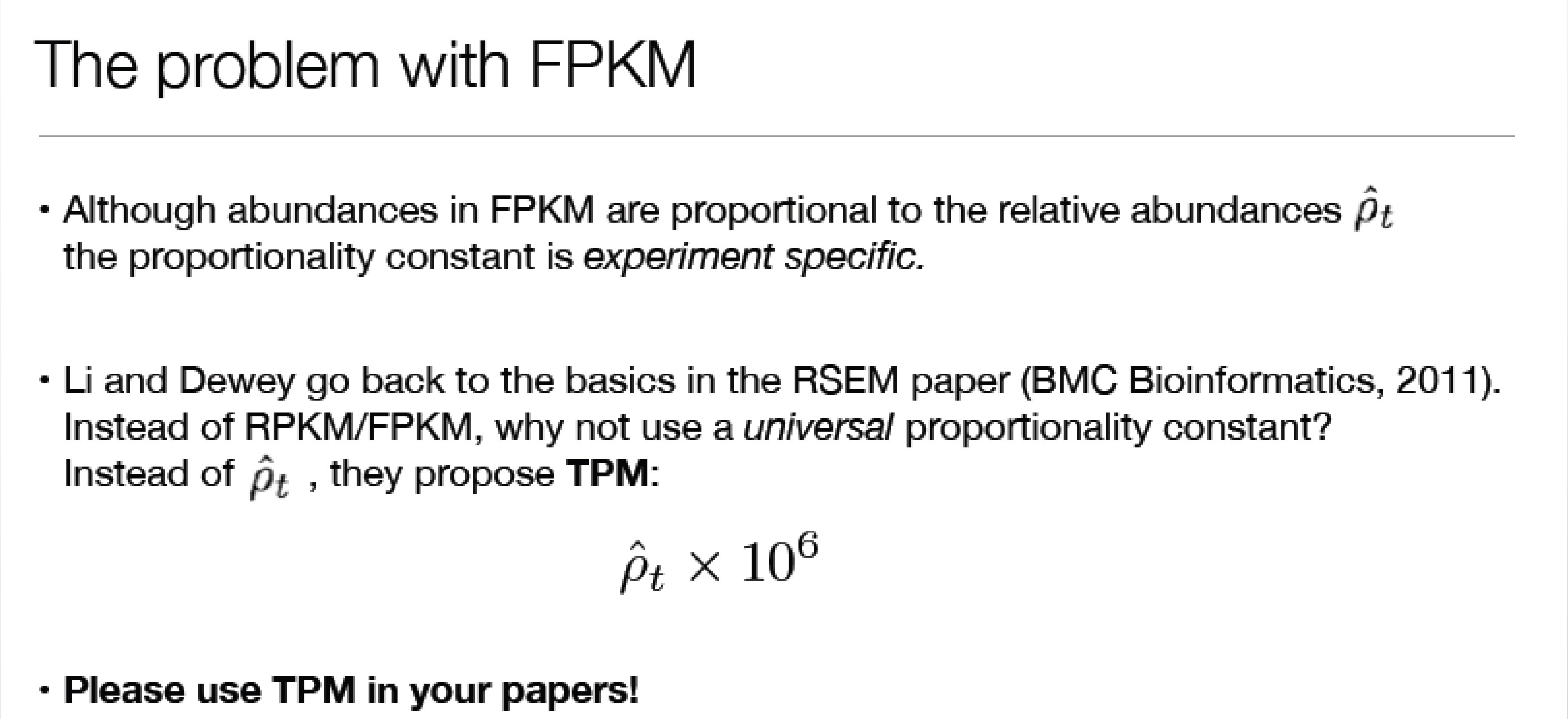
Sailfish reports both TPM and RPKM

Transcript	Length	TPM	RPKM
AB000402	2213	10.22	8.81
AB000718	1438	83.29	71.83
AB001420	2145	0.39	0.34
AB002397	9294	0.50	0.43

It makes sense to report both, so that appropriate comparisons can be made. But, which one should you use?



Sailfish reports both TPM and RPKM



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**:

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

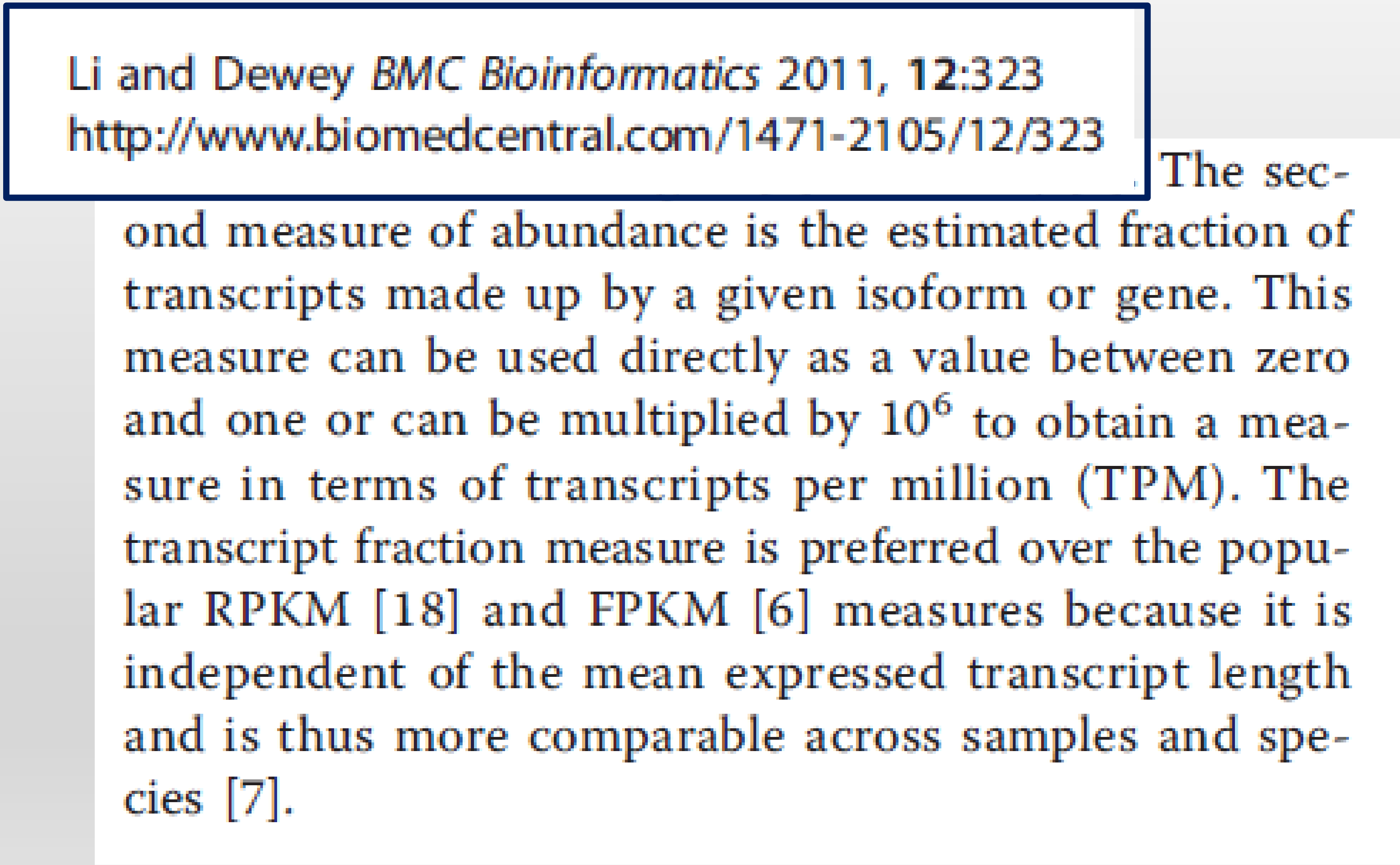
- **Please use TPM in your papers!**

Lior Pachter recommends TPM

Lior Pachter
CSHL Genome Informatics

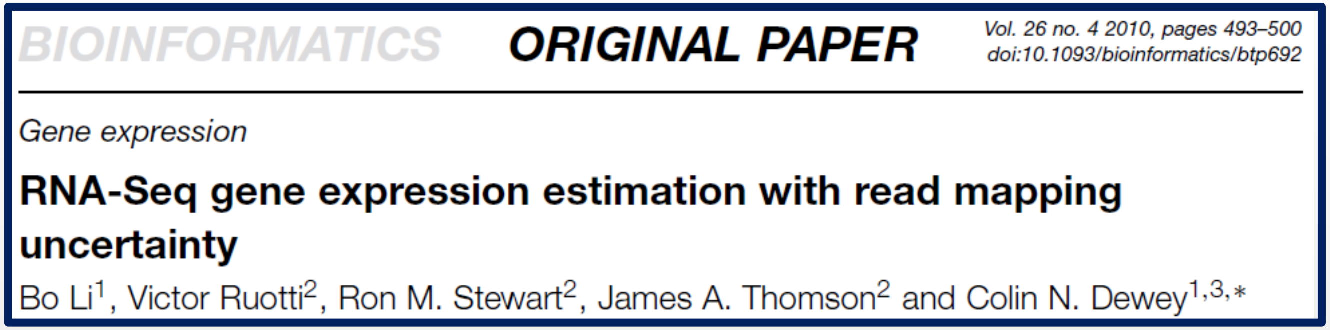
http://liorpachter.files.wordpress.com/2013/11/lior-pachter-genome-informatics-2013-keynote.pdf

So, what is the difference?



Li and Dewey *BMC Bioinformatics* 2011, **12**:323
http://www.biomedcentral.com/1471-2105/12/323

The second measure of abundance is the estimated fraction of transcripts made up by a given isoform or gene. This measure can be used directly as a value between zero and one or can be multiplied by 10^6 to obtain a measure in terms of transcripts per million (TPM). The transcript fraction measure is preferred over the popular RPKM [18] and FPKM [6] measures because it is independent of the mean expressed transcript length and is thus more comparable across samples and species [7].

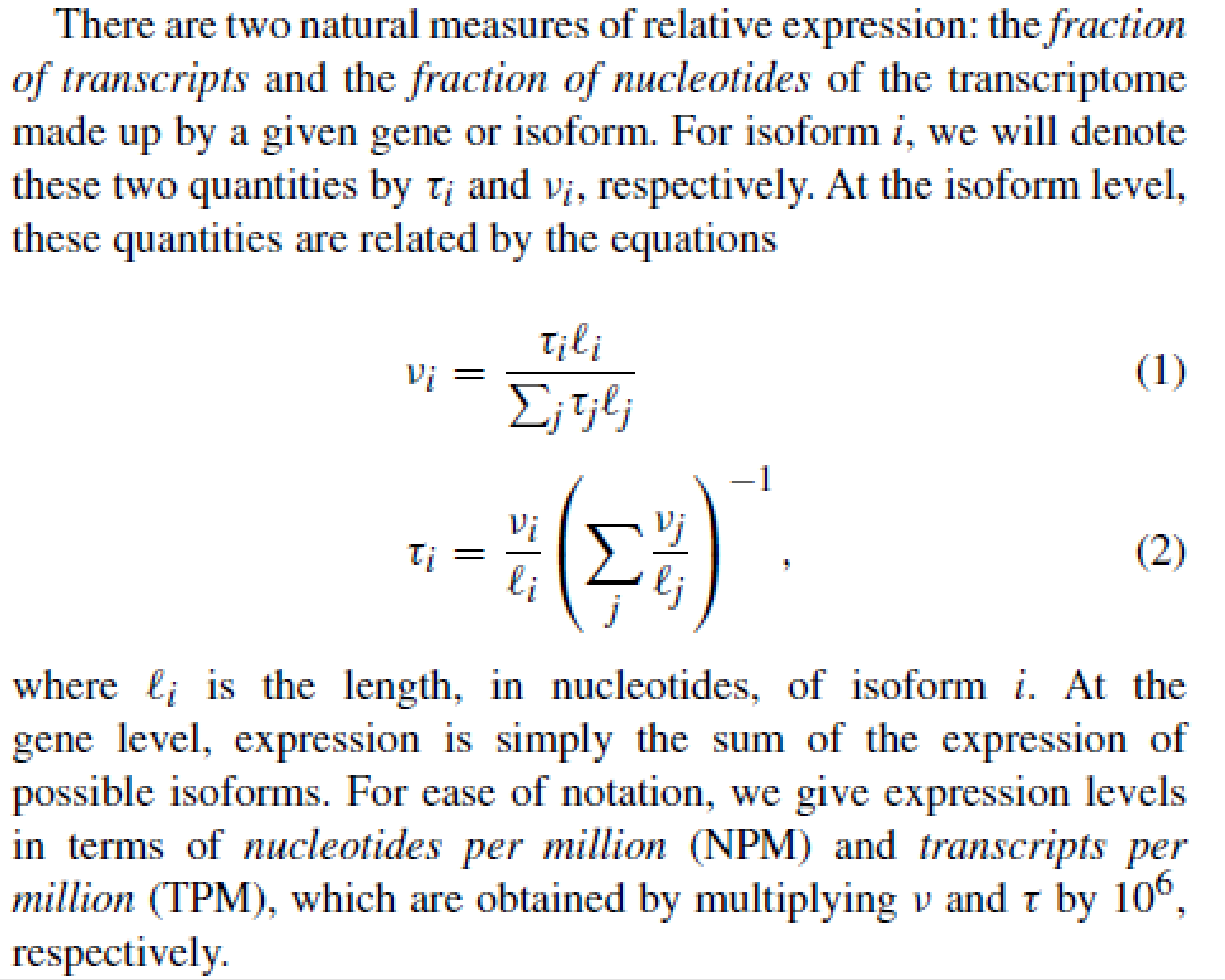


BIOINFORMATICS ORIGINAL PAPER Vol. 26 no. 4 2010, pages 493–500
doi:10.1093/bioinformatics/btp692

Gene expression

RNA-Seq gene expression estimation with read mapping uncertainty

Bo Li[1], Victor Ruotti[2], Ron M. Stewart[2], James A. Thomson[2] and Colin N. Dewey[1,3,*]



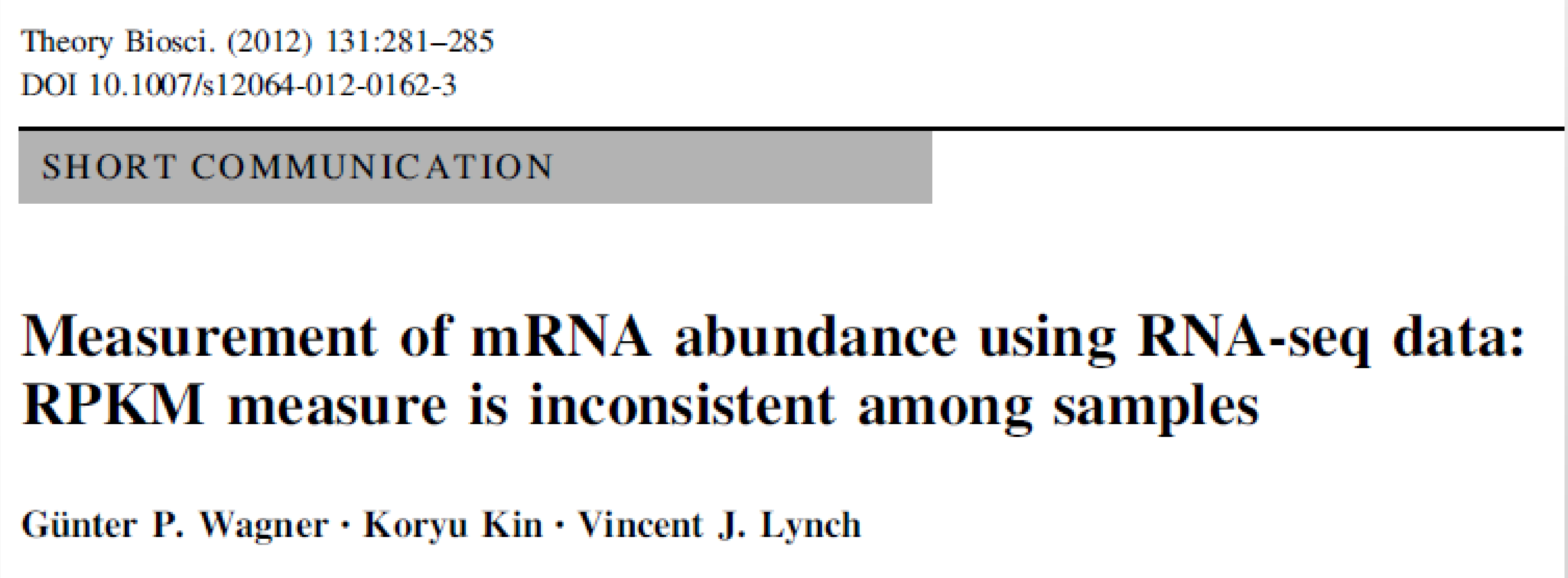
There are two natural measures of relative expression: the *fraction of transcripts* and the *fraction of nucleotides* of the transcriptome made up by a given gene or isoform. For isoform i, we will denote these two quantities by τ_i and ν_i, respectively. At the isoform level, these quantities are related by the equations

$$\nu_i = \frac{\tau_i \ell_i}{\sum_j \tau_j \ell_j} \tag{1}$$

$$\tau_i = \frac{\nu_i}{\ell_i} \left(\sum_j \frac{\nu_j}{\ell_j} \right)^{-1}, \tag{2}$$

where ℓ_i is the length, in nucleotides, of isoform i. At the gene level, expression is simply the sum of the expression of possible isoforms. For ease of notation, we give expression levels in terms of *nucleotides per million* (NPM) and *transcripts per million* (TPM), which are obtained by multiplying ν and τ by 10^6, respectively.

RPKM (or FPKM) vs. TPM – a toy example



Theory Biosci. (2012) 131:281–285
DOI 10.1007/s12064-012-0162-3

SHORT COMMUNICATION

Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples

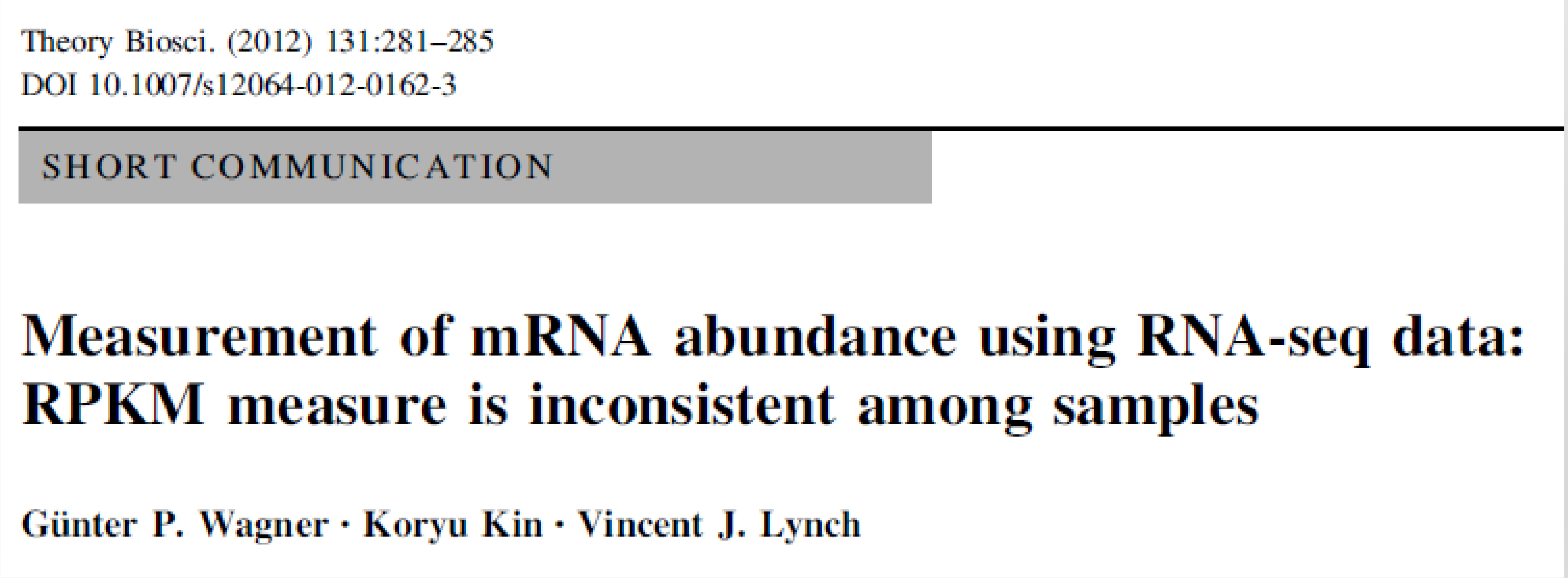
Günter P. Wagner · Koryu Kin · Vincent J. Lynch

This is a paper specifically on TPM. It is not especially well-written but it explains TPM vs. RPKM in detail.
The example is my own, but it illustrates their equations.

RPKM (or FPKM) vs. TPM – a toy example

100 reads
and 4 genes.

	Sample A	Sample B	length
Gene 01	50	10	5 kb.
Gene 02	10	50	2 kb.
Gene 03	20	20	1 kb.
Gene 04	20	20	1 kb.



Theory Biosci. (2012) 131:281–285
DOI 10.1007/s12064-012-0162-3

SHORT COMMUNICATION

Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples

Günter P. Wagner · Koryu Kin · Vincent J. Lynch

RPKM (or FPKM) vs. TPM – a toy example

100 reads
and 4 genes.

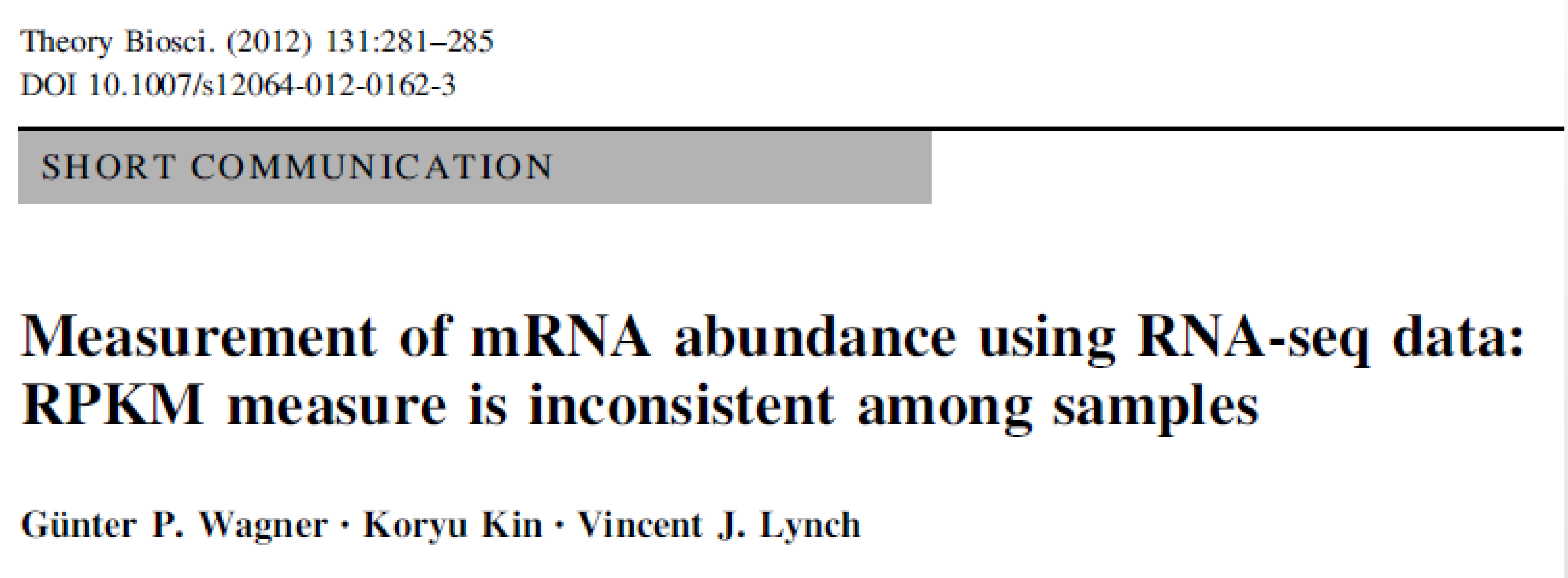
	Sample A	Sample B	length
Gene 01	50	10	5 kb.
Gene 02	10	50	2 kb.
Gene 03	20	20	1 kb.
Gene 04	20	20	1 kb.

Reads per kb.

~ RPK~~M~~

	Sample A	Sample B	length
Gene 01	10	2	5 kb.
Gene 02	5	25	2 kb.
Gene 03	20	20	1 kb.
Gene 04	20	20	1 kb.

This is proportional to the number of reads and corrected for length.
Thus, **RPKM is a measure of read density or molar abundance**.

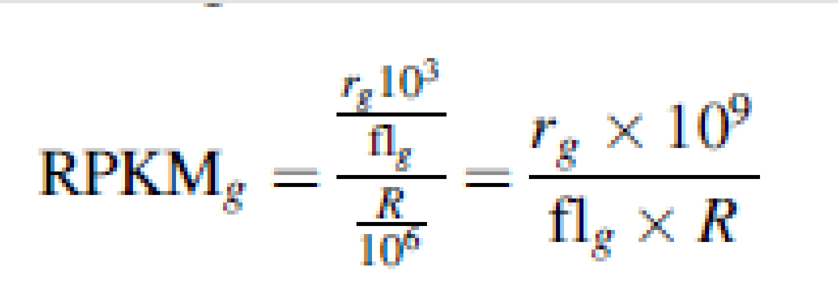


Theory Biosci. (2012) 131:281–285
DOI 10.1007/s12064-012-0162-3

SHORT COMMUNICATION

Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples

Günter P. Wagner · Koryu Kin · Vincent J. Lynch



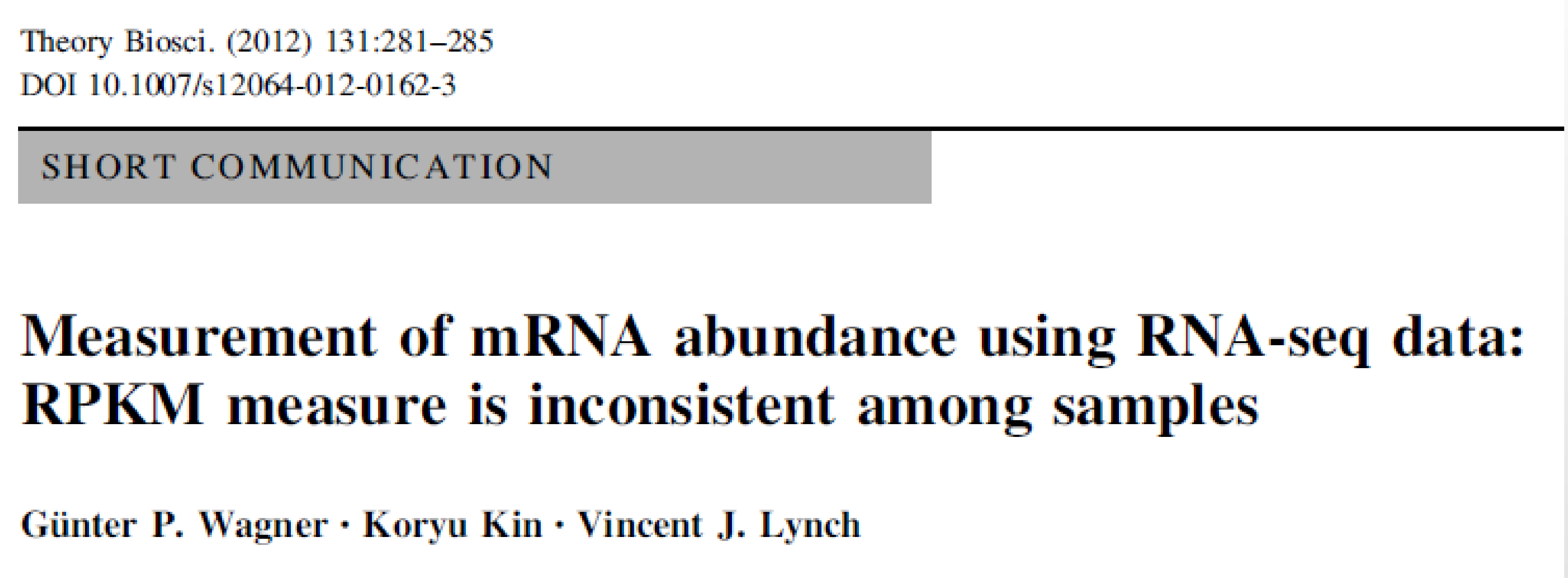
$$\mathrm{RPKM}_g = \frac{\frac{r_g 10^3}{\mathrm{fl}_g}}{\frac{R}{10^6}} = \frac{r_g \times 10^9}{\mathrm{fl}_g \times R}$$

$\boldsymbol{r_g}$ is the number of reads mapping to g.

$\boldsymbol{fl_g}$ is the length of the feature g.

$\boldsymbol{R}$ is the total number of reads.

“Reads per kilobase per million reads.”

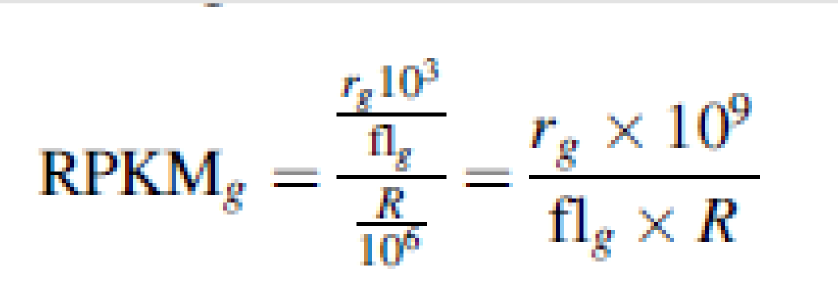


Theory Biosci. (2012) 131:281–285
DOI 10.1007/s12064-012-0162-3

SHORT COMMUNICATION

Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples

Günter P. Wagner · Koryu Kin · Vincent J. Lynch

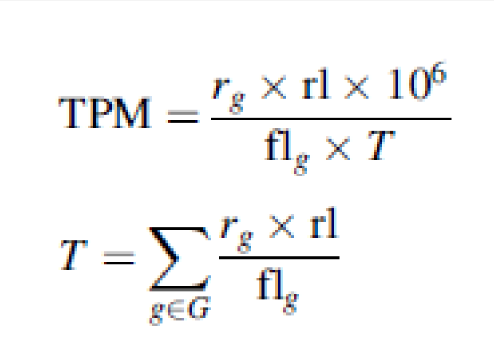


$$\mathrm{RPKM}_g = \frac{\frac{r_g 10^3}{\mathrm{fl}_g}}{\frac{R}{10^6}} = \frac{r_g \times 10^9}{\mathrm{fl}_g \times R}$$

r_g is the number of reads mapping to g.

fl_g is the length of the feature g.

R is the total number of reads.

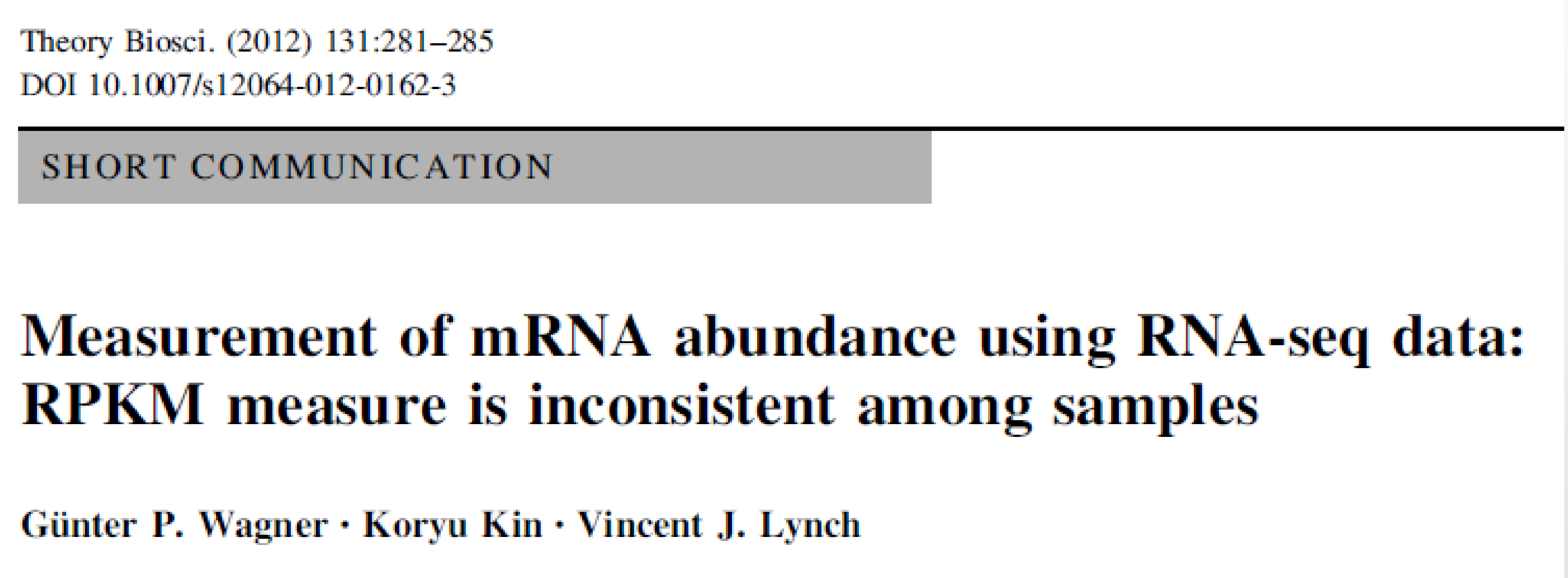


$$\mathrm{TPM} = \frac{r_g \times \mathrm{rl} \times 10^6}{\mathrm{fl}_g \times T}$$

$$T = \sum_{g \in G} \frac{r_g \times \mathrm{rl}}{\mathrm{fl}_g}$$

T is a constant of proportionality, which varies by experiment.

rl is the read length (a constant).

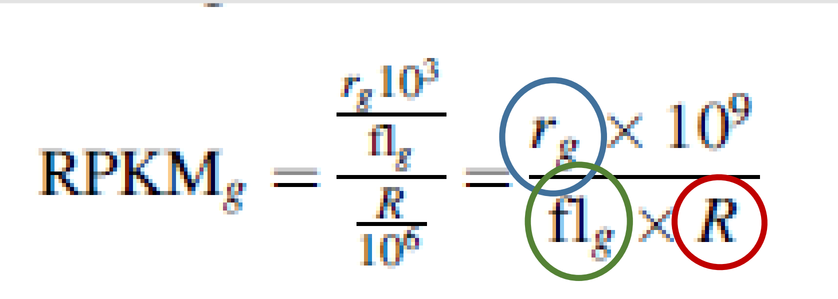


Theory Biosci. (2012) 131:281–285
DOI 10.1007/s12064-012-0162-3

SHORT COMMUNICATION

Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples

Günter P. Wagner · Koryu Kin · Vincent J. Lynch

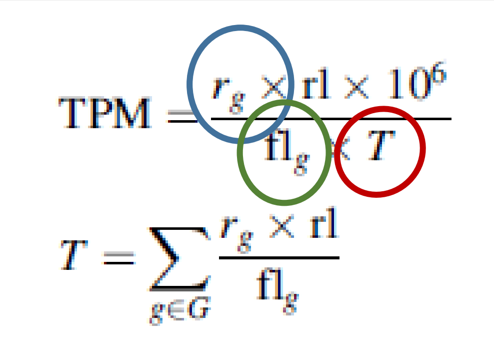


$$\mathrm{RPKM}_g = \frac{\frac{r_g 10^3}{\mathrm{fl}_g}}{\frac{R}{10^6}} = \frac{r_g \times 10^9}{\mathrm{fl}_g \times R}$$

r_g is the number of reads mapping to g.

fl_g is the length of the feature g.

R is the total number of reads.

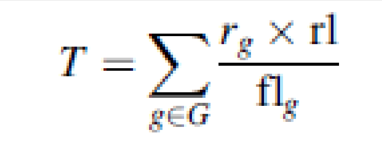


$$\mathrm{TPM} = \frac{r_g \times \mathrm{rl} \times 10^6}{\mathrm{fl}_g \times T}$$

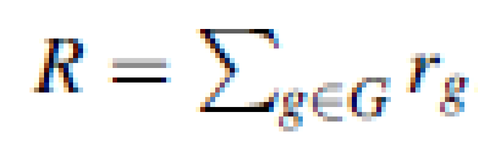
$$T = \sum_{g \in G} \frac{r_g \times \mathrm{rl}}{\mathrm{fl}_g}$$

T is a constant of proportionality, which varies by experiment.

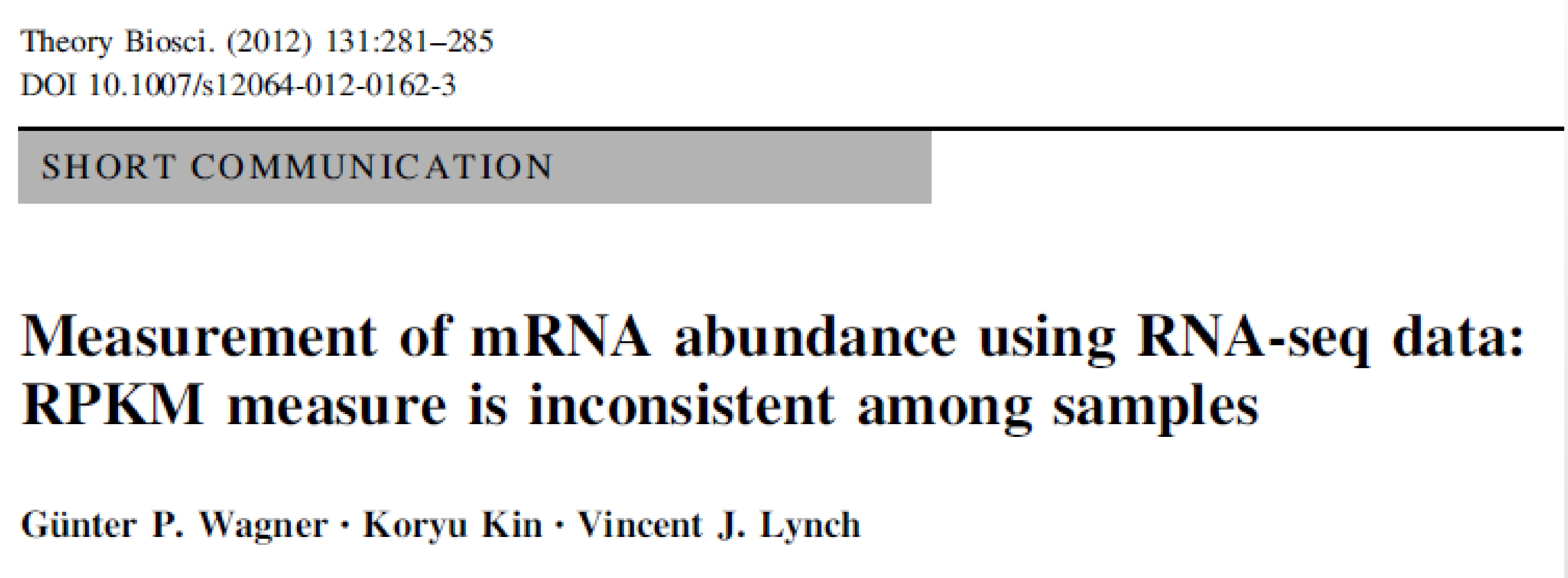
rl is the read length (a constant).



$$T = \sum_{g \in G} \frac{r_g \times \mathrm{rl}}{\mathrm{fl}_g}$$



$$R = \sum\nolimits_{g \in G} r_g$$

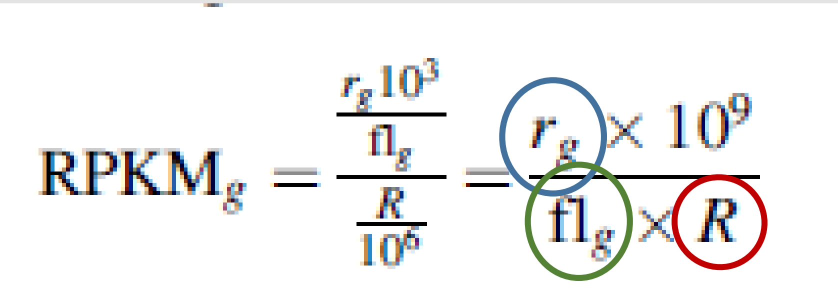


Theory Biosci. (2012) 131:281–285
DOI 10.1007/s12064-012-0162-3

SHORT COMMUNICATION

Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples

Günter P. Wagner · Koryu Kin · Vincent J. Lynch

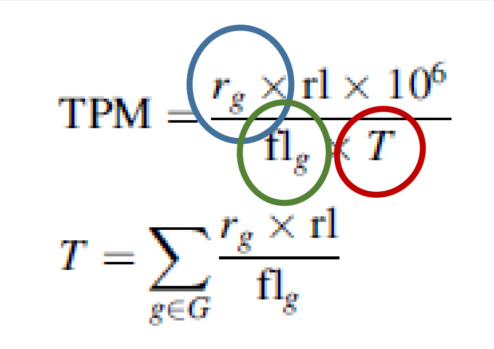


$$\mathrm{RPKM}_g = \frac{\frac{r_g 10^3}{\mathrm{fl}_g}}{\frac{R}{10^6}} = \frac{r_g \times 10^9}{\mathrm{fl}_g \times R}$$

r_g is the number of reads mapping to g.

fl_g is the length of the feature g.

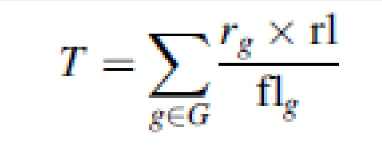
R is the total number of reads.



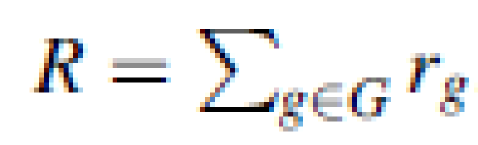
$$\mathrm{TPM} = \frac{r_g \times \mathrm{rl} \times 10^6}{\mathrm{fl}_g \times T}$$

$$T = \sum_{g \in G} \frac{r_g \times \mathrm{rl}}{\mathrm{fl}_g}$$

T and ***R*** differ in that T is weighted average, with reads from shorter features counting more.



$$T = \sum_{g \in G} \frac{r_g \times \mathrm{rl}}{\mathrm{fl}_g}$$



$$R = \sum\nolimits_{g \in G} r_g$$

100 reads
and 4 genes.

	Sample A	Sample B	length
Gene 01	50	10	5 kb.
Gene 02	10	50	2 kb.
Gene 03	20	20	1 kb.
Gene 04	20	20	1 kb.

Reads per kb.

~ RPK~~M~~

	Sample A	Sample B	length
Gene 01	10	2	5 kb.
Gene 02	5	25	2 kb.
Gene 03	20	20	1 kb.
Gene 04	20	20	1 kb.

Calculation of T

$\boldsymbol{r_g} \times \boldsymbol{rl} / \boldsymbol{fl_g}$

	Sample A	Sample B	length
Gene 01	1.0	0.2	5 kb.
Gene 02	0.5	2.5	2 kb.
Gene 03	2.0	2.0	1 kb.
Gene 04	2.0	2.0	1 kb.
Sum	5.5	6.7	

Here we are summing a value that is proportional to RPKM over all gene features.

Reads per kb.

~ RPK~~M~~

	Sample A	Sample B	length
Gene 01	10	2	5 kb.
Gene 02	5	25	2 kb.
Gene 03	20	20	1 kb.
Gene 04	20	20	1 kb.

Calculation of T

$\boldsymbol{r_g} \times \boldsymbol{rl} / \boldsymbol{fl_g}$

	Sample A	Sample B	length
Gene 01	1.0	0.2	5 kb.
Gene 02	0.5	2.5	2 kb.
Gene 03	2.0	2.0	1 kb.
Gene 04	2.0	2.0	1 kb.
Sum (T)	5.5	6.7	

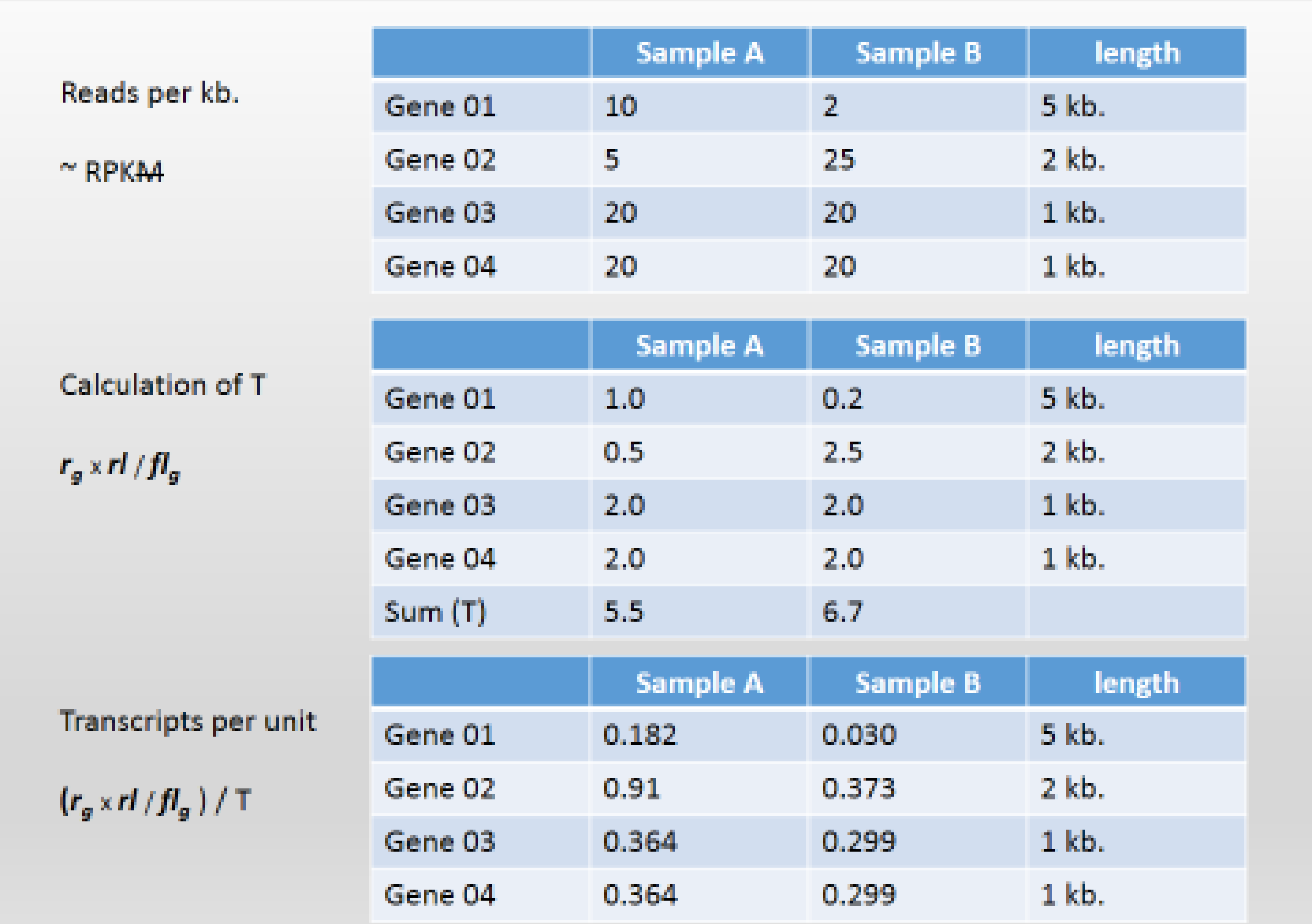
Genes 03 and 04 have the same RPKM but different TPM in these two samples.

Transcripts per unit

$(\boldsymbol{r_g} \times \boldsymbol{rl} / \boldsymbol{fl_g}) / \mathrm{T}$

	Sample A	Sample B	length
Gene 01	0.182	0.030	5 kb.
Gene 02	0.91	0.373	2 kb.
Gene 03	0.364	0.299	1 kb.
Gene 04	0.364	0.299	1 kb.

Within one sample, RPKM and TPM are proportional.
The ratio varies between samples based on the size and the distribution of sizes.



Reads per kb.

~ RPKM

	Sample A	Sample B	length
Gene 01	10	2	5 kb.
Gene 02	5	25	2 kb.
Gene 03	20	20	1 kb.
Gene 04	20	20	1 kb.

Calculation of T

$r_g \times rl / fl_g$

	Sample A	Sample B	length
Gene 01	1.0	0.2	5 kb.
Gene 02	0.5	2.5	2 kb.
Gene 03	2.0	2.0	1 kb.
Gene 04	2.0	2.0	1 kb.
Sum (T)	5.5	6.7	

Transcripts per unit

$(r_g \times rl / fl_g) / \mathrm{T}$

	Sample A	Sample B	length
Gene 01	0.182	0.030	5 kb.
Gene 02	0.91	0.373	2 kb.
Gene 03	0.364	0.299	1 kb.
Gene 04	0.364	0.299	1 kb.

RPKM (or FPKM) vs. TPM – a toy example

100 reads
and 4 genes.

	Sample A	Sample B	length
Gene 01	50	10	5 kb.
Gene 02	10	50	2 kb.
Gene 03	20	20	1 kb.
Gene 04	20	20	1 kb.

100 reads and 4
much shorter genes.

	Sample C	Sample D	length
Gene 05	50	10	0.5 kb.
Gene 06	10	50	0.2 kb.
Gene 07	20	20	0.1 kb.
Gene 08	20	20	0.1 kb.

RPKM (or FPKM) vs. TPM – a toy example

Reads per kb.

~ RPK~~M~~

	Sample A	Sample B	length
Gene 01	10	2	5 kb.
Gene 02	5	25	2 kb.
Gene 03	20	20	1 kb.
Gene 04	20	20	1 kb.

Reads per kb.

	Sample C	Sample D	length
Gene 05	100	20	0.5 kb.
Gene 06	50	250	0.2 kb.
Gene 07	200	200	0.1 kb.
Gene 08	200	200	0.1 kb.

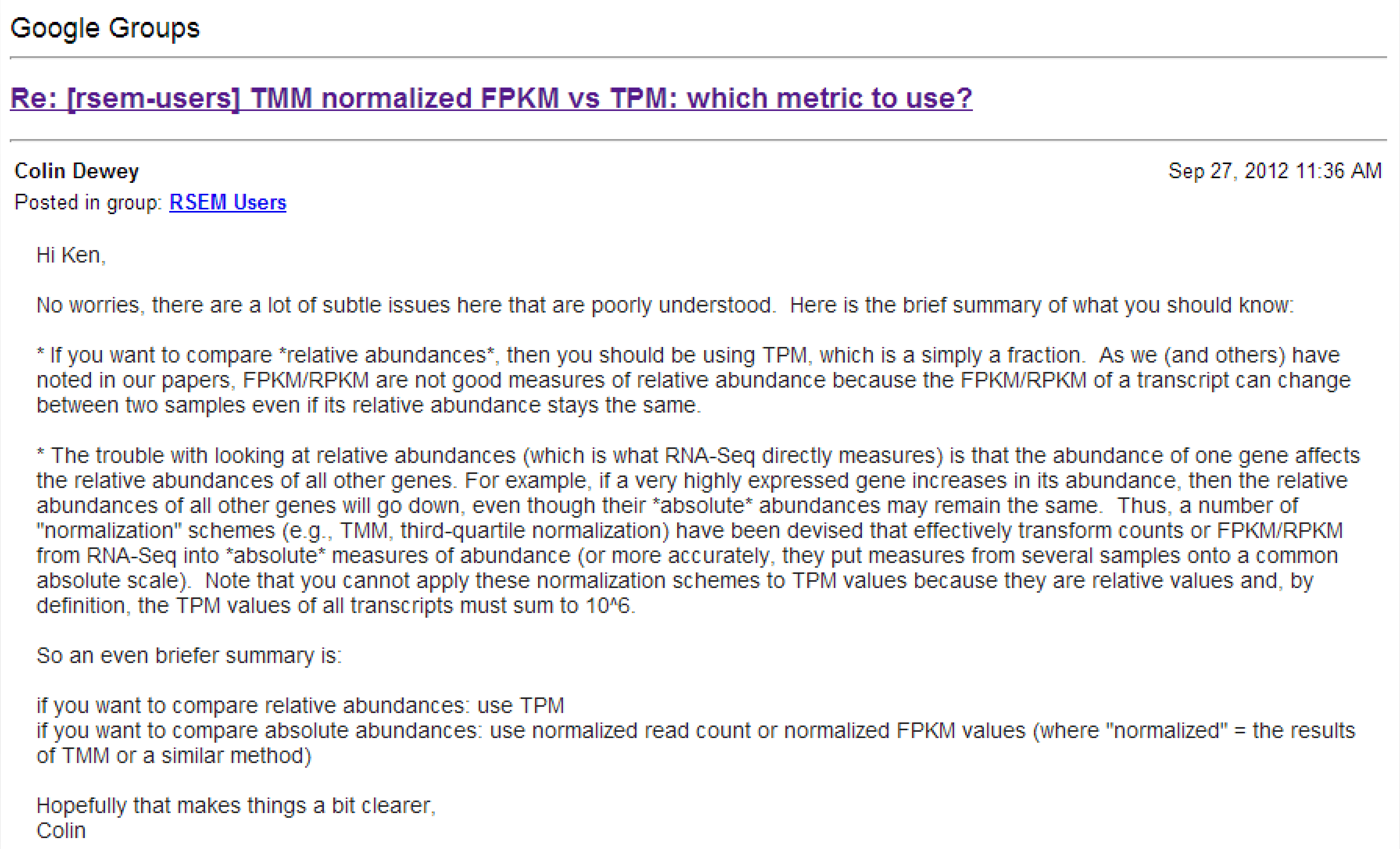
Transcripts per unit

These values are exactly the same.

$(\boldsymbol{r_g} \times \boldsymbol{rl} / \boldsymbol{fl_g}) / \mathrm{T}$

	Sample C	Sample D	length
Gene 01	0.182	0.030	0.5 kb.
Gene 02	0.91	0.373	0.2 kb.
Gene 03	0.364	0.299	0.1 kb.
Gene 04	0.364	0.299	0.1 kb.

If you're going to use another normalization method, then RPKM may make more sense.



Google Groups

Re: [rsem-users] TMM normalized FPKM vs TPM: which metric to use?

Colin Dewey
Sep 27, 2012 11:36 AM
Posted in group: RSEM Users

Hi Ken,

No worries, there are a lot of subtle issues here that are poorly understood. Here is the brief summary of what you should know:

* If you want to compare *relative abundances*, then you should be using TPM, which is a simply a fraction. As we (and others) have noted in our papers, FPKM/RPKM are not good measures of relative abundance because the FPKM/RPKM of a transcript can change between two samples even if its relative abundance stays the same.

* The trouble with looking at relative abundances (which is what RNA-Seq directly measures) is that the abundance of one gene affects the relative abundances of all other genes. For example, if a very highly expressed gene increases in its abundance, then the relative abundances of all other genes will go down, even though their *absolute* abundances may remain the same. Thus, a number of "normalization" schemes (e.g., TMM, third-quartile normalization) have been devised that effectively transform counts or FPKM/RPKM from RNA-Seq into *absolute* measures of abundance (or more accurately, they put measures from several samples onto a common absolute scale). Note that you cannot apply these normalization schemes to TPM values because they are relative values and, by definition, the TPM values of all transcripts must sum to 10^6.

So an even briefer summary is:

if you want to compare relative abundances: use TPM
if you want to compare absolute abundances: use normalized read count or normalized FPKM values (where "normalized" = the results of TMM or a similar method)

Hopefully that makes things a bit clearer,
Colin

https://groups.google.com/forum/print/msg/rsem-users/GRyJfEOK1BQ/l8_hEtsYVK8J

What about different length isoforms within the same gene?

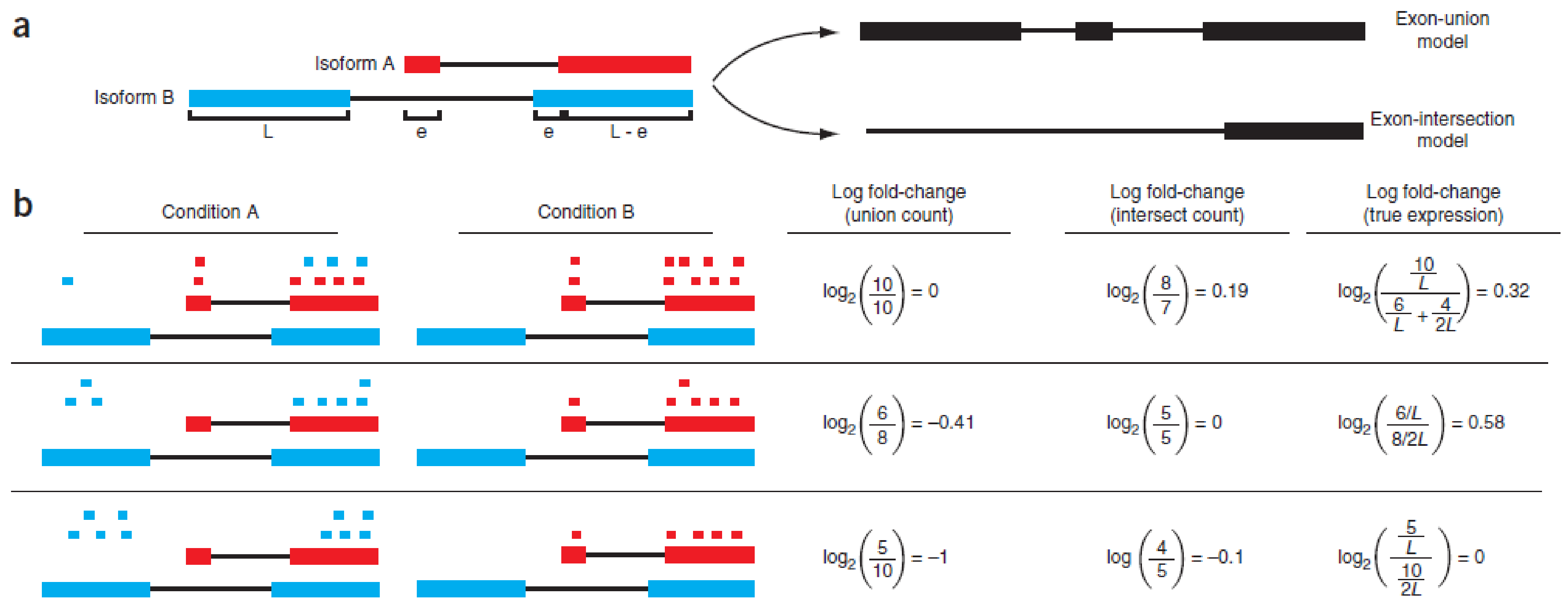

a
Isoform A
Isoform B
L
e
e
L - e
Exon-union model
Exon-intersection model
b
Condition A
Condition B
Log fold-change (union count)
Log fold-change (intersect count)
Log fold-change (true expression)
$\log_2\left(\frac{10}{10}\right) = 0$
$\log_2\left(\frac{8}{7}\right) = 0.19$
$\log_2\left(\frac{\frac{10}{L}}{\frac{6}{L}+\frac{4}{2L}}\right) = 0.32$
$\log_2\left(\frac{6}{8}\right) = -0.41$
$\log_2\left(\frac{5}{5}\right) = 0$
$\log_2\left(\frac{6/L}{8/2L}\right) = 0.58$
$\log_2\left(\frac{5}{10}\right) = -1$
$\log\left(\frac{4}{5}\right) = -0.1$
$\log_2\left(\frac{\frac{5}{L}}{\frac{10}{2L}}\right) = 0$

Figure 1 Changes in fragment count for a gene does not necessarily equal a change in expression. (**a**) Simple read-counting schemes sum the fragments incident on a gene's exons. The exon-union model counts reads falling on any of a gene's exons, whereas the exon-intersection model counts only reads on constitutive exons. (**b**) Both of the exon-union and exon-intersection counting schemes may incorrectly estimate a change in expression in genes with multiple isoforms. The true expression is estimated by the sum of the length-normalized isoform read counts. The discrepancy between a change in the union or intersection count and a change in gene expression is driven by a change in the abundance of the isoforms with respect to one another. In the top row, the gene generates the same number of reads in conditions A and B, but in condition B, all of the reads come from the shorter of the two isoforms, and thus the true expression for the gene is higher in condition B. The intersection count scheme underestimates the true change in gene expression, and the union scheme fails to detect the change entirely. In the middle row, the intersection count fails to detect a change driven by a shift in the dominant isoform for the gene. The union scheme detects a shift in the wrong direction. In the bottom row, the gene's expression is constant, but the isoforms undergo a complete switch between conditions A and B. Both simplified counting schemes register a change in count that does not reflect a change in gene expression.

Trapnell et al. 2013 Nature Biotechnology. "**Differential analysis of gene regulation at transcript resolution with RNA-seq" Cuffdiff 2**

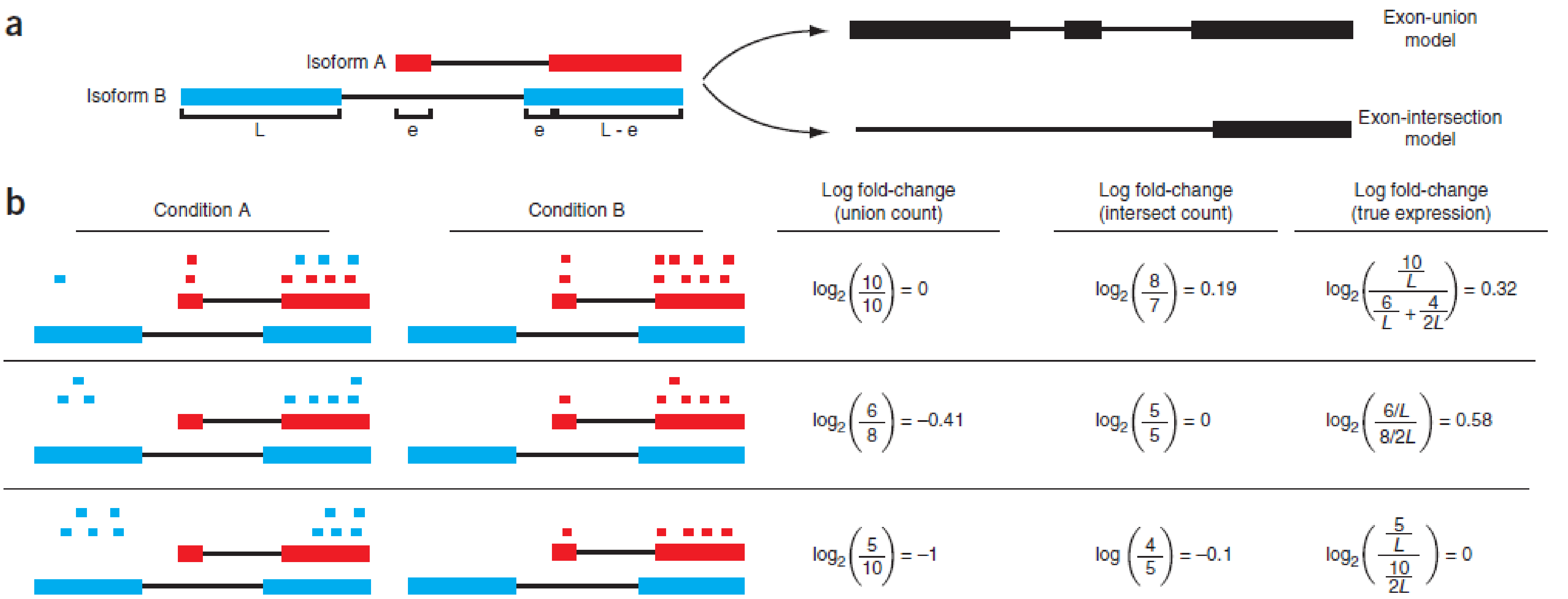

a
Isoform A
Isoform B
L
e
e
L - e
Exon-union model
Exon-intersection model
b
Condition A
Condition B
Log fold-change (union count)
Log fold-change (intersect count)
Log fold-change (true expression)
$\log_2\left(\frac{10}{10}\right) = 0$
$\log_2\left(\frac{8}{7}\right) = 0.19$
$\log_2\left(\frac{\frac{10}{L}}{\frac{6}{L}+\frac{4}{2L}}\right) = 0.32$
$\log_2\left(\frac{6}{8}\right) = -0.41$
$\log_2\left(\frac{5}{5}\right) = 0$
$\log_2\left(\frac{6/L}{8/2L}\right) = 0.58$
$\log_2\left(\frac{5}{10}\right) = -1$
$\log\left(\frac{4}{5}\right) = -0.1$
$\log_2\left(\frac{\frac{5}{L}}{\frac{10}{2L}}\right) = 0$

Figure 1 Changes in fragment count for a gene does not necessarily equal a change in expression. (**a**) Simple read-counting schemes sum the fragments incident on a gene's exons. The exon-union model counts reads falling on any of a gene's exons, whereas the exon-intersection model counts only reads on constitutive exons. (**b**) Both of the exon-union and exon-intersection counting schemes may incorrectly estimate a change in expression in genes with multiple isoforms. The true expression is estimated by the sum of the length-normalized isoform read counts. The discrepancy between a change in the union or intersection count and a change in gene expression is driven by a change in the abundance of the isoforms with respect to one another. In the top row, the gene generates the same number of reads in conditions A and B, but in condition B, all of the reads come from the shorter of the two isoforms, and thus the true expression for the gene is higher in condition B. The intersection count scheme underestimates the true change in gene expression, and the union scheme fails to detect the change entirely. In the middle row, the intersection count fails to detect a change driven by a shift in the dominant isoform for the gene. The union scheme detects a shift in the wrong direction. In the bottom row, the gene's expression is constant, but the isoforms undergo a complete switch between conditions A and B. Both simplified counting schemes register a change in count that does not reflect a change in gene expression.

Trapnell et al. 2013 Nature Biotechnology. "**Differential analysis of gene regulation at transcript resolution with RNA-seq" Cuffdiff 2**

This is wrong!
Or, at least, misleading

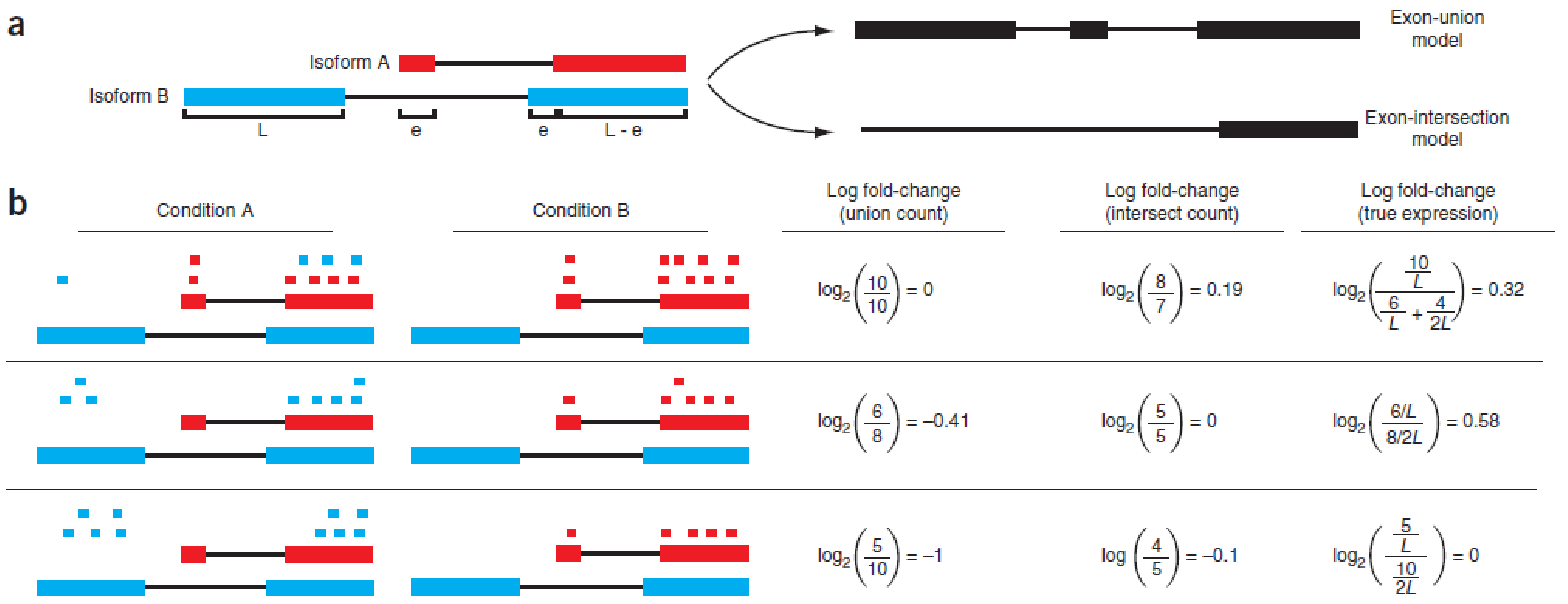

a
Isoform A
Isoform B
L
e
e
L - e
Exon-union model
Exon-intersection model
b
Condition A
Condition B
Log fold-change (union count)
Log fold-change (intersect count)
Log fold-change (true expression)
$\log_2\left(\frac{10}{10}\right) = 0$
$\log_2\left(\frac{8}{7}\right) = 0.19$
$\log_2\left(\frac{\frac{10}{L}}{\frac{6}{L}+\frac{4}{2L}}\right) = 0.32$
$\log_2\left(\frac{6}{8}\right) = –0.41$
$\log_2\left(\frac{5}{5}\right) = 0$
$\log_2\left(\frac{6/L}{8/2L}\right) = 0.58$
$\log_2\left(\frac{5}{10}\right) = –1$
$\log\left(\frac{4}{5}\right) = –0.1$
$\log_2\left(\frac{\frac{5}{L}}{\frac{10}{2L}}\right) = 0$

Figure 1 Changes in fragment count for a gene does not necessarily equal a change in expression. (**a**) Simple read-counting schemes sum the fragments incident on a gene's exons. The exon-union model counts reads falling on any of a gene's exons, whereas the exon-intersection model counts only reads on constitutive exons. (**b**) Both of the exon-union and exon-intersection counting schemes may incorrectly estimate a change in expression in genes with multiple isoforms. The true expression is estimated by the sum of the length-normalized isoform read counts. The discrepancy between a change in the union or intersection count and a change in gene expression is driven by a change in the abundance of the isoforms with respect to one another. In the top row, the gene generates the same number of reads in conditions A and B, but in condition B, all of the reads come from the shorter of the two isoforms, and thus the true expression for the gene is higher in condition B. The intersection count scheme underestimates the true change in gene expression, and the union scheme fails to detect the change entirely. In the middle row, the intersection count fails to detect a change driven by a shift in the dominant isoform for the gene. The union scheme detects a shift in the wrong direction. In the bottom row, the gene's expression is constant, but the isoforms undergo a complete switch between conditions A and B. Both simplified counting schemes register a change in count that does not reflect a change in gene expression.

Trapnell et al. 2013 Nature Biotechnology. "**Differential analysis of gene regulation at transcript resolution with RNA-seq" Cuffdiff 2**

This is wrong!
Or, at least, misleading

In the absence of errors, the intersection count will give you exactly the "true expression"

Lior's "thought experiment" (Skip this if you like)

The example (Fig. 1 in Trapnell et al, slide 46 in the GI talk) conflates differences in the **expected** values obtained from different methods (in general, raw count vs. isoform deconvolution -- the methods compared are "union count", "intersection count" and "true expression," which are their terms) with differences due to **stochastic variation** in the number of reads for different regions by putting down arbitrary counts and using them in an example. The fluctuations that they introduce into their example (presumably to make it more realistic) work in favor of their argument but are arbitrary.

They describe a hypothetical gene with two isoforms, short and long, with lengths L and $2L$.
Then they consider three cases, and two conditions per case.

In general, the reads due to any segment will be proportional to its length and the abundance of the transcripts that contain that segment. This proportionality will be affected by errors, edge effects, polymorphisms, differential representation of different regions, etc.. However, for the sake of this discussion the consideration of those factors can be deferred. Another (major) source of error comes from the statistics of (few) counts, but the expected representation can be considered independently.

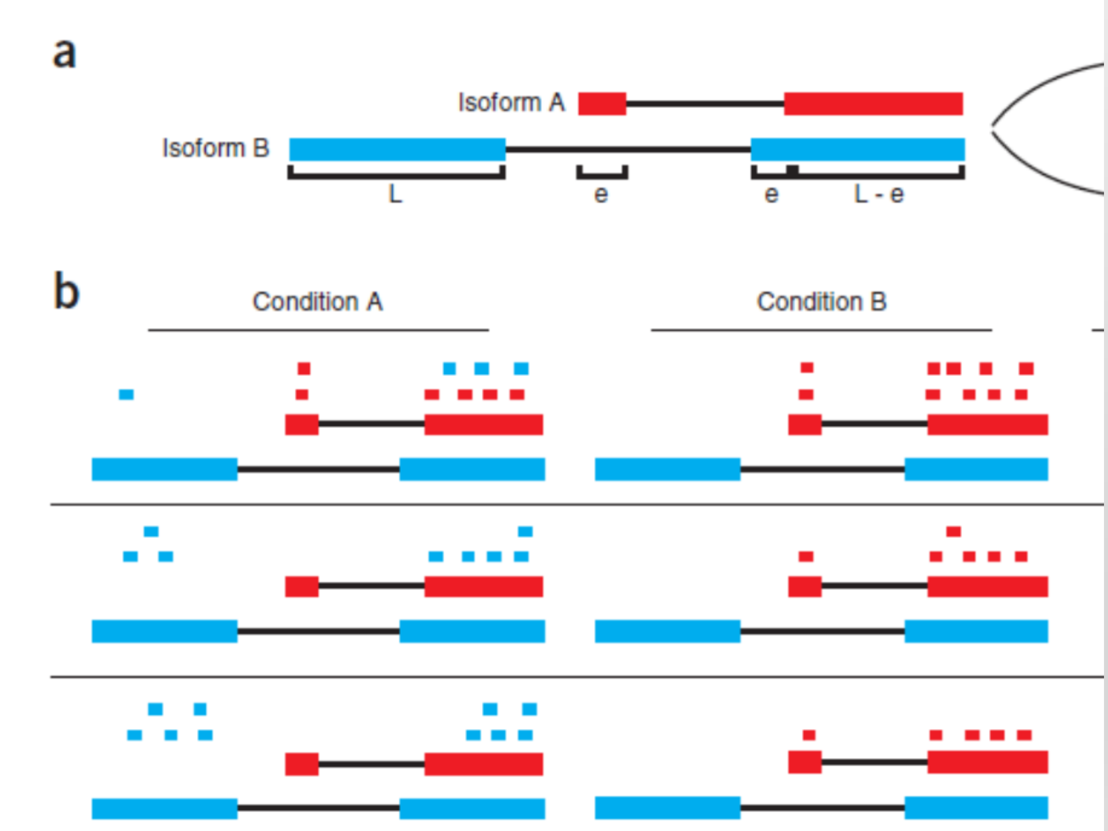
In their example, the expected count of reads from the shared exon will be proportional to its length, which is L-e, so we can say that the expected number of reads is $\mathbf{a}*(L-e)$, where $\mathbf{a}$ is some constant of proportionality.
Let f_L be the (molar) fraction of total transcripts represented by the long isoform and f_S be the (molar) fraction of total transcripts represented by the short isoform. $f_L + f_S = 1$
The expected counts of the leftmost exon unique to the long isoform (length L) will be $\mathbf{a}*f_L*L$ and for the portion of the rightmost exon unique to the long isoform, $\mathbf{a}*f_L*e$.
Together reads that come only from the long isoform will be $\mathbf{a}*f_L*(L + e)$.
Likewise, reads that come only from the small exon unique to the short isoform will be $\mathbf{a}*f_S*e$.

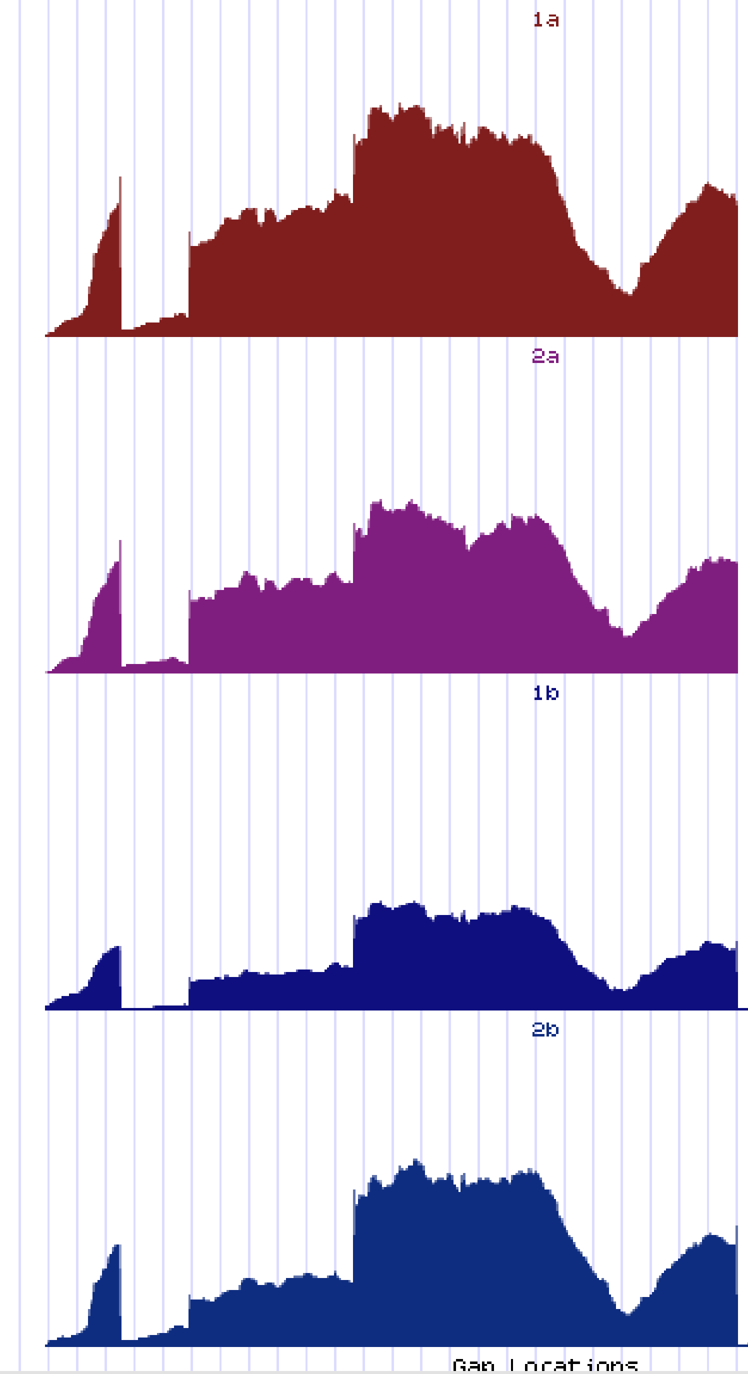
My main point is that the ratio between "intersect count" and "true expression" should be the same.
Expected reads by the intersect count will be $a * (L-e)$, and the expression level will be simply $\mathbf{a}$ after correcting for length. Expression by the isoform deconvolution method illustrated here ("true expression") will be:

$$(a * f_S * e) + (a * f_S * (L-e)) / L + ((a * f_L * (L + e)) + (a * f_L * (L-e))) / 2L = (a * f_S) + (a * f_L) = \mathbf{a}$$

These results are the same.

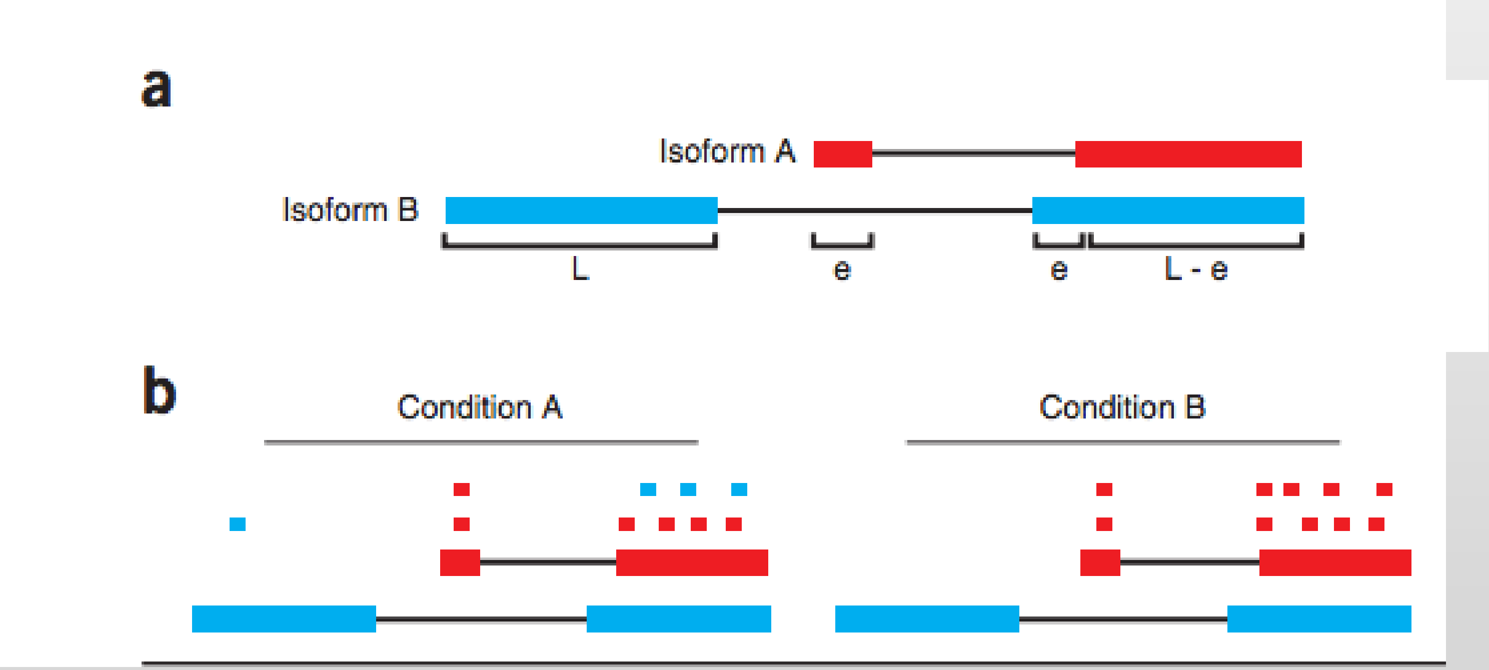
There is significant variation across an mRNA but much of that is reproducible, so comparing counts mapping to the same region across samples will give less error.

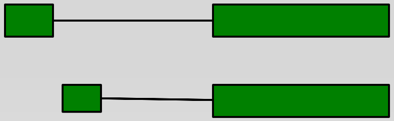




Segments

- Segments can correct for incomplete and erroneous annotations



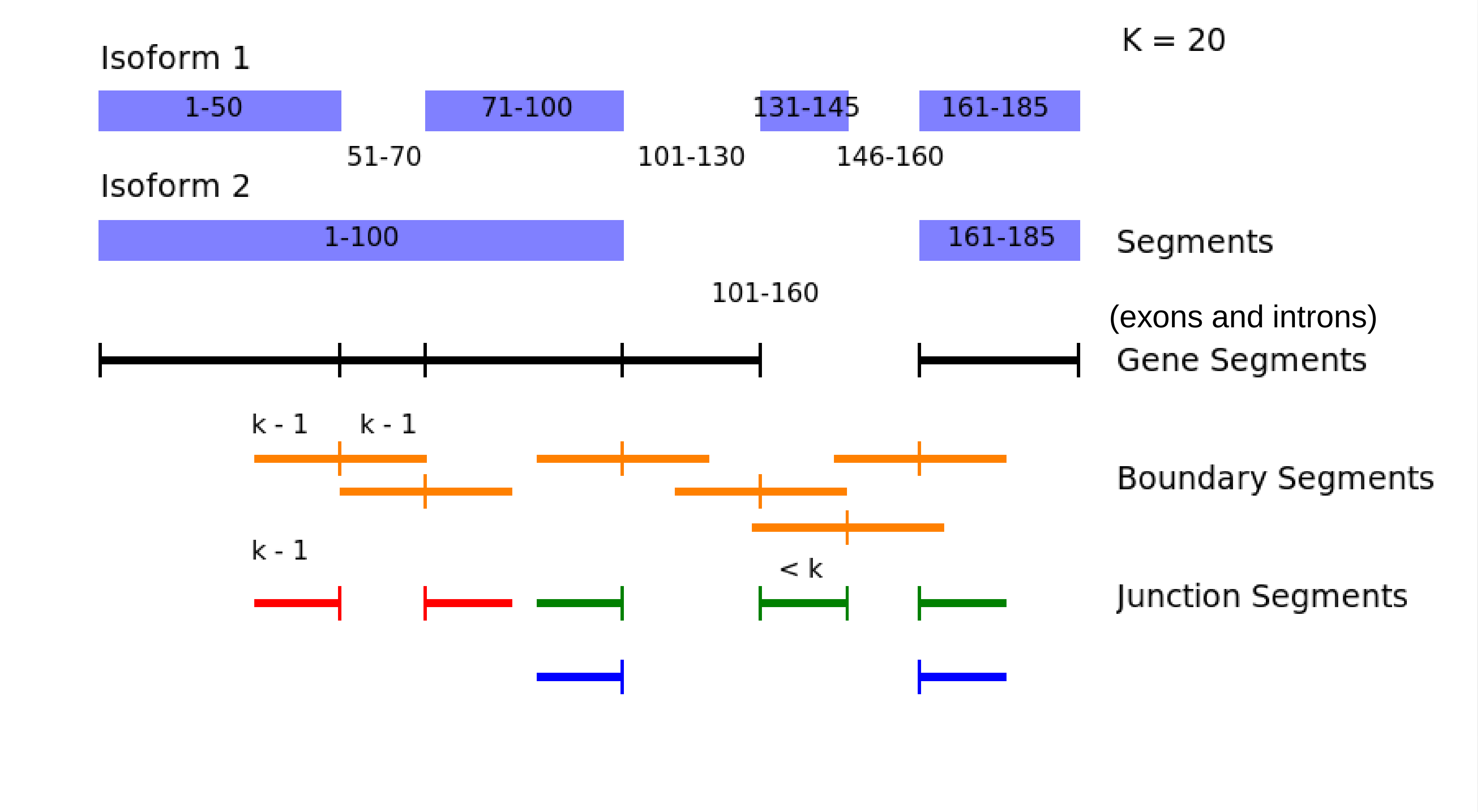


- Transcripts that are unknown to the annotation can cause errors.

Figure from Trapnell et al. 2013 Nat. Biotech.

We’re working on using gene segments with Sailfish. We will want to use KPKM (k-mers per kilobase per million reads) or KPKK.

Introns are never smaller than 25


K = 20
Isoform 1
1-50
71-100
131-145
161-185
51-70
101-130
146-160
Isoform 2
1-100
161-185
Segments
101-160
(exons and introns)
Gene Segments
k - 1
k - 1
Boundary Segments
k - 1
< k
Junction Segments