Tutorial: Additional information for gene expression analysis

BIOTECH-7005-BIOINF-3000

Zhipeng Qu

School of Biological Sciences,
The University of Adelaide

October 18th, 2022

Outline

- 1 Multiple mapping issue
- 2 Gene count normalisation
- 3 Over-representation analysis

Short reads can be mapped to multiple features (genes/transcripts)

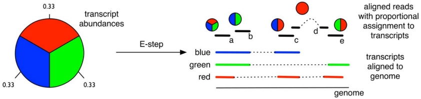
- Identical/similar sequences in different genes (e.g. gene family, repetitive elements)
- Different transcription isoforms from same gene

Species	**Aligner**	**Read length**	**multiple mapping rate (%)**
Human	STAR	PE100	4.88
Mouse	STAR	PE100	15.72
Rat	STAR	PE75	12.07
Arabidopsis	STAR	PE150	1.41
Rice	Tophat2	PE150	43.7
Soybean	Tophat2	PE150	26.4

Strategies for handling multiple mapping

- Use uniquely mapping reads only
- Simple “rescue” method. Uniformly divide each multi-mapping read to all of the positions it maps to. In other words, a read mapping to 10 positions will count as 10% of a read at each position.
- “Rescue” method using Expectation-Maximization model
  1. E-step (Expectation) Give transcript abundances, estimate the probability of each read mapping to each transcript
  2. M-step (Maximization) Update the abundances by redistributing the reads
  3. Go to step 1 (E-step) until convergence

“Rescue” method using Expectation-Maximization model


0.33
transcript abundances
0.33
0.33
E-step
aligned reads with proportional assignment to transcripts
a
b
c
d
e
blue
green
red
transcripts aligned to genome
genome

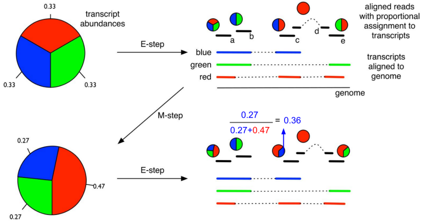
$f_{\text{blue}} = (0.33+0.5+0.5)/5 = 0.27$

$f_{\text{green}} = (0.33+0.5+0.5)/5 = 0.27$

$f_{\text{red}} = (0.33+0.5+1+0.5)/5 = 0.47$

Pachter L. Models for transcript quantification from RNA-Seq. arXiv. 2011

“Rescue” method using Expectation-Maximization model



$f_{\text{blue}} = (0.33+0.5+0.5)/5 = 0.27$

$f_{\text{green}} = (0.33+0.5+0.5)/5 = 0.27$

$f_{\text{red}} = (0.33+0.5+1+0.5)/5 = 0.47$

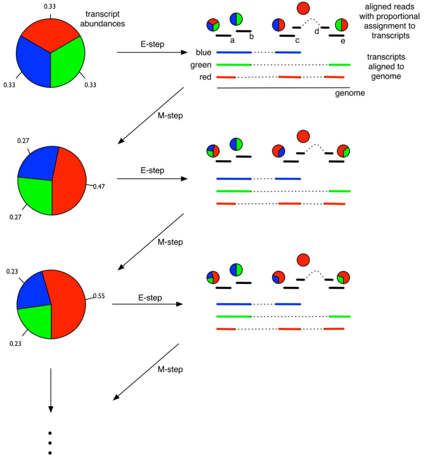
$f_{\text{blue}} = (0.27+0.5+0.36)/5 = 0.23$

$f_{\text{green}} = (0.27+0.5+0.36)/5 = 0.23$

$f_{\text{red}} = (0.47+0.64+1+0.64)/5 = 0.55$

Pachter L. Models for transcript quantification from RNA-Seq. arXiv. 2011

“Rescue” method using Expectation-Maximization model



$f_{\text{blue}} = (0.33+0.5+0.5)/5 = 0.27$

$f_{\text{green}} = (0.33+0.5+0.5)/5 = 0.27$

$f_{\text{red}} = (0.33+0.5+1+0.5)/5 = 0.47$

$f_{\text{blue}} = (0.27+0.5+0.36)/5 = 0.23$

$f_{\text{green}} = (0.27+0.5+0.36)/5 = 0.23$

$f_{\text{red}} = (0.47+0.64+1+0.64)/5 = 0.55$

...

Pachter L. Models for transcript quantification from RNA-Seq. arXiv. 2011

2 Gene count normalisation: RPKM and TPM

RNA-Seq is a relative abundance measurement of RNA expression level

- Short reads are RNA fragments randomly picked and sequenced from library
- Additional information, such as levels of “spike-in” transcripts, are required for absolute measurements
- Normalization of read count is needed to compare gene/transcript abundance
  1. RPKM/FPKM (Reads/Fragments Per Kilobase Million)
  2. TPM (Transcripts Per Million)

RPKM

Gene ID	Length	Rep1	Rep2	Rep3
A	2 kb	10	12	30
B	4 kb	20	25	60
C	1 kb	5	8	15
D	10 kb	0	0	1

We assume:

1) The genome has 4 genes
2) The RNA-Seq dataset has three replicates

https://www.youtube.com/watch?v=TTUrtCY2k-w

RPKM

Gene ID	Length	Rep1	Rep2	Rep3
A	2 kb	10	12	30
B	4 kb	20	25	60
C	1 kb	5	8	15
D	10 kb	0	0	1

Replicate 3 has much more reads than the other two replicates

[1] https://www.youtube.com/watch?v=TTUrtCY2k-w

RPKM

Gene ID	Length	Rep1	Rep2	Rep3
A	2 kb	10	12	30
B	4 kb	20	25	60
C	1 kb	5	8	15
D	10 kb	0	0	1

Gene B is twice as long as gene A, which might explain why it always gets twice as many reads

`https://www.youtube.com/watch?v=TTUrtCY2k-w`

RPKM

Gene ID	Length	Rep1	Rep2	Rep3
A	2 kb	10	12	30
B	4 kb	20	25	60
C	1 kb	5	8	15
D	10 kb	0	0	1

		Rep1	Rep2	Rep3
	Total reads:	35	45	106
“Per Million” scaling factors →	Tens of reads:	3.5	4.5	10.6

1) In this example, we scale the total read counts by 10 instead of 1,000,000
2) Million (1,000,000) was chosen just because it made the numbers look nice (Standard RNA-Seq datasets usually have multiple million reads)

https://www.youtube.com/watch?v=TTUrtCY2k-w

RPKM

Gene ID	Length	Rep1	Rep2	Rep3
A	2 kb	10	12	30
B	4 kb	20	25	60
C	1 kb	5	8	15
D	10 kb	0	0	1

Count table

Total reads:	35	45	106
Tens of reads:	3.5	4.5	10.6

“Per Million” scaling factors → Tens of reads

Gene ID	Length	Rep1	Rep2	Rep3
A	2 kb	2.86	2.67	2.83
B	4 kb	5.71	5.56	5.66
C	1 kb	1.43	1.78	1.42
D	10 kb	0	0	0.09

RPM table

https://www.youtube.com/watch?v=TTUrtCY2k-w

RPKM

Gene ID	Length	Rep1	Rep2	Rep3
A	2 kb	2.86	2.67	2.83
B	4 kb	5.71	5.56	5.66
C	1 kb	1.43	1.78	1.42
D	10 kb	0	0	0.09

RPM table

Scale Per Kilobase

Gene ID	Length	Rep1	Rep2	Rep3
A	2 kb	1.43	1.33	1.42
B	4 kb	1.43	1.39	1.42
C	1 kb	1.43	1.78	1.42
D	10 kb	0	0	0.009

RPKM table

https://www.youtube.com/watch?v=TTUrtCY2k-w

RPKM summary

Gene ID	Length	Rep1	Rep2	Rep3
A	2 kb	10	12	30
B	4 kb	20	25	60
C	1 kb	5	8	15
D	10 kb	0	0	1

Count table

Read count was:

1) Normalized for differences in sequencing depth
2) Normalized for gene length

Gene ID	Length	Rep1	Rep2	Rep3
A	2 kb	1.43	1.33	1.42
B	4 kb	1.43	1.39	1.42
C	1 kb	1.43	1.78	1.42
D	10 kb	0	0	0.009

RPKM table

`https://www.youtube.com/watch?v=TTUrtCY2k-w`

TPM

Gene ID	Length	Rep1	Rep2	Rep3
A	2 kb	10	12	30
B	4 kb	20	25	60
C	1 kb	5	8	15
D	10 kb	0	0	1

Count table

Scale Per Kilobase

Gene ID	Length	Rep1	Rep2	Rep3
A	2 kb	5	6	15
B	4 kb	5	6.25	15
C	1 kb	5	8	15
D	10 kb	0	0	0.1

RPK table

https://www.youtube.com/watch?v=TTUrtCY2k-w

TPM

Gene ID	Length	Rep1	Rep2	Rep3
A	2 kb	5	6	15
B	4 kb	5	6.25	15
C	1 kb	5	8	15
D	10 kb	0	0	0.1

RPK table

Total reads:	15	20.25	45.1
Tens of reads:	1.5	2.025	4.51

“Per Million” scaling factors →

In this example, we scale the total read counts by 10 instead of 1,000,000

https://www.youtube.com/watch?v=TTUrtCY2k-w

TPM

Gene ID	Length	Rep1	Rep2	Rep3
A	2 kb	5	6	15
B	4 kb	5	6.25	15
C	1 kb	5	8	15
D	10 kb	0	0	0.1

RPK table

Total reads:	15	20.25	45.1
Tens of reads:	1.5	2.025	4.51

“Per Million” scaling factors →

Gene ID	Length	Rep1	Rep2	Rep3
A	2 kb	3.33	2.96	3.326
B	4 kb	3.33	3.09	3.326
C	1 kb	3.33	3.95	3.326
D	10 kb	0	0	0.02

TPM table

https://www.youtube.com/watch?v=TTUrtCY2k-w

TPM summary

Gene ID	Length	Rep1	Rep2	Rep3
A	2 kb	10	12	30
B	4 kb	20	25	60
C	1 kb	5	8	15
D	10 kb	0	0	1

Count table

Read count was:

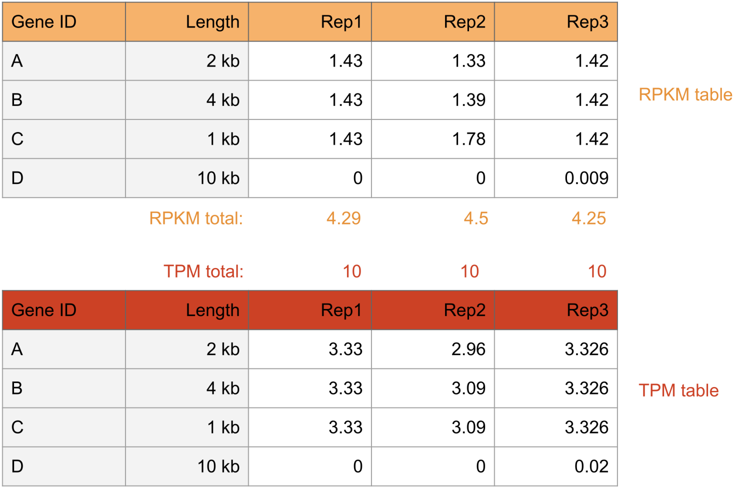
1) Normalized for **gene length**
2) Normalized for **differences in sequencing depth**

Gene ID	Length	Rep1	Rep2	Rep3
A	2 kb	3.33	2.96	3.326
B	4 kb	3.33	3.09	3.326
C	1 kb	3.33	3.09	3.326
D	10 kb	0	0	0.02

TPM table

`https://www.youtube.com/watch?v=TTUrtCY2k-w`

RPKM vs TPM



Gene ID	Length	Rep1	Rep2	Rep3
A	2 kb	1.43	1.33	1.42
B	4 kb	1.43	1.39	1.42
C	1 kb	1.43	1.78	1.42
D	10 kb	0	0	0.009

RPKM table

RPKM total: 4.29 4.5 4.25

TPM total: 10 10 10

Gene ID	Length	Rep1	Rep2	Rep3
A	2 kb	3.33	2.96	3.326
B	4 kb	3.33	3.09	3.326
C	1 kb	3.33	3.09	3.326
D	10 kb	0	0	0.02

TPM table

https://www.youtube.com/watch?v=TTUrtCY2k-w

RPKM vs TPM

Gene ID	Length	Rep1	Rep2	Rep3
A	2 kb	1.43	1.33	1.42
B	4 kb	1.43	1.39	1.42
C	1 kb	1.43	1.78	1.42
D	10 kb	0	0	0.009

RPKM table

RPKM total: 4.29 4.5 4.25

TPM total: 10 10 10

Gene ID	Length	Rep1	Rep2	Rep3
A	2 kb	3.33	2.96	3.326
B	4 kb	3.33	3.09	3.326
C	1 kb	3.33	3.09	3.326
D	10 kb	0	0	0.02

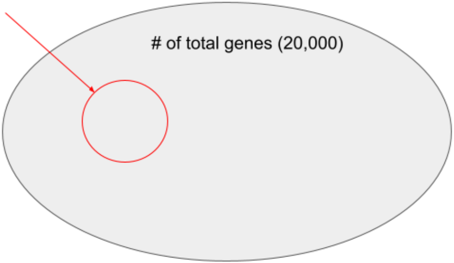
TPM table

https://www.youtube.com/watch?v=TTUrtCY2k-w

3 Over-representation analysis

Over-representation analysis

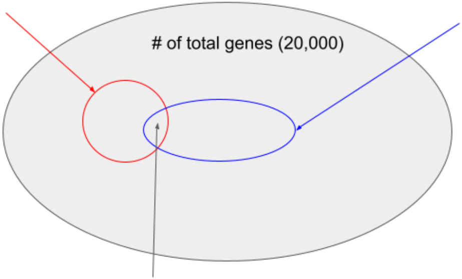
Genes in pathway “cell cycle” (20)
Proportion in # total genes: 20/20000 = 0.001



# of total genes (20,000)

Over-representation analysis

Genes in pathway “cell cycle” (20)
Proportion in # total genes: 20/20000 = 0.001



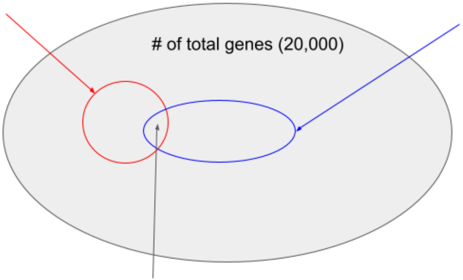
# of total genes (20,000)

DE genes (100)

DE genes in pathway “cell cycle”: 10
Proportion in #DE genes: 10/100=0.1 >> 0.001

Over-representation analysis

Genes in pathway “cell cycle” (20)
Proportion in # total genes: 20/20000 = 0.001



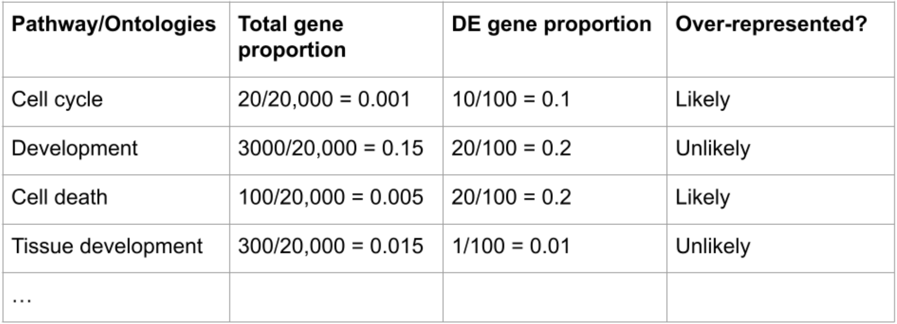
# of total genes (20,000)

DE genes (100)

DE genes in pathway “cell cycle”: 10
Proportion in #DE genes: 10/100=0.1 >> 0.001

What we can conclude: It is more likely that expression of genes in pathway “cell cycle” were perturbed between comparisons. We say “cell cycle” genes were over-represented in DE genes

Over-representation analysis



Pathway/Ontologies	Total gene proportion	DE gene proportion	Over-represented?
Cell cycle	20/20,000 = 0.001	10/100 = 0.1	Likely
Development	3000/20,000 = 0.15	20/100 = 0.2	Unlikely
Cell death	100/20,000 = 0.005	20/100 = 0.2	Likely
Tissue development	300/20,000 = 0.015	1/100 = 0.01	Unlikely
…			

We can use statistical test, such as Hypergeometric test to determine the statistical significance of this kind of over-representation analysis.

Thank you!