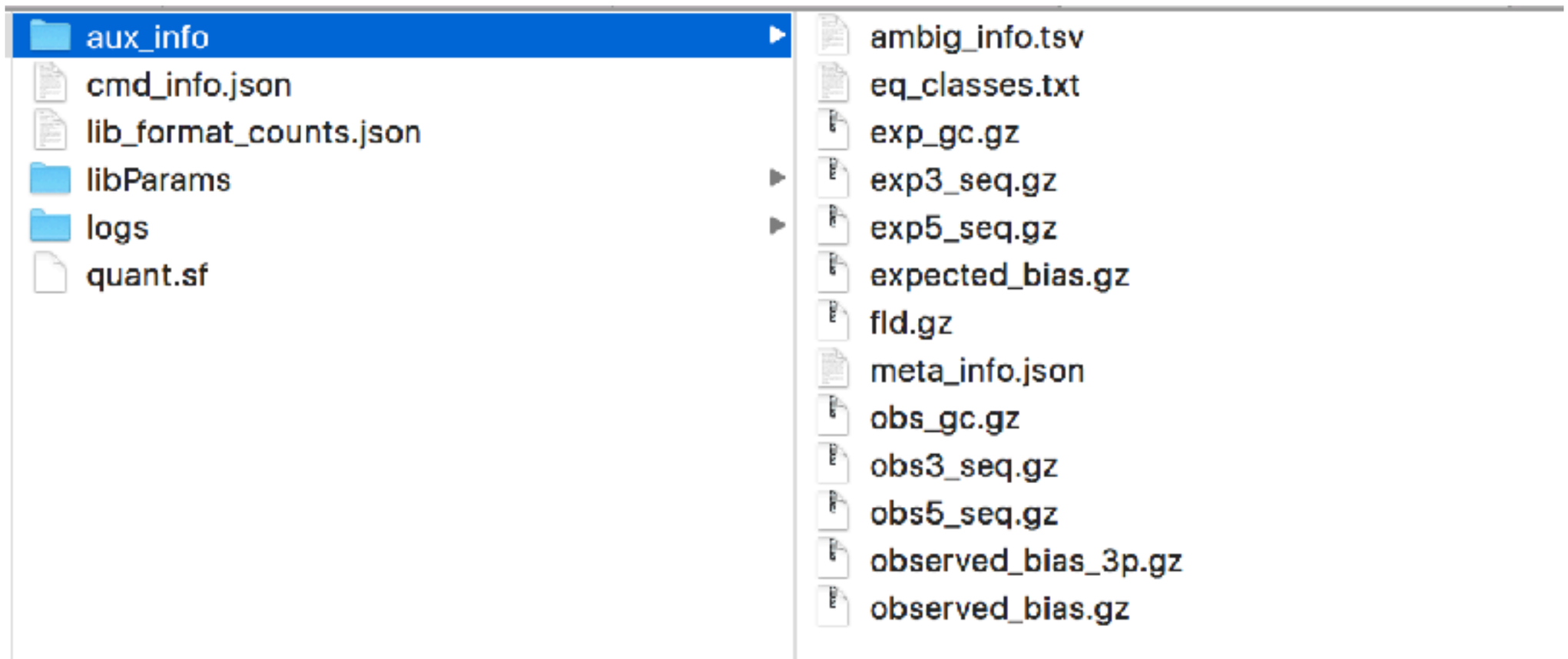
RNA-seq abundance units and import into R

Charlotte Soneson
CSAMA, Brixen, July 10 2018

Salmon output (per sample)



How to get the output of Salmon into R?

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Search:

Home | Install | Help | Developers | About

Home » Bioconductor 3.7 » Software Packages » tximport

tximport

platforms all | downloads top 5% | posts 28 / 1 / 3 / 7 | in Bioc 2 years
build ok

DOI: 10.18129/B9.bioc.tximport

Import and summarize transcript-level estimates for transcript- and gene-level analysis

Bioconductor version: Release (3.7)

Imports transcript-level abundance, estimated counts and transcript lengths, and summarizes into matrices for use with downstream gene-level analysis packages. Average transcript length, weighted by sample-specific transcript abundance estimates, is provided as a matrix which can be used as an offset for different expression of gene-level counts.

Author: Michael Love [cre,aut], Charlotte Soneson [aut], Mark Robinson [aut], Rob Patro [ctb], Andrew Parker Morgan [ctb], Ryan C. Thompson [ctb], Matt Shirley [ctb]

Maintainer: Michael Love <michaelisaiahlove at gmail.com>

Citation (from within R, enter `citation("tximport")`):

> Soneson C, Love MI, Robinson MD (2015). "Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences." *F1000Research*, **4**. doi: 10.12688/f1000research.7563.1.

Installation

To install this package, start R and enter:

Documentation »

Bioconductor

- Package vignettes and manuals.
- Workflows for learning and use.
- Course and conference material.
- Videos.
- Community resources and tutorials.

R / CRAN packages and documentation

Support »

Please read the posting guide. Post questions about Bioconductor to one of the following locations:

- Support site - for questions about Bioconductor packages
- Bioc-devel mailing list - for package developers

How to get the output of Salmon into R?

```
> library(tximport)
> salmon_files
                    SRR1039508                                SRR1039509
"salmon/SRR1039508/quant.sf" "salmon/SRR1039509/quant.sf"
                    SRR1039512                                SRR1039513
"salmon/SRR1039512/quant.sf" "salmon/SRR1039513/quant.sf"
                    SRR1039516                                SRR1039517
"salmon/SRR1039516/quant.sf" "salmon/SRR1039517/quant.sf"
                    SRR1039520                                SRR1039521
"salmon/SRR1039520/quant.sf" "salmon/SRR1039521/quant.sf"
> head(tx2gene)
                tx            gene
1 ENST00000415118 ENSG00000223997
2 ENST00000434970 ENSG00000237235
3 ENST00000448914 ENSG00000228985
4 ENST00000604642 ENSG00000270961
5 ENST00000603326 ENSG00000271317
6 ENST00000604950 ENSG00000270783
```

How to get the Salmon output into R?

Transcript level counts:

```
> txi <- tximport(files = salmon_files, type = "salmon", txOut = TRUE)
reading in files with read_tsv
1 2 3 4 5 6 7 8
> names(txi)
[1] "abundance"           "counts"              "length"
[4] "countsFromAbundance"
> dim(txi$counts)
[1] 180253      8
> head(txi$counts)
                 SRR1039508 SRR1039509 SRR1039512 SRR1039513 SRR1039516
ENST00000415118          0          0          0          0          0
ENST00000434970          0          0          0          0          0
ENST00000448914          0          0          0          0          0
ENST00000604642          0          0          0          0          0
ENST00000603326          0          0          0          0          0
ENST00000604950          0          0          0          0          0
                 SRR1039517 SRR1039520 SRR1039521
ENST00000415118          0          0          0
ENST00000434970          0          0          0
ENST00000448914          0          0          0
ENST00000604642          0          0          0
ENST00000603326          0          0          0
ENST00000604950          0          0          0
```

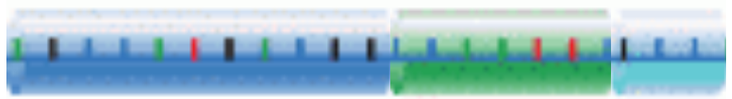
How to get the Salmon output into R?

Gene level counts:

```
> txi <- tximport(files = salmon_files, type = "salmon", txOut = FALSE, tx2gene = tx
2gene)
reading in files with read_tsv
1 2 3 4 5 6 7 8
summarizing abundance
summarizing counts
summarizing length
> names(txi)
[1] "abundance"           "counts"              "length"
[4] "countsFromAbundance"
> dim(txi$counts)
[1] 39293     8
> head(txi$counts)
                SRR1039508 SRR1039509 SRR1039512 SRR1039513 SRR1039516
ENSG00000000003  698.49149  463.02512  895.68652  420.45024  1154.6804
ENSG00000000005    0.00000    0.00000    0.00000    0.00000     0.0000
ENSG00000000419  465.99976  515.59630  625.00020  365.68360   590.0994
ENSG00000000457  334.62080  287.17914  385.61940  230.77805   377.4769
ENSG00000000460   99.96656   82.14327   74.90531   56.78863   128.8831
ENSG00000000938    0.00000    0.00000    2.00000    0.00000     1.0000
                SRR1039517 SRR1039520 SRR1039521
ENSG00000000003  1078.4641   780.3976  589.22033
ENSG00000000005     0.0000     0.0000    0.00000
ENSG00000000419   797.9870   419.6755  510.91963
ENSG00000000457   473.0888   321.4662  301.07969
ENSG00000000460   128.6009   107.5006   90.86691
ENSG00000000938     0.0000     0.0000    0.00000
```

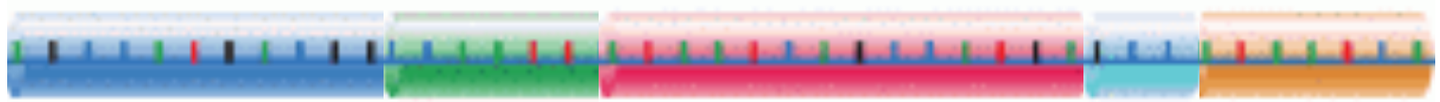
Impact of isoform composition on gene-level counts

T1



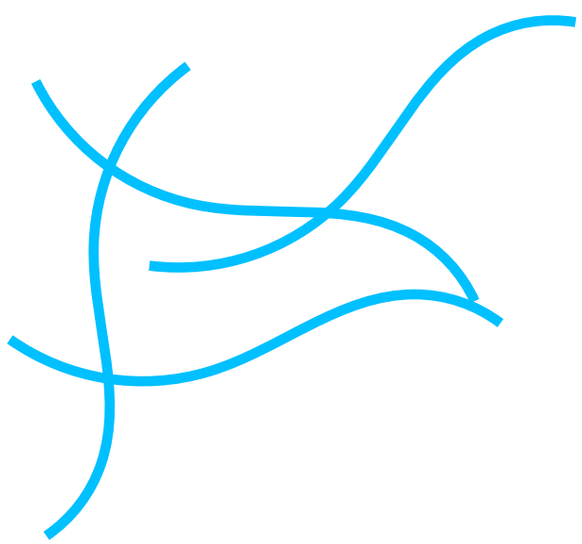
length = **L**

T2

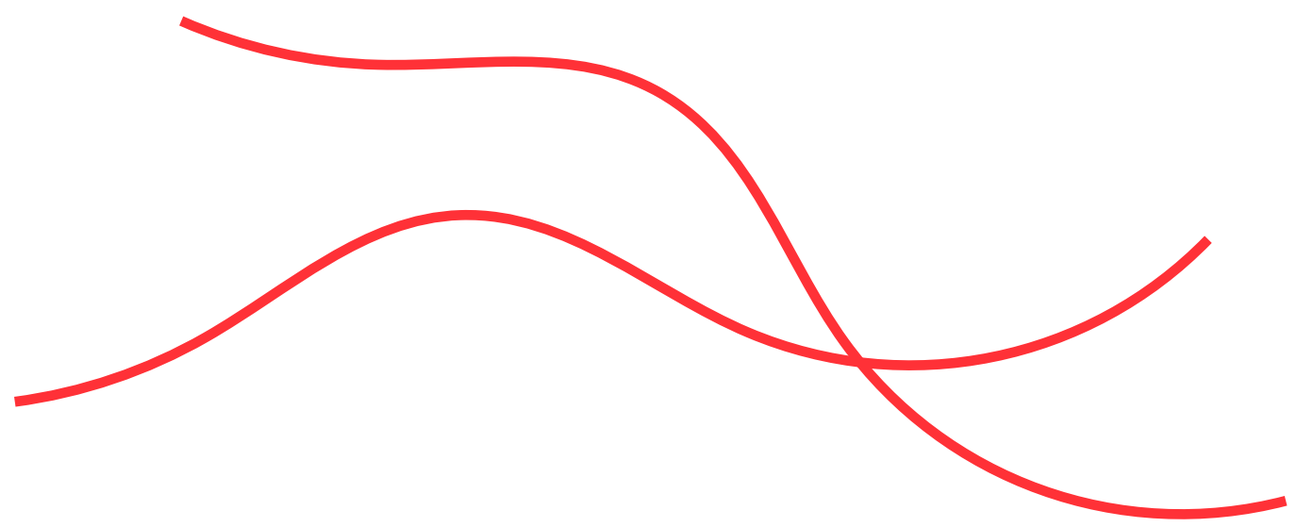


length = **2L**

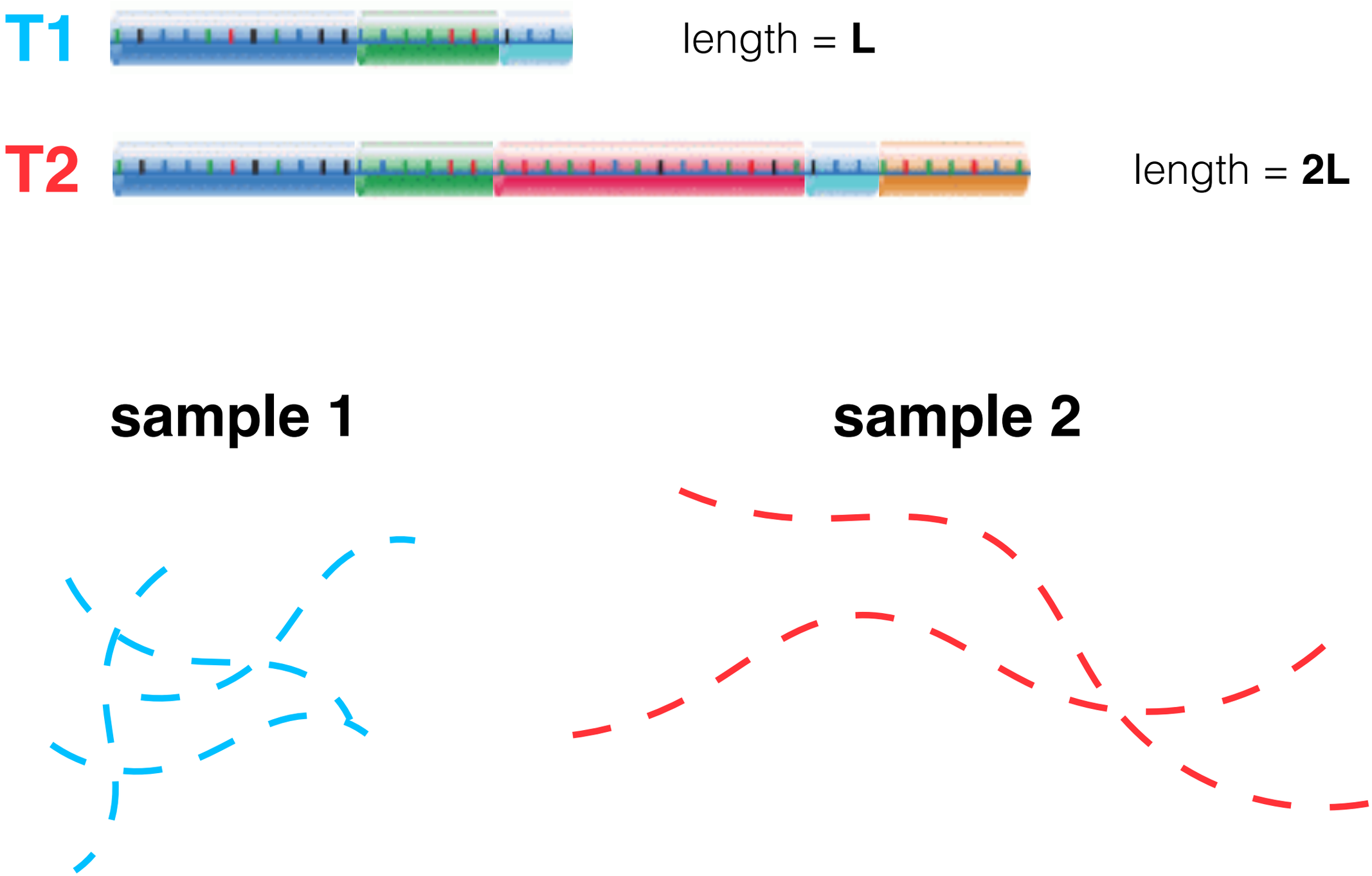
sample 1



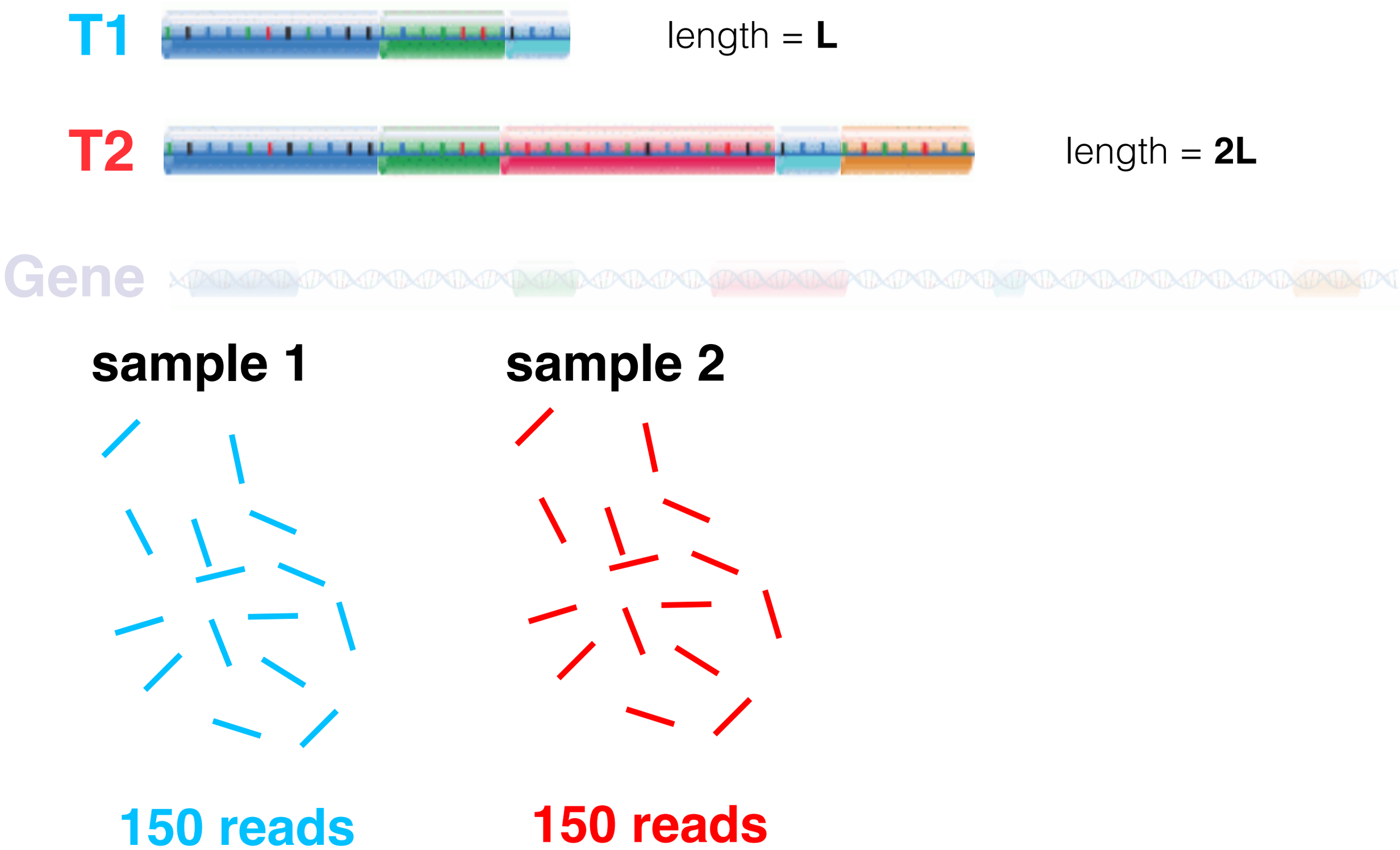
sample 2



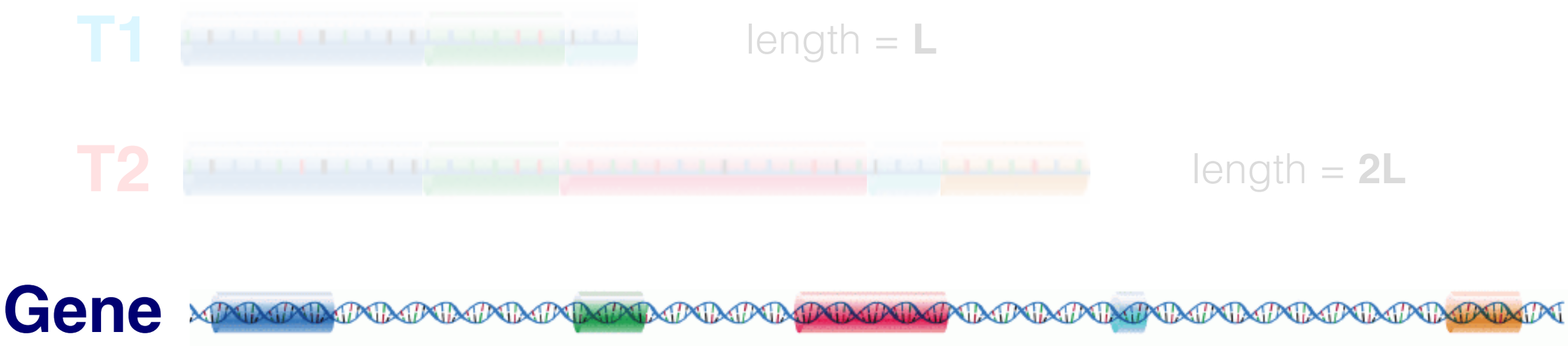
Impact of isoform composition on gene-level counts



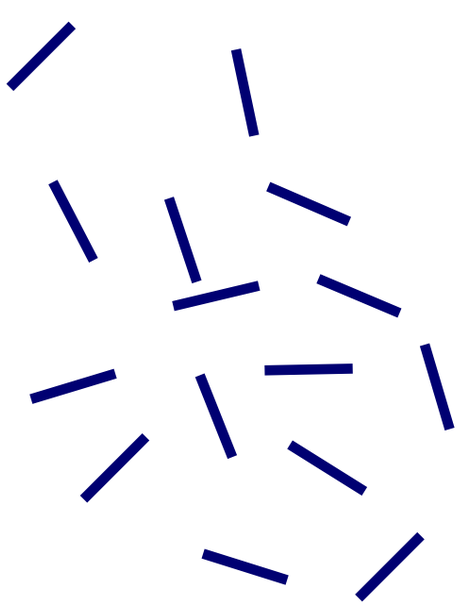
Impact of isoform composition on gene-level counts



Impact of isoform composition on gene-level counts

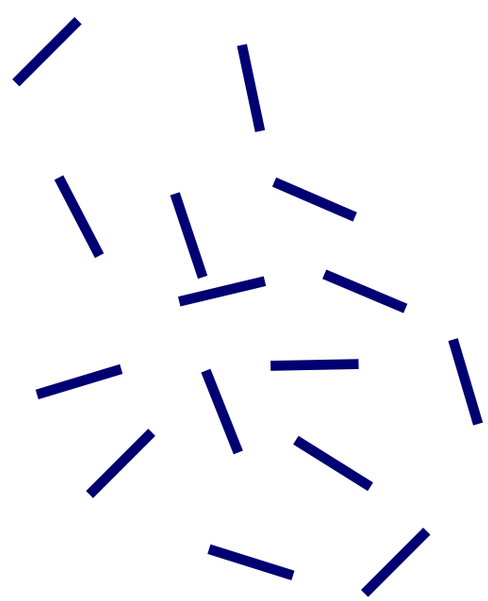


sample 1



150 reads

sample 2

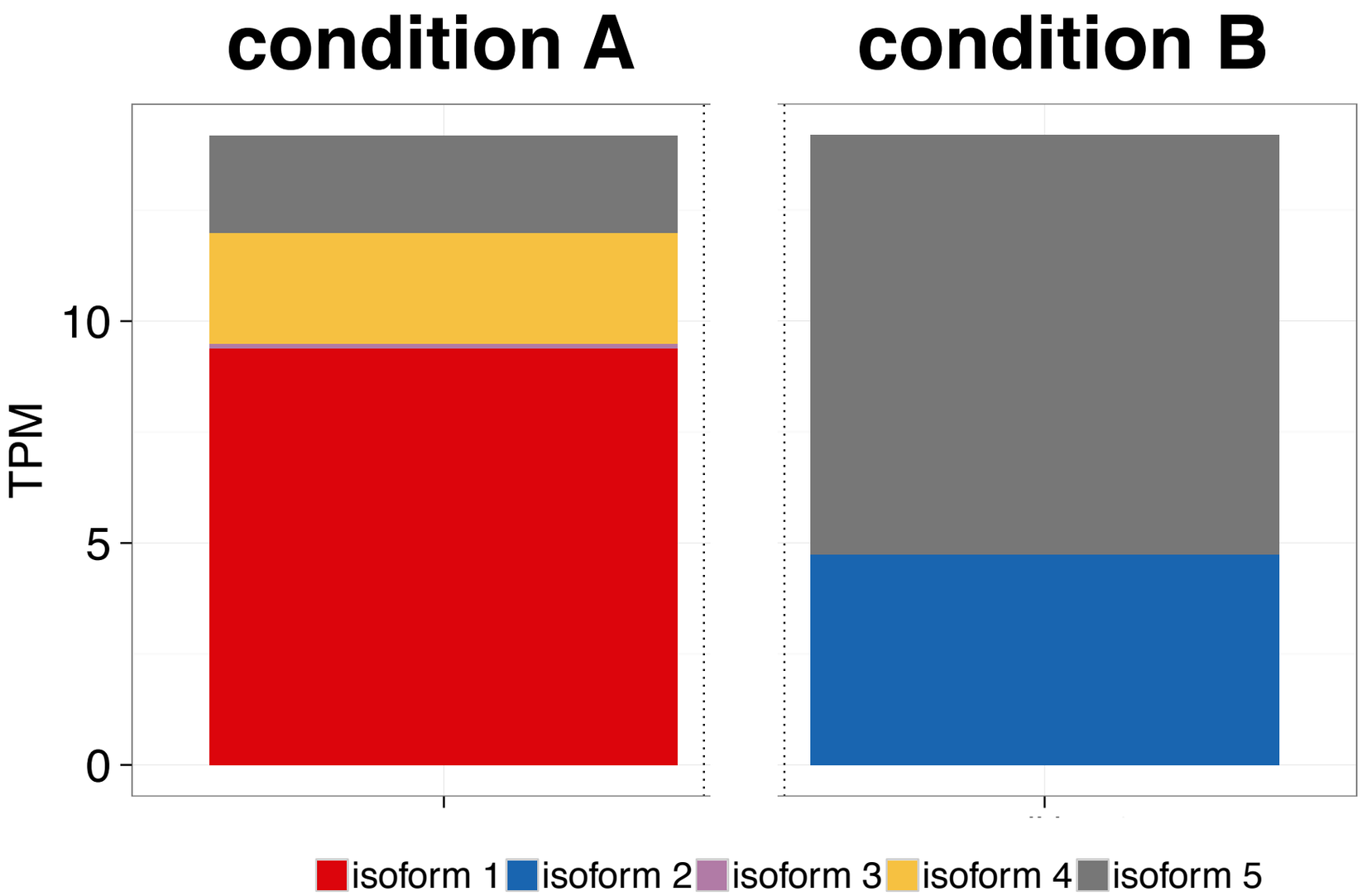


150 reads

Gene	S1	S2
Count	150	150

Impact of differential isoform usage on gene-level counts

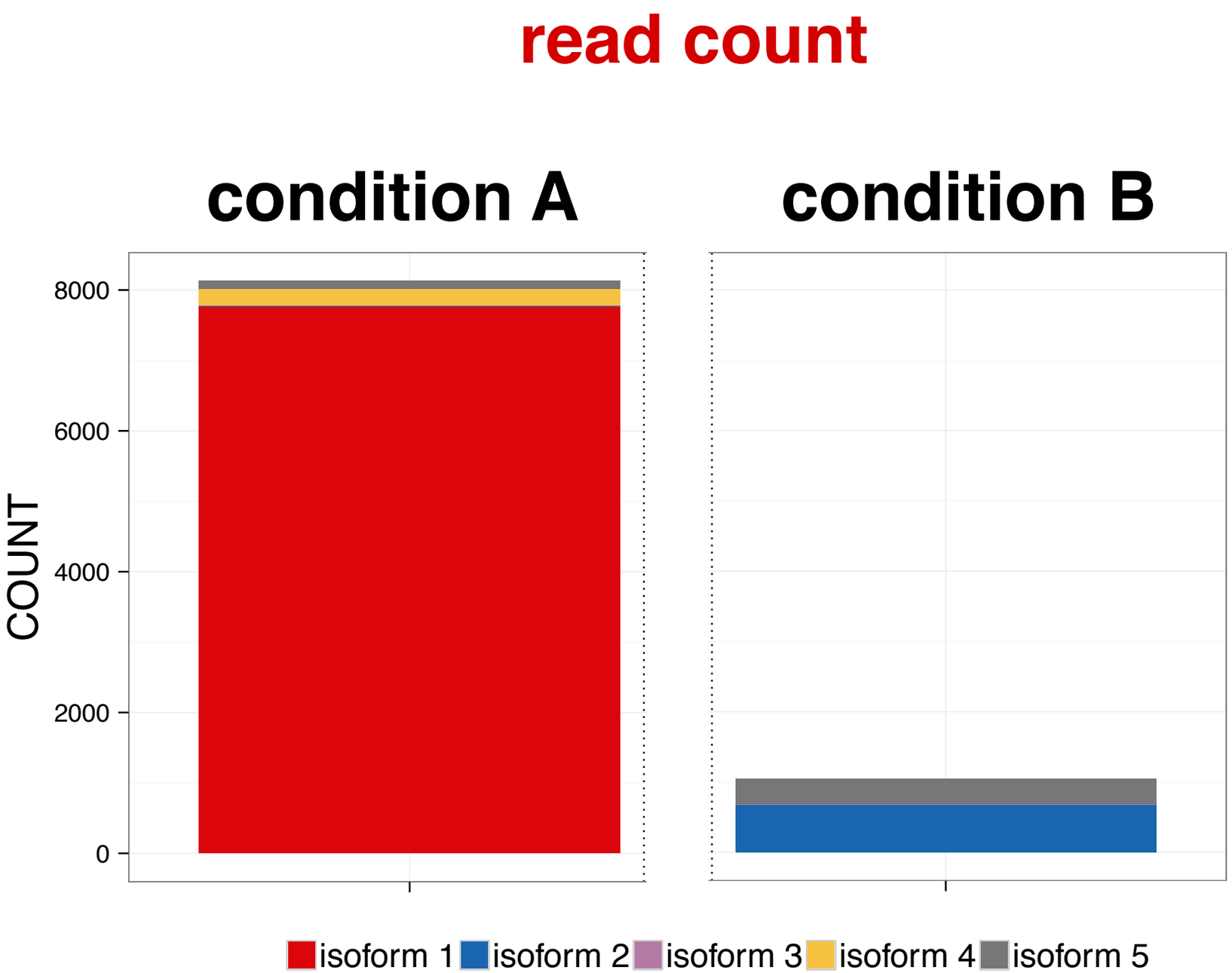
true abundance



Lengths:

isoform 1:	12'232 bp
isoform 2:	1'733 bp
isoform 3:	891 bp
isoform 4:	1'404 bp
isoform 5:	543 bp

Impact of differential isoform usage on gene-level counts



Lengths:

isoform 1:	12'232 bp
isoform 2:	1'733 bp
isoform 3:	891 bp
isoform 4:	1'404 bp
isoform 5:	543 bp

The isoform composition affects the observed read count for a gene

Differential isoform usage* can lead to false positives and false negatives in differential **gene** expression analyses

*differences in isoform composition between groups

What can we do?

- Consider another abundance unit that better reflects the underlying abundances (“number of transcript molecules”)
- Include “adjustment” of gene counts in the statistical model to reflect underlying isoform composition

What can we do?

How can we get such values?

- Consider another abundance unit that better reflects the underlying abundances ("number of transcript molecules")

- Include "adjustment" of gene counts in the statistical model to reflect underlying isoform composition

How could such adjustment be done?

Abundance units

read count for transcript *i*

c_i

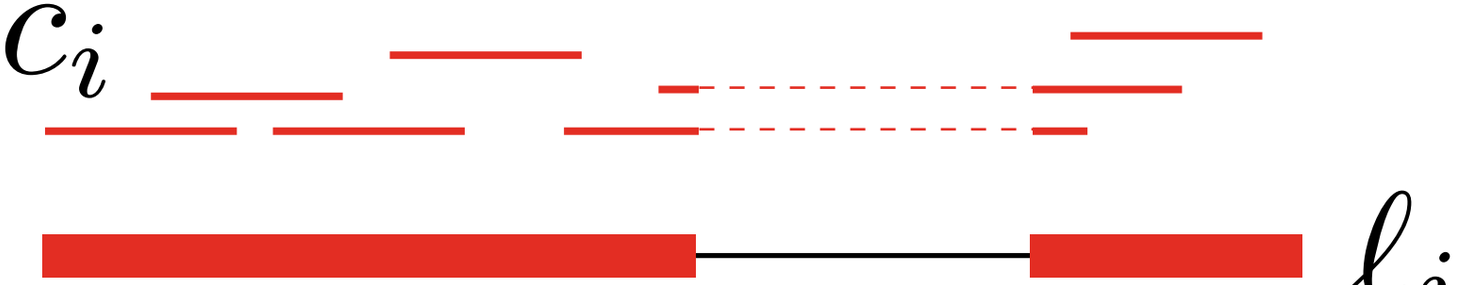
ℓ_i

length of transcript *i*

Abundance units

read count for transcript *i*

c_i



ℓ_i

fragment length

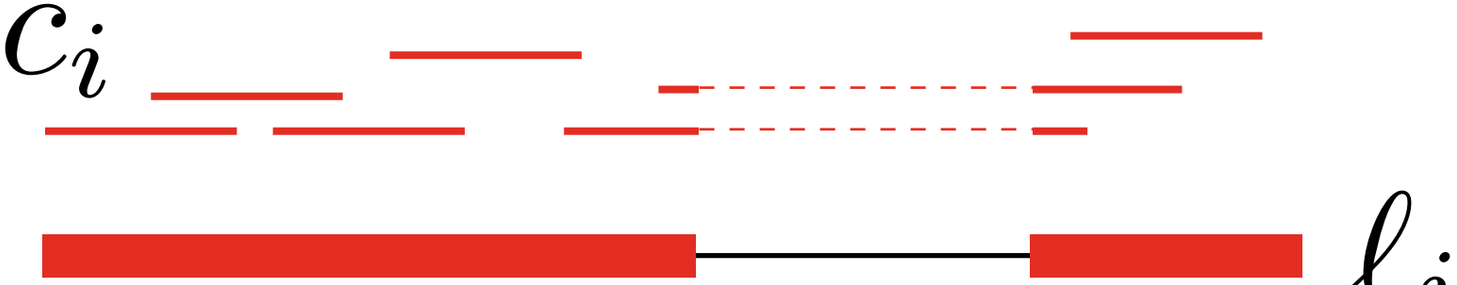
length of transcript *i*

$$t_i = \frac{c_i r}{\ell_i}$$

Abundance units

read count for transcript *i*

c_i



ℓ_i

fragment length

length of transcript *i*

$$t_i = \frac{c_i r}{\ell_i}$$

$$TPM_i = 10^6 \cdot \frac{t_i}{\sum_k t_k}$$

Abundance units

read count for transcript *i*

c_i

ℓ_i

fragment length

length of transcript *i*

$$t_i = \frac{c_i r}{\ell_i}$$

$$TPM_i = 10^6 \cdot \frac{t_i}{\sum_k t_k}$$

library size

$$RPKM_i = 10^9 \cdot \frac{c_i}{\ell_i \boxed{\sum_k c_k}} = 10^9 \cdot \frac{t_i}{\sum_k (t_k \ell_k)}$$

Abundance units

read count for transcript *i*

c_i

ℓ_i

length of transcript *i*

fragment length

$$t_i = \frac{c_i r}{\ell_i}$$

$$TPM_i \propto RPKM_i$$

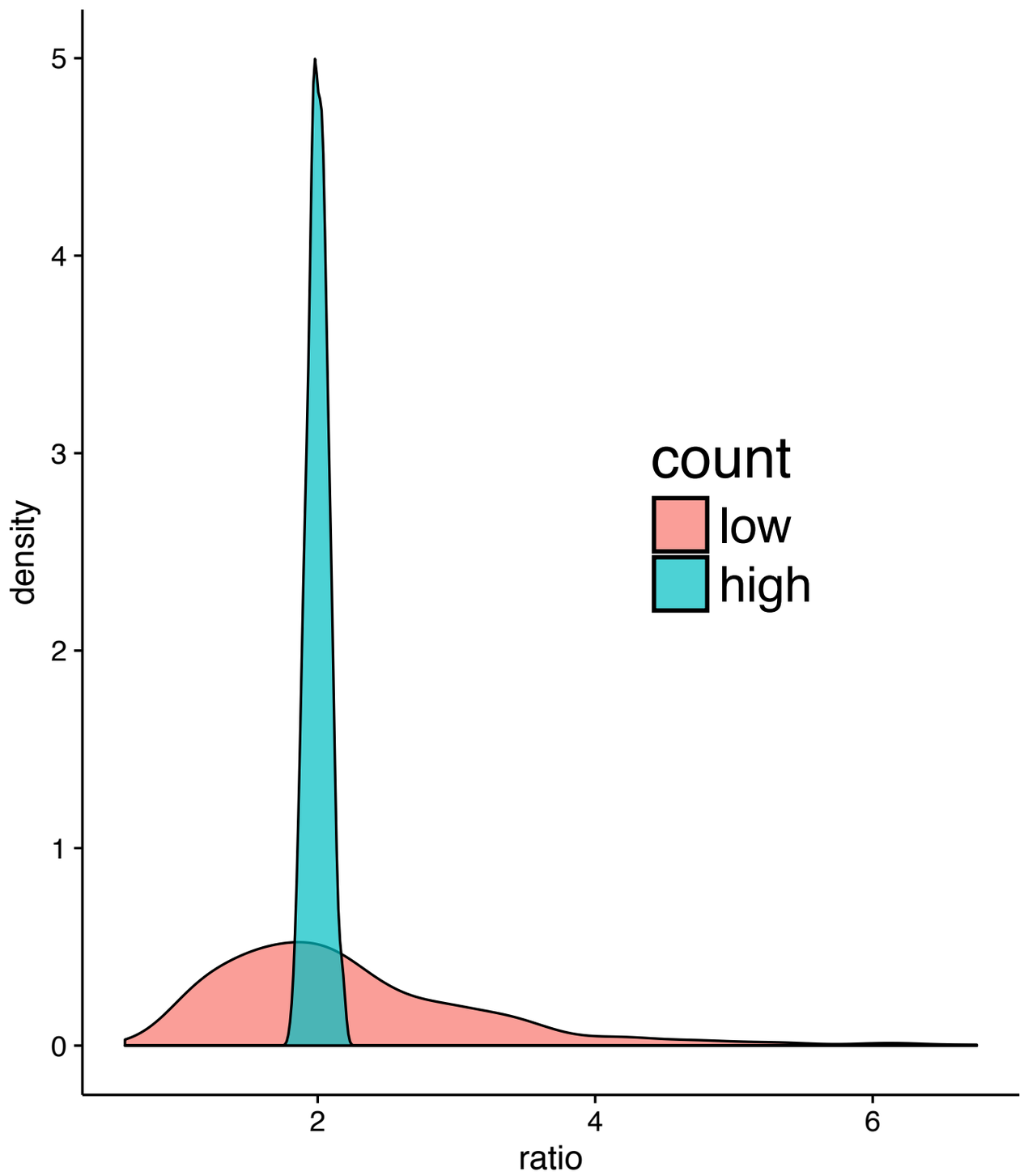
$$\sum_i TPM_i = 10^6$$

$$TPM_i = 10^6 \cdot \frac{t_i}{\sum_k t_k}$$

library size

$$RPKM_i = 10^9 \cdot \frac{c_i}{\ell_i \boxed{\sum_k c_k}} = 10^9 \cdot \frac{t_i}{\sum_k (t_k \ell_k)}$$

Why not only relative abundances?



- Ex: ratio between two Poisson distributed variables
- Low:
  mean = 20 vs
  mean = 10
- High:
  mean = 2000 vs
  mean = 1000

Getting TPM estimates with tximport

```
> txi <- tximport(files = salmon_files, type = "salmon", txOut = FALSE, tx2gene = tx
2gene)
reading in files with read_tsv
1 2 3 4 5 6 7 8
summarizing abundance
summarizing counts
summarizing length
> head(txi$abundance)
                 SRR1039508 SRR1039509 SRR1039512 SRR1039513 SRR1039516
ENSG00000000003  26.951817  19.629238 28.3308210  23.246920  36.716880
ENSG00000000005   0.000000   0.000000  0.0000000   0.000000   0.000000
ENSG00000000419  38.518878  46.108530 42.3467400  43.380940  40.212570
ENSG00000000457   7.705755   6.049874  6.6344630   7.253561   5.883886
ENSG00000000460   2.697636   2.456616  1.7890904   2.576508   2.941266
ENSG00000000938   0.000000   0.000000  0.0673626   0.000000   0.121150
                 SRR1039517 SRR1039520 SRR1039521
ENSG00000000003  29.094257  34.831930  24.209444
ENSG00000000005   0.000000   0.000000   0.000000
ENSG00000000419  45.723287  39.296450  44.809122
ENSG00000000457   7.090778   8.341518   7.700865
ENSG00000000460   2.168365   3.593189   2.193913
ENSG00000000938   0.000000   0.000000   0.000000
```