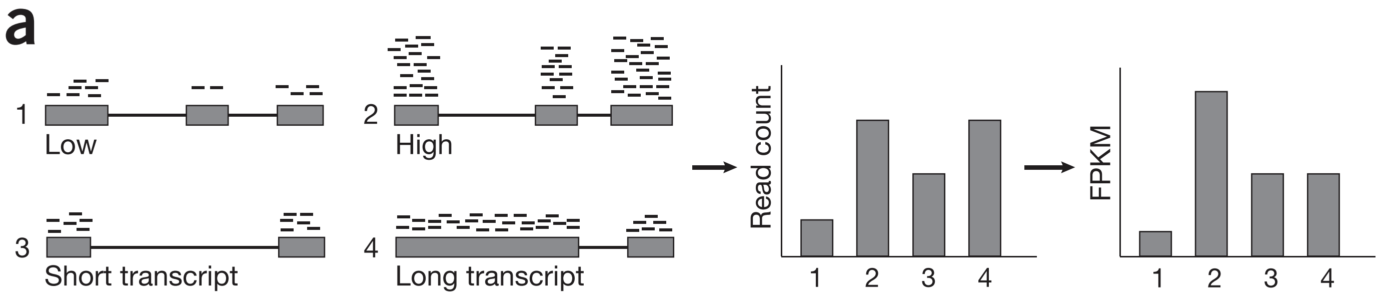
Differential Expressed Genes – Normalization

RNA-seq to estimate gene expression, read counts need to be properly normalized to extract meaningful expression estimates

There are two main sources of systematic variability that require normalization.

1. RNA fragmentation during library construction causes longer transcripts to generate more reads compared to shorter transcripts present at the same abundance in the sample.
2. The variability in the number of reads produced for each run causes fluctuations in the number of fragments mapped across samples



To account for these issues, the reads per kilobase of transcript per million mapped reads (RPKM) metric normalizes a transcript’s read count by both its length and the total number of mapped reads in the sample When data originate from paired-end sequencing, the analogous fragments per kilo- base of transcript per million mapped reads (FPKM)

There’s a new RNA-seq metric on the block…

- We used to report RPKM (Reads Per Kilobase Million) or FPKM (Fragments Per Kilobase Million)
    - These normalized read counts for:
        - 1) The sequencing depth (that’s the “Million” part)
            - Sequencing runs with more depth will have more reads mapping to each gene.
        - 2) The length of the gene (that’s the “Kilobase” part)
            - Longer genes will have more reads mapping to them.
- Now they want us to use TPM – Transcripts per million

To understand the differences between TPM and RPKM and FPKM, we'll work through the math using an imaginary RNA-seq data with three replicates (Rep1, 2 and 3) for a genome with 4 genes (A, B, C and D).

Gene Name	Rep1 Counts	Rep2 Counts	Rep3 Counts
A (2kb)	10	12	30
B (4kb)	20	25	60
C (1kb)	5	8	15
D (10kb)	0	0	1

Gene Name	Rep1 Counts	Rep2 Counts	Rep3 Counts
A (2kb)	10	12	30
B (4kb)	20	25	60
C (1kb)	5	8	15
D (10kb)	0	0	1

Rep3 has way more reads than the other replicates, regardless of the gene.

Gene B is twice as long as gene A, and that might explain why it always gets twice as many reads, regardless of replicate.

Gene Name	Rep1 Counts	Rep2 Counts	Rep3 Counts
A (2kb)	10	12	30
B (4kb)	20	25	60
C (1kb)	5	8	15
D (10kb)	0	0	1

RPKM – step 1: normalize for read depth.

Gene Name	Rep1 Counts	Rep2 Counts	Rep3 Counts
A (2kb)	10	12	30
B (4kb)	20	25	60
C (1kb)	5	8	15
D (10kb)	0	0	1
Total reads:	35	45	106
Tens of reads:	3.5	4.5	10.6

For the purpose of this 4 gene example, we're scaling the total read counts by 10 instead of 1,000,000.

Originally, 1,000,000 was picked just because it made the numbers look nice (i.e. they didn't require too many decimal places)

RPKM – step 1: normalize for read depth.

Gene Name	Rep1 Counts	Rep2 Counts	Rep3 Counts
A (2kb)	10	12	30
B (4kb)	20	25	60
C (1kb)	5	8	15
D (10kb)	0	0	1
Total reads:	35	45	106
Tens of reads:	3.5	4.5	10.6

RPM - scaled using the "per million" factors.

Gene Name	Rep1 RPM	Rep2 RPM	Rep3 RPM
A (2kb)	2.86	2.67	2.83
B (4kb)	5.71	5.56	5.66
C (1kb)	1.43	1.78	1.43
D (10kb)	0	0	0.09

RPKM – step 2: normalize for gene length.

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

Reads are scaled for depth (M) and gene length (K).

Gene Name	Rep1 RPKM	Rep2 RPKM	Rep3 RPKM
A (2kb)	1.43	1.33	1.42
B (4kb)	1.43	1.39	1.42
C (1kb)	1.43	1.78	1.42
D (10kb)	0	0	0.009

RPKM Summary

BEFORE

Gene Name	Rep1 Counts	Rep2 Counts	Rep3 Counts
A (2kb)	10	12	30
B (4kb)	20	25	60
C (1kb)	5	8	15
D (10kb)	0	0	1

Read counts were…

1) Normalized for differences in sequencing depth.
2) Normalized for gene size.

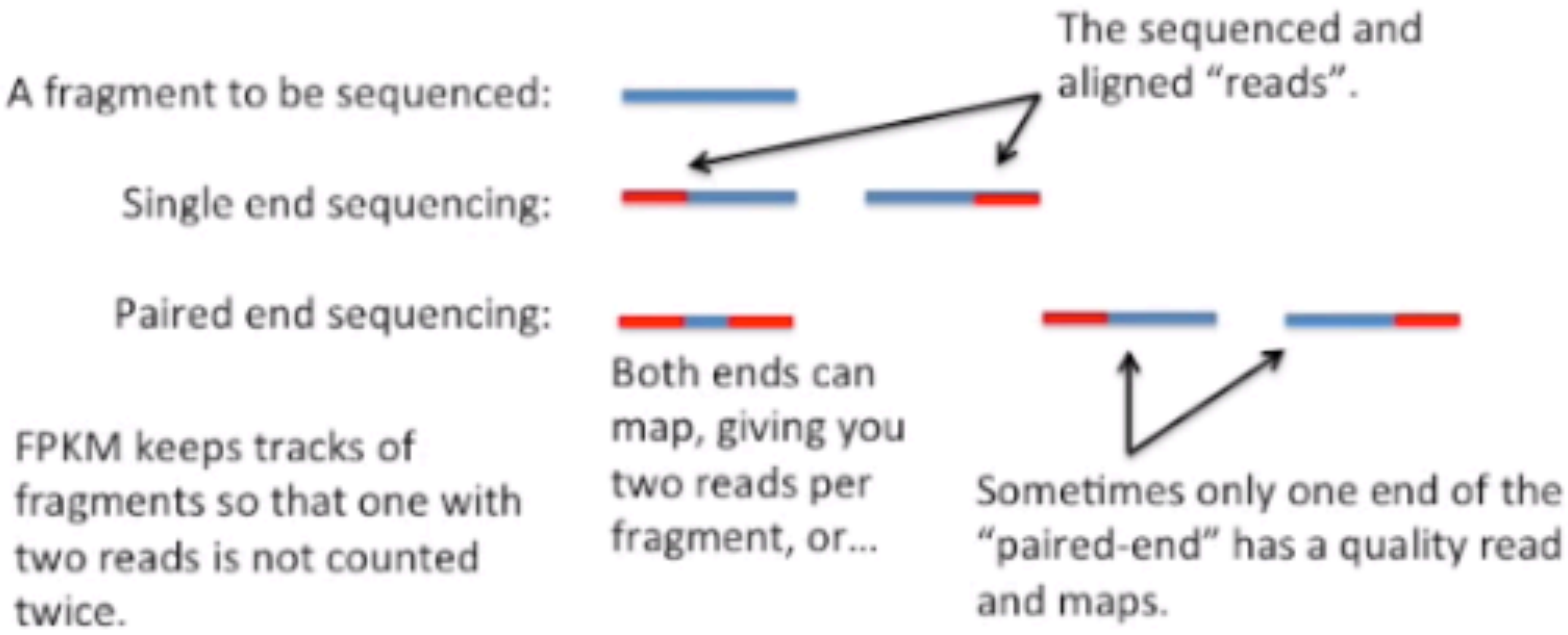
AFTER

Gene Name	Rep1 RPKM	Rep2 RPKM	Rep3 RPKM
A (2kb)	1.43	1.33	1.42
B (4kb)	1.43	1.39	1.42
C (1kb)	1.43	1.78	1.42
D (10kb)	0	0	0.009

RPKM and FPKM – two very closely related terms…

RPKM = Reads Per Kilobase Million
FPKM = Fragments per Kilobase Million

RPKM is for single end RNA-seq.
FPKM is very similar to RPKM, but for paired end RNA-seq.



The sequenced and aligned “reads”.

A fragment to be sequenced:

Single end sequencing:

Paired end sequencing:

Both ends can map, giving you two reads per fragment, or…

Sometimes only one end of the “paired-end” has a quality read and maps.

FPKM keeps tracks of fragments so that one with two reads is not counted twice.

TPM – step 1: normalize for gene length

Original data:

Gene Name	Rep1 Counts	Rep2 Counts	Rep3 Counts
A (2kb)	10	12	30
B (4kb)	20	25	60
C (1kb)	5	8	15
D (10kb)	0	0	1

RPK – scaled by gene length:

Gene Name	Rep1 RPK	Rep2 RPK	Rep3 RPK
A (2kb)	5	6	15
B (4kb)	5	6.25	15
C (1kb)	5	8	15
D (10kb)	0	0	0.1

TPM – step 2: normalize for sequencing depth

Gene Name	Rep1 RPK	Rep2 RPK	Rep3 RPK
A (2kb)	5	6	15
B (4kb)	5	6.25	15
C (1kb)	5	8	15
D (10kb)	0	0	0.1
Total RPK:	15	20.25	45.1
Tens of RPK:	1.5	2.025	4.51

Again, for this 4 gene example, we are only dividing by 10 to get the "per million" scaling factors.

TPM – step 2: normalize for sequencing depth

Gene Name	Rep1 RPK	Rep2 RPK	Rep3 RPK
A (2kb)	5	6	15
B (4kb)	5	6.25	15
C (1kb)	5	8	15
D (10kb)	0	0	0.1

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

TPM – scaled by gene length and sequencing depth (M):

Gene Name	Rep1 TPM	Rep2 TPM	Rep3 TPM
A (2kb)	3.33	2.96	3.326
B (4kb)	3.33	3.09	3.326
C (1kb)	3.33	3.95	3.326
D (10kb)	0	0	0.02

RPKM vs TPM

RPKM

Gene Name	Rep1 RPKM	Rep2 RPKM	Rep3 RPKM
A (2kb)	1.43	1.33	1.42
B (4kb)	1.43	1.39	1.42
C (1kb)	1.43	1.78	1.42
D (10kb)	0	0	0.009

Both TPM RPKM (and FPKM) correct for biases in gene length and sequencing depth. But....

... the sums of each column are very different.

TPM

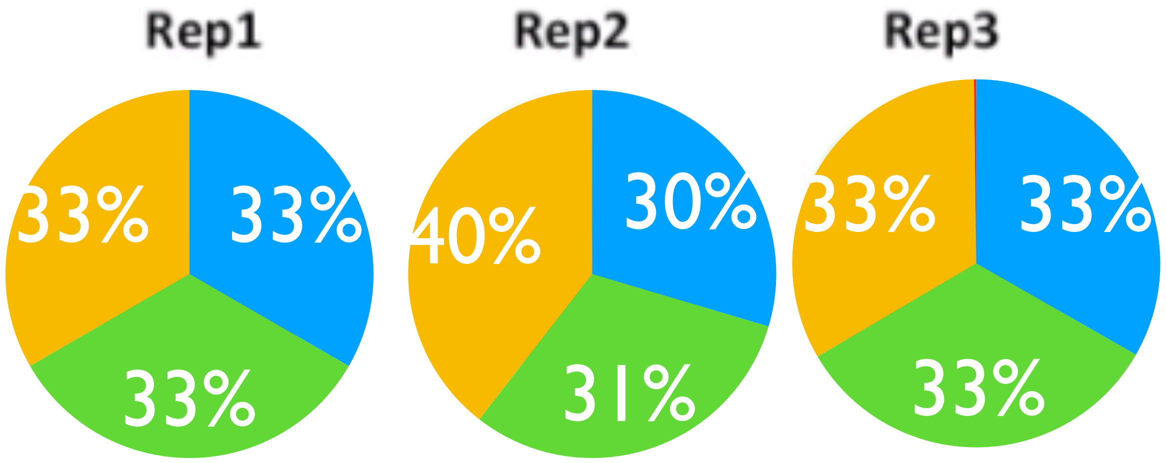
Gene Name	Rep1 TPM	Rep2 TPM	Rep3 TPM
A (2kb)	3.33	2.96	3.326
B (4kb)	3.33	3.09	3.326
C (1kb)	3.33	3.95	3.326
D (10kb)	0	0	0.02

RPKM vs TPM

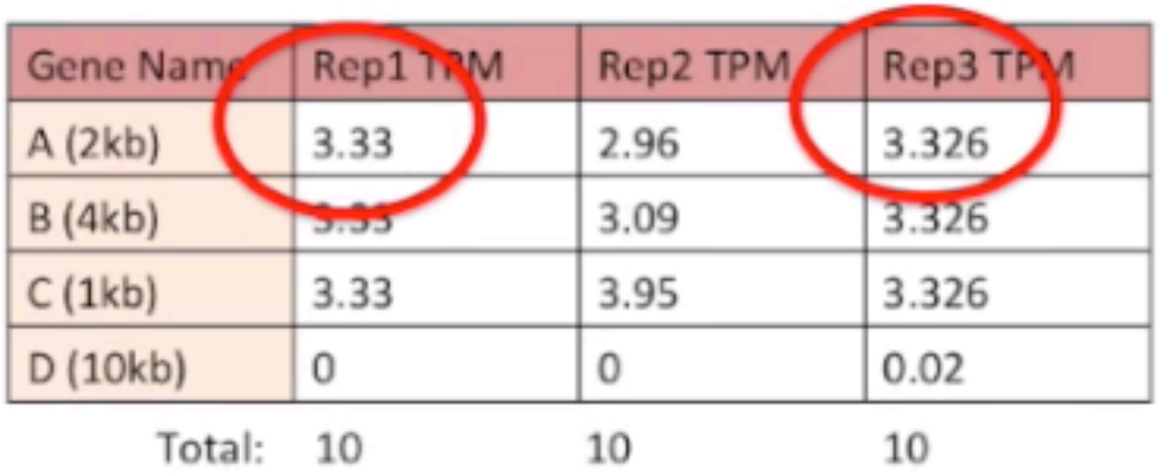
Consider 3 pies, each the same size (10).

A 3.33 sized slice is the same in each pie, and is always larger than 3.32.

TPM makes it clear that in Rep1, more of its total reads mapped to gene A than in Rep3.



TPM



Gene Name	Rep1 TPM	Rep2 TPM	Rep3 TPM
A (2kb)	3.33	2.96	3.326
B (4kb)	3.33	3.09	3.326
C (1kb)	3.33	3.95	3.326
D (10kb)	0	0	0.02
Total:	10	10	10

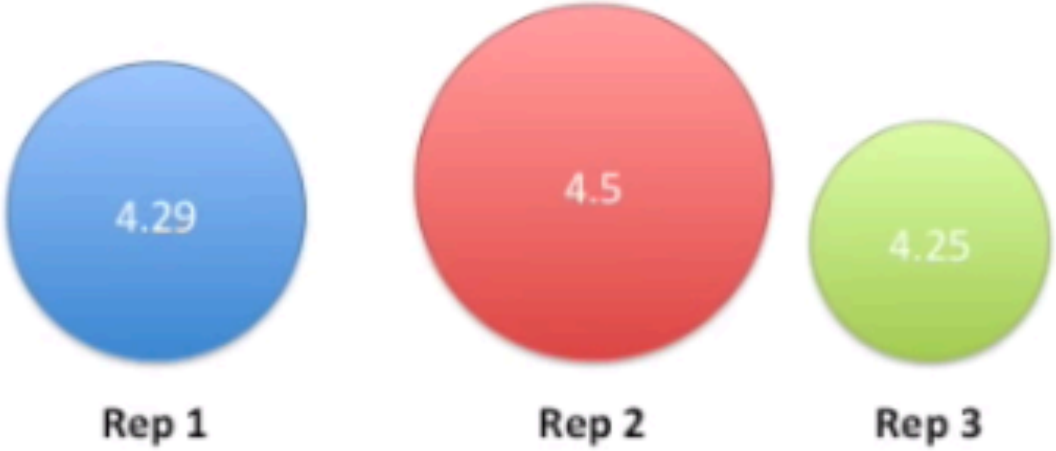
RPKM vs TPM

RPKM

Gene Name	Rep1 RPKM	Rep2 RPKM	Rep3 RPKM
A (2kb)	1.43	1.33	1.42
B (4kb)	1.43	1.39	1.42
C (1kb)	1.43	1.78	1.42
D (10kb)	0	0	0.009
Total:	4.29	4.5	4.25

With RPKM, it is harder to compare the proportion of total reads because each replicate has different total (each pie has a different size)

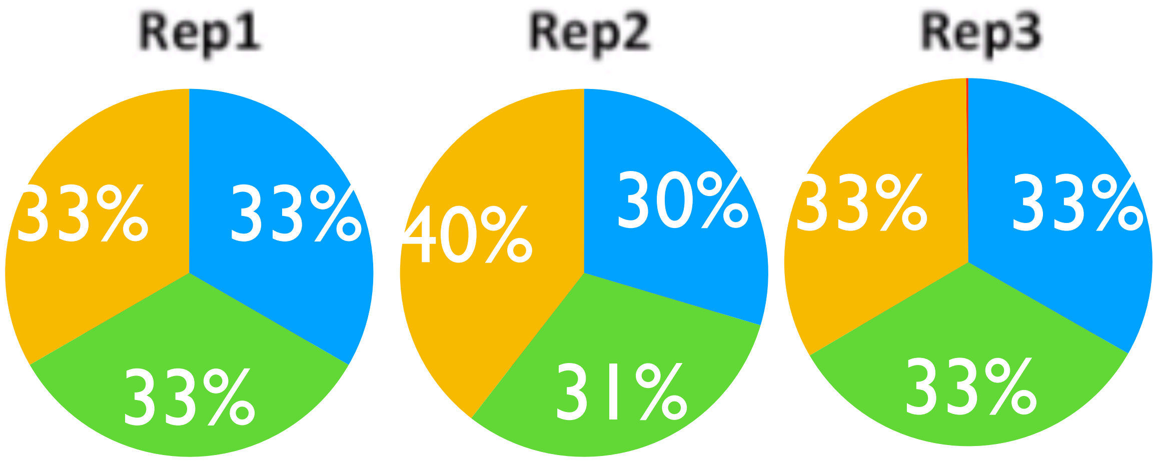
A 1.43 size slice represents a different proportion of reads in in different pies.



Main point: With TPM, everyone gets the same sized pie.

using TPM because the numbers can clearly tell you what proportion of reads mapped to what in each sample.

And since RNA-seq is all about comparing relative proportions of reads, this metric seems more appropriate.



TPM

Gene Name	Rep1 TPM	Rep2 TPM	Rep3 TPM
A (2kb)	3.33	2.96	3.326
B (4kb)	3.33	3.09	3.326
C (1kb)	3.33	3.95	3.326
D (10kb)	0	0	0.02
Total:	10	10	10