BIO214 Lecture 5

Bioinformatics-II

Genomic Data Normalization-1

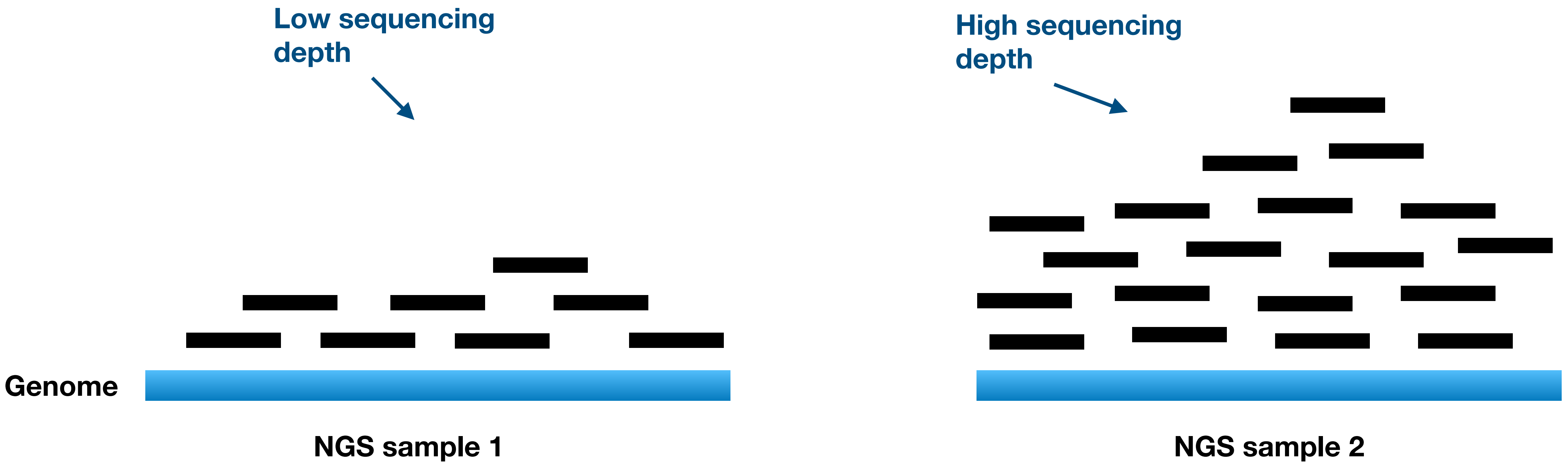
Zhen Wei; 2023-Feb-14

Outline

- Account for sequencing depth
- RPKM, FPKM, TPM
- Z-score and quantile normalization
- MA normalization
- Transformation

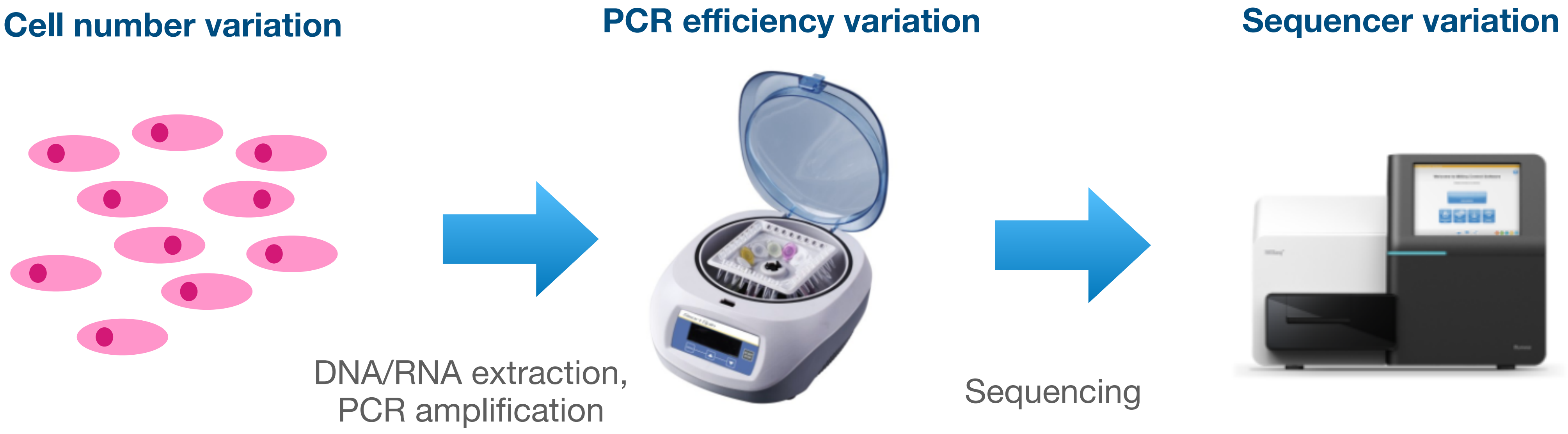
Account for sequencing depth

Sequencing depth



- **Sequencing depth** can be understood as the mean read coverage over the genome / transcriptome of an aligned NGS library.
- Sequencing depth changes a lot across sequencing samples
- As a type of technical variation, sequencing depth is often estimated in order to normalize read count.

What causes sequencing depth variation?



Cell number variation

PCR efficiency variation

Sequencer variation

DNA/RNA extraction,
PCR amplification

Sequencing

- **Initial # of cells in the sample**
  NGS library is constructed with different amount of starting cells.

- **PCR amplification efficiency**
  Variation in PCR temperature and cycle # can affect the fragment amplification rate.

- **NGS platform**
  The fragment detection rate varies across sequencing lanes and platforms.

How to estimate sequencing depth from read counts?

	Sample 1	Sample 2	Sample 3
Gene A	16	5	28
Gene B	13	3	15
Gene C	7	0	9
Gene D	28	12	21
Estimated Sequencing depth	16+13+7+28 = 64	5+3+0+12 = 20	28+15+9+21 = 73

- Sequencing depth is often estimated by the location estimators (e.g. mean or median) over read counts in a sequencing sample.
- A commonly used estimation is by summing up all counts within a sample.

How to normalize sequencing depth in gene expression quantification?

	Sample 1	Sample 2	Sample 3
Gene A	16	5	28
Gene B	13	3	15
Gene C	7	0	9
Gene D	28	12	21
Sequencing depth	64	20	73

Dividing each column by its size factor

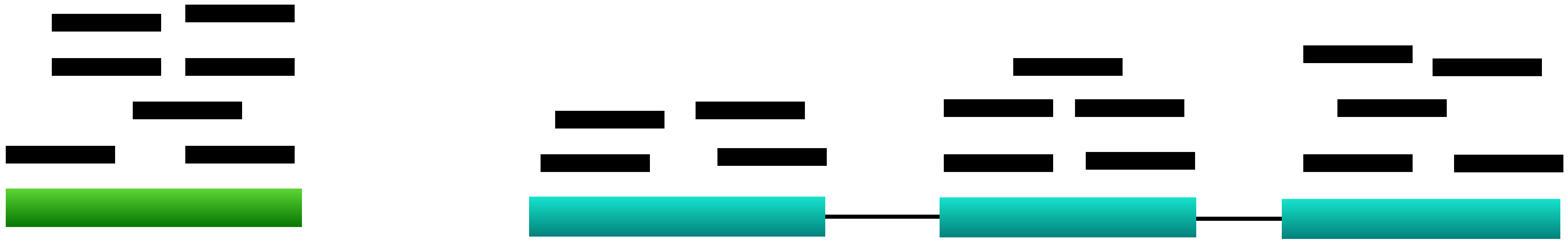
Expression matrix normalized by sequencing depth

	Sp 1	Sp 2	Sp 3
Gene A	16/64	5/20	28/73
Gene B	13/64	3/20	15/73
Gene C	7/64	0/20	9/73
Gene D	28/64	12/20	21/73

- A natural way to adjust sequencing depth is to divide counts by the size factors.

RPKM, FPKM, TPM

Effect of feature length



Read count in gene A = 7

Read count in gene B = 14

Gene A: width 200 bp

Gene B: width 800 bp

So, the # of transcript copies expressed
by gene B is 2 times that of gene A?

- Longer genes express longer transcripts, thereby producing more RNA fragments to be sequenced.
- The gene lengths (calculated over exonic regions) also need to be normalized when quantifying gene expression.

Feature specific normalization factors

	length (bp)	Sp 1	Sp 2	Sp 3
Gene A	500	16	5	28
Gene B	700	13	3	15
Gene C	150	7	0	9
Gene D	900	28	12	21

Dividing both length and sequencing depth

	Sp 1	Sp 2	Sp 3
Gene A	16/(500*64)	5/(500*20)	28/(500*73)
Gene B	13/(700*64)	3/(700*20)	15/(700*73)
Gene C	7/(150*64)	0/(150*20)	9/(150*73)
Gene D	28/(900*64)	12/(900*20)	21/(900*73)

- We can normalize over multiple size factors at once by dividing the product of size factors (in this case the sequencing depth and the feature length).

RPKM, FPKM, TPM

Three popular normalization strategies for gene expression quantification are:

1. **RPKM** (reads per kilobase of transcript per million reads mapped)

$$\text{RPKM} = \frac{\text{Read Count}}{\text{Gene length} \times \sum_{\forall genes} \text{Read Count}} \times 10^9$$

2. **FPKM** (Fragments per kilobase of transcript per million reads mapped)

$$\text{FPKM} = \frac{\text{Fragment Count}}{\text{Gene length} \times \sum_{\forall genes} \text{Fragment Count}} \times 10^9$$

Sum over all genes within a sample

3. **TPM** (Transcripts per million)

$$\text{TPM} = \frac{\text{Read Count}}{\text{Gene length} \times \sum_{\forall genes} (\text{Read Count/Gene length})} \times 10^6$$

Sequencing depth estimated on the length normalized count, ensuring sample wised sum of TPM = constant

RPKM is viewing RNA-Seq experiment as a pool of dice rolls

Outcomes of rolling an unfair dice 218 times

	1	2	3	4	5	6
Event Count	16	5	28	101	23	45

=

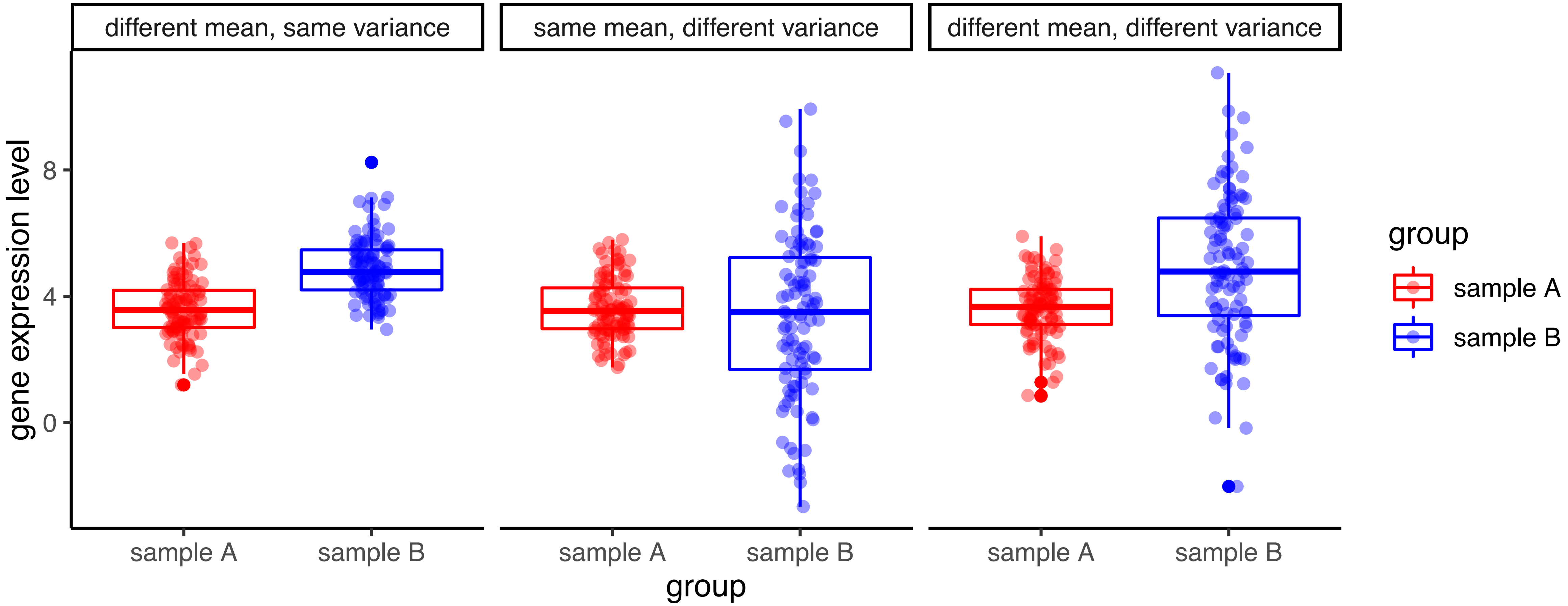
Read counts of a sequencing sample

	Gene A	Gene B	Gene C	Gene D	Gene E	Gene F
Read Count	16	5	28	101	23	45

- Essentially, the RPKM liked measures are making empirical estimation on the probabilities of getting each facet of a biased dice.

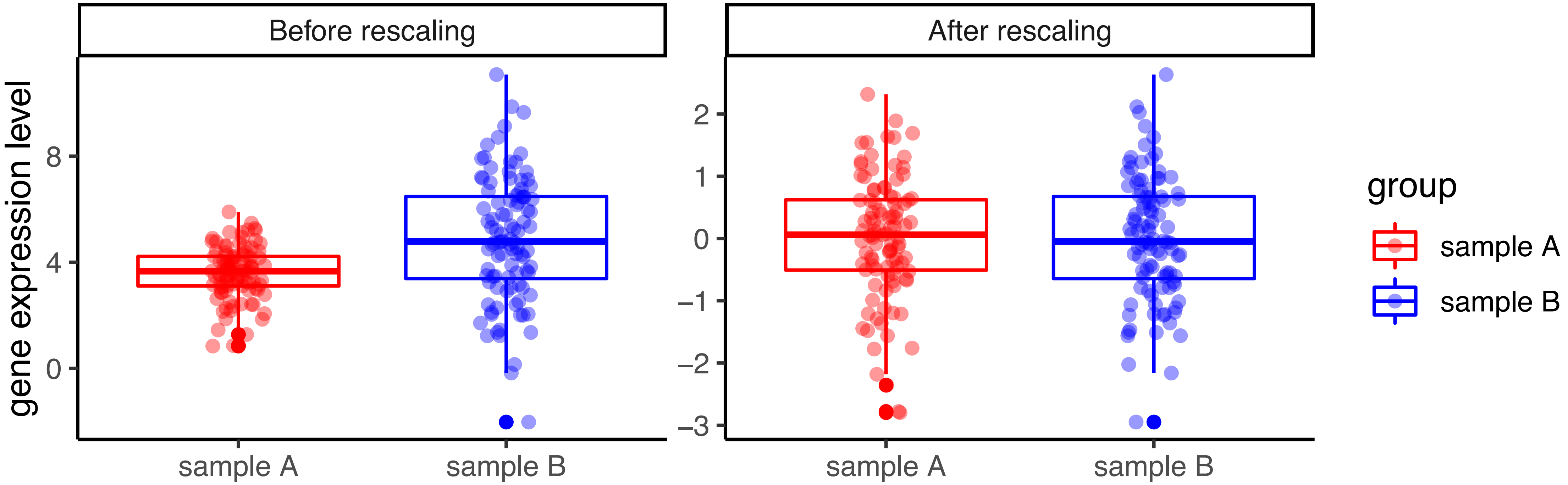
Z-score and quantile normalization

What about the difference in variances?



- The 2 libraries can be different in both means and variances, normalizing only over sequencing depths (means) cannot account for the dispersion level difference.

Z-score normalization



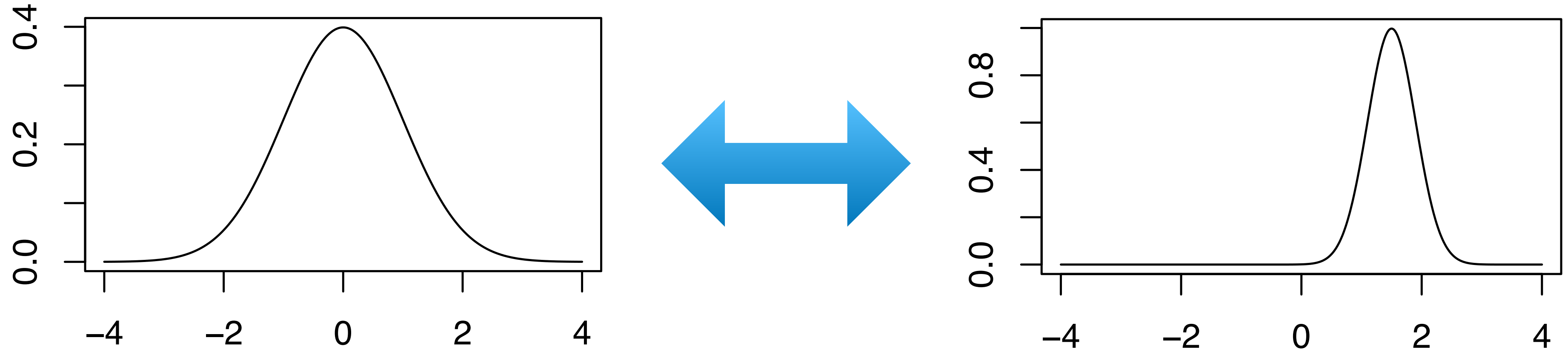
- The **z-score normalization** is defined by:

$$Z = \frac{X - \text{mean}(X)}{\text{sd}(X)}$$

- mean is the sample mean; sd is the sample standard deviation.
- The process transforms any data variable into 0 mean and unit variance (sd = 1).
- z-scores are also useful to be computed within genes (row z-scores).
- Rescaling is often crucial for downstream analysis, such as clustering and PCA.

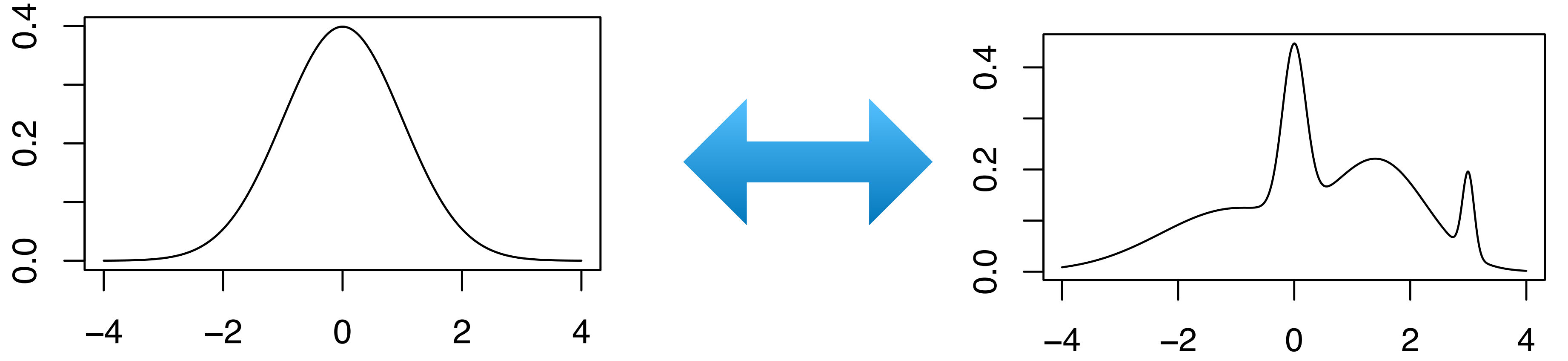
How to account for the shape difference?

2 distributions have only mean & dispersion difference:



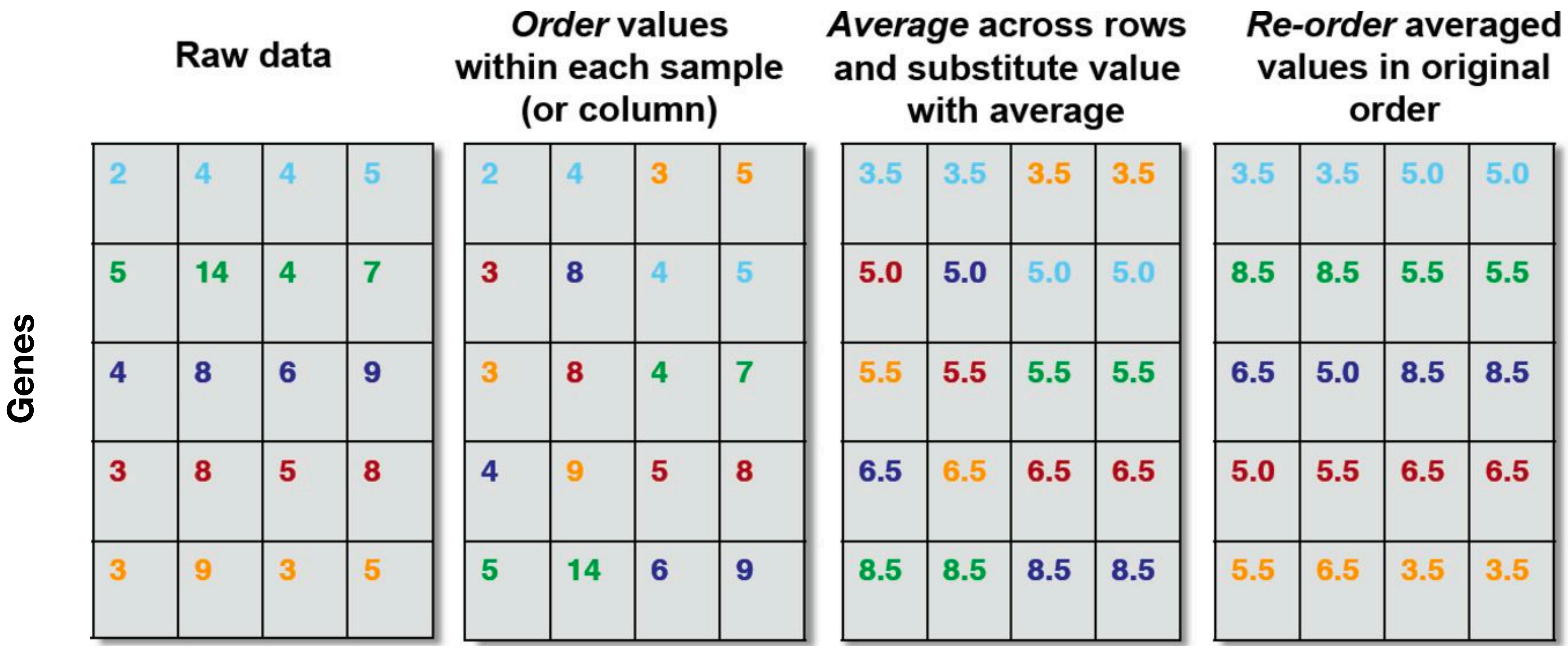
Can be transformed
by rescaling

2 distributions have shape difference:



Cannot be transformed
by rescaling

Quantile normalization



Genes

Raw data

2	4	4	5
5	14	4	7
4	8	6	9
3	8	5	8
3	9	3	5

Order values within each sample (or column)

2	4	3	5
3	8	4	5
3	8	4	7
4	9	5	8
5	14	6	9

Average across rows and substitute value with average

3.5	3.5	3.5	3.5
5.0	5.0	5.0	5.0
5.5	5.5	5.5	5.5
6.5	6.5	6.5	6.5
8.5	8.5	8.5	8.5

Re-order averaged values in original order

3.5	3.5	5.0	5.0
8.5	8.5	5.5	5.5
6.5	5.0	8.5	8.5
5.0	5.5	6.5	6.5
5.5	6.5	3.5	3.5

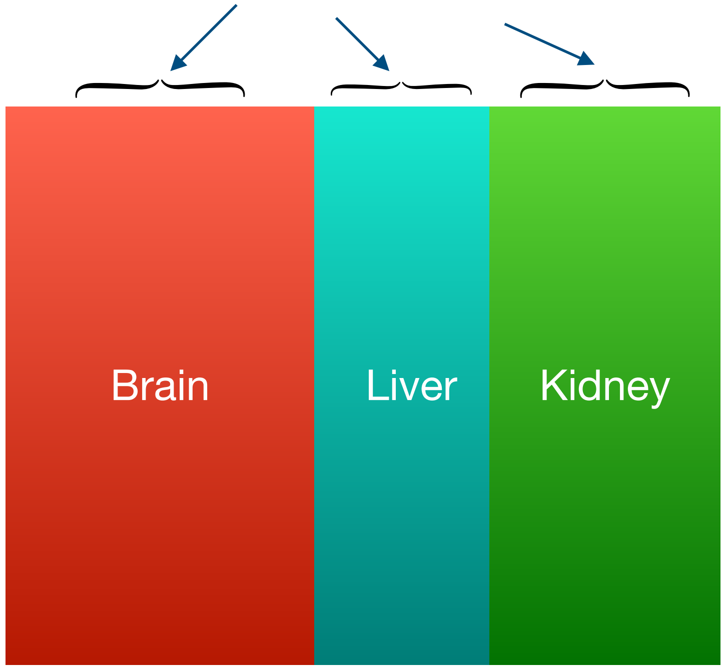
- **Quantile normalization** (QN) can enforce identical distributions across any sequencing samples.

- QN steps: 1. order column (sample) values. 2. substitute values with row (gene) averages. 3. return to the original order.

- The procedure can effectively remove batch effect in genomic data.

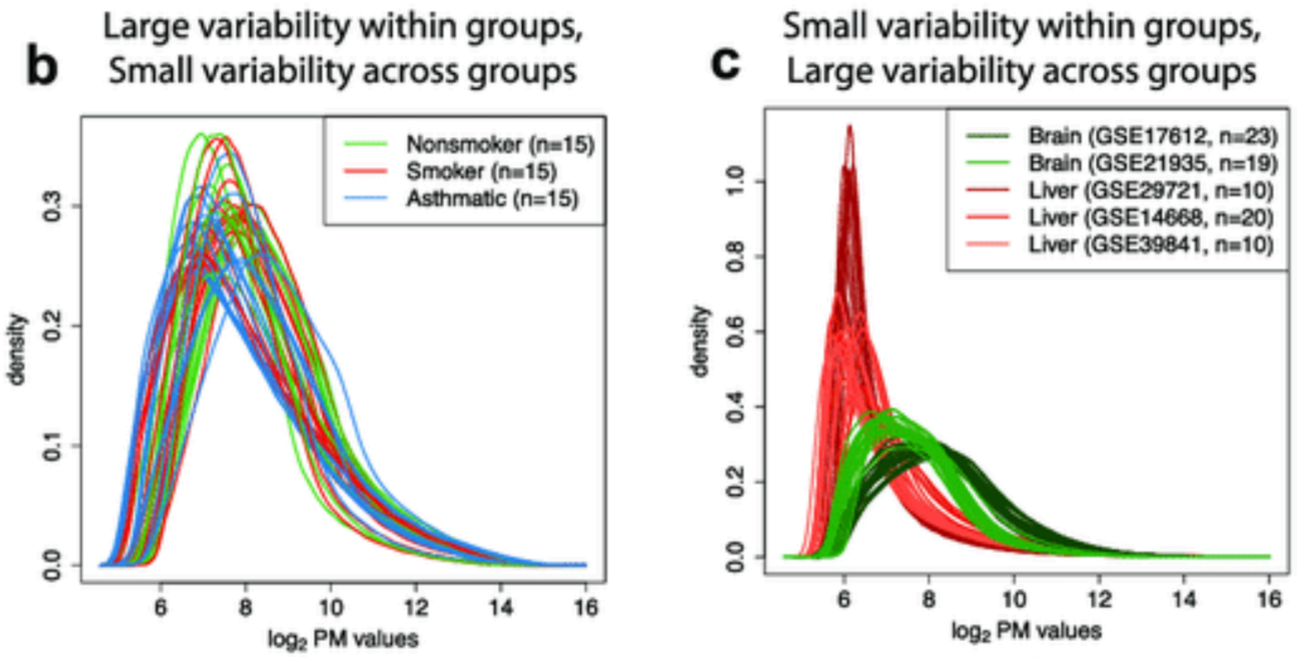
The importance of performing QN within biological groups

Run QN within each tissue or biological condition, not cross them.



Should apply QN

Should not apply QN

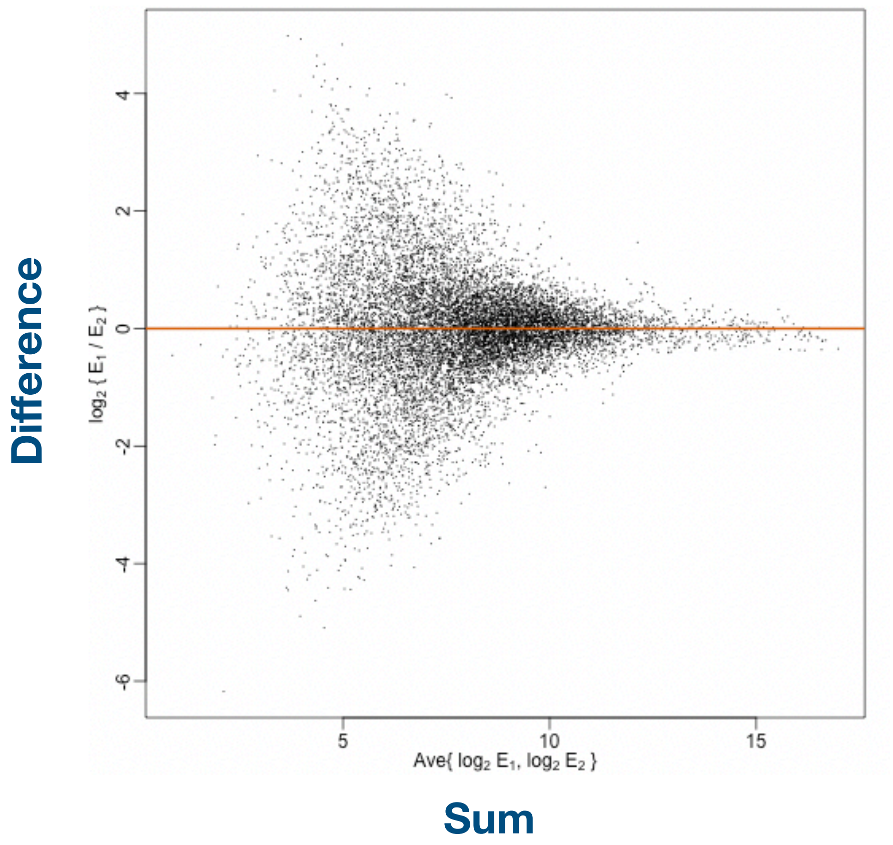


What if some conditions, such as brain and liver, do have significant biological differences in their distributions of expression level?

- Perform QN across biological groups may distort meaningful biological signal.
- QN should be ideally performed within major biological conditions (e.g. tissues and cell types).

MA normalization

MA-plot: check for reproducibility



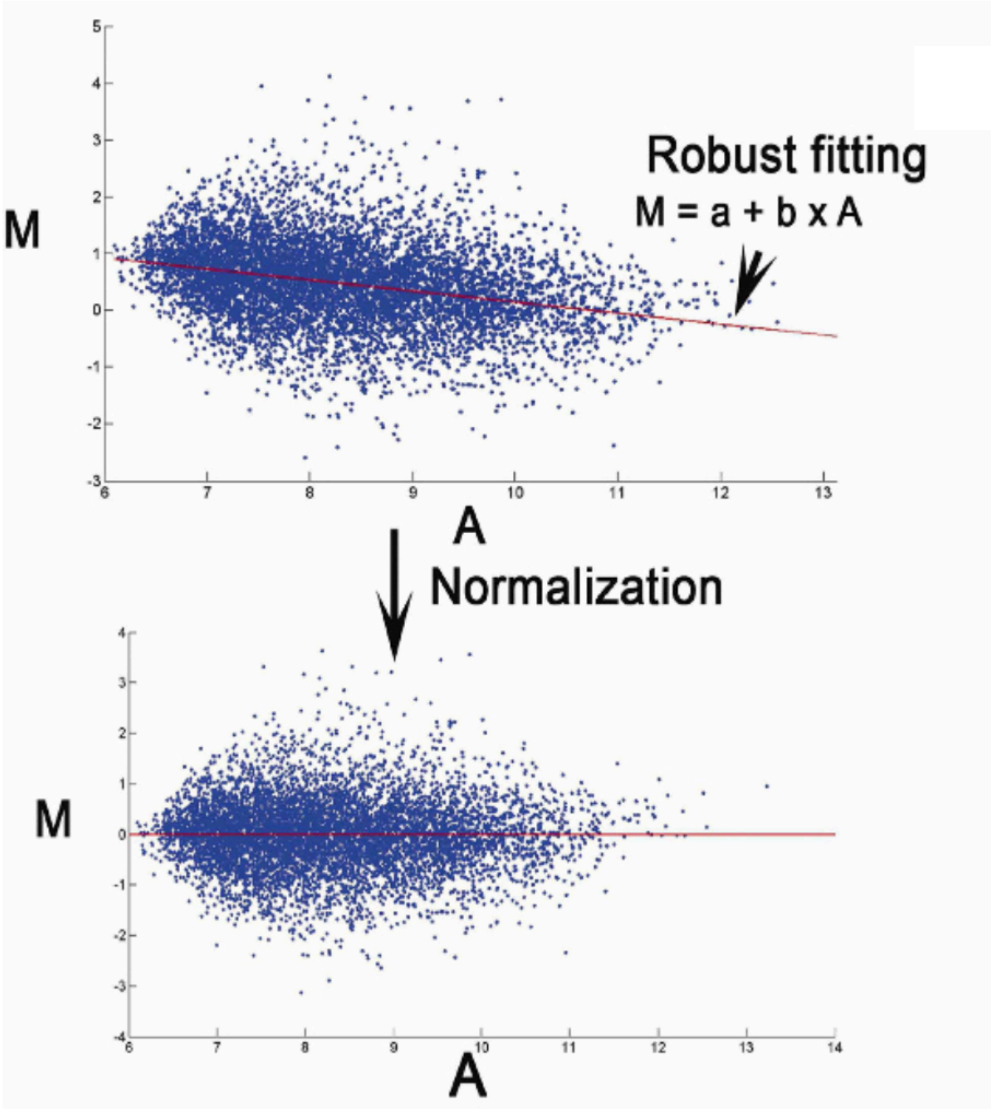
How do we know if 2 sequencing samples, say they are biological or technical replicates, are well reproduced?

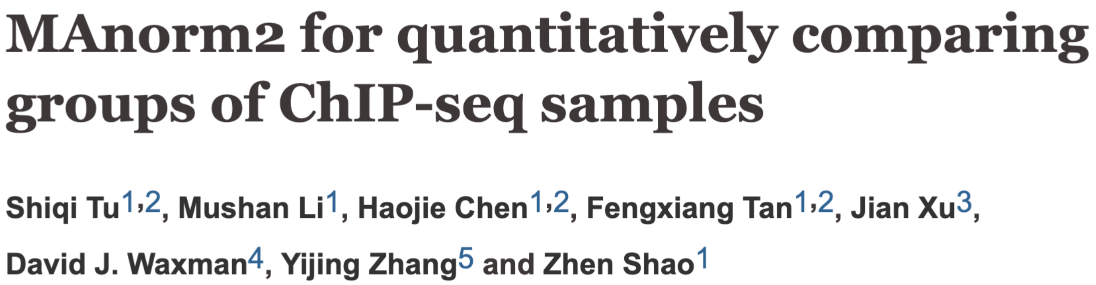
- Correlation coefficient (just a number).

- **MA-plot** is a graphic technique for reproducibility assessment; its x axis is $(\log(E_1) + \log(E_2))/2$ (average of the log expressions), its y axis is $\log(E_1/E_2)$ (expression log fold change).

○ We expect the points to be centered around a horizontal line on MA-plot.

MA-normalization





MAnorm2 for quantitatively comparing groups of ChIP-seq samples

Shiqi Tu[1,2], Mushan Li[1], Haojie Chen[1,2], Fengxiang Tan[1,2], Jian Xu[3], David J. Waxman[4], Yijing Zhang[5] and Zhen Shao[1]

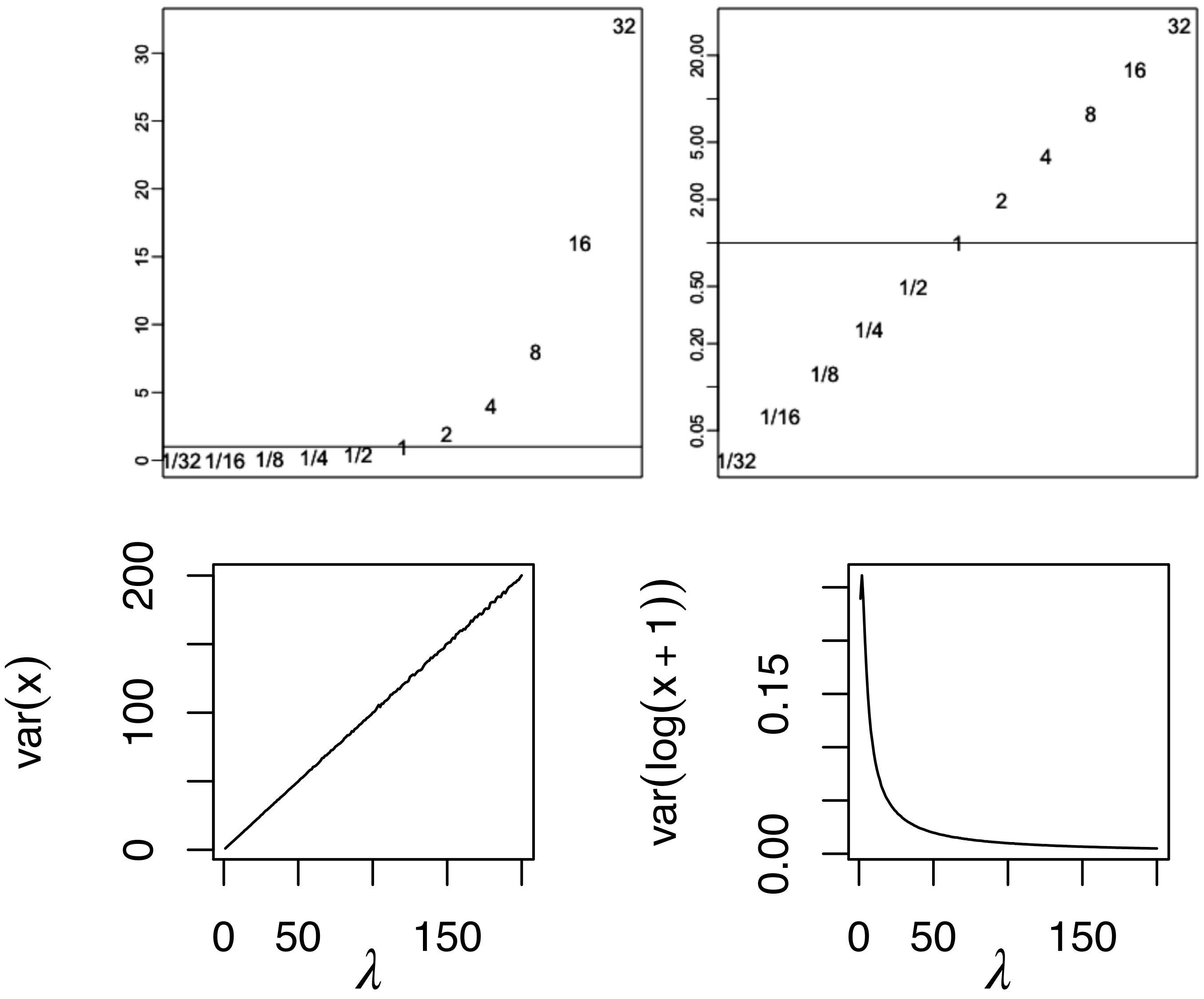
CSH PRESS GENOME RESEARCH

One can correct the genomics data by **MA-normalization**:

1. Choose a reference sample, typically computed by gene-wise averages.
2. Generate an MA-plot for each sample by comparing it to the reference sample, and fit a linear regression to each plot.
3. Normalize each sample by subtracting the fitted values to account for deviations from the expected horizontal line passing origin.

Log transformation

Log transformation



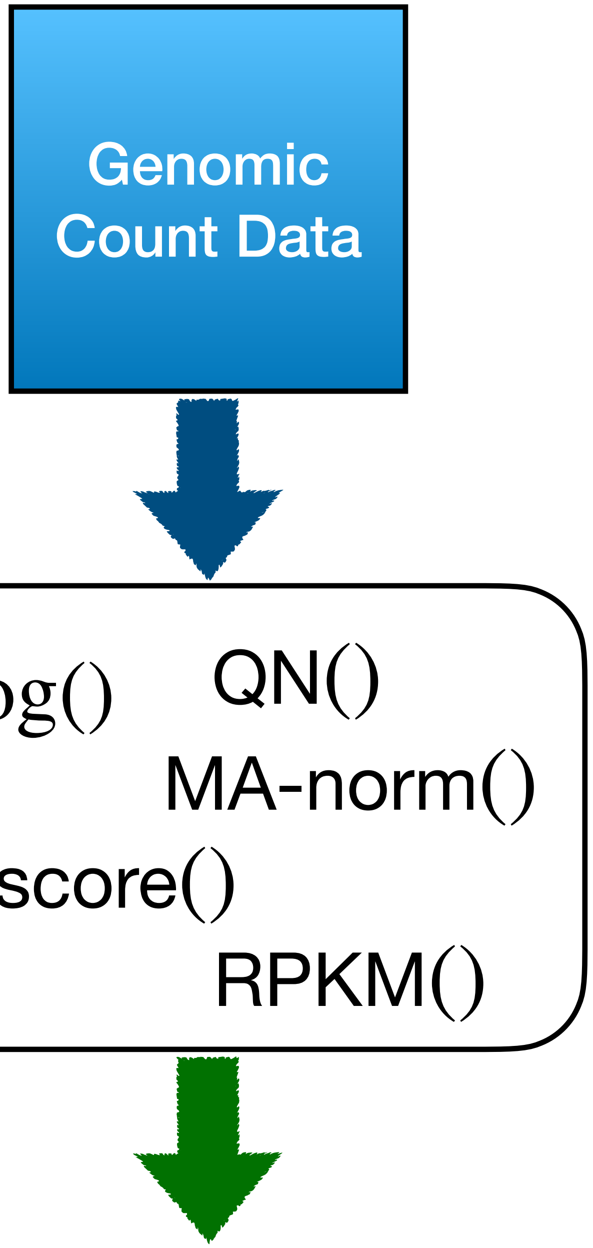
Ratio becomes symmetrical on the log scale (y axis).

$x \sim$ Poisson(λ), after taking the logarithm, mean (λ) and variance are no longer highly dependent.

Thus, log is also called the variance stabilizing transformation.

- **Count** and **ratio** data types are often beneficial from log transformation.
- log(count + 1) and log fold changes are commonly used in genomic data visualization and data analysis.
- log is also a mathematically natural transformation for ratio and count.

Trial and error are encouraged



Pick and combine using your normalization tool box

- No single normalization pipeline is guaranteed to perform well for all data.
- A suitable normalization procedure need to be selected for the specific genomic data type and end application.

Performance evaluation on down stream analysis, e.g. AUROC, p-values of GO enrichment