



# BIO214 Lecture 4

## Bioinformatics-II

### *Quantification of Genome Mapping Results*

Zhen Wei; 2023-Feb-14

# Outline

- Overview of NGS pipeline
- The aim of quantification
- Read count methods
- Isoform level quantification
- Ratio based quantities

# **Overview of NGS pipeline**

# NGS pipeline

Key words at each step:

**Raw Reads (fastq)**

Dastq,  
Phred scores

**Quality Control**

fastqc,  
Adapter contamination,  
Low quality ends

**Trimming**

Adaptor trimming,  
Quality trimming

**Genome Alignment**

Reference genome (fasta),  
Aligners (Hisat2/Tophat2/Bowtie2),  
Alignment results (SAM/BAM)

**Quantification**

IGV visualization  
Transcript annotation (GFF/GTF),  
Read count

**Data Analysis**

Normalization,  
Differential analysis,  
Clustering, PCA,  
and more ...

Quantification is knowing the distribution & counts of reads over genomic features.

# SAM format

**Header section**

```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
```

**Alignment section**

```
r001 99 ref 7 30 8M2I4M1D3M = 37 39 TTAGATAAAGGATACTG *
r002 0 ref 9 30 3S6M1P1I4M * 0 0 AAAAGATAAGGATA *
r003 0 ref 9 30 5S6M * 0 0 GCCTAAGCTAA * SA:Z:ref,29,-,6H5M,17,0;
r004 0 ref 16 30 6M14N5M * 0 0 ATAGCTTCAGC *
r003 2064 ref 29 17 6H5M * 0 0 TAGGC * SA:Z:ref,9,+,5S6M,30,1;
r001 147 ref 37 30 9M = 7 -39 CAGCGGCAT * NM:i:1
```

**QUAL** (read quality; \* meaning such information is not available)

**QNAME** (query template name, aka. read ID)

**FLAG** (indicates alignment information about the read, e.g. paired, aligned, etc.)

**RNAME** (reference sequence name, e.g. chromosome /transcript id)

**POS** (1-based position)

**MAPQ** (mapping quality)

**CIGAR** (summary of alignment, e.g. insertion, deletion)

**RNEXT** (reference sequence name of the primary alignment of the NEXT read; for paired-end sequencing, NEXT read is the paired read; corresponding to the RNAME column)

**PNEXT** (Position of the primary alignment of the NEXT read in the template; corresponding to the POS column)

**TLEN** (the number of bases covered by the reads from the same fragment. In this particular case, it's 45 - 7 + 1 = 39 as highlighted in Panel A). Sign: plus for leftmost read, and minus for rightmost read)

**SEQ** (read sequence)

Optional fields in the format of TAG:TYPE:VALUE

**SAM** is a text-based format. It stores the genome mapping results reported by the aligner. The alignment section within is a table. Each row in the table is a read alignment record, several key columns are:

- **FLAG**: the alignment information (aligned or not, aligned in pairs, multiple alignment ect.)
- **RNAME**: ID for the aligned chromosome / transcript.
- **POS**: the aligned position on the chromosome / transcript (based on read 5' start).
- **MAPQ**: the mapping quality in terms of alignment score
- **CIGAR**: summary of alignment events (e.g. insertion, deletion)

# GTF/GFF format

- **Annotations**

Genome annotations are the genomic experiments conducted earlier than the current one.

The commonly used genome annotations are genes, transcripts, exons, introns, CDS, and various epigenetic markers.

- **How can we represent and store genome annotations?**

Genome annotations are defined in genomic intervals, which contain the information of the locations (start, end, width) on chromosomes, chromosome numbers, and strands (+, -, ; \* for unknown strand).

The gene annotations are often stored under the formats of **GTF**, **GFF**, and **BED**.

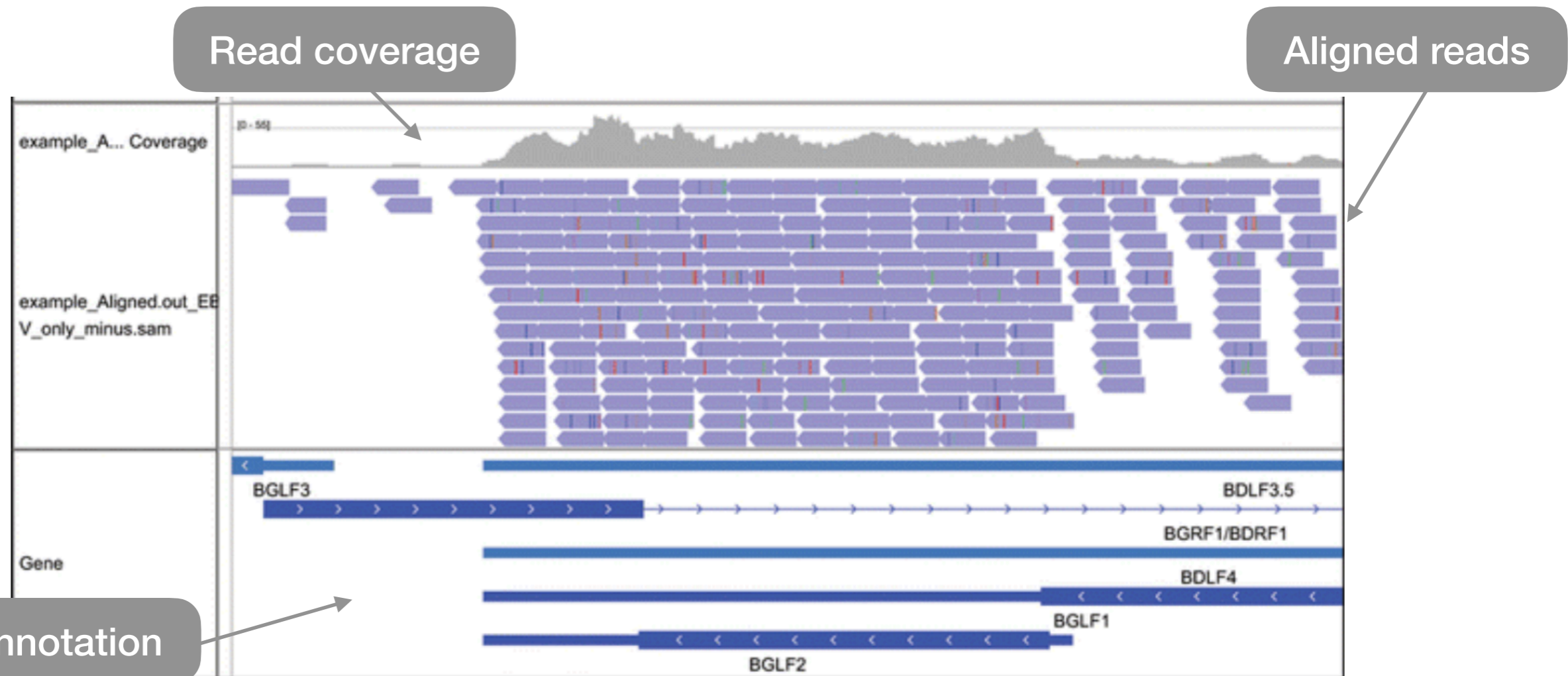
1	#gff-version 3								
2	chr1	BLAST	exon	1300	1500	.	+	.	ID=exon0001; PARENT=Gene1
3	chr1	BLAST	exon	1050	1500	.	+	.	ID=exon0002; PARENT=Gene1
4	chr1	BLAST	exon	3000	3902	.	+	.	ID=exon0003; PARENT=Gene1
5	chr1	BLAST	exon	5000	5500	.	+	.	ID=exon0004; PARENT=Gene1
6	chr1	BLAST	exon	7000	9000	.	+	.	ID=exon0005; PARENT=Gene1

GFF columns: Chromosome, Source, Feature type, Start position, End position, Score, Strand

Reading Frame - 0, 1 or 2 indicating which base of the feature is the first base of the codon

Semicolon separated attribute: ID (feature name); PARENT (meta-feature name)

# How to see the aligned reads?



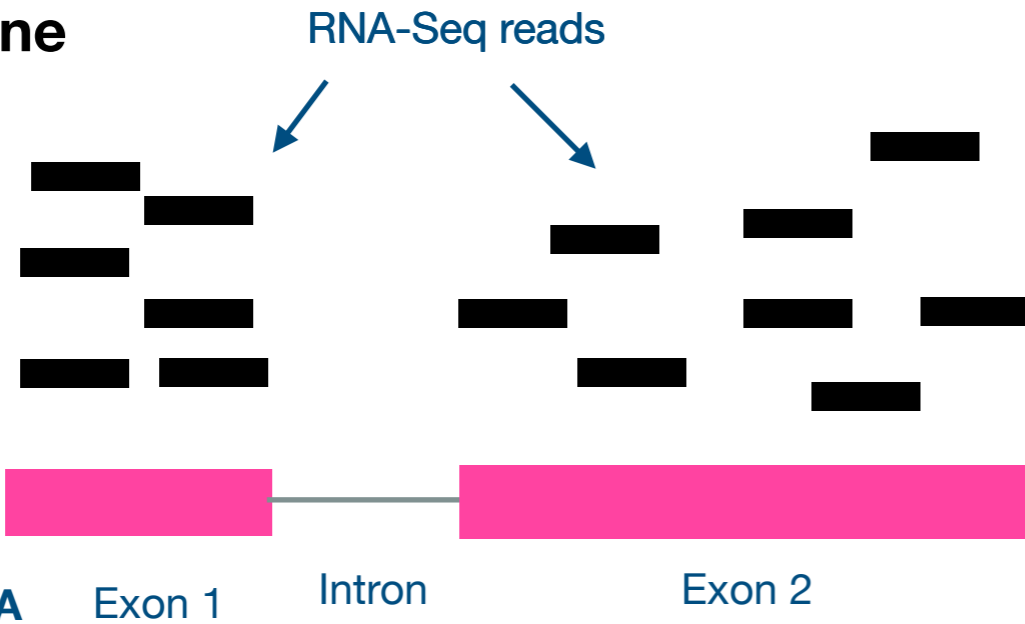
- **IGV (Integrative Genome Viewer)** is a genome browser developed by the Broad Institute, which accepts BAM/SAM files as input.
- It automatically computes **read coverage** by stacking the alignments along genome coordinates.
- One of the best ways to check and understand a high throughput genomic experiment is through the visualization of aligned reads against gene annotation in IGV.

# **The aim of quantification**

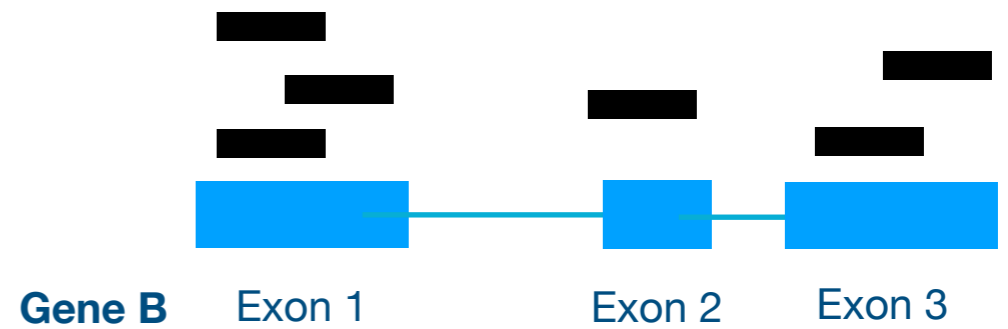


# What biological questions can be answered after genome mapping?

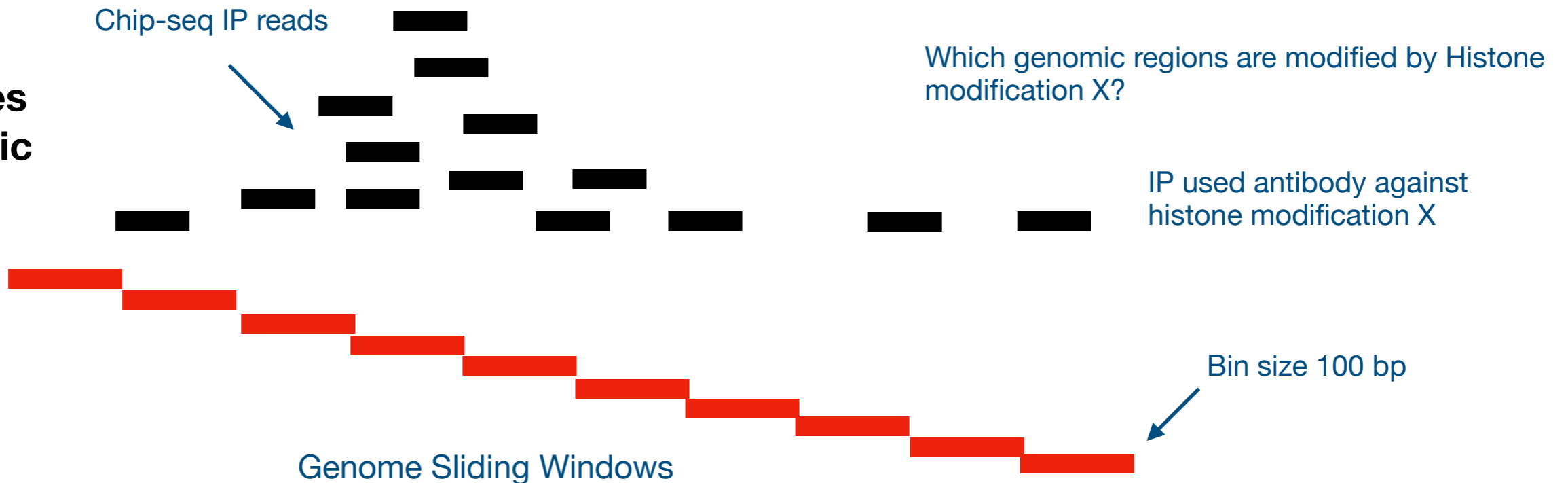
**Quantify gene expression levels:**



How many (relative) copies of transcripts are expressed by Gene A and Gene B, respectively?



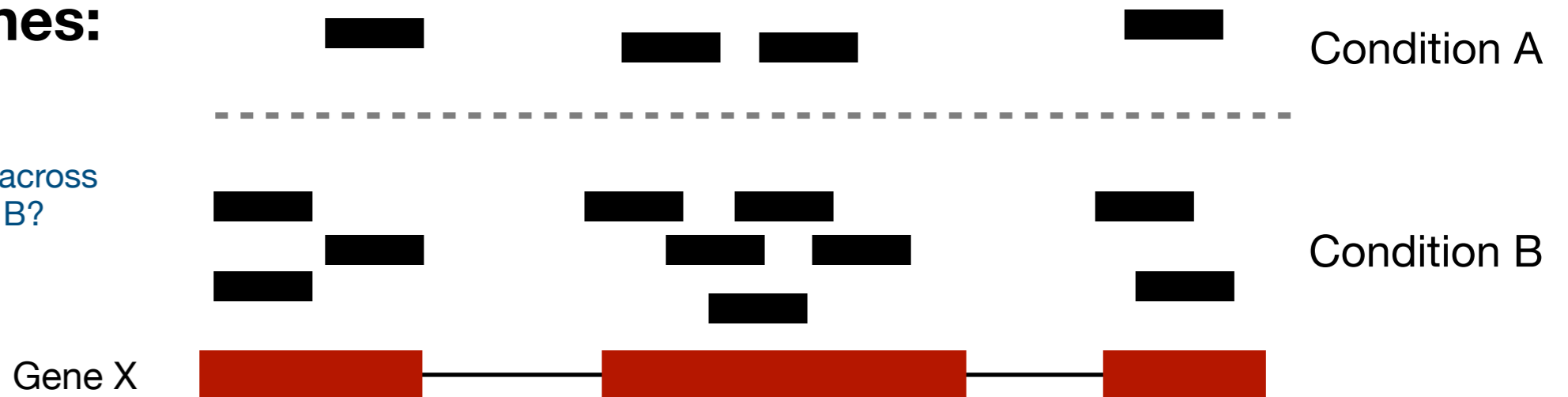
**Identify binding sites of epigenetic markers:**



# What biological questions can be answered after genome mapping?

## Identify differentially expressed genes:

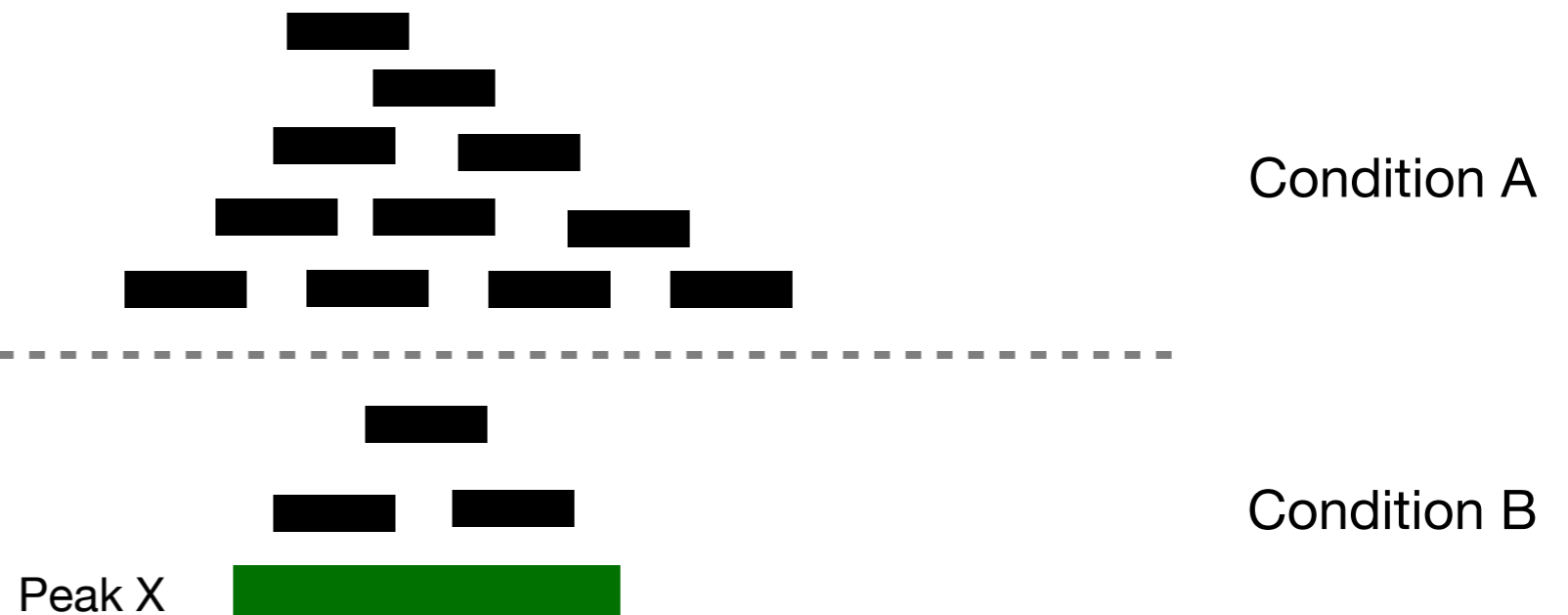
Is gene X significantly differentially expressed across condition A & condition B?



## Identify differentially enriched peaks:



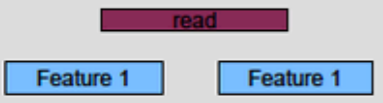

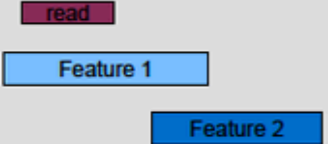
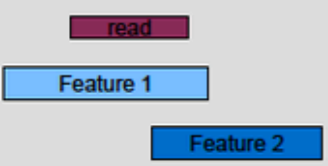

Is peak X significantly differentially modified across condition A & condition B?

Accurate quantification is necessary for differential analysis



# **Read count methods**

# Read count methods over genomic ranges

	Union	IntersectionStrict	IntersectionNotEmpty
	Feature I	Feature I	Feature I
	Feature I	No hit	Feature I
	Feature I	No hit	Feature I
	Feature I	Feature I	Feature I
	Feature I	Feature I	Feature I
	No hit	Feature 1	Feature I
	No hit	No hit	No hit

3 major modes are implemented in *HTSeq Count* (or equivalently R *summarizeOverlaps*):

1. **Union**<sup>\*</sup>, a read belongs to the feature if any overlap exist between read & feature.

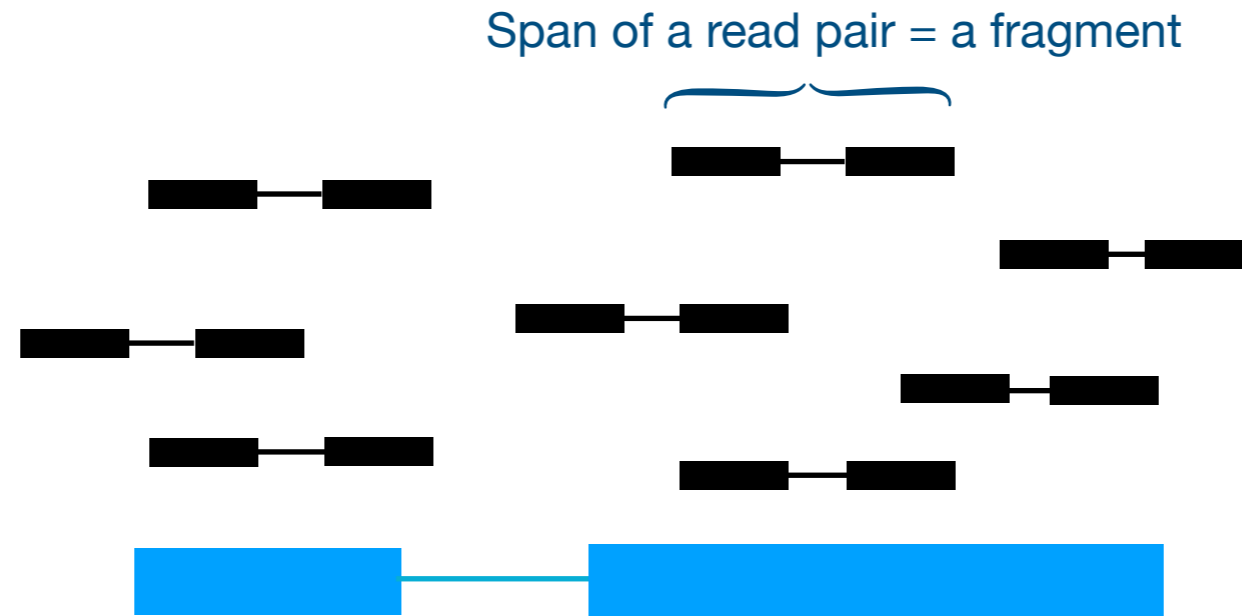
\* Can ensure sensitivity, should be used for bin count in peak calling.

2. **IntersectionStrict**<sup>\*</sup>, a read belongs to the feature if it falls “within” a feature. i.e. only compatible reads are counted.

\* Can ensure specificity, should be used for transcript quantification.

3. **IntersectionNotEmpty**, a loosely defined union mode, reads mapped to > 1 features are still counted to the compatible feature.

# How to count paired end reads?



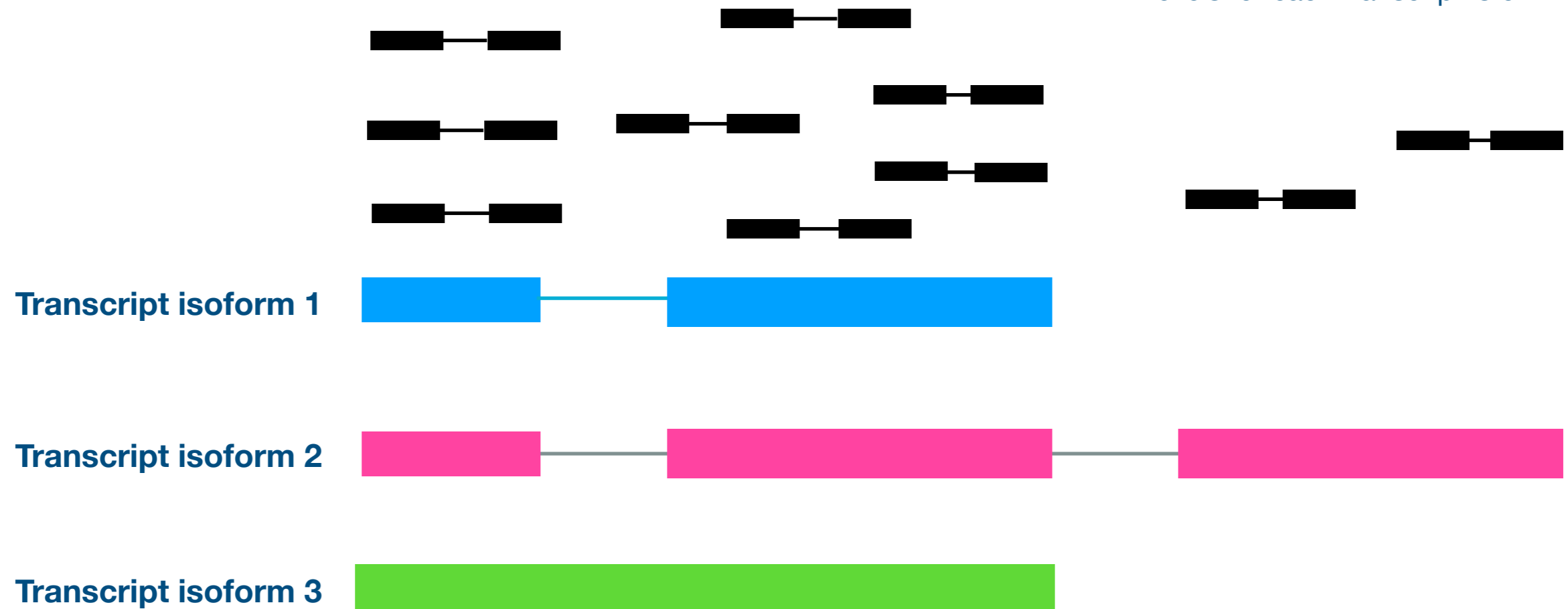
## Fragment count vs read count

- Illumina paired-end sequencing library generates reads from both ends of a DNA/cDNA fragment.
- The paired reads are expected to be aligned concordantly by genome mapping software, which allows the determination of the range of the fragment on the genome.
- To quantify PE NGS library, fragment count is often used instead of read count, as it better reflects the underlying biology.
- In practice, fragment count is approximately half of the corresponding read count.

# **Isoform level quantification**

# The challenge of transcript isoform

What are the count / expression levels for each transcript isoform?



- Alternative splicing can result in genes expressing multiple transcript isoforms.
- The read coverage of such genes can be convolved by signals originating from multiple transcript isoforms.
- To estimate isoform-specific expression levels, an EM (Expectation-Maximization) algorithm can be used.

# Isoform level quantification: EM algorithm

	Tx isoform 1	Tx isoform 2	Tx isoform 3
<b>Read 1</b>	1 0.8	0 0	1 0.2
<b>Read 2</b>	0 0	1 0.6	1 0.4
<b>Read 3</b>	0 0	0 0	1 1
<b>Read 4</b>	1 0.7	0 0	1 0.3
<b>Estimated abundance</b>	0.8+0+0+0.7	0+0.6+0+0	0.2+0.4+1+0.3

Red number:  
Probabilities reads  
coming from each  
transcript

**EM algorithm** is an iterative procedure for estimating the expression levels of transcripts, given a compatibility matrix between reads and transcripts. The goal of the EM algorithm is to estimate the "probability" of reads coming from each transcript.

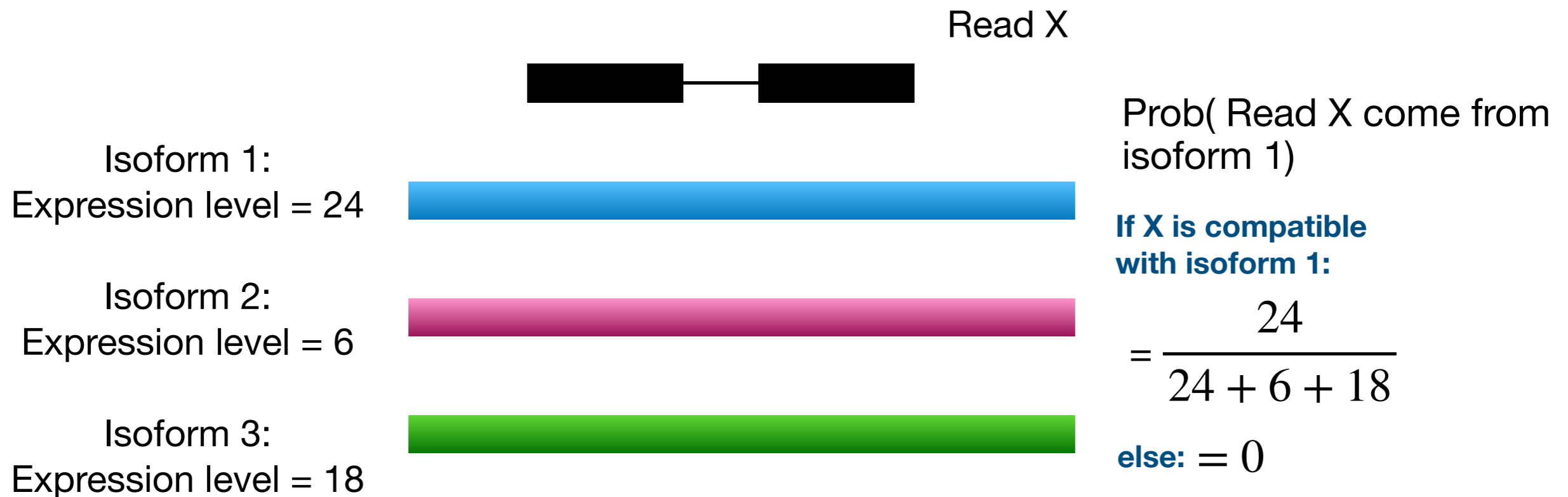
The algorithm works as follows:

1. Initialize with some random expression level estimates.
2. E-step: Estimate the probability of reads being assigned to different transcripts, given the compatibility matrix and the current expression level estimates.
3. M-step: Update the expression level estimates by summing the read probabilities (column sums).
4. Repeat steps 2 and 3 until the expression level estimates converge.



# E-step: how to calculate

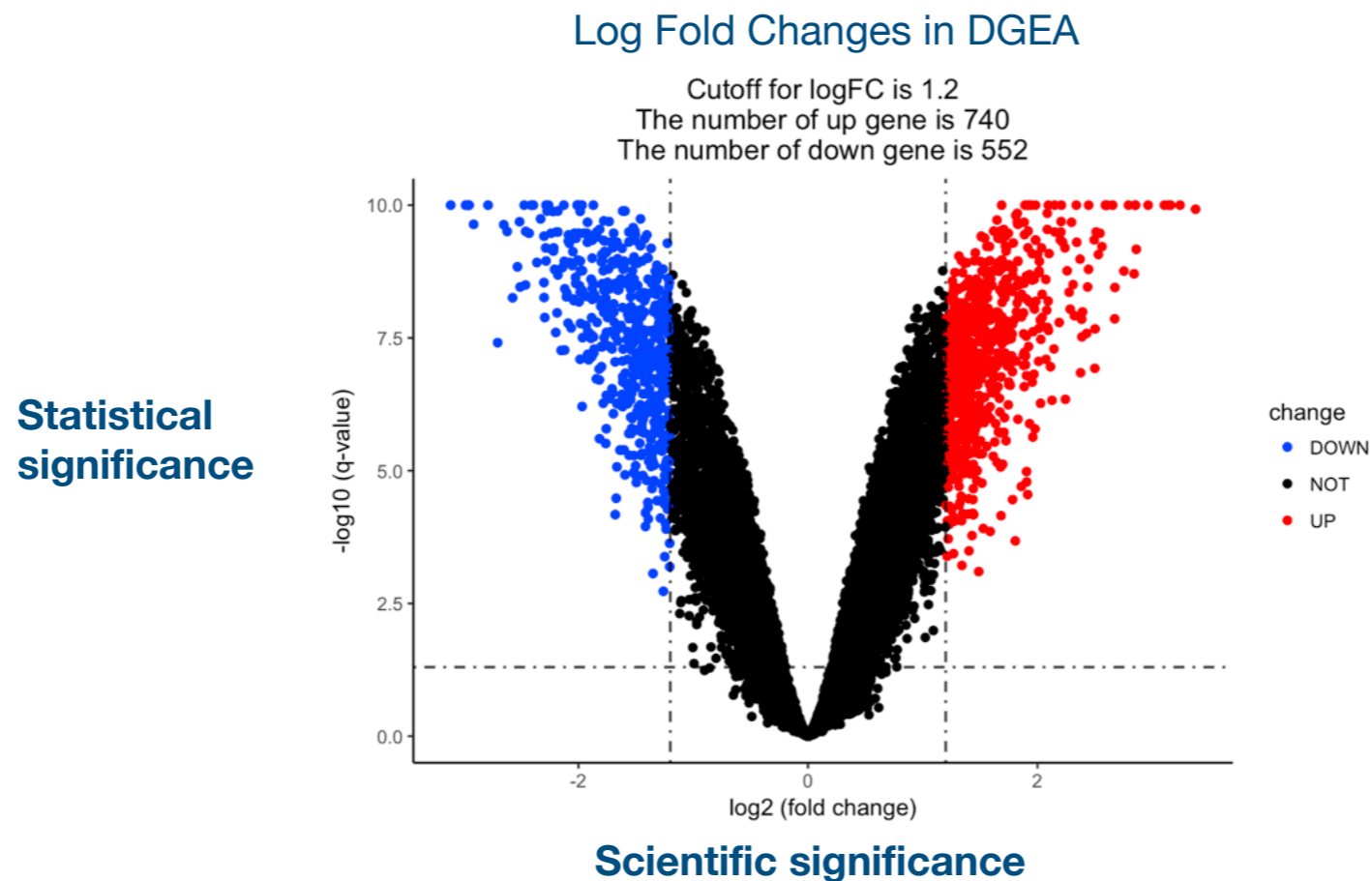
Prob( read -> transcript | transcript expression levels ) ?



- Assuming uniform generation, the probability of a read coming from a transcript is the fraction of that transcript's expression level among the sum of the expression levels of all transcripts compatible with the read.
- The EM algorithm is commonly used to estimate transcript expression levels and is implemented in many RNA-Seq quantification software such as Kallisto, salmon, and alpine.

# **Ratio based quantities**

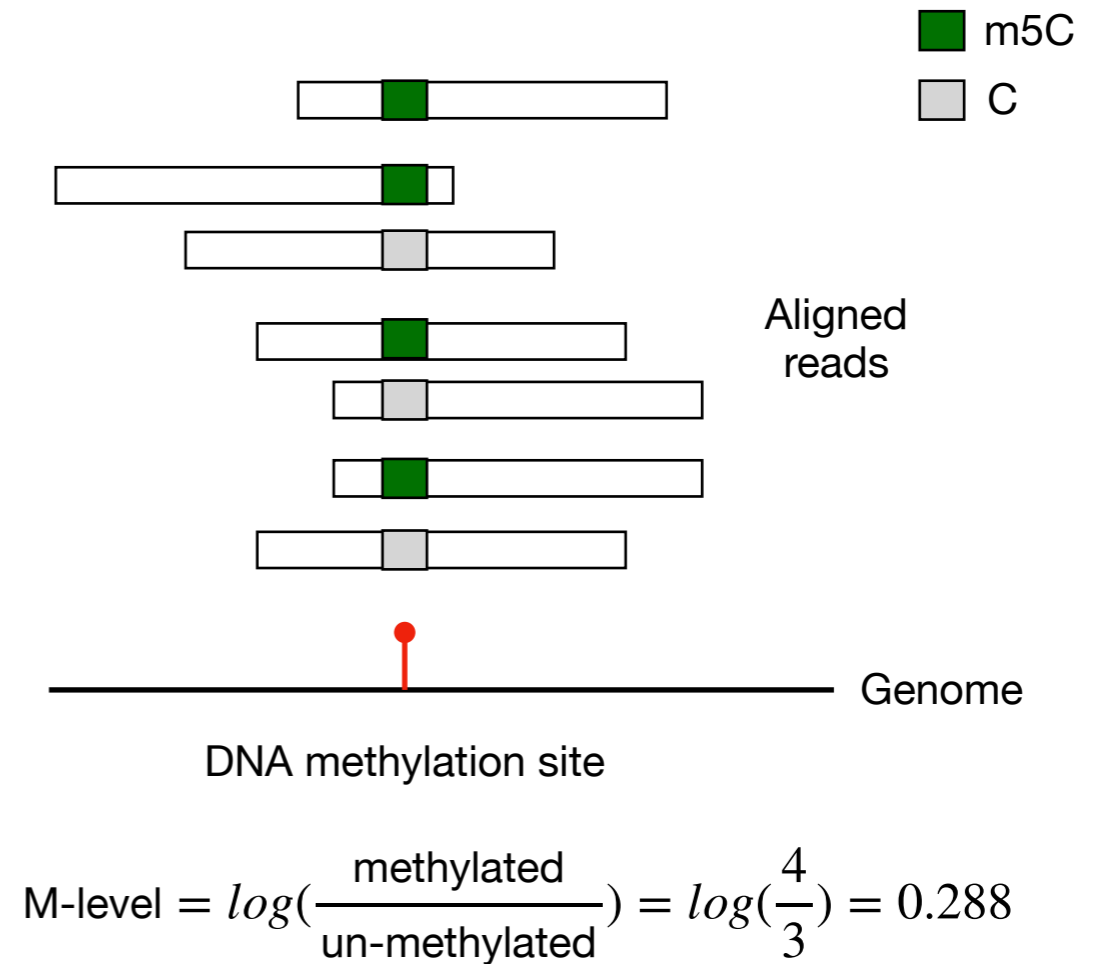
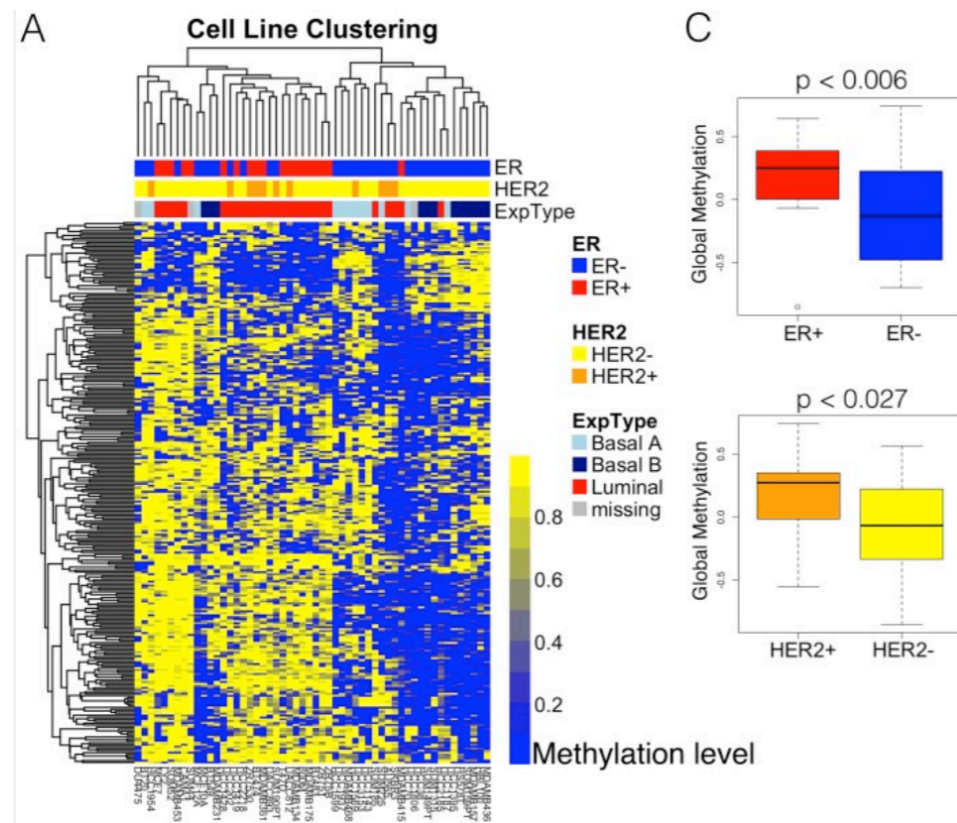
# Ratio based quantities



- The log of ratio between read counts is often used in functional genomics and epigenetics to represent meaningful quantities.
- For instance, the **log fold change** estimate is used to measure how much a gene's expression level has changed across two conditions.
- **log odds** estimate is used to measure the abundance of an epigenetic site in a given condition.

# Ratio based quantities

Clustering analysis based on DNA methylation level



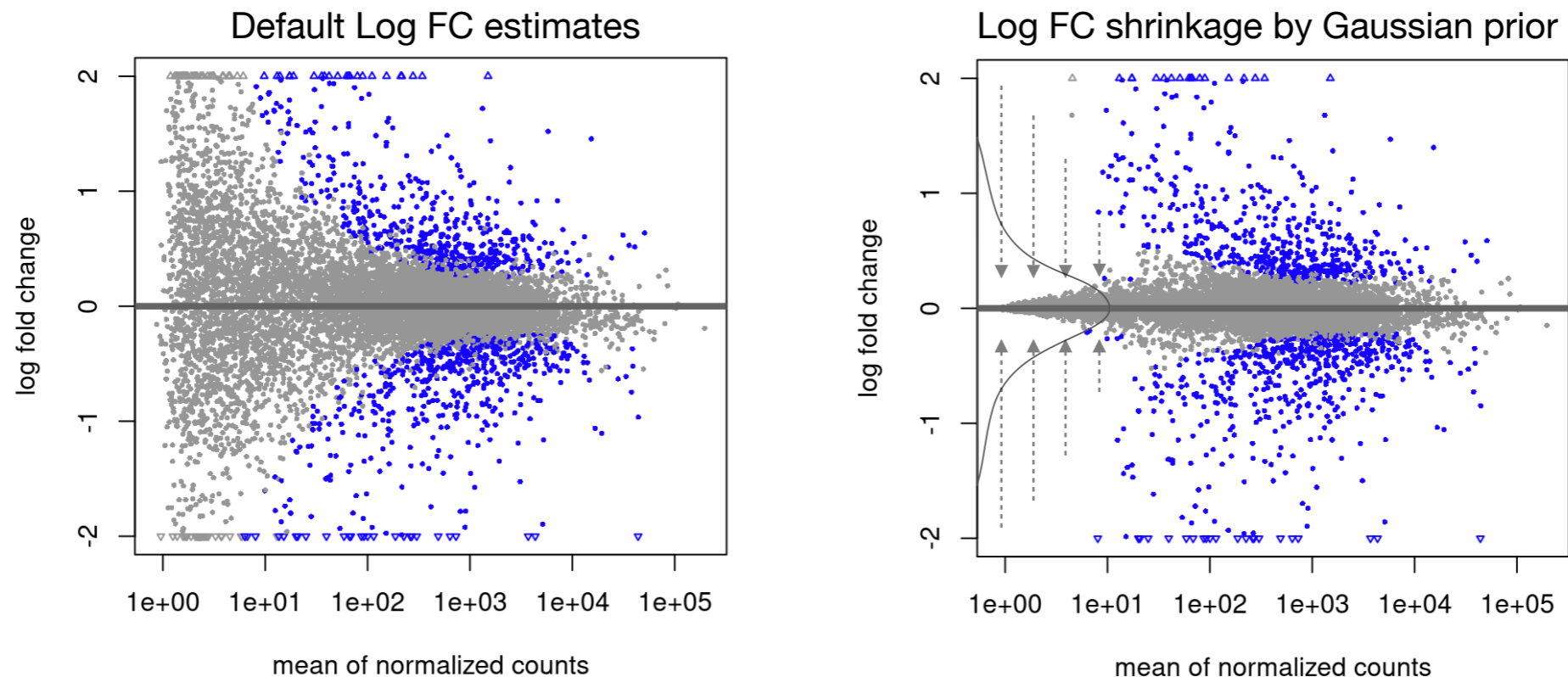
## Log Odds

- **Methylation level / M-level**:  $\log(\text{methylated read count} / \text{unmethylated read count})$  over methylation sites in bisulfite sequencing
- **DBP enrichment level**:  $\log(\text{IP read count} / \text{input read count})$  over peaks in CHIP-Seq

## Log Fold Changes

- **Differential gene expression effect size**:  $\log(\text{treatment read count} / \text{control read count})$  over genes in RNA-Seq

# Shrinkage estimator for ratio



- One critical challenge of log fold change estimates is the high estimation noise (standard error) when counts are small (typically  $\leq 10$ ).
- Therefore, low-count genes or epigenetic sites are often filtered out or treated as missing values in down stream analysis.
- A bayesian solution to reduce statistical noise in low count regions is **empirical Bayes shrinkage**, which is implemented by R packages such as DESeq2, ashR, and apeglm ...