1222·2022 800 ANNI

UNIVERSITÀ DEGLI STUDI DI PADOVA

CSAMA 2022 - BRIXEN/BRESSANONE

INTRODUCTION TO RNA-SEQ

Davide Risso
@drisso1893
@drisso

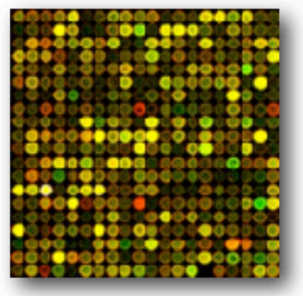
Some slides of this lecture reuse material from “DNA sequencing” by Ben Langmead (Johns Hopkins University) released under CC-BY licence

Some slides of this lecture reuse material developed by Gabriele Sales (University of Padova) courtesy of the author

SEQUENCING TECHNOLOGIES

Sanger DNA sequencing

1977-1990s



DNA Microarrays

Since mid-1990s

2nd-generation DNA sequencing

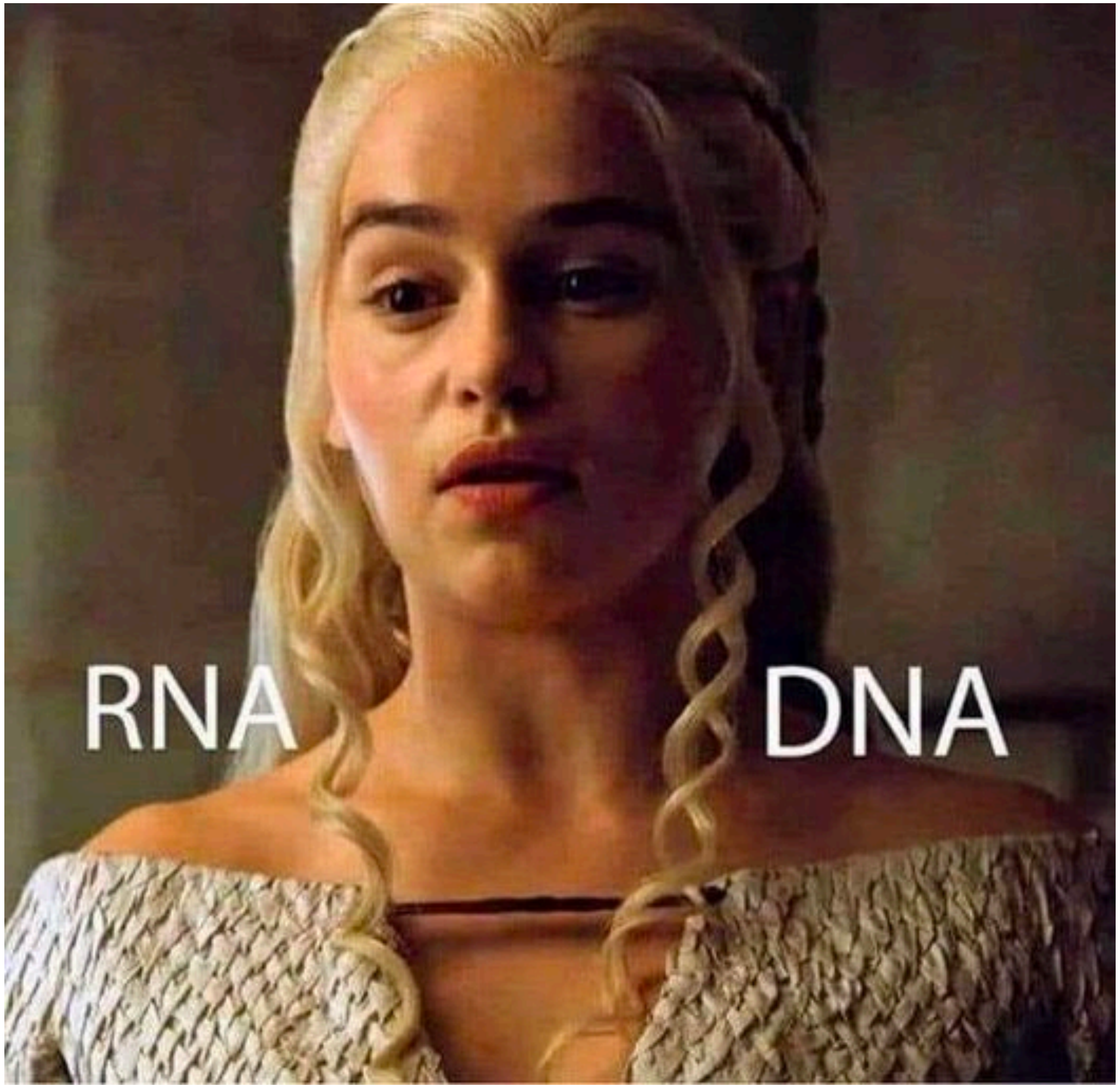
Since ~2007

3rd-generation & single-molecule DNA sequencing

Since ~2010

- **Second generation**
  - Millions of reads per sample
  - Each read ~100-300 bp
  - Very low error rates
- **Third generation**
  - Much longer reads: up to full RNA molecule
  - Not as many reads
  - Much higher error rates

RNA SEQUENCING

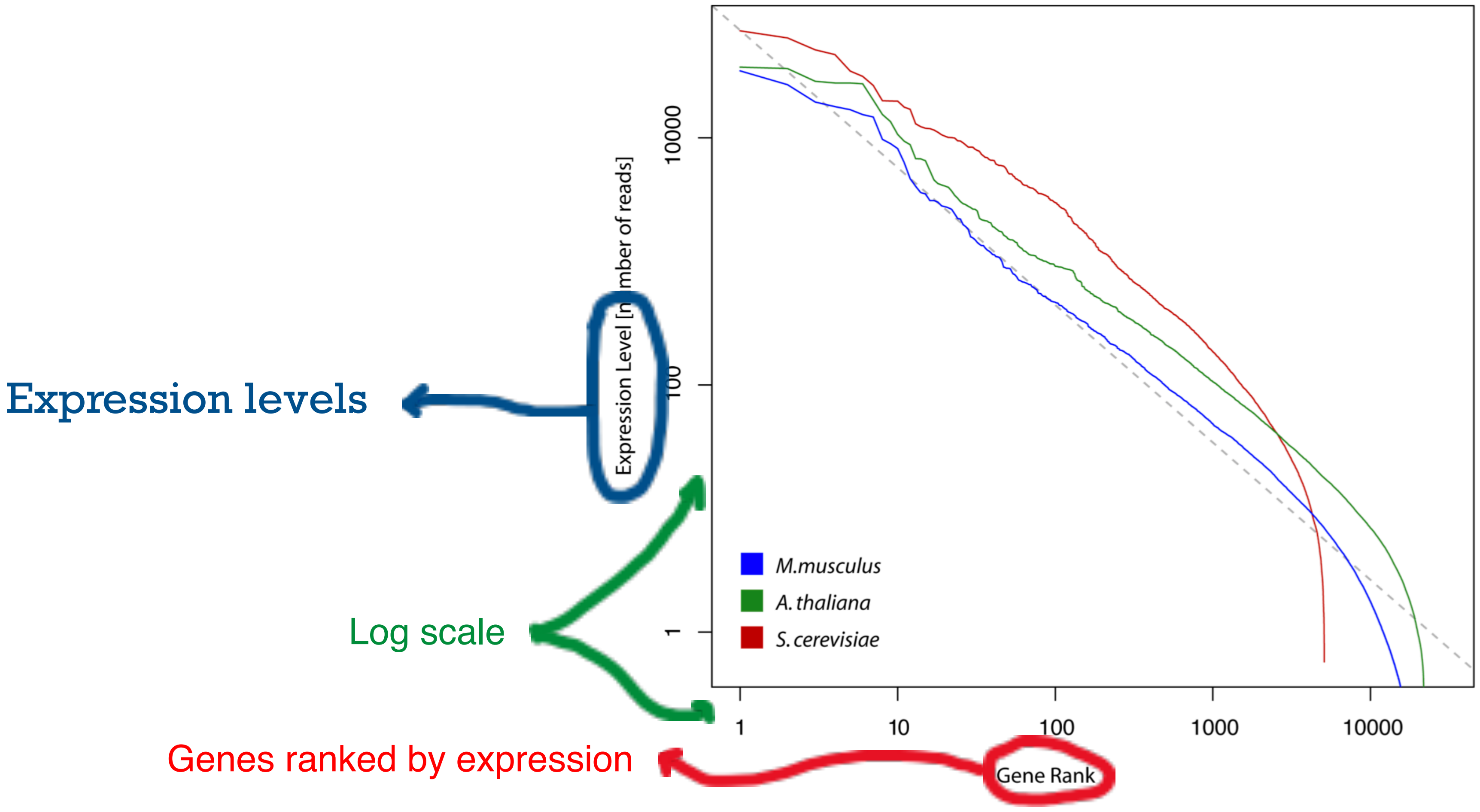


https://bioinformaticsonline.com/file/view/42693/dna-rna-meme

RNA-SEQ

- Compared to DNA sequencing, RNA sequencing is more challenging:

1. While with DNA sequencing it is reasonable to assume a uniform coverage of the genome, this is not the case for the transcriptome.



Expression levels

Log scale

Genes ranked by expression

Few genes with many reads and many genes with few reads (Zipf Law).

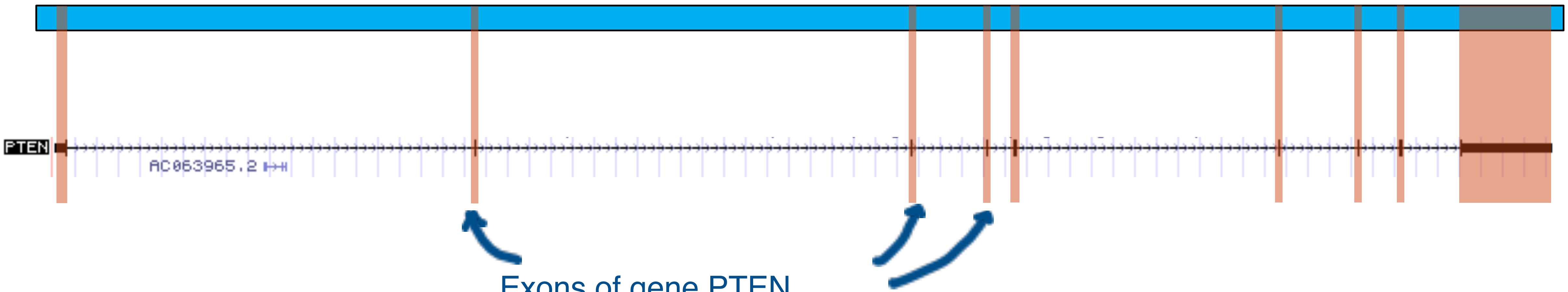
Griebel *et al*, NAR 2012

RNA-SEQ

- Compared to DNA sequencing, RNA sequencing is more challenging:

2. A read does not necessarily correspond to a contiguous genomic region.

Chromosome 10, Homo sapiens



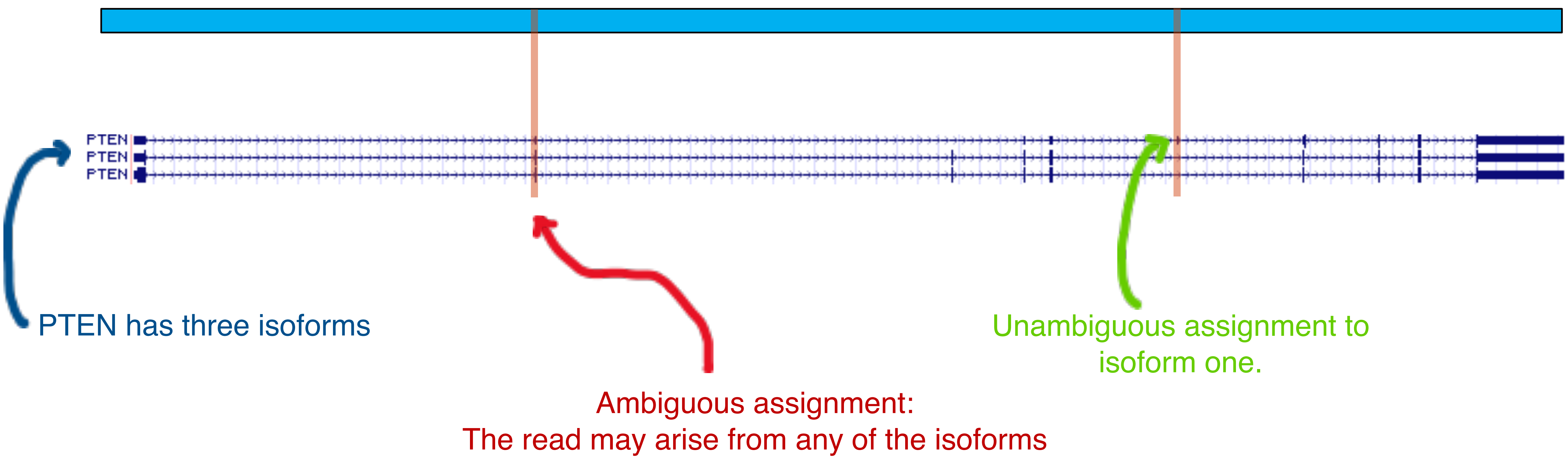
Exons of gene PTEN

RNA-SEQ

- Compared to DNA sequencing, RNA sequencing is more challenging:

3. A read can be associated to more than one transcript.

Chromosome 10, Homo sapiens



PTEN has three isoforms

Ambiguous assignment:
The read may arise from any of the isoforms

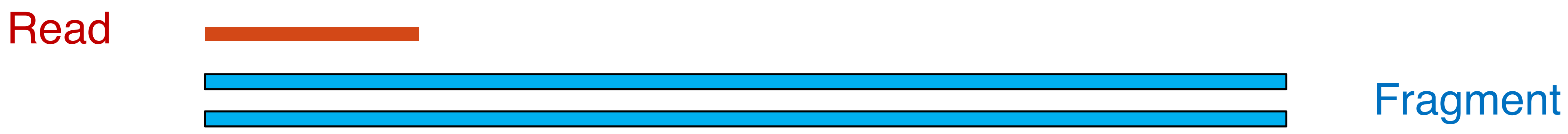
Unambiguous assignment to
isoform one.

LIBRARY PREPARATION PROTOCOLS

- One advantage of Illumina sequencing is its versatility.
- Different types of libraries can be used depending on the biological question at hand.

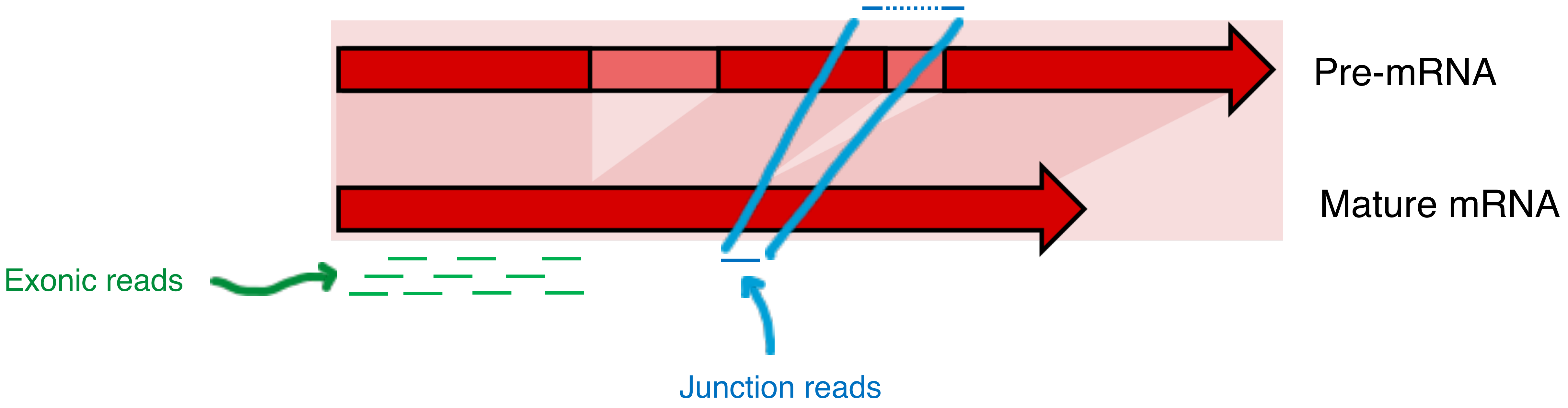
Single-end sequencing.

We sequence only one of the two ends of each fragment of cDNA.



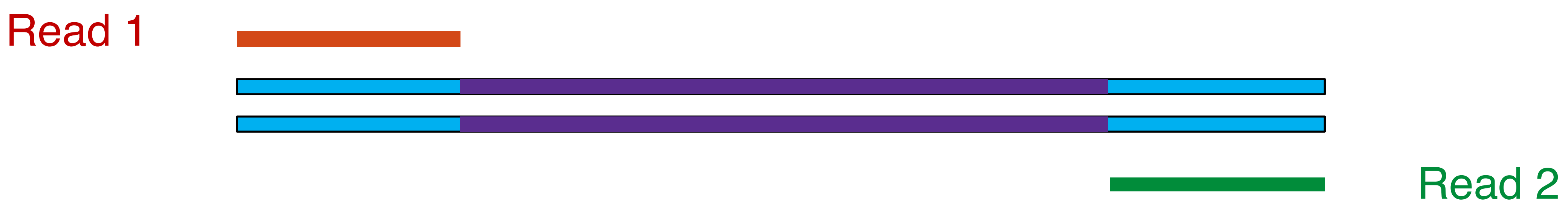
ALTERNATIVE SPLICING

- Single-end sequencing provide short-range information (100-200 bp), while alternative splicing can involve long exons.
- To quantify isoform expression levels, we need reads that map to exon-exon junctions.
- Only a small fraction of reads will map to *splice junctions*.



PAIRED-END SEQUENCING

- **Paired-end** sequencing allows us to simultaneously measure both ends of each fragment.

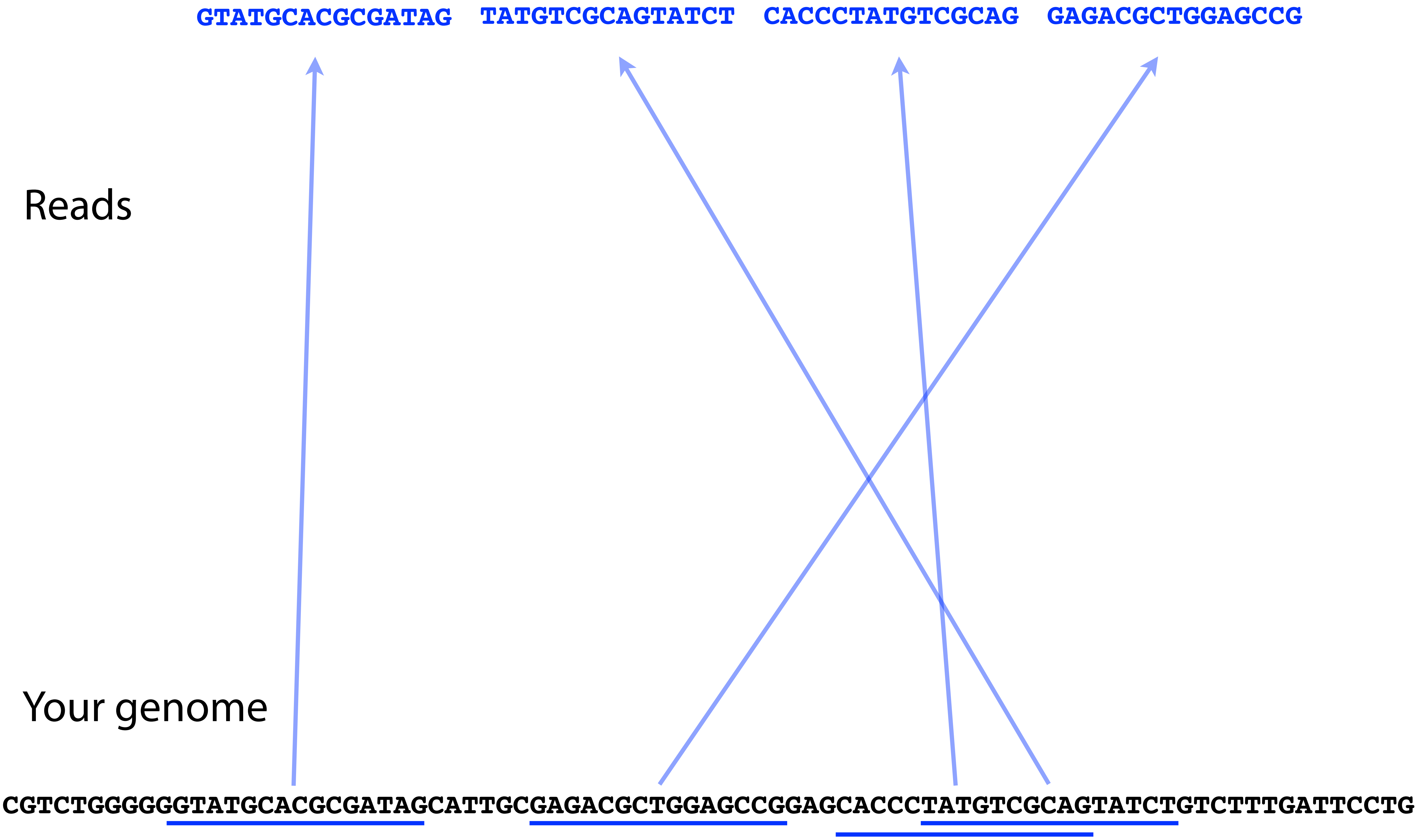


- Often the two reads do not “touch” each other
  - We simply ignore the internal sequence.
  - However, we can infer the relative position of the reads from the average fragment length.

ALIGNMENT

READ ALIGNMENT

- Since the technology allows to sequence only **short reads**, it is not straightforward to understand where the reads come from in the genome.
- A necessary step, called **alignment**, maps the reads to their origin in either the genome or the transcriptome.
- Once we have aligned the reads, we need to quantify gene expression by "counting" how many reads mapped to a given gene.
- The **counts** are our estimate of the **gene expression level**.

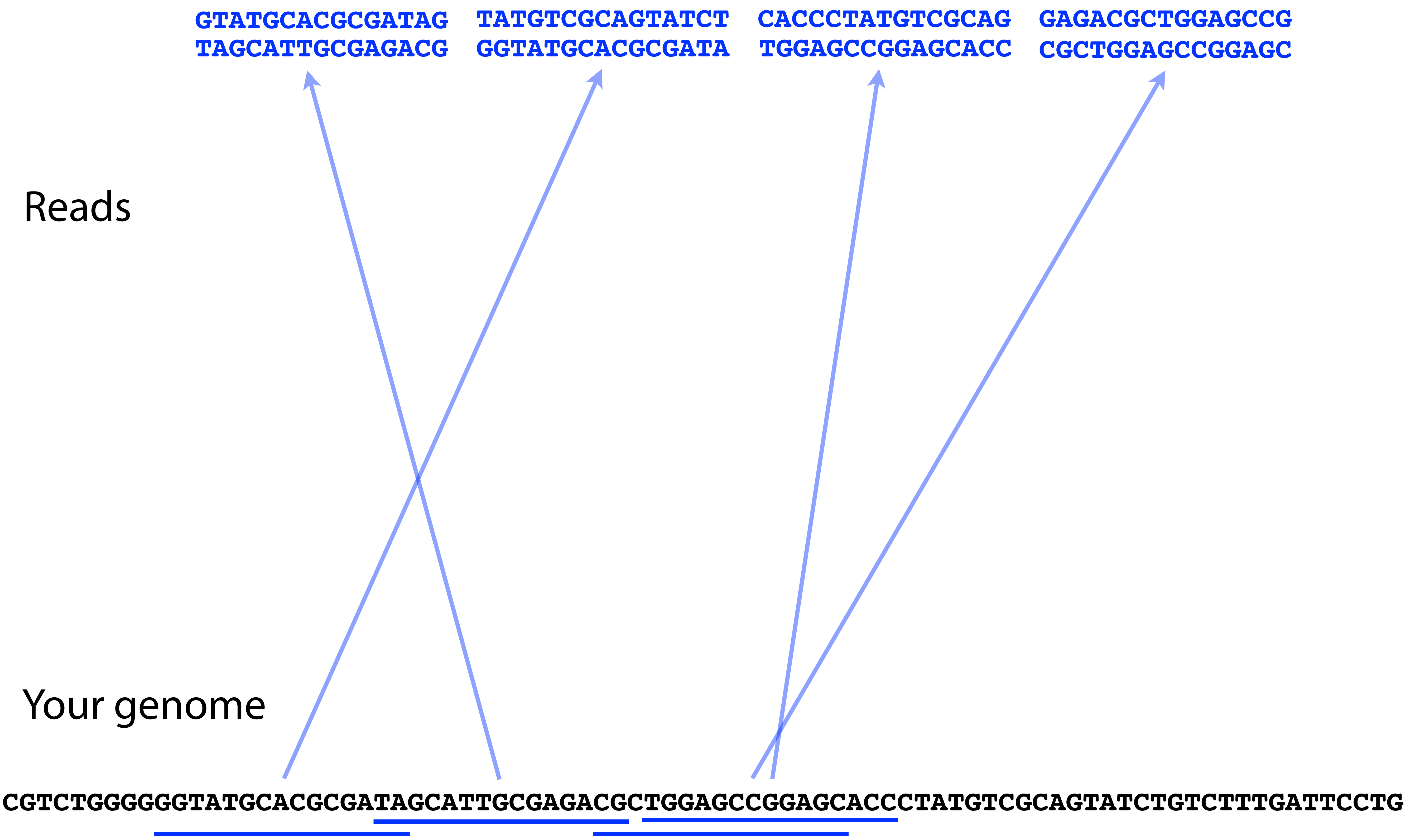


GTATGCACGCGATAG TATGTCGCAGTATCT CACCCTATGTCGCAG GAGACGCTGGAGCCG

Reads

Your genome

CGTCTGGGGGGTATGCACGCGATAGCATTGCGAGACGCTGGAGCCGGAGCACCCTATGTCGCAGTATCTGTCTTTGATTCCTG



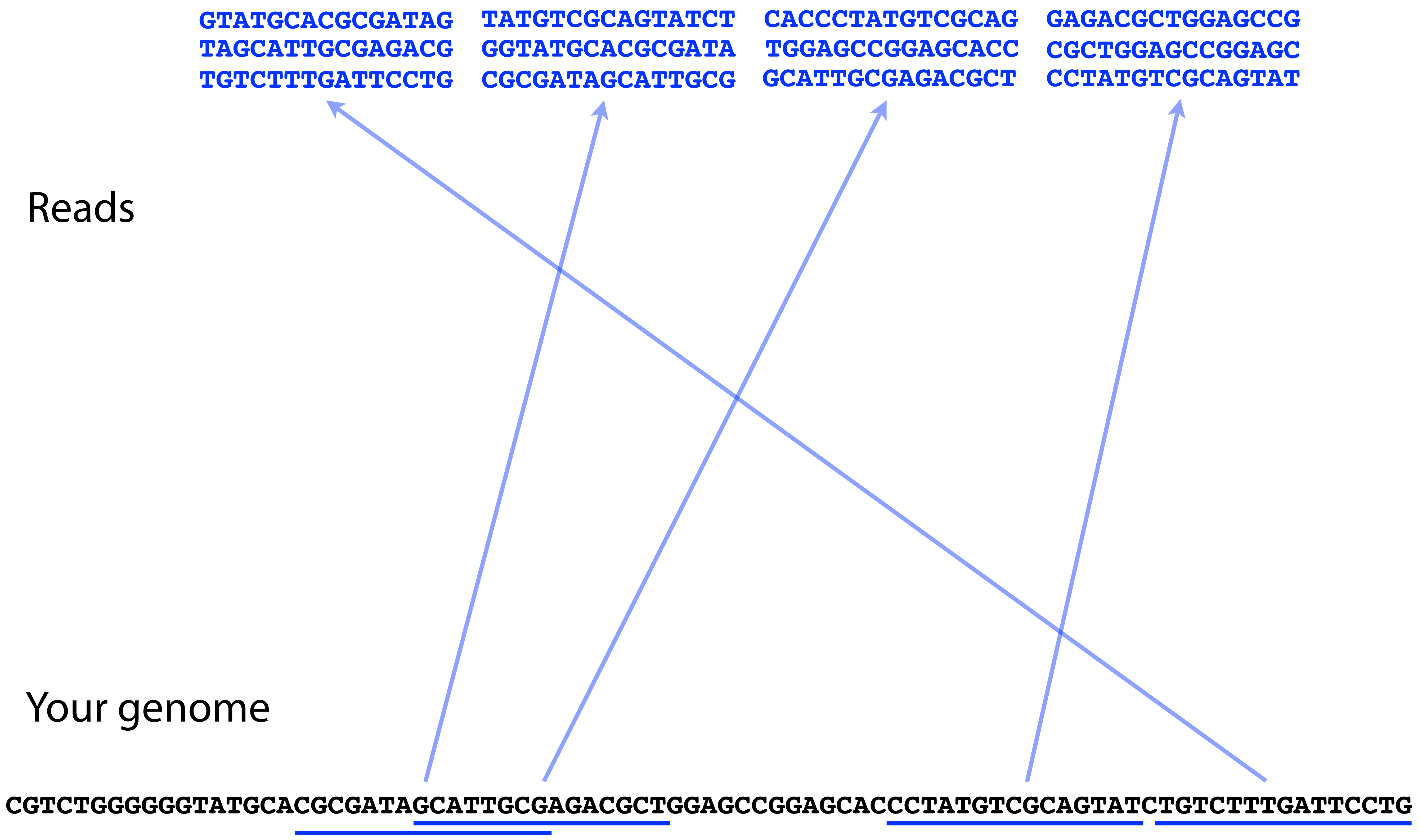
GTATGCACGCGATAG TATGTCGCAGTATCT CACCCTATGTCGCAG GAGACGCTGGAGCCG

TAGCATTGCGAGACG GGTATGCACGCGATA TGGAGCCGGAGCACC CGCTGGAGCCGGAGC

Reads

Your genome

CGTCTGGGGGGTATGCACGCGATAGCATTGCGAGACGCTGGAGCCGGAGCACCCTATGTCGCAGTATCTGTCTTTGATTCCTG



GTATGCACGCGATAG
TAGCATTGCGAGACG
TGTCTTTGATTCCTG

TATGTCGCAGTATCT
GGTATGCACGCGATA
CGCGATAGCATTGCG

CACCCTATGTCGCAG
TGGAGCCGGAGCACC
GCATTGCGAGACGCT

GAGACGCTGGAGCCG
CGCTGGAGCCGGAGC
CCTATGTCGCAGTAT

Reads

Your genome

CGTCTGGGGGGTATGCACGCGATAGCATTGCGAGACGCTGGAGCCGGAGCACCCTATGTCGCAGTATCTGTCTTTGATTCCTG

Reads

GTATGCACGCGATAG
TAGCATTGCGAGACG
TGTCTTTGATTCCTG
GACGCTGGAGCCGGA
TATCGCACCTACGTT
CACGGGAGCTCTCCA
GTATGCACGCGATAG
GCGAGACGCTGGAGC
CCTACGTTCAATATT
GACGCTGGAGCCGGA
TATCGCACCTACGTT
CACGGGAGCTCTCCA

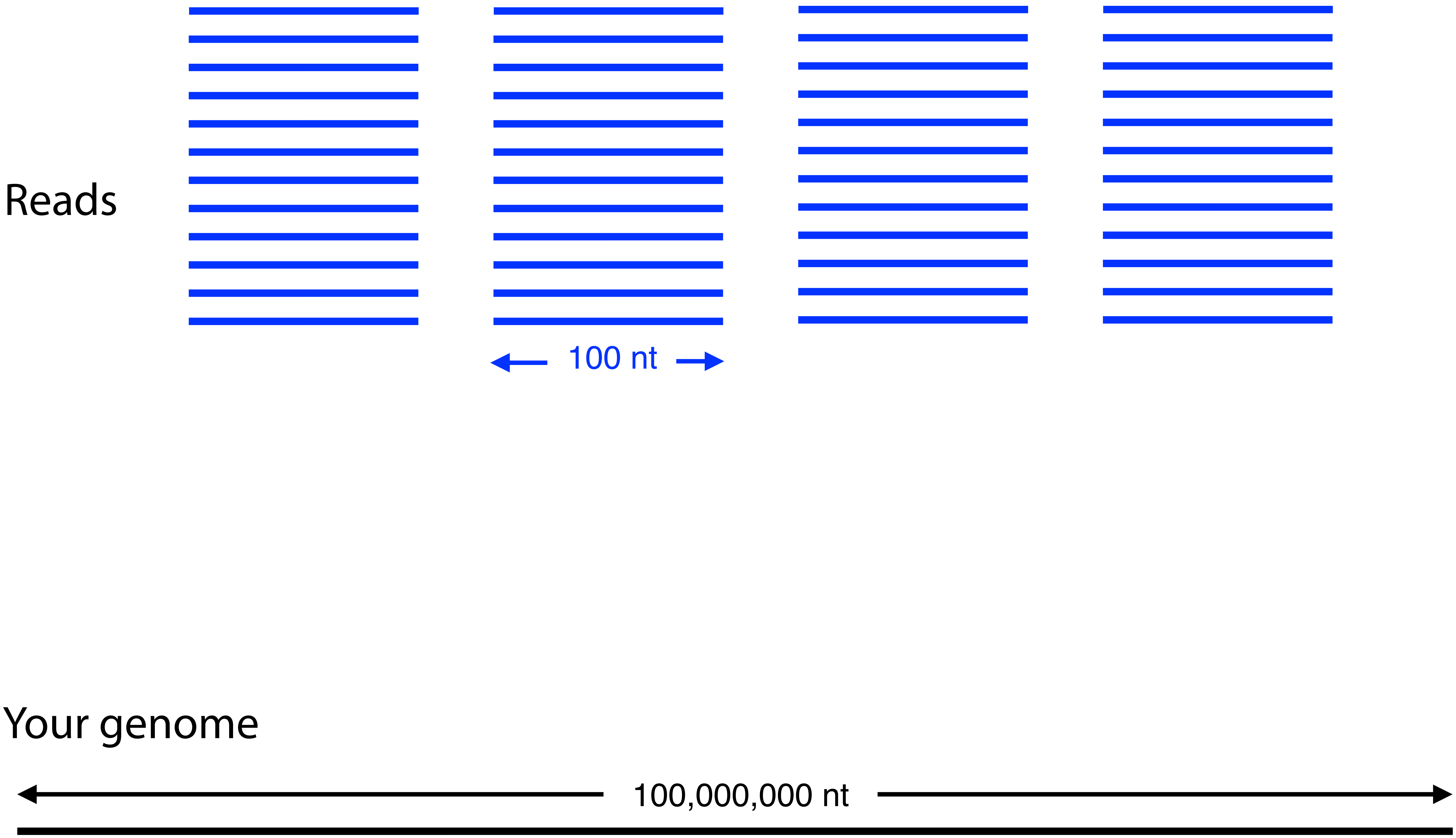
TATGTCGCAGTATCT
GGTATGCACGCGATA
CGCGATAGCATTGCG
GCACCCTATGTCGCA
CAATATTCGATCATG
TGCATTTGGTATTTT
ACCTACGTTCAATAT
CTATCACCCTATTAA
GCACCTACGTTCAAT
GCACCCTATGTCGCA
CAATATTCGATCATG
TGCATTTGGTATTTT

CACCCTATGTCGCAG
TGGAGCCGGAGCACC
GCATTGCGAGACGCT
GTATCTGTCTTTGAT
GATCACAGGTCTATC
CGTCTGGGGGGTATG
TATTTATCGCACCTA
CTGTCTTTGATTCCT
GTCTGGGGGGTATGC
GTATCTGTCTTTGAT
GATCACAGGTCTATC
CGTCTGGGGGGTATG

GAGACGCTGGAGCCG
CGCTGGAGCCGGAGC
CCTATGTCGCAGTAT
CCTCATCCTATTATT
ACCCTATTAACCACT
CACGCGATAGCATTG
CCACTCACGGGAGCT
ACTCACGGGAGCTCT
AGCCGGAGCACCCTA
CCTCATCCTATTATT
ACCCTATTAACCACT
CACGCGATAGCATTG

Your genome

CGTCTGGGGGGTATGCACGCGATAGCATTGCGAGACGCTGGAGCCGGAGCACCCTATGTCGCAGTATCTGTCTTTGATTCCTG

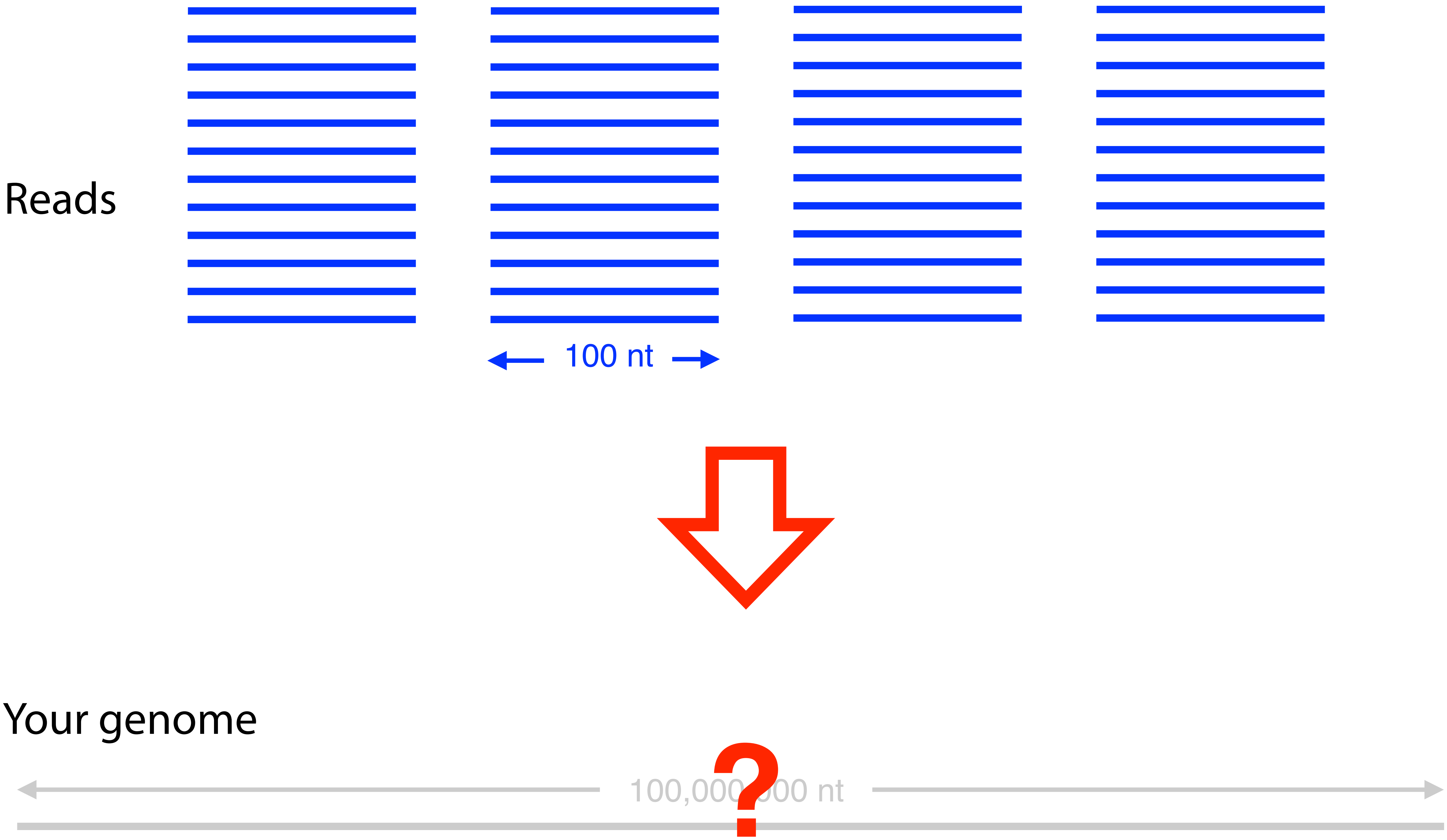


Reads

100 nt

Your genome

100,000,000 nt



Reads

100 nt

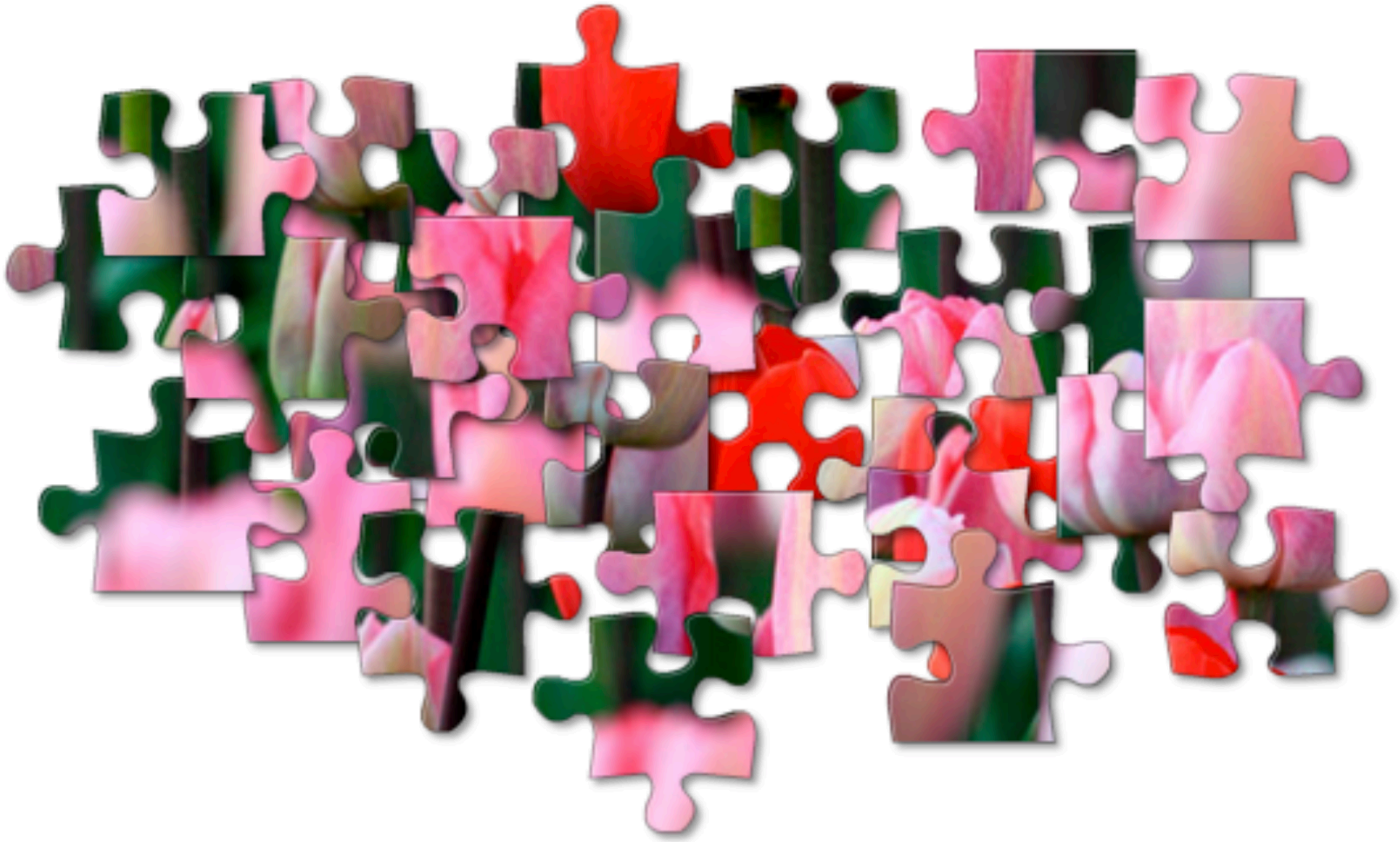
Your genome

100,000,000 nt

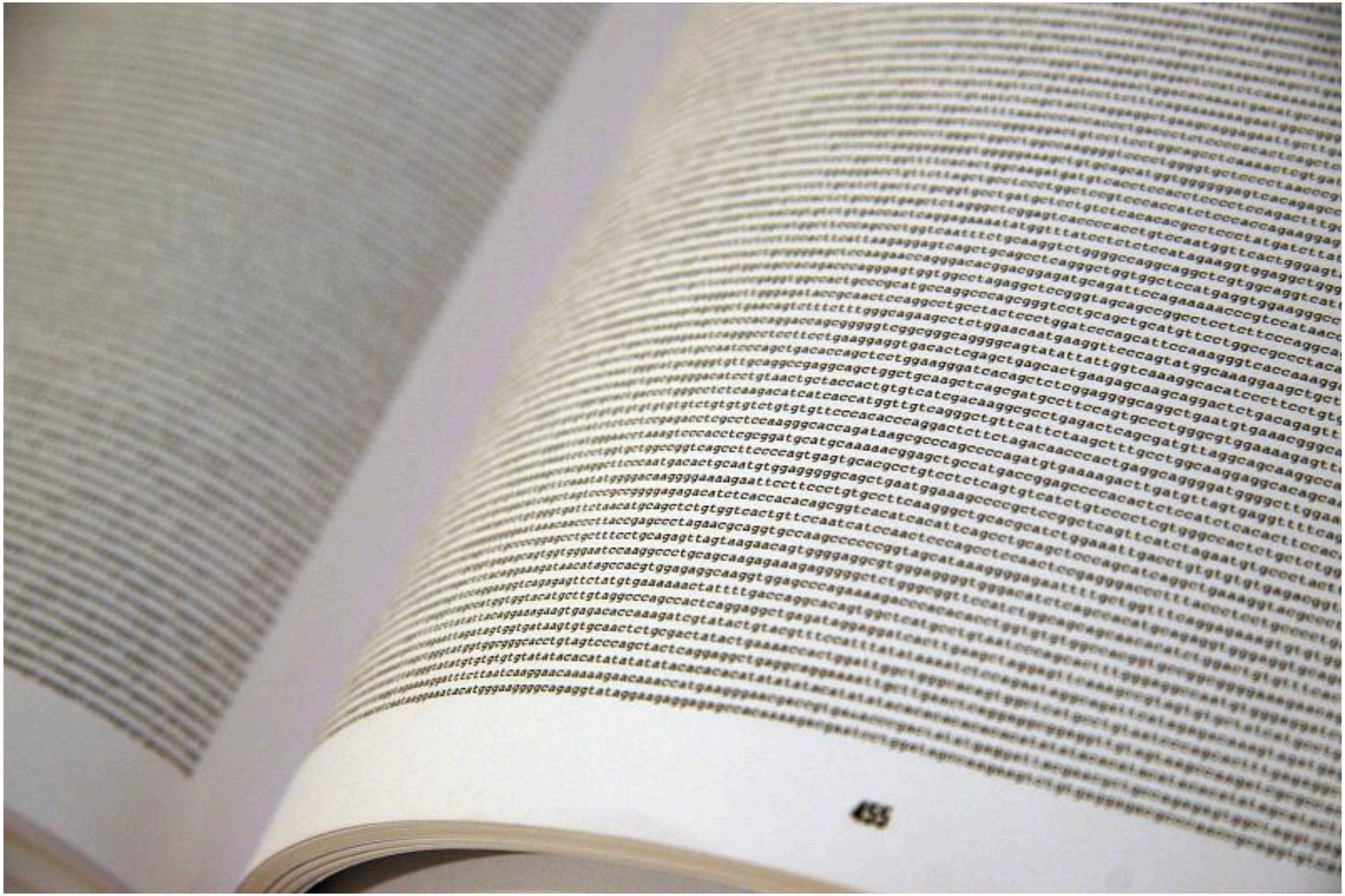
?

READS ARE CHARACTER STRINGS

```
GTATGCACGCGATAG
TAGCATTGCGAGACG
TGTCTTTGATTCCTG
GACGCTGGAGCCGGA
TATCGCACCTACGTT
CACGGGAGCTCTCCA
GTATGCACGCGATAG
GCGAGACGCTGGAGC
CCTACGTTCAATATT
GACGCTGGAGCCGGA
TATCGCACCTACGTT
CACGGGAGCTCTCCA
TATGTCGCAGTATCT
GGTATGCACGCGATA
CGCGATAGCATTGCG
GCACCCTATGTCGCA
CAATATTCGATCATG
TGCATTTGGTATTTT
ACCTACGTTCAATAT
CTATCACCCTATTAA
GCACCTACGTTCAAT
GCACCCTATGTCGCA
CAATATTCGATCATG
TGCATTTGGTATTTT
CACCCTATGTCGCAG
TGGAGCCGGAGCACC
GCATTGCGAGACGCT
GTATCTGTCTTTGAT
GATCACAGGTCTATC
CGTCTGGGGGGTATG
TATTTATCGCACCTA
CTGTCTTTGATTCCT
GTCTGGGGGGTATGC
GTATCTGTCTTTGAT
GATCACAGGTCTATC
CGTCTGGGGGGTATG
```

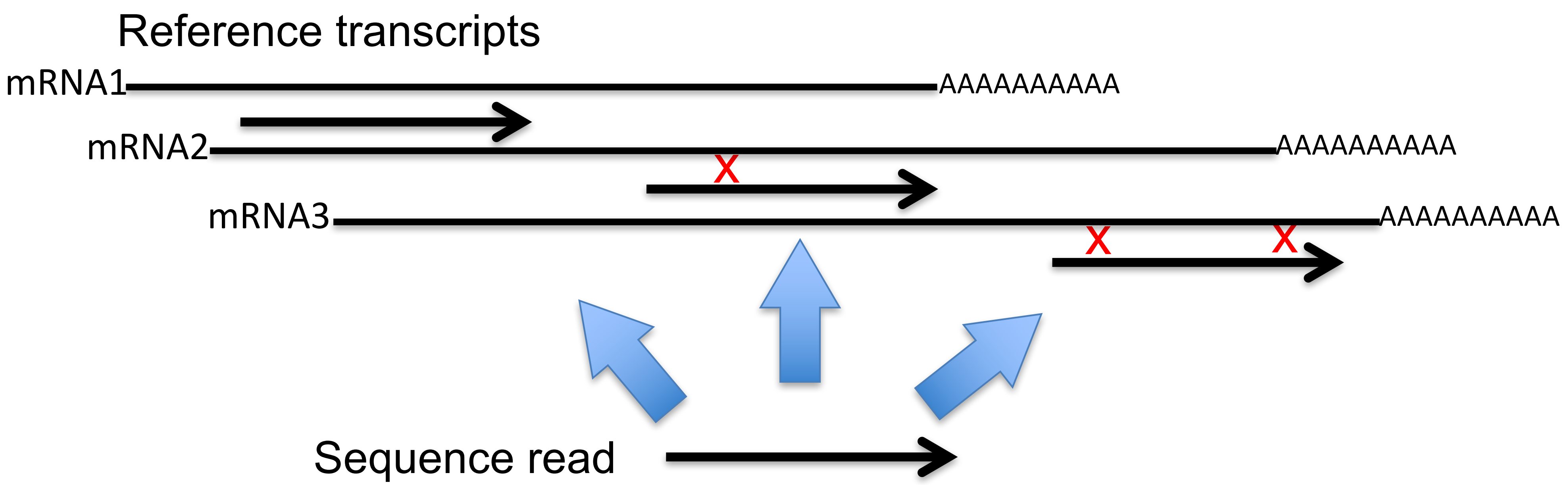


- Reads are character strings
- The character sequence is the only information that we have on the origin of the reads.
- Like a jigsaw puzzle, we need to reconstruct the picture from individual pieces.



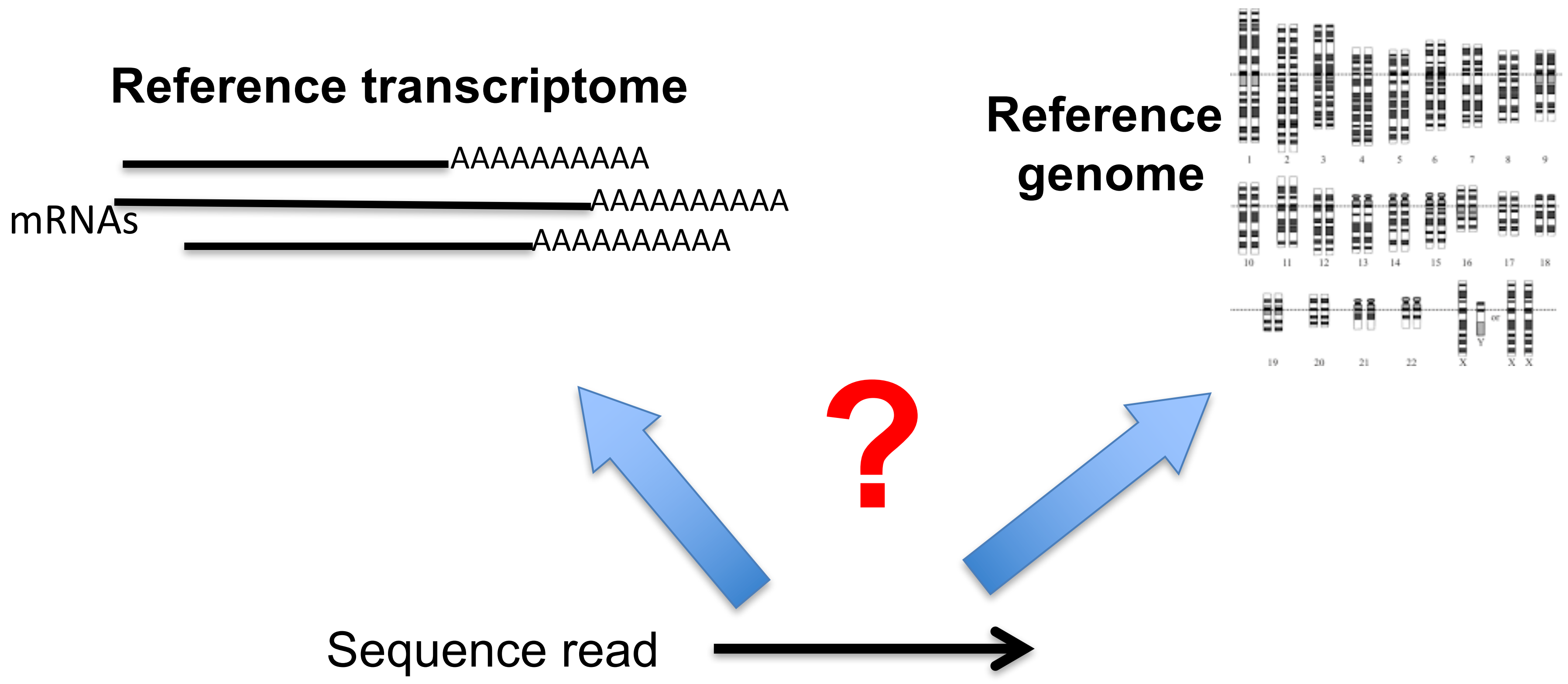
- Many algorithms have been developed in the computer science literature to solve this problem.

DIFFICULTIES



- An alignment algorithm must support mismatches.
- Mismatches are due to either **sequencing errors** or **mutations**.

GENOME OR TRANSCRIPTOME ALIGNMENT?

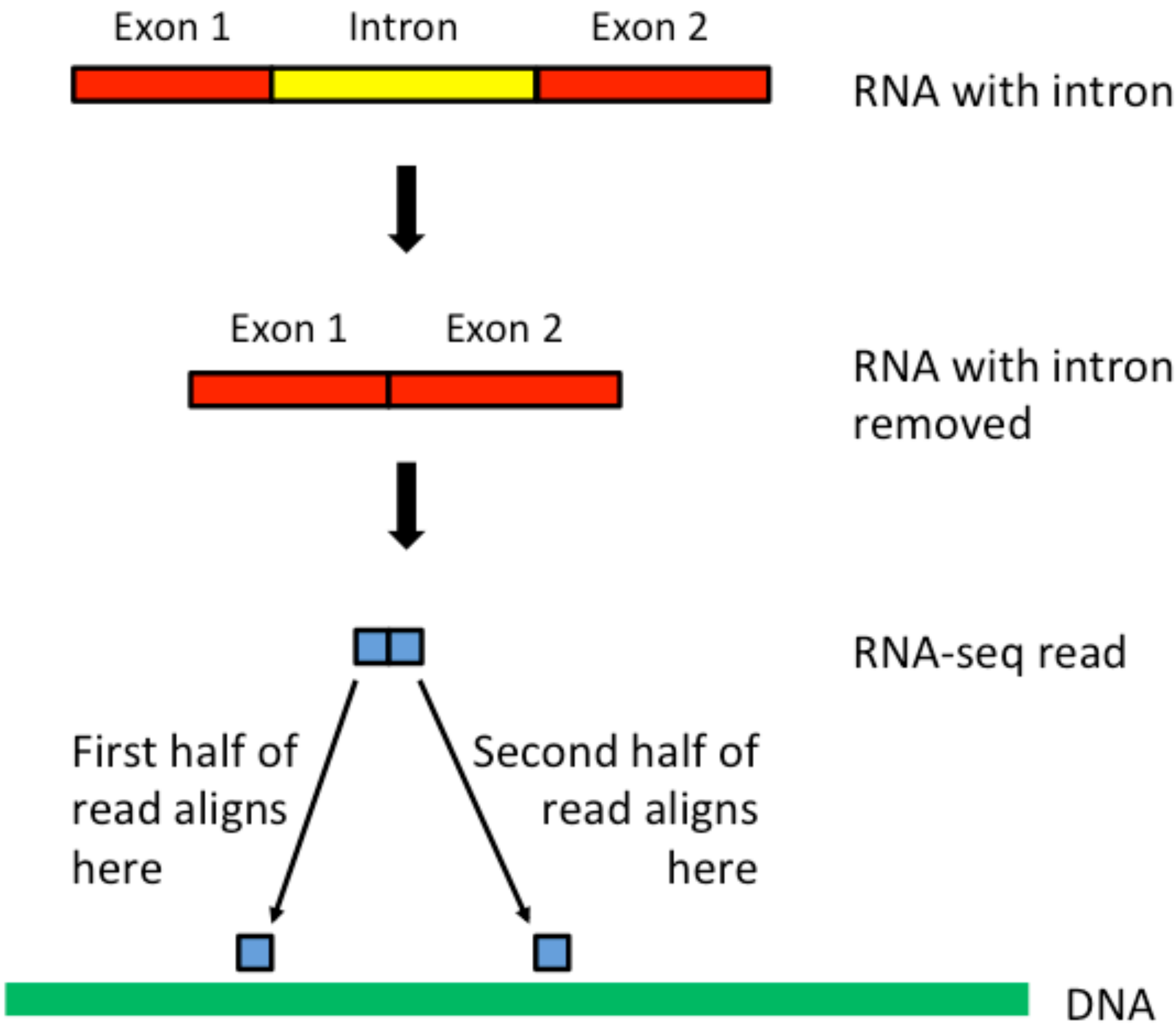


- Reads can be aligned either to the genome or the transcriptome, i.e., the set of all transcripts.
- Only about **5-10% of the genome is transcribed**; hence transcript alignment is faster computationally.
- However, because of **alternative splicing**, many transcripts share large portions of their sequences, leading to **multiply mapped reads** (o **multi-reads**), i.e., reads that map to more than one transcript.
- On the other hand, mapping reads to the genome is complicated by splicing, i.e. reads consist of non adjacent regions in the genome.

GENOME ALIGNMENT

- We do not have time to go into the algorithmic details, but many modern software packages (e.g., BWA, Bowtie) use the **Burrows-Wheeler transformation** to speed up the search for matching sequences.
- They also implement a **backtracking algorithm** to allow for mismatches.
- More details:
  - https://langmead-lab.org/teaching-materials/
  - https://kingsfordlab.cbd.cmu.edu/teaching/

GENOME ALIGNMENT (WITH SPLICING)

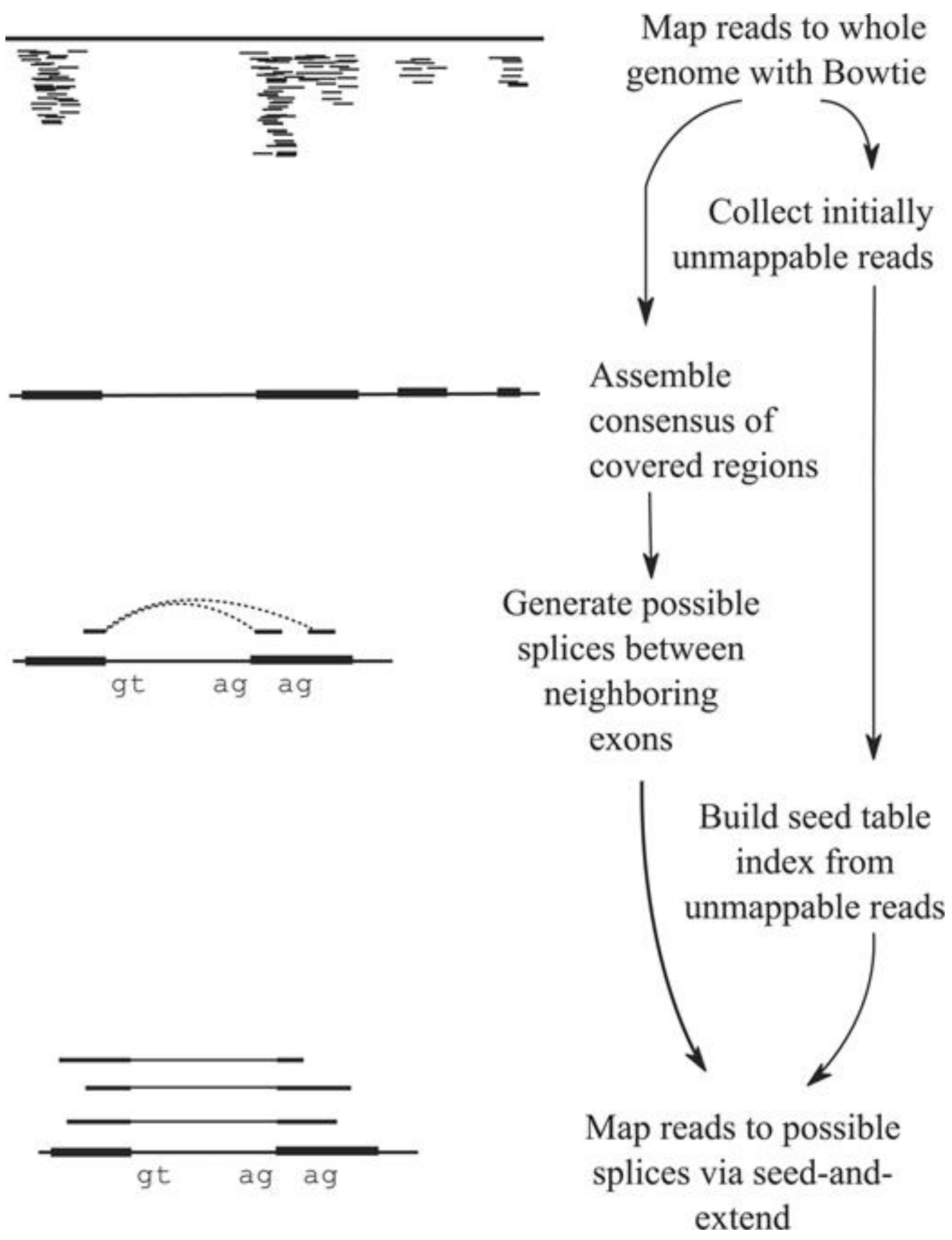


https://discoveringthegenome.org/discovering-genome/rna-sequencing-up-close-data/spliced-alignment

TOPHAT

One strategy is that employed by TopHat

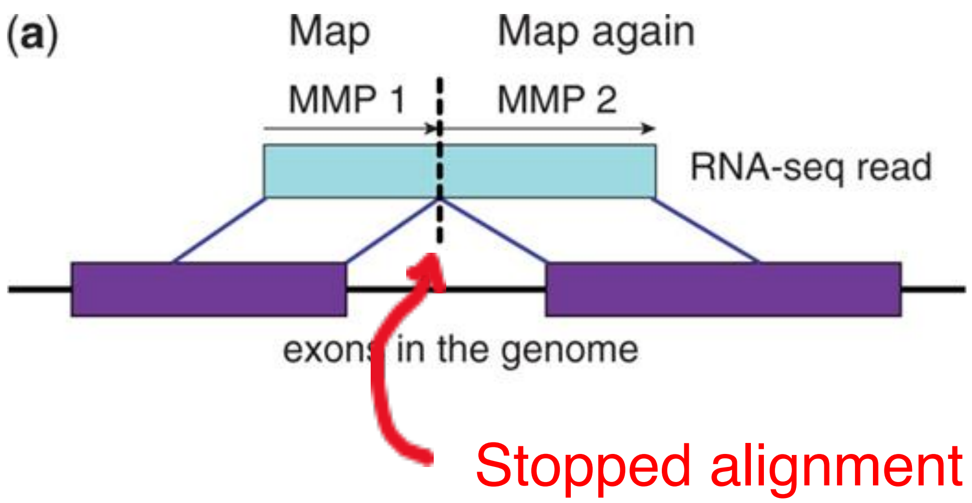
- In the first step it aligns the reads to the genome.
- It collects all the non-aligned reads (potentially caused by splicing).
- It groups the genomic regions covered by alignments in "islands".
- It enumerates all possible canonical splicing patterns (GT-AG) among islands.
- Non-aligned reads are compared to potential splicing sites.


Map reads to whole genome with Bowtie
Collect initially unmappable reads
Assemble consensus of covered regions
Generate possible splices between neighboring exons
gt
ag
ag
Build seed table index from unmappable reads
Map reads to possible splices via seed-and-extend
gt
ag
ag

Trapnell *et al*, 2009

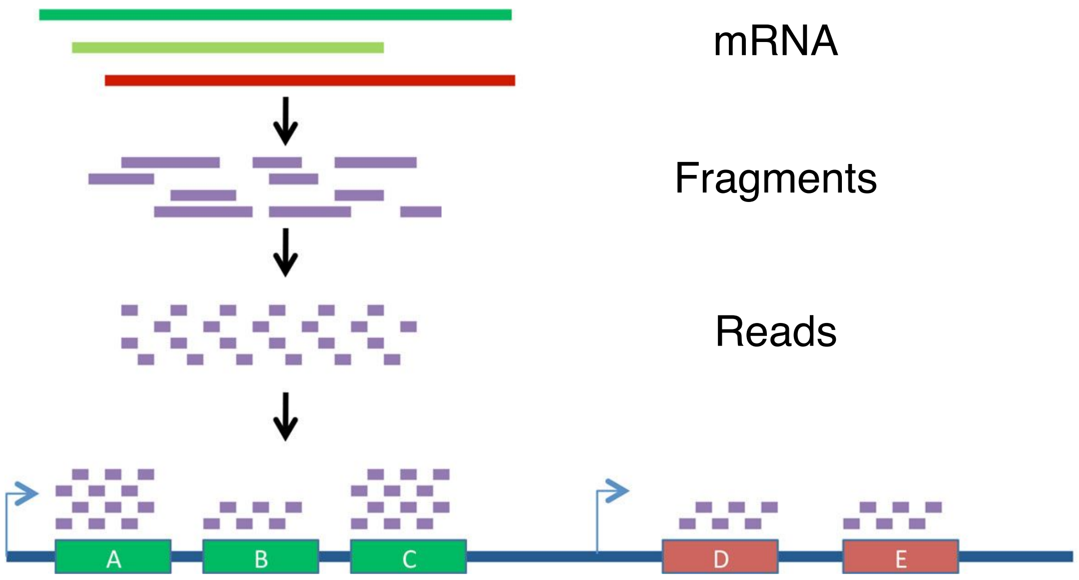
STAR

- An alternative approach is STAR.
- It searches for the Maximal Mappable Prefix (MMP) of each read against the genome.
  - In (a) the first part of the read corresponds to an exon
  - The alignment stops at the exon-intron boundary
  - The mapping is resumed for the read part not yet mapped.
- Very efficient search based on a pre-computed suffix array.



Dobin *et al*, 2013

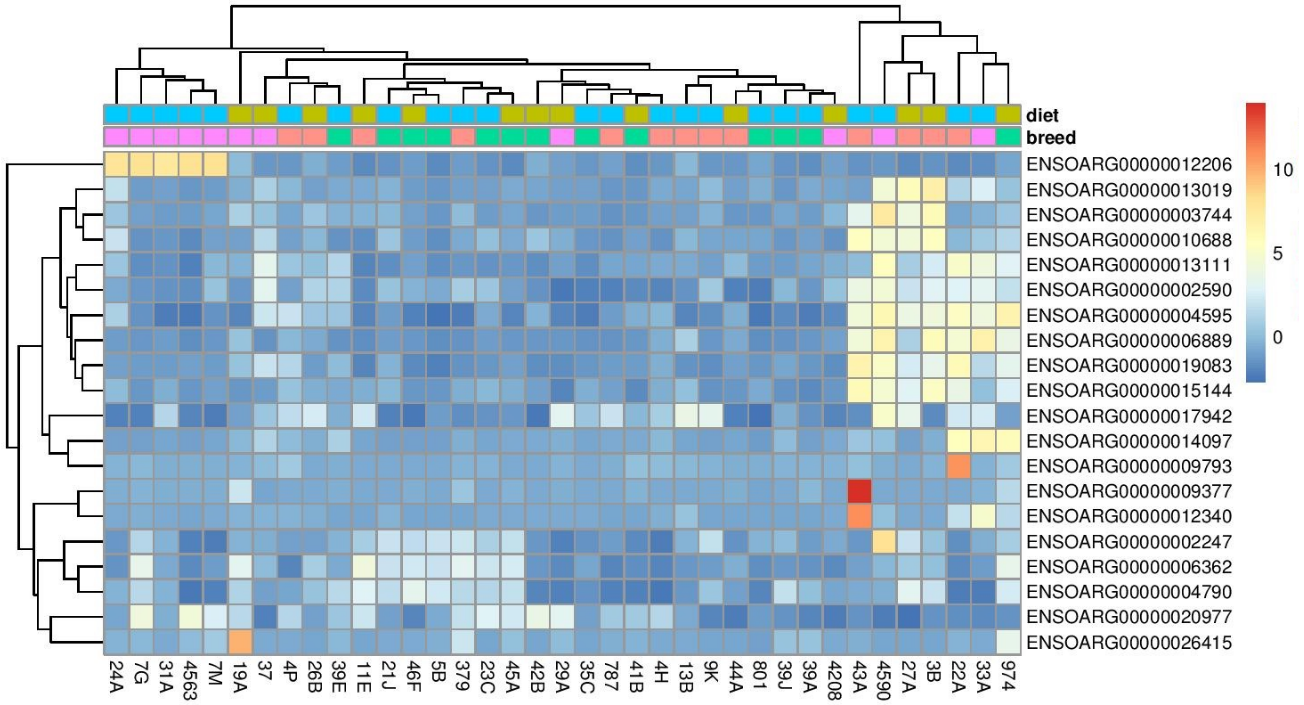
QUANTIFICATION



mRNA

Fragments

Reads

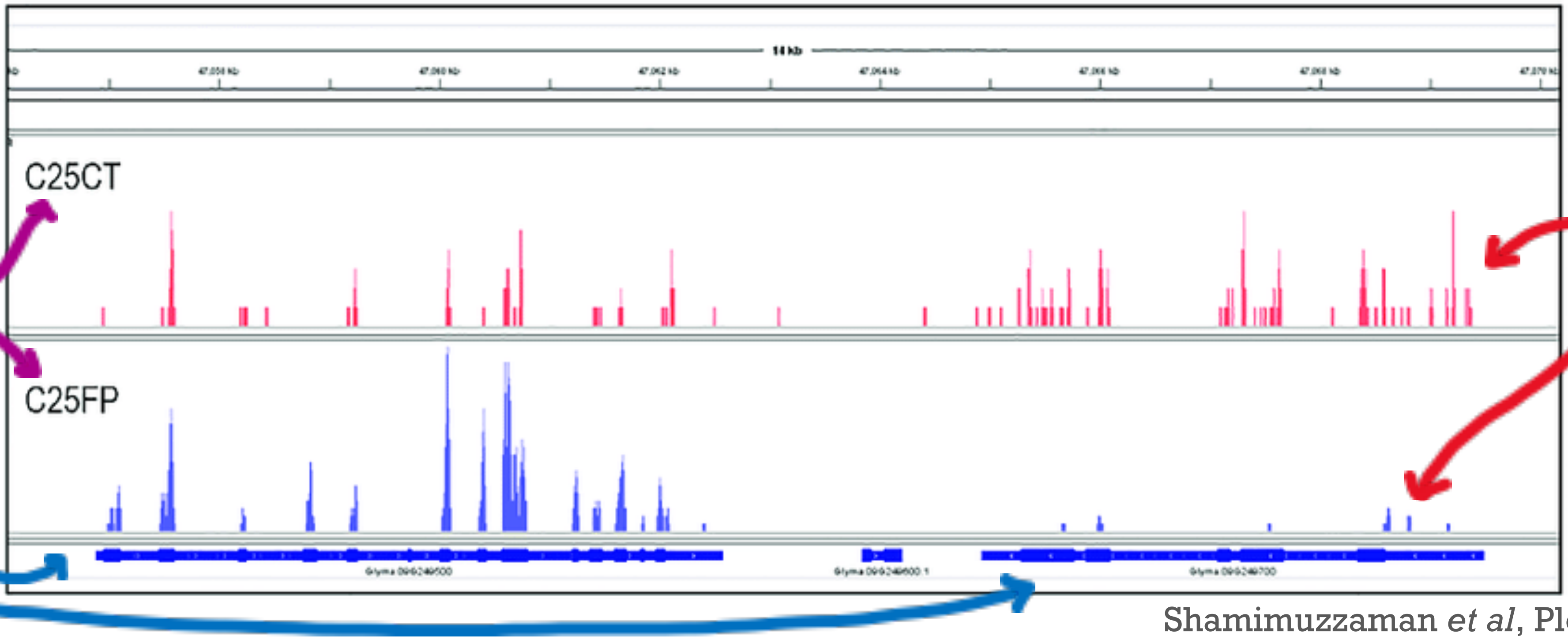


Mangul *et al*, BMC Genomics 2014

DIRECT COUNTING

The simplest method we can think of:

1. Align the reads to the genome
2. Identify regions corresponding to exons
3. Count the number of reads mapped to each exon
4. Sum the counts for all exons of a given gene



Samples

Coverage

Genes

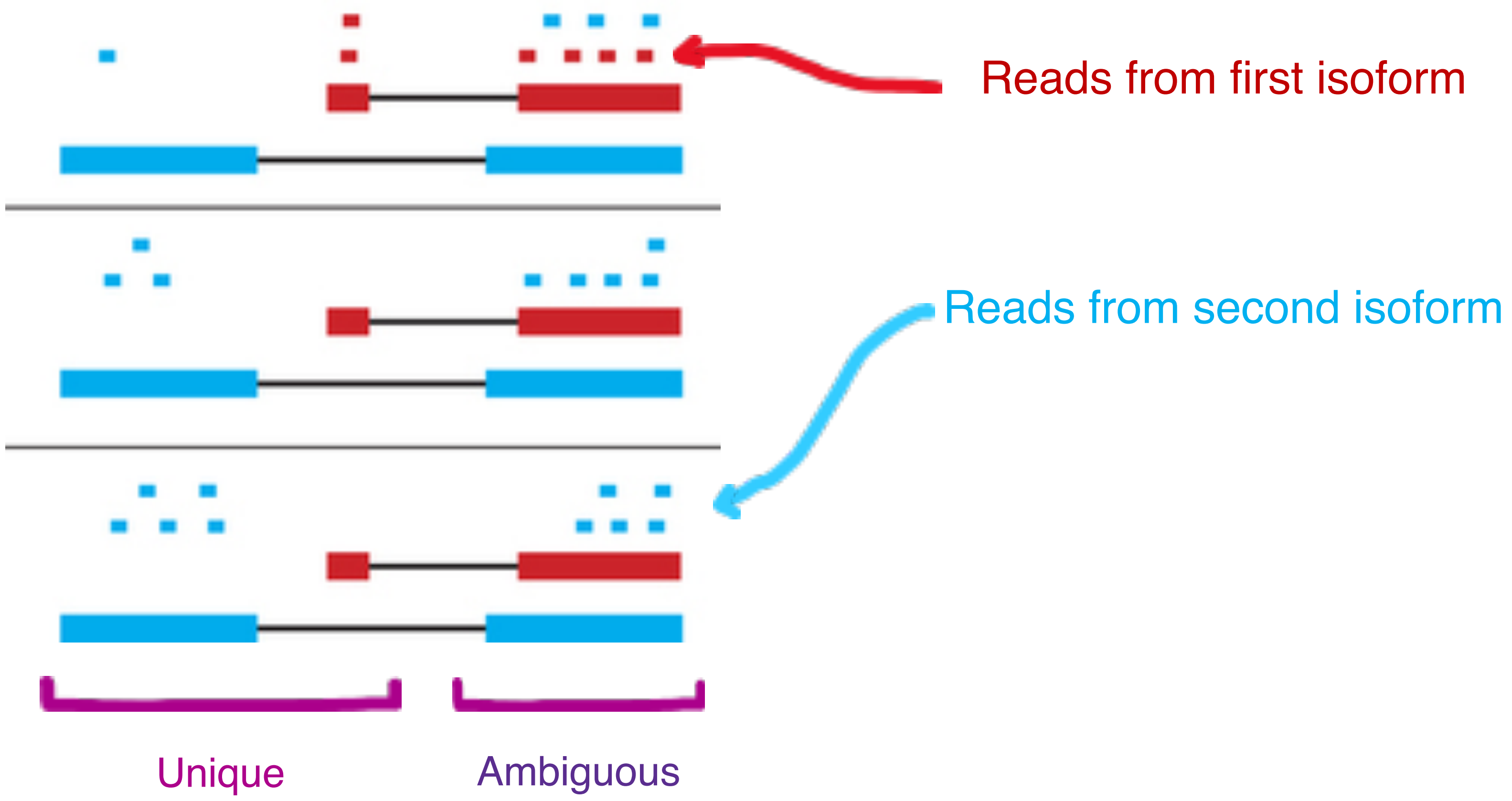
Shamimuzzaman *et al*, Plos ONE. 2018

AMBIGUOUS READS

This simple strategy is not sufficient to deal with alternative splicing.

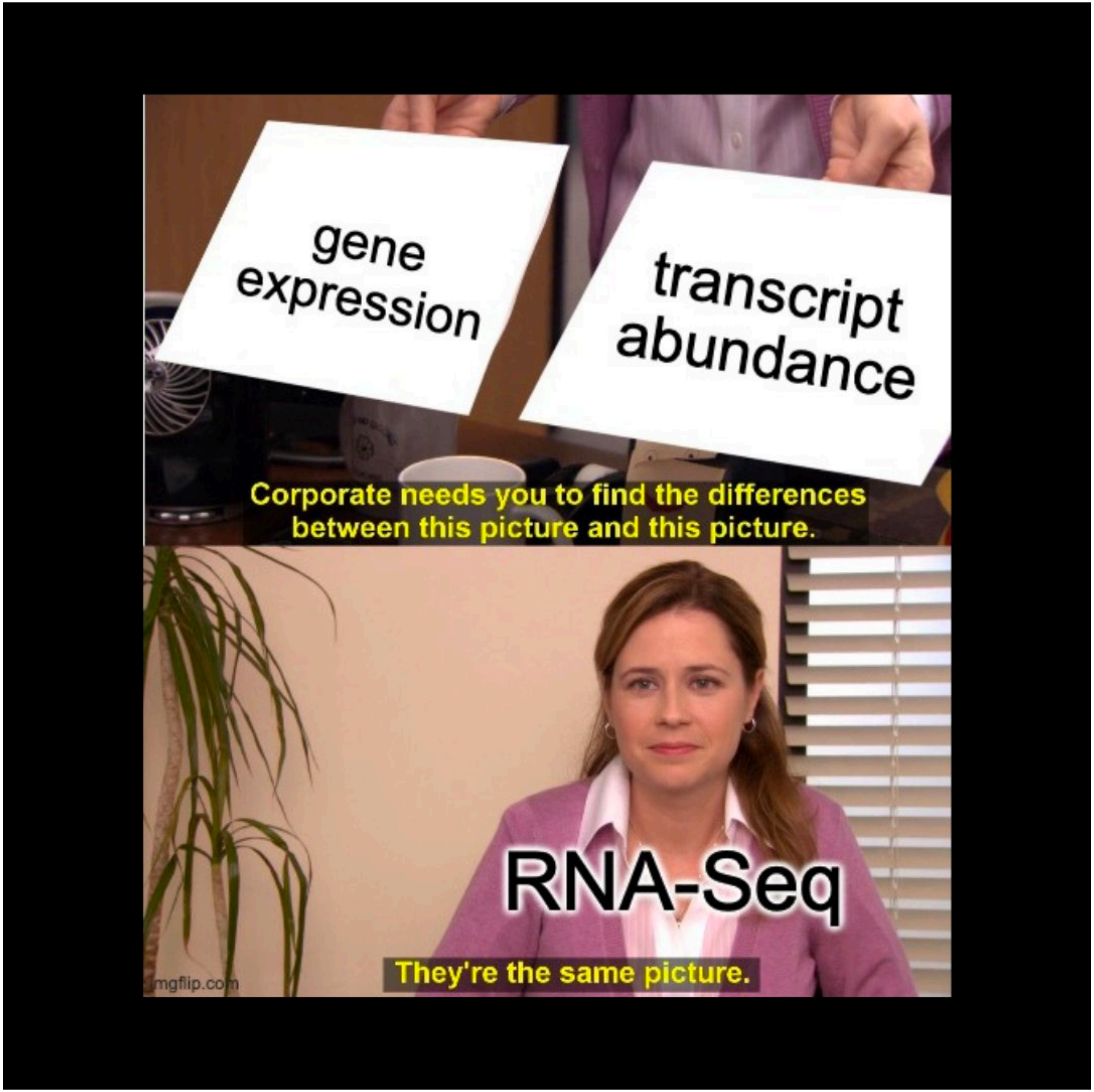
- A read can be aligned to an exon shared by more transcripts.
- In quantifying transcript expression to which isoform do we assign the read?

Three hypothetical examples



Trapnell *et al*, Nature Biotech. 2012

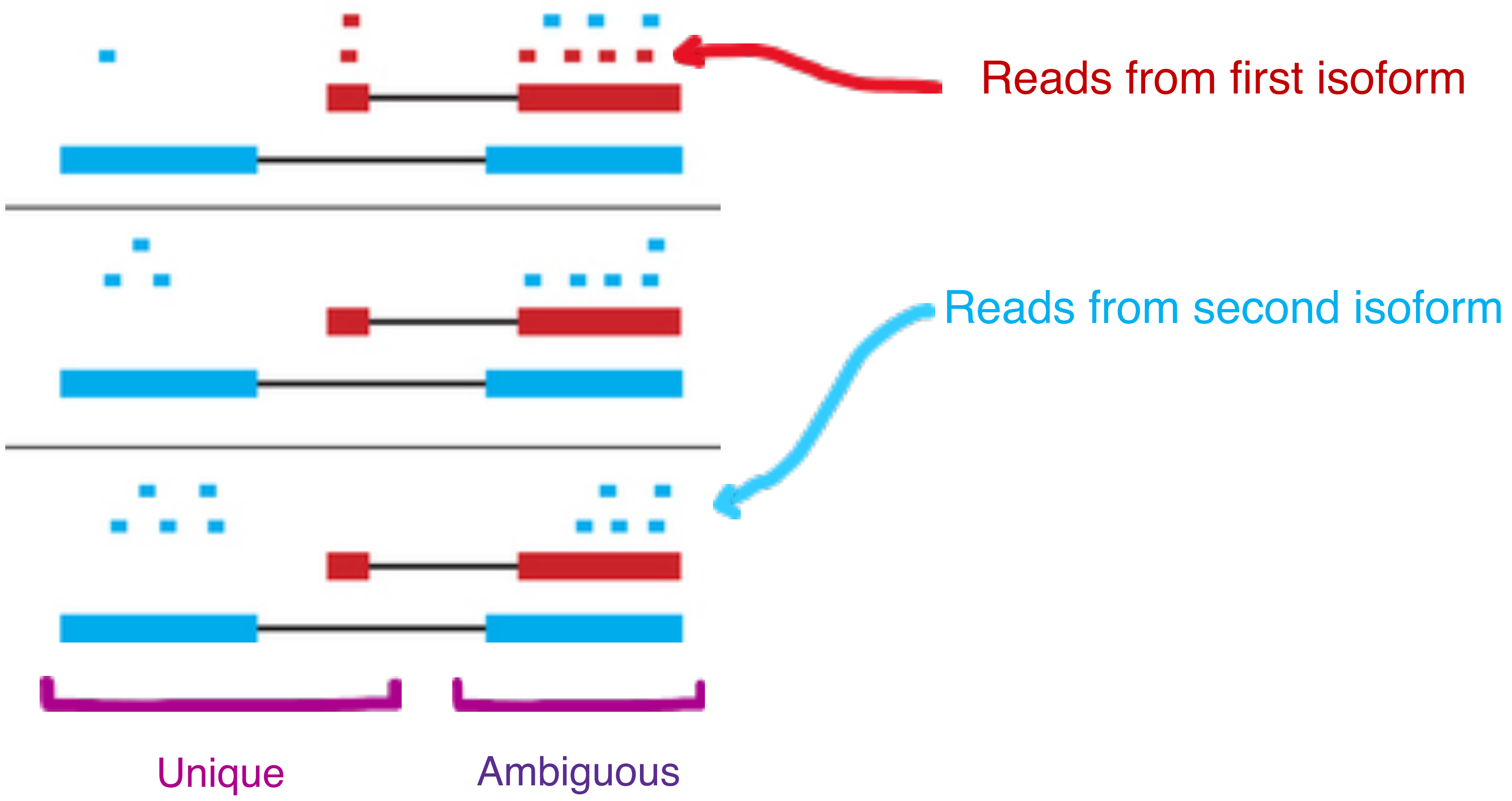
GENE- OR TRANSCRIPT-LEVEL SUMMARIES?


gene expression
transcript abundance
Corporate needs you to find the differences between this picture and this picture.
RNA-Seq
They're the same picture.
imgflip.com

A STATISTICAL MODEL

- A more proper solution is to develop a statistical approach.
- We define and estimate a set of parameters, some latent, that allow us to fully leverage the information present in the data to infer gene expression.

Three hypothetical examples



Trapnell *et al*, Nature Biotech. 2012

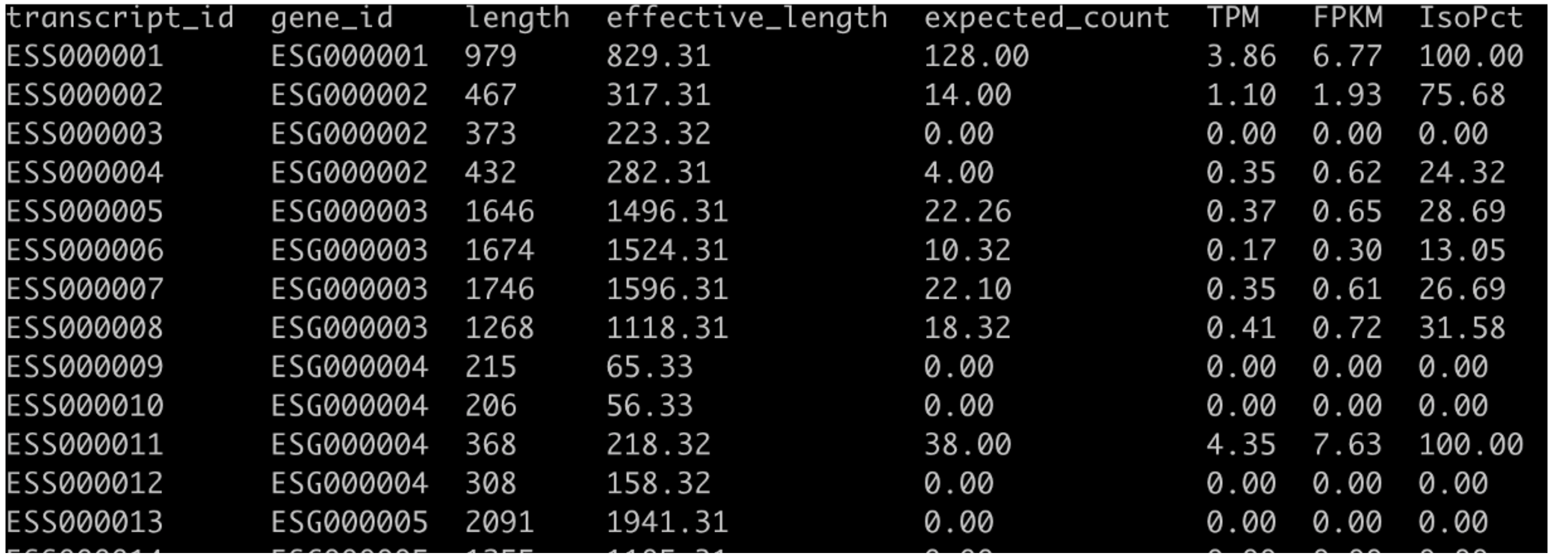
RNA-SEQ BY EXPECTATION-MAXIMIZATION (RSEM)

- An example of such approach is RSEM

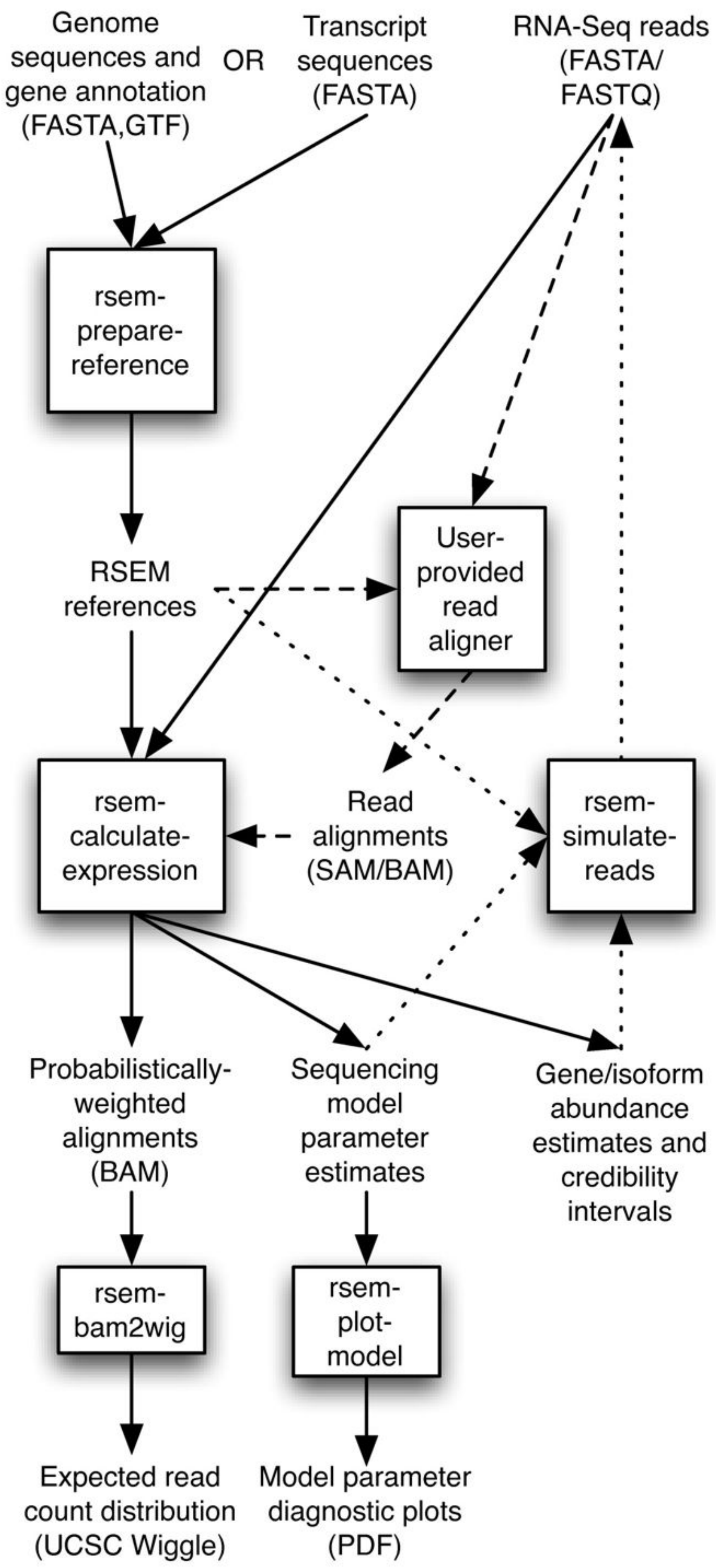
- Available as open-source software:

https://deweylab.github.io/RSEM/

- It starts from a set of aligned reads (typically aligned to the transcriptome).



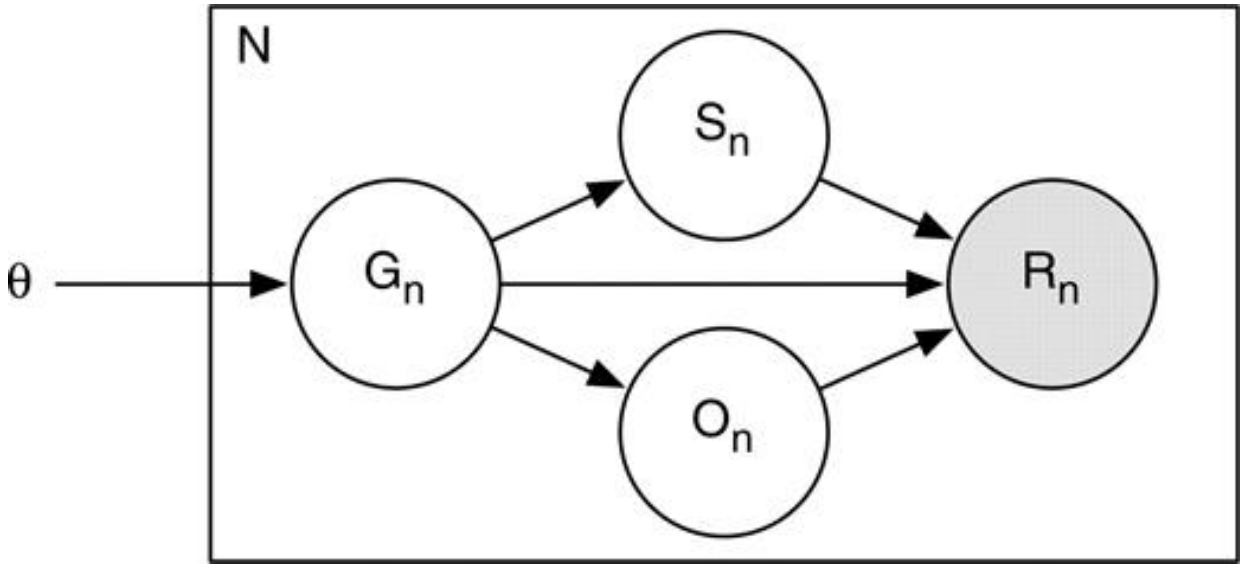
```
transcript_id  gene_id    length  effective_length  expected_count  TPM   FPKM  IsoPct
ESS000001      ESG000001  979     829.31            128.00          3.86  6.77  100.00
ESS000002      ESG000002  467     317.31            14.00           1.10  1.93  75.68
ESS000003      ESG000002  373     223.32            0.00            0.00  0.00  0.00
ESS000004      ESG000002  432     282.31            4.00            0.35  0.62  24.32
ESS000005      ESG000003  1646    1496.31           22.26           0.37  0.65  28.69
ESS000006      ESG000003  1674    1524.31           10.32           0.17  0.30  13.05
ESS000007      ESG000003  1746    1596.31           22.10           0.35  0.61  26.69
ESS000008      ESG000003  1268    1118.31           18.32           0.41  0.72  31.58
ESS000009      ESG000004  215     65.33             0.00            0.00  0.00  0.00
ESS000010      ESG000004  206     56.33             0.00            0.00  0.00  0.00
ESS000011      ESG000004  368     218.32            38.00           4.35  7.63  100.00
ESS000012      ESG000004  308     158.32            0.00            0.00  0.00  0.00
ESS000013      ESG000005  2091    1941.31           0.00            0.00  0.00  0.00
```



Genome sequences and gene annotation (FASTA,GTF)
OR
Transcript sequences (FASTA)
RNA-Seq reads (FASTA/ FASTQ)
rsem-prepare-reference
RSEM references
User-provided read aligner
rsem-calculate-expression
Read alignments (SAM/BAM)
rsem-simulate-reads
Probabilistically-weighted alignments (BAM)
Sequencing model parameter estimates
Gene/isoform abundance estimates and credibility intervals
rsem-bam2wig
rsem-plot-model
Expected read count distribution (UCSC Wiggle)
Model parameter diagnostic plots (PDF)

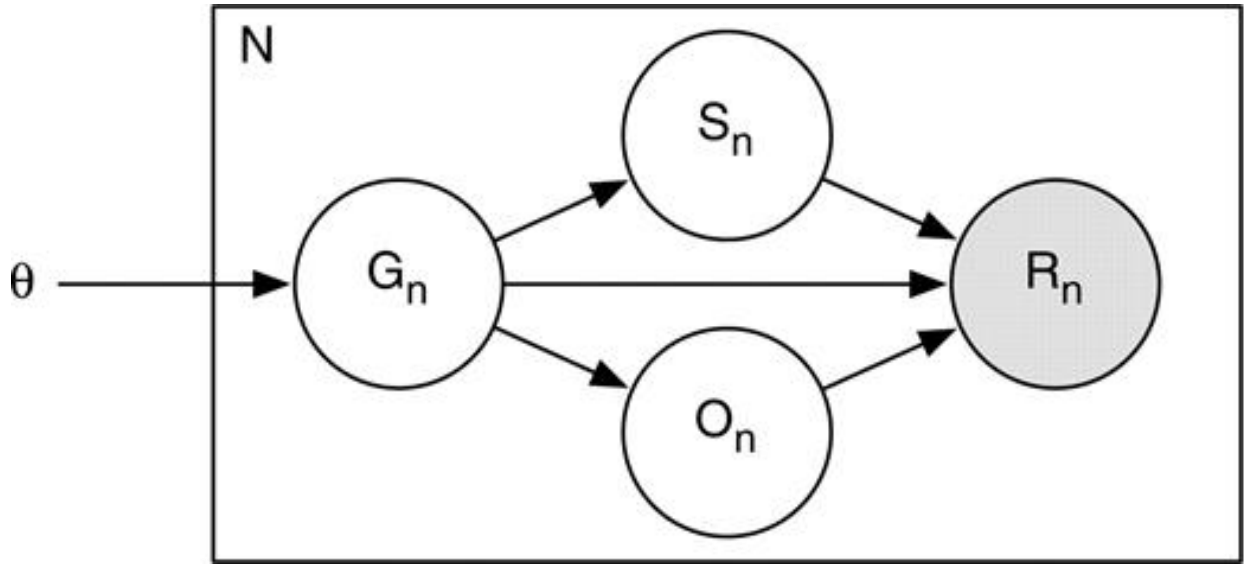
Li *et al*, Bioinformatics 2011

RSEM GENERATIVE MODEL



- We focus on the initial, simpler version of RSEM (Li et al. 2009)
- R_n represents the observed reads ($n = 1,\ldots,N$) and is the only **observed quantity.**
- $\theta = [\theta_1, \ldots, \theta_M]$ is the vector of transcript abundances, which we want to estimate.
- There are several latent variables:
  - G_n: the isoform that generates R_n.
  - S_n: the position in the isoform.
  - O_n: the strand.

LIKELIHOOD



$$P(g, s, o, r \mid \theta) = \prod_{n=1}^{N} P(g_n \mid \theta) P(s_n \mid g_n) P(o_n \mid g_n) P(r_n \mid g_n, s_n, o_n) .$$

- We only observe R_n and we cannot directly compute the likelihood.
- RSEM uses an Expectation-Maximization (EM) algorithm to maximize the likelihood.

E STEP

- Assume that we know θ.
- We define the indicator Z:

$$Z_{nijk} = 1 \iff (G_n, S_n, O_n) = (i, j, k)$$

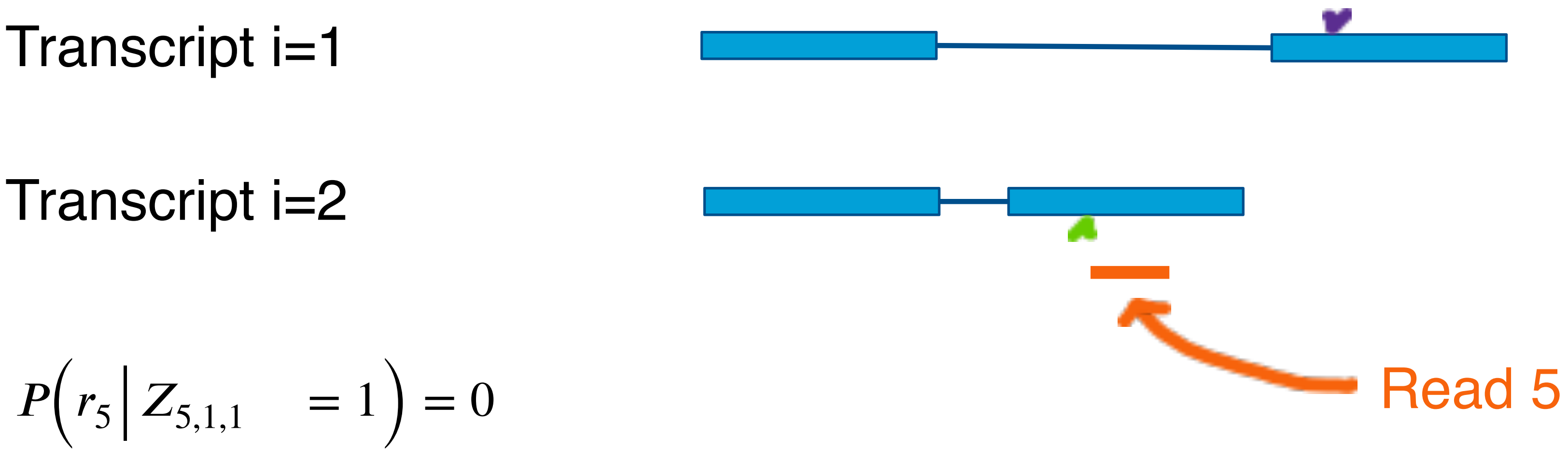
- Compute the probability that read *n* comes from transcript *i*.

Transcript length

$$P(Z_{nij} = 1 \mid r, \theta^{(t)}) = \frac{(\theta_i^{(t)} / l_i)\overbrace{P(r_n \mid Z_{nij} = 1)}^{\text{Read alignments}}}{\sum_{i',j'} (\theta_{i'}^{(t)} / l_{i'}) P(r_n \mid Z_{ni'j'} = 1)}$$

Transcript abundance

E STEP — EXAMPLE



Transcript i=1

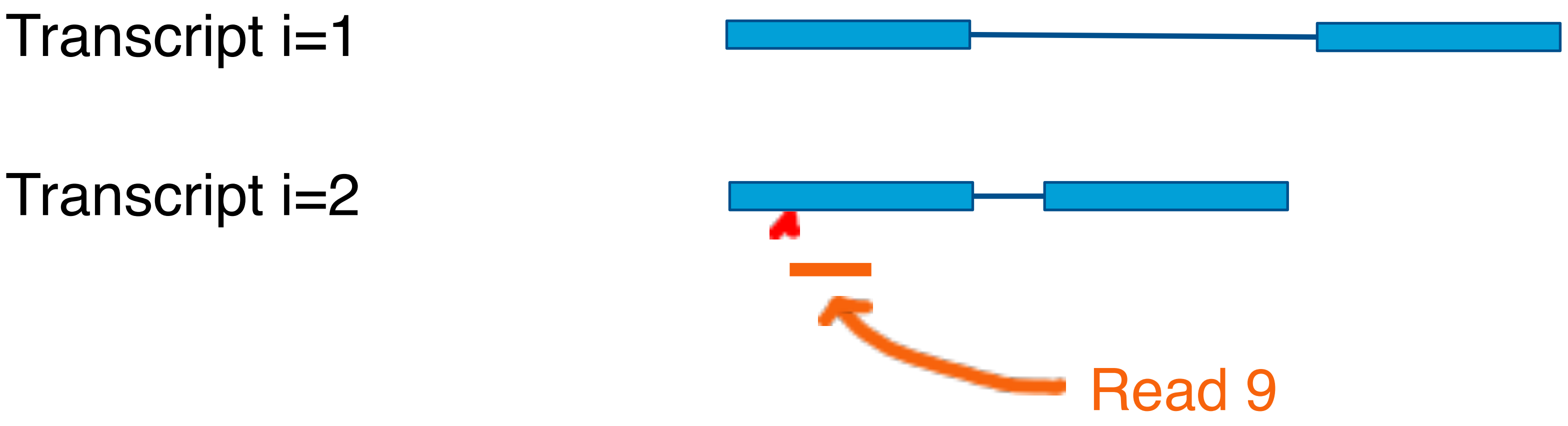
Transcript i=2

Read 5

$$P\left(r_5 \middle| Z_{5,1,1} = 1\right) = 0$$
$$P\left(r_5 \middle| Z_{5,1,100} = 1\right) = 0$$
$$P\left(r_5 \middle| Z_{5,2,1} = 1\right) = 0$$
$$P\left(r_5 \middle| Z_{5,2,75} = 1\right) = 1$$

$$\boxed{P(Z_{nij} = 1 | r, \theta^{(t)}) \;=\; \frac{(\theta_i^{(t)}/\ell_i)P(r_n|Z_{nij} = 1)}{\sum_{i',j'}(\theta_{i'}^{(t)}/\ell_{i'})P(r_n|Z_{ni'j'} = 1)}}$$

E STEP — EXAMPLE



$$P\left(r_9 \middle| Z_{9,1,1} = 1\right) = 0$$
$$P\left(r_9 \middle| Z_{9,1,20} = 1\right) = 0.5$$
$$P\left(r_9 \middle| Z_{9,2,20} = 1\right) = 0.5$$

$$\boxed{P(Z_{nij} = 1 | r, \theta^{(t)}) = \frac{(\theta_i^{(t)}/\ell_i)P(r_n|Z_{nij} = 1)}{\sum_{i',j'}(\theta_{i'}^{(t)}/\ell_{i'})P(r_n|Z_{ni'j'} = 1)}}$$

M STEP

- Assume you have a current estimate of the probabilities (from the E step)

- We look for the values of θ that explain the most of those probabilities.

$$P(Z_{nij} = 1 | r, \theta^{(t)}) \quad (1)$$

Estimated count for transcript i, based on (1)

$$\theta_i^{(t+1)} = \frac{C_i \,\big|\, r, \theta^{(t)}}{N}$$

... dipends on the estimate at the previous iteration (t)

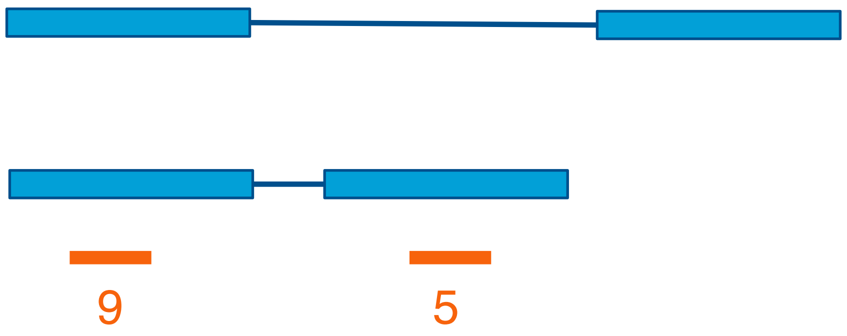
Estimate at iteration t+1

Normalization factor

M STEP — EXAMPLE

Transcript i=1

Transcript i=2



$$P\left(r_5 \,\middle|\, Z_{5,2,75} \;= 1\right) = 1$$

$$C_1 = 0.5$$

$$P\left(r_9 \,\middle|\, Z_{9,1,20} = 1\right) \;= 0.5$$

$$P\left(r_9 \,\middle|\, Z_{9,2,20} = 1\right) \;= 0.5$$

$$C_2 = 1 + 0.5 = 1.5$$

Assumptions:

1. No sequencing errors
2. All transcripts have the same length

$$\boxed{\theta_i^{(t+1)} = \frac{C_i \,\middle|\, r, \theta^{(t)}}{N}}$$

CONVERGENCE

- The E and M steps are alternated until convergence.
- I.e., at each step until the estimates of $\theta^{(t)}$ and $\theta^{(t+1)}$ are so close that are almost indistinguishable.
- By default the relative difference is set to 10^{-3}.

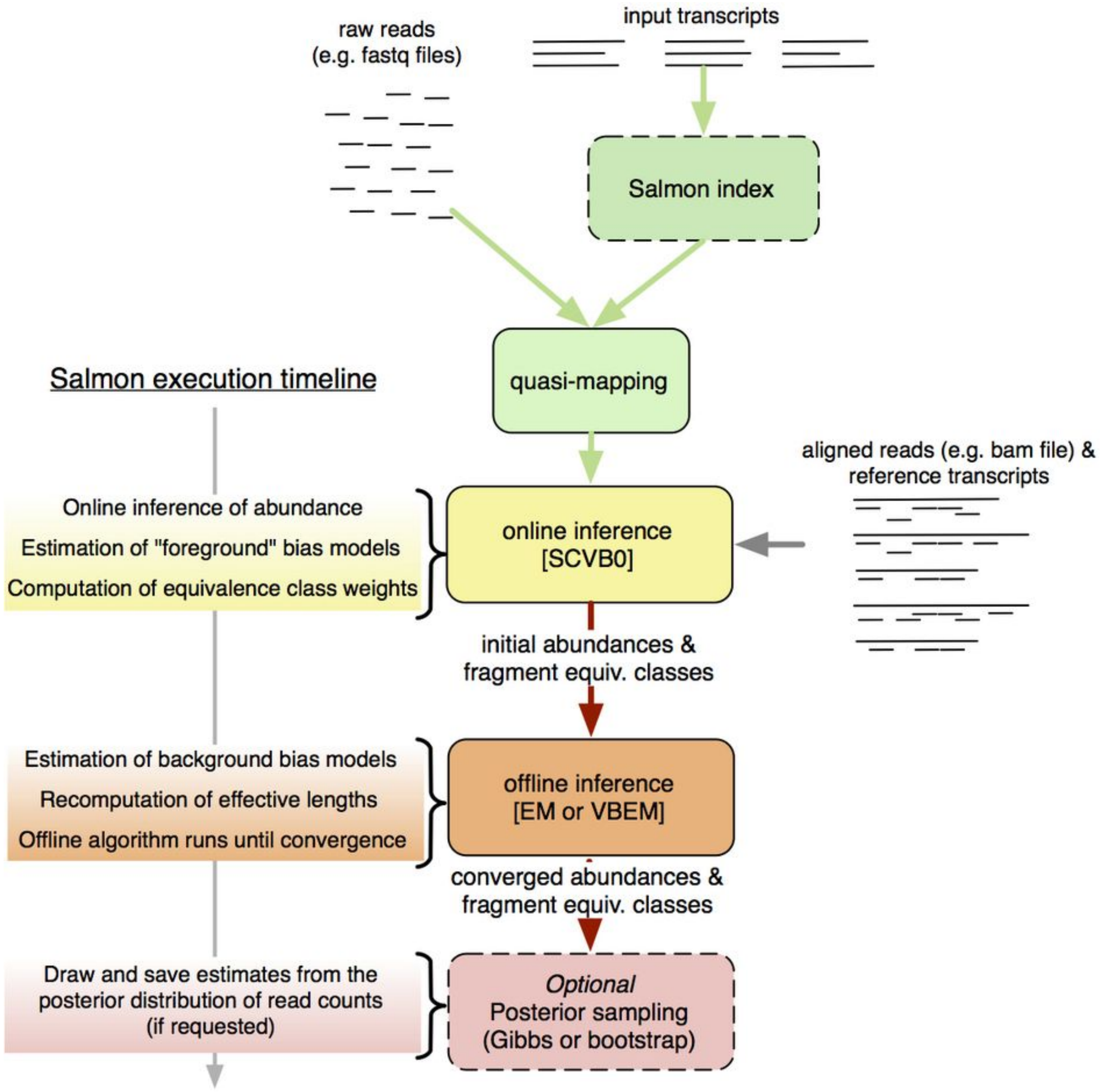
PSEUDO-ALIGNMENT

SALMON

- An alternative faster approach
- Available as open-source software:

https://github.com/COMBINE-lab/salmon

- It uses quasi-mapping to speed up computations
  - It can process 600M paired-end reads in 20 minutes.


raw reads
(e.g. fastq files)
input transcripts
Salmon index
quasi-mapping
Salmon execution timeline
aligned reads (e.g. bam file) &
reference transcripts
Online inference of abundance
Estimation of "foreground" bias models
Computation of equivalence class weights
online inference
[SCVB0]
initial abundances &
fragment equiv. classes
Estimation of background bias models
Recomputation of effective lengths
Offline algorithm runs until convergence
offline inference
[EM or VBEM]
converged abundances &
fragment equiv. classes
Draw and save estimates from the
posterior distribution of read counts
(if requested)
Optional
Posterior sampling
(Gibbs or bootstrap)

Patro *et al*, Nature Methods 2017

QUASI-MAPPING

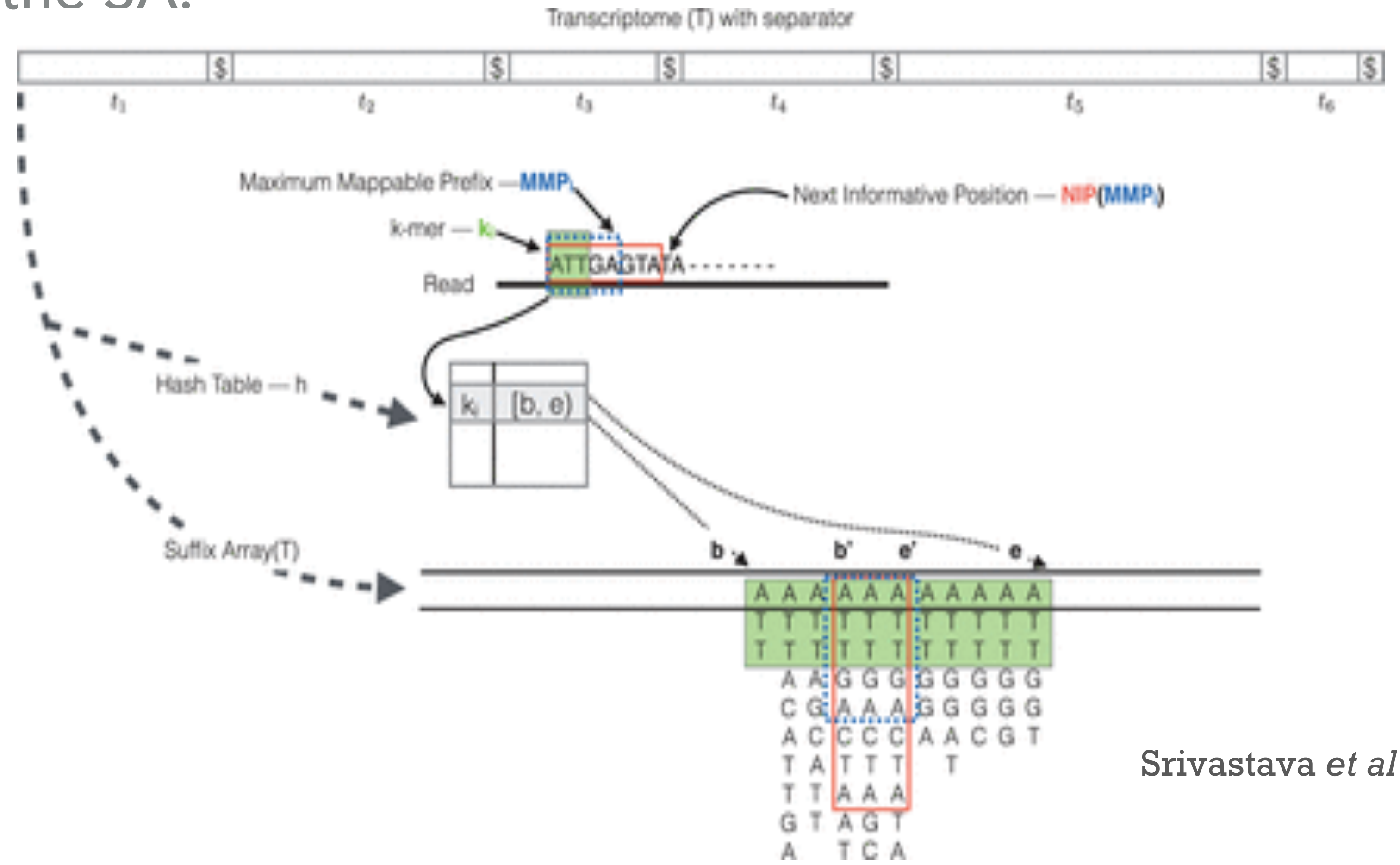
- Alignment is the step with the main computational cost:
  - High computational time
  - High memory consumption
- In some cases we do need the full read alignments
  - E.g., variant calling (SNPs).
- If we are only interested in expression quantification, it is possible to leverage alternative algorithms that do not require the full mapping.
- There are several alternative strategies called *quasi-mapping* or *pseudo-alignment*.

QUASI-MAPPING

- We start from the sequences of all transcripts.
  - We concatenate the sequences.
  - Separated by a special character (e.g., «$»).

We construct two structures:

  - A *suffix array* SA, similar to STAR.
  - A table (*hash map*) that maps all the sequences of a fixed length (*k-mers*) to the positions in the SA.



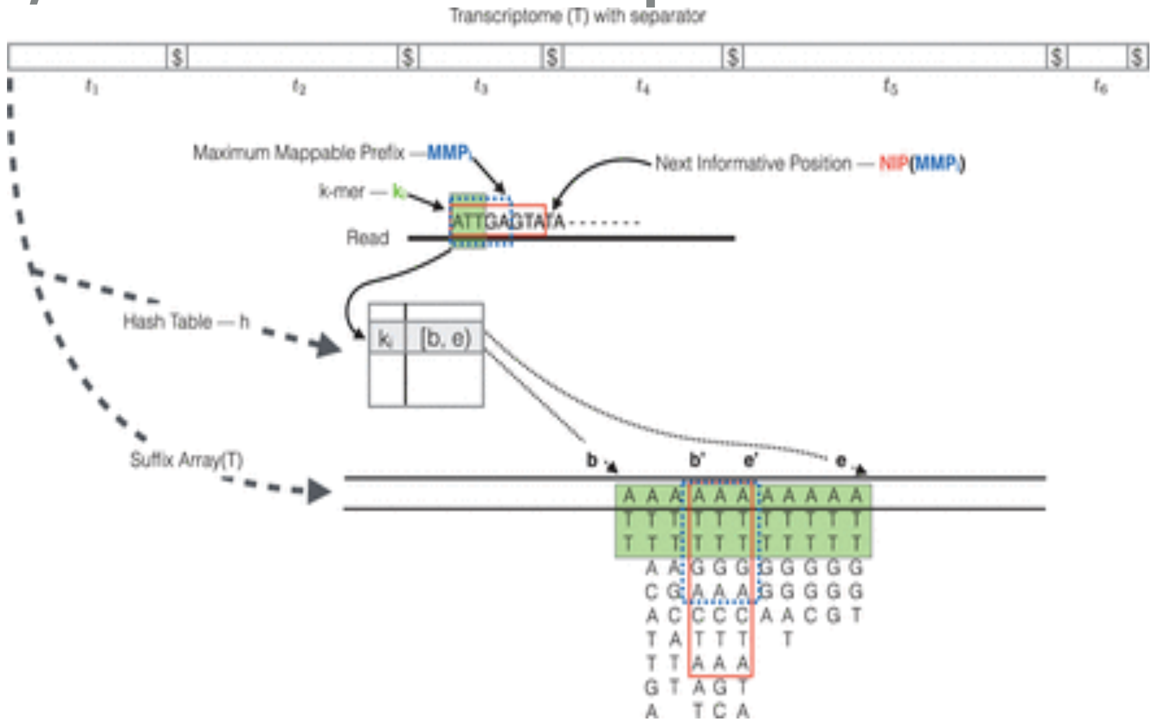
Transcriptome (T) with separator
$
t_1
t_2
t_3
t_4
t_5
t_6
Maximum Mappable Prefix — MMP$_i$
k-mer — k_i
Next Informative Position — NIP(MMP$_i$)
ATTGAGTATA
Read
Hash Table — h
k_i
[b, e)
Suffix Array(T)
b
b'
e'
e

Srivastava *et al*, Bioinformatics 2016

SEARCH PHASE

- Given a read R
  - We select the first *k* nucleotides.
  - We search for them in the hash map.
  - We find the corresponding interval in the SA.
  - We expand the search to the following positions, until we find exact matches.
- Every time that we find a mismatch the procedure starts again from the next position in the read.
- Once the process is complete, we have a map:
  - $R_i \rightarrow T_j, P_k$
  - $R_i \rightarrow T_l, P_m$

Read i is compatible with transcript j at position k


Transcriptome (T) with separator
$
t1
t2
t3
t4
t5
t6
Maximum Mappable Prefix — MMPi
Next Informative Position — NIP(MMPi)
k-mer — ki
ATTGAGTATA
Read
Hash Table — h
ki
[b, e)
Suffix Array(T)
b
b'
e'
e

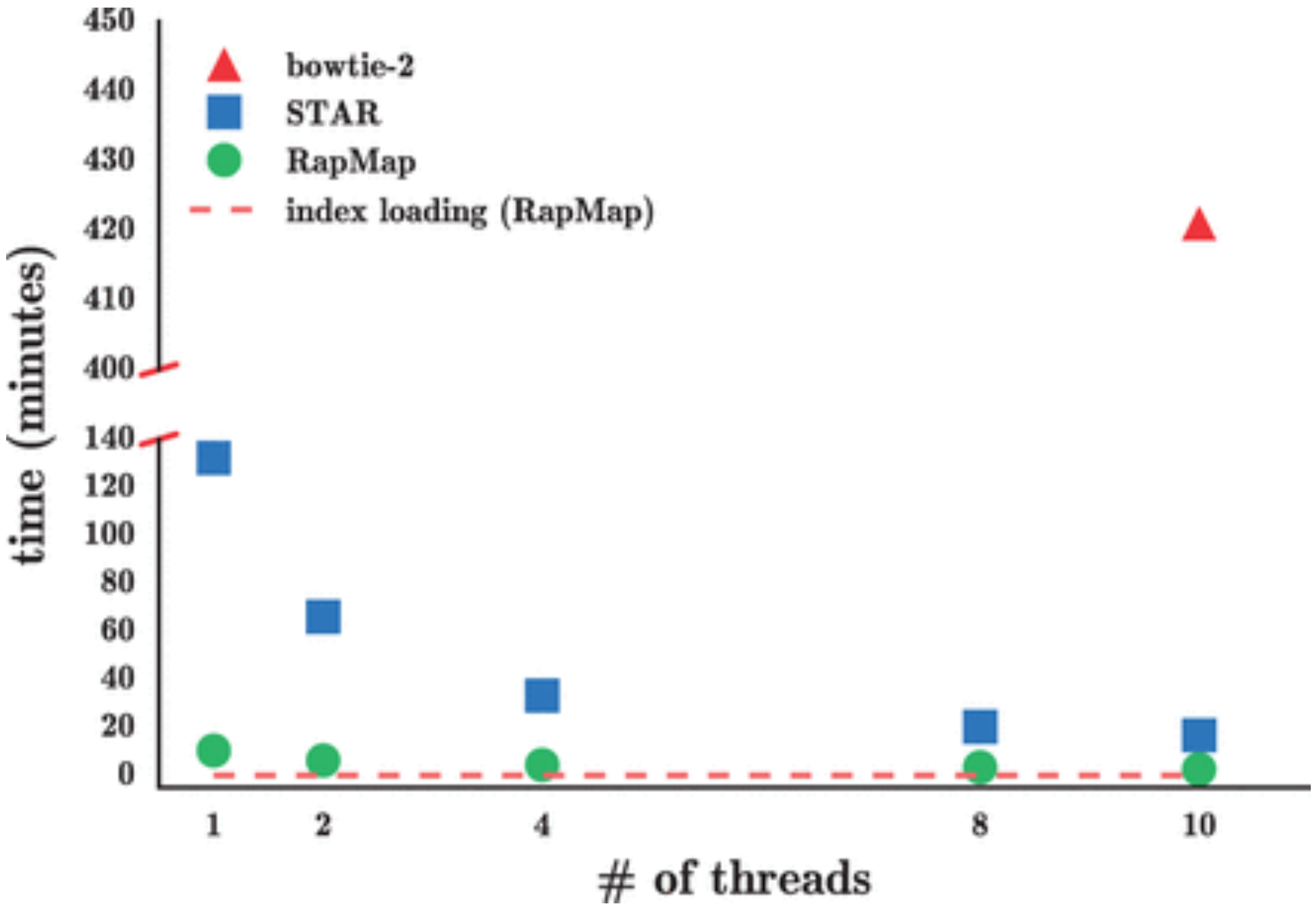
CONSENSUS PHASE

- Given the map:
  - R_1 → T_5, P_{10}
  - R_1 → T_7, P_{50}
  - R_1 → T_5, P_{100}
  - R_1 → T_9, P_{110}
- The only transcript compatible with all positions is T_5.
  - We take the intersection of all transcripts associated to R.
- This procedure is computationally very efficient.

COMPUTATIONAL TIME

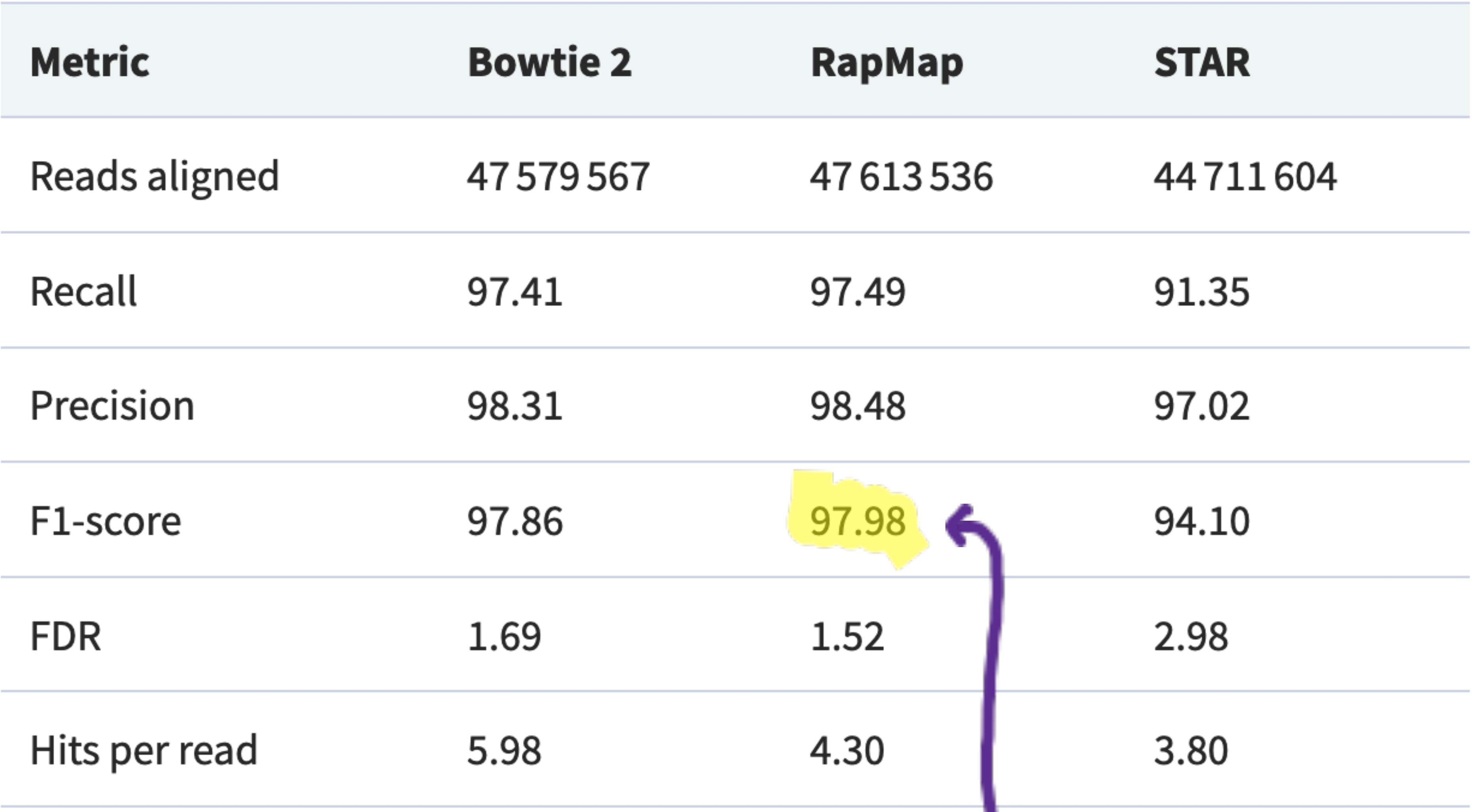
Quasi-mapping is much faster than full mapping.

- «RapMap» indicates quasi-mapping here.



Srivastava *et al*, Bioinformatics 2016

ACCURACY



Metric	Bowtie 2	RapMap	STAR
Reads aligned	47 579 567	47 613 536	44 711 604
Recall	97.41	97.49	91.35
Precision	98.31	98.48	97.02
F1-score	97.86	97.98	94.10
FDR	1.69	1.52	2.98
Hits per read	5.98	4.30	3.80

Srivastava *et al*, Bioinformatics 2016

As accurate as mapping

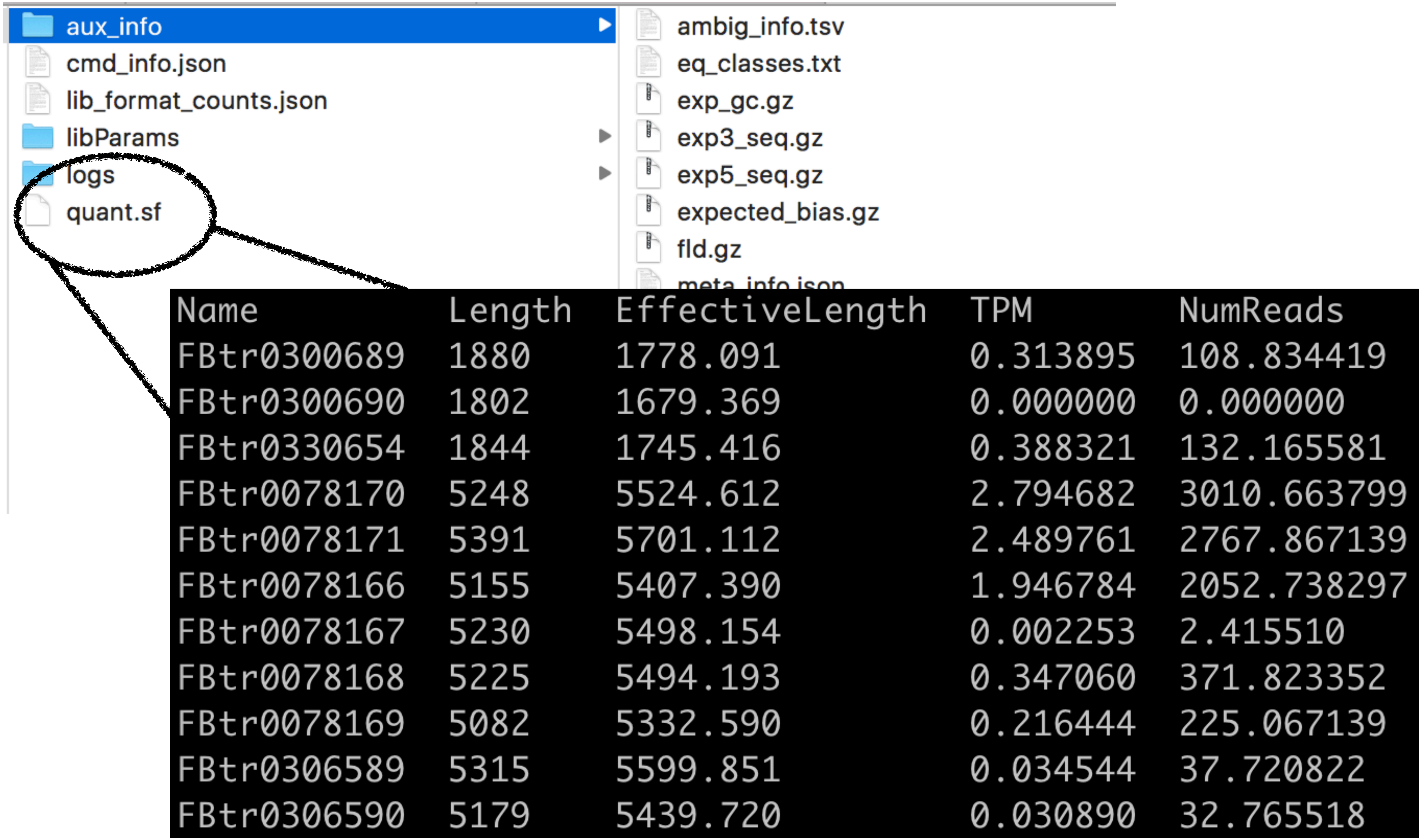
EXPRESSION QUANTIFICATION

- Once we have the quasi-alignment results, we need to quantify each transcript expression.
- Salmon uses a statistical model conceptually similar to that of RSEM.
- Compared to the simplified version that we considered, it models:
  - Fragment size.
  - Positional bias and transcript coverage.
  - 3′ and 5′ bias.
  - GC-content.
  - Strand-specificity.

See Patro et al. (2017) for details

OUTPUT

- Salmon yields several files for each sample.



```
Name          Length   EffectiveLength   TPM        NumReads
FBtr0300689   1880     1778.091          0.313895   108.834419
FBtr0300690   1802     1679.369          0.000000   0.000000
FBtr0330654   1844     1745.416          0.388321   132.165581
FBtr0078170   5248     5524.612          2.794682   3010.663799
FBtr0078171   5391     5701.112          2.489761   2767.867139
FBtr0078166   5155     5407.390          1.946784   2052.738297
FBtr0078167   5230     5498.154          0.002253   2.415510
FBtr0078168   5225     5494.193          0.347060   371.823352
FBtr0078169   5082     5332.590          0.216444   225.067139
FBtr0306589   5315     5599.851          0.034544   37.720822
FBtr0306590   5179     5439.720          0.030890   32.765518
```

NORMALIZED EXPRESSION

- Both RSEM and Salmon return, in addition to expected counts, two expression measures:
  - FPKM
  - TPM
- They are both attempts at *normalizing* gene expression.
- Intuitively, the number of reads for each gene depends, in addition to its gene expression, on:

1. Sequencing depth. E.g., if we sequence twice as many total reads, we will have on average double counts.
2. Transcript length. I.e., the longer the transcript the more reads we are likely to sequence.

FPKM

- Acronym of:

fragments per kilobase of exon model per million mapped reads

Number of reads

Transcript length

Sequencing depth

- I.e., for each transcript *i*:

$$fpkm_i = \frac{r_i\,10^9}{l_i\,R} = \frac{\frac{r_i\,10^3}{l_i}}{\frac{R}{10^6}}$$

$$R = \sum_{i \in T} r_i$$

TPM

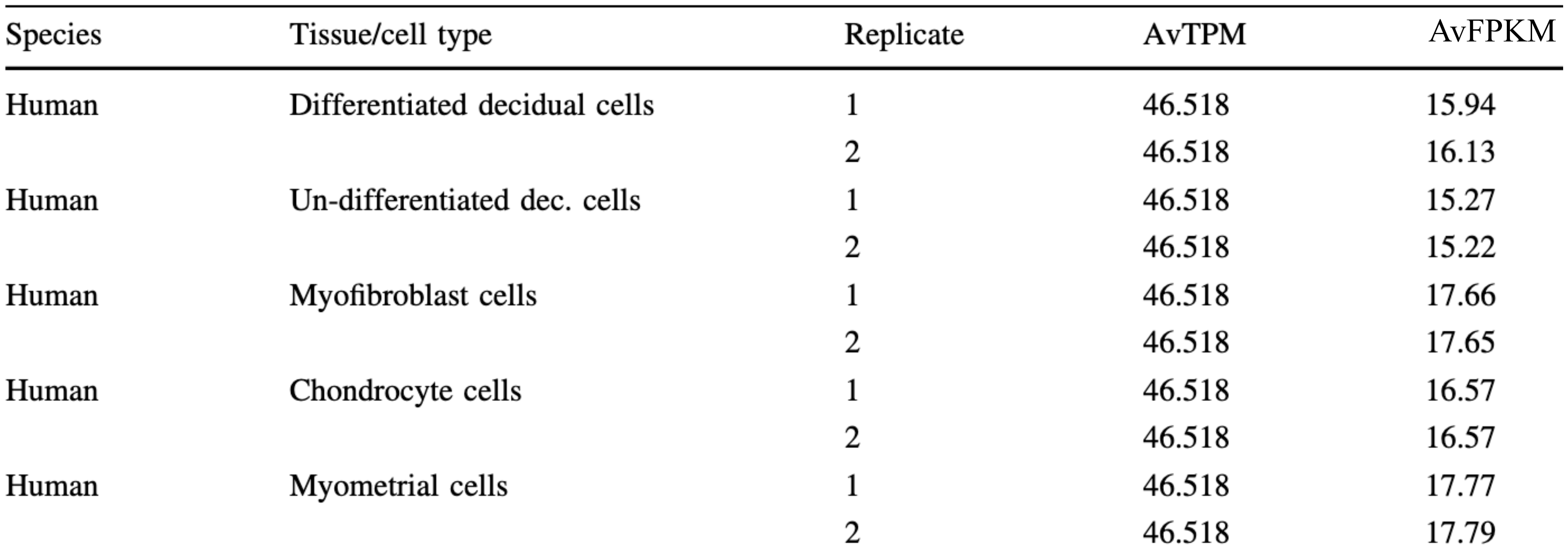
Acronym of transcripts per million.

$$tpm_i = \frac{r_i \times L \times 10^6}{l_i \times R}$$

Read length

$$R = \sum_{i \in T} \frac{r_i \times L}{l_i}$$

Constant average value across experiments



Species	Tissue/cell type	Replicate	AvTPM	AvFPKM
Human	Differentiated decidual cells	1	46.518	15.94
		2	46.518	16.13
Human	Un-differentiated dec. cells	1	46.518	15.27
		2	46.518	15.22
Human	Myofibroblast cells	1	46.518	17.66
		2	46.518	17.65
Human	Chondrocyte cells	1	46.518	16.57
		2	46.518	16.57
Human	Myometrial cells	1	46.518	17.77
		2	46.518	17.79

Wagner *et al*, Theory Biosci. 2012

DATA REPRESENTATION

- At the end of the quantification process, the data can be represented as a numeric matrix, which contains non-negative integers.

	Exp 1	Exp 2	…	Exp n
Gene 1				
Gene 2				
…				
Gene p				

- Very often $n \ll p$

DATA REPRESENTATION

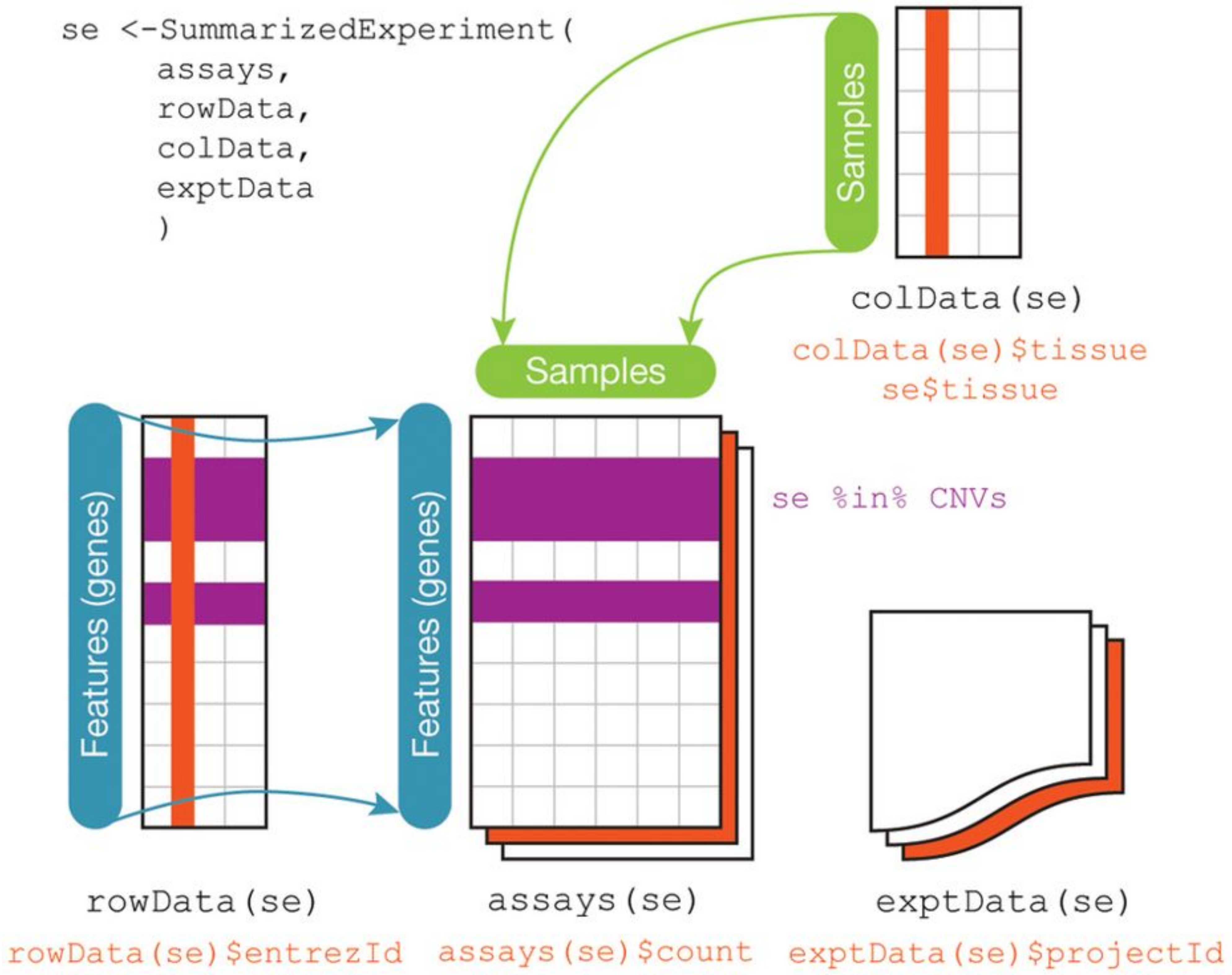
- Columns correspond to **statistical units** (samples, individuals, cell lines, …)

- Rows correspond to features (genes, transcripts)

	Exp 1	Exp 2	…	Exp *n*
Gene 1				
Gene 2				
…				
Gene *p*				

- Furthermore, we often have additional information on genes and/or samples, often referred to **metadata**.

DATA REPRESENTATION IN R/BIOCONDUCTOR



```
se <-SummarizedExperiment(
     assays,
     rowData,
     colData,
     exptData
     )
```

THANKS FOR YOUR ATTENTION!