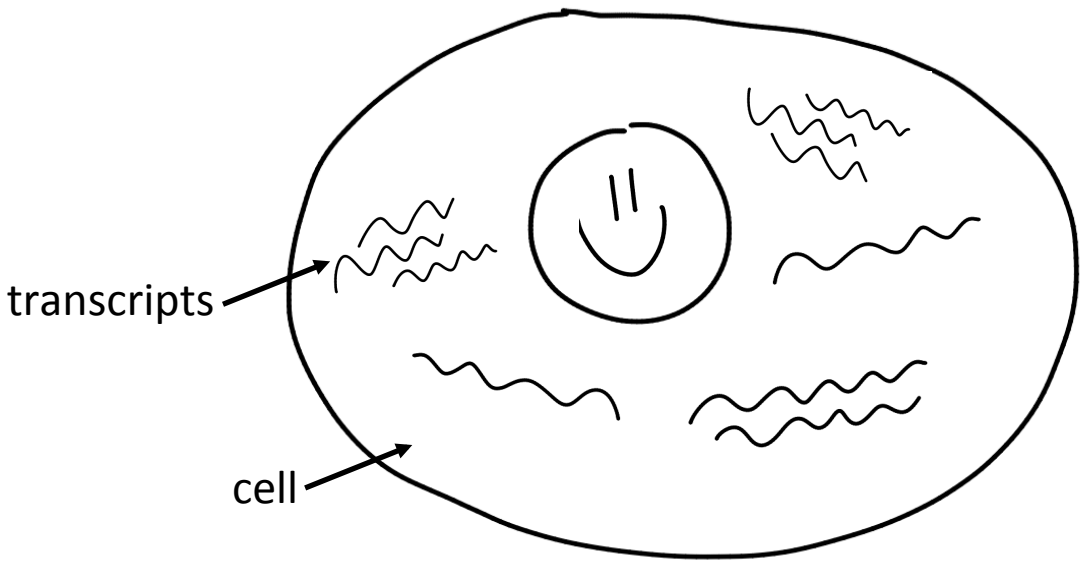
RNA-seq:
What? Why? How?

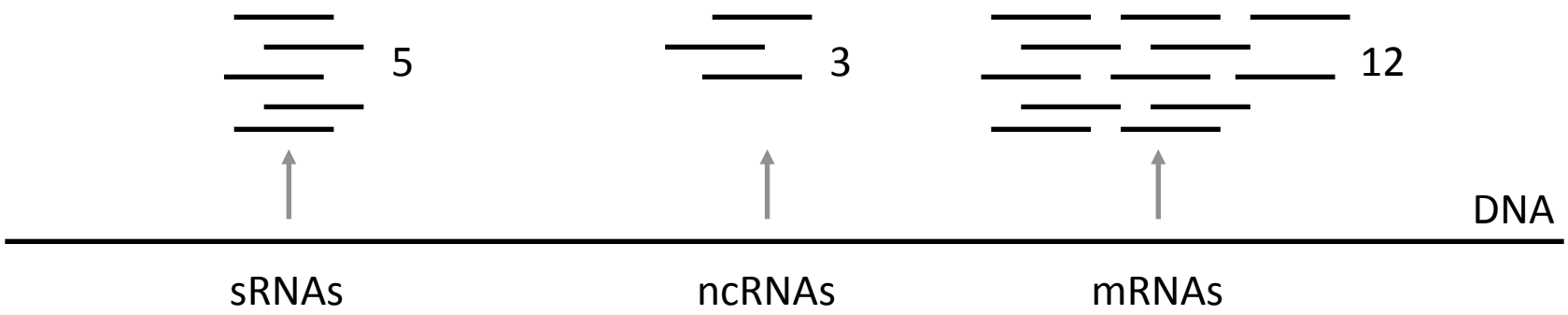
Dariya Sydykova

What is RNA-seq?

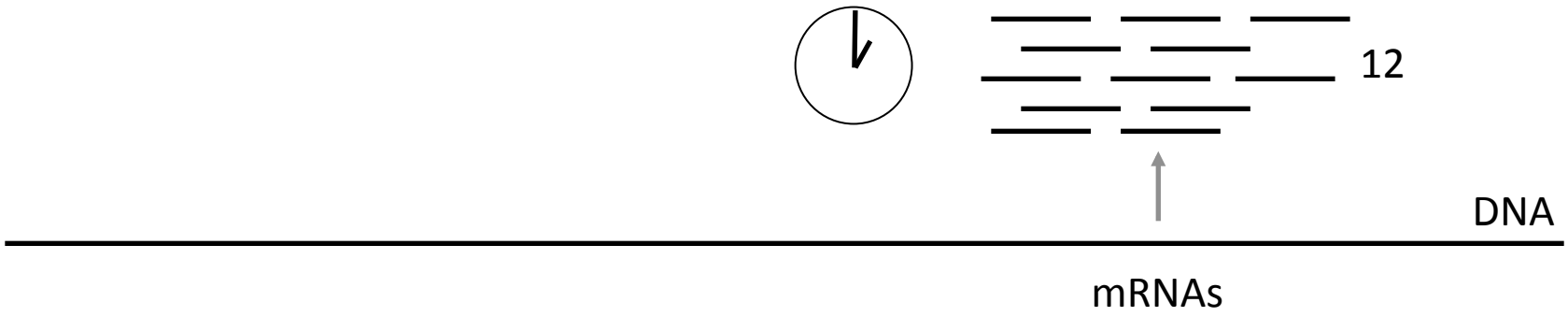
RNA-seq is a deep-sequencing technology that captures and quantifies the complete set of transcripts in a cell or the cell's transcriptome.



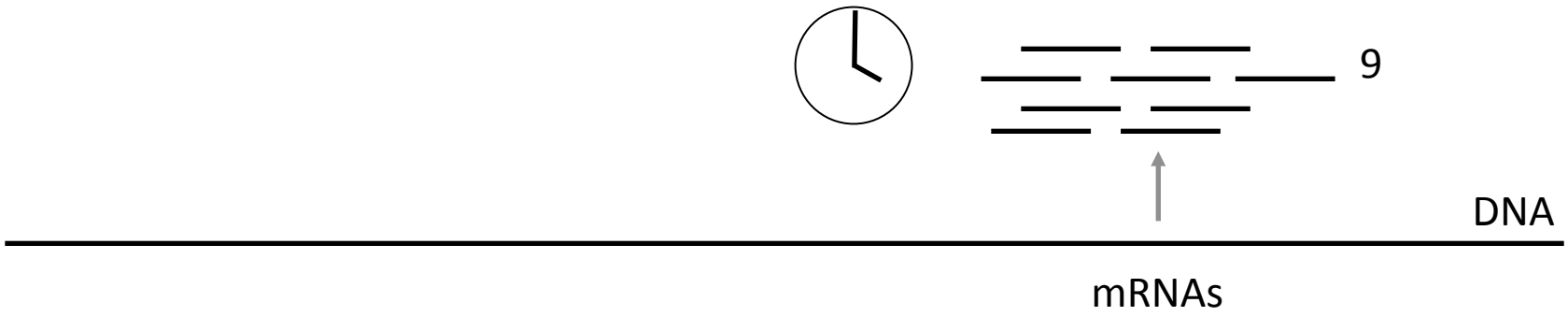
Why do we do RNA-seq?



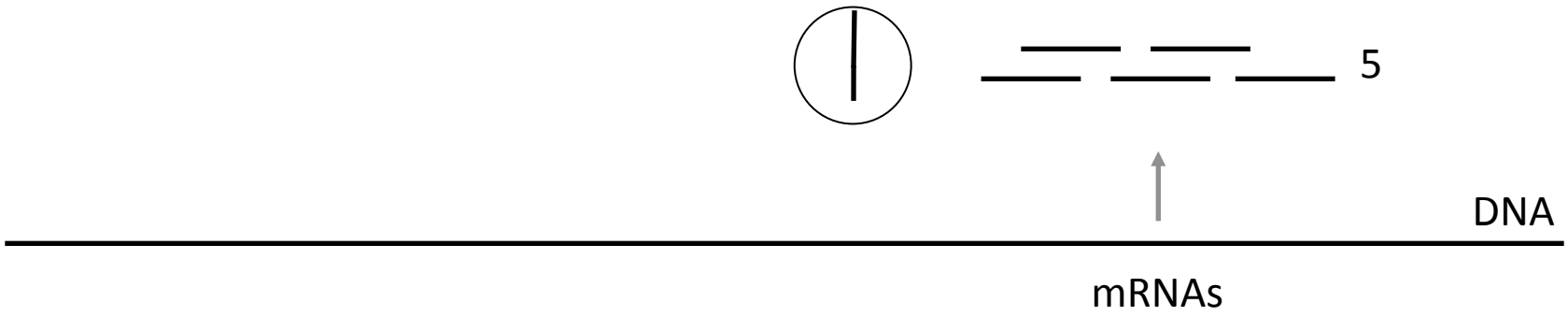
Why do we do RNA-seq?



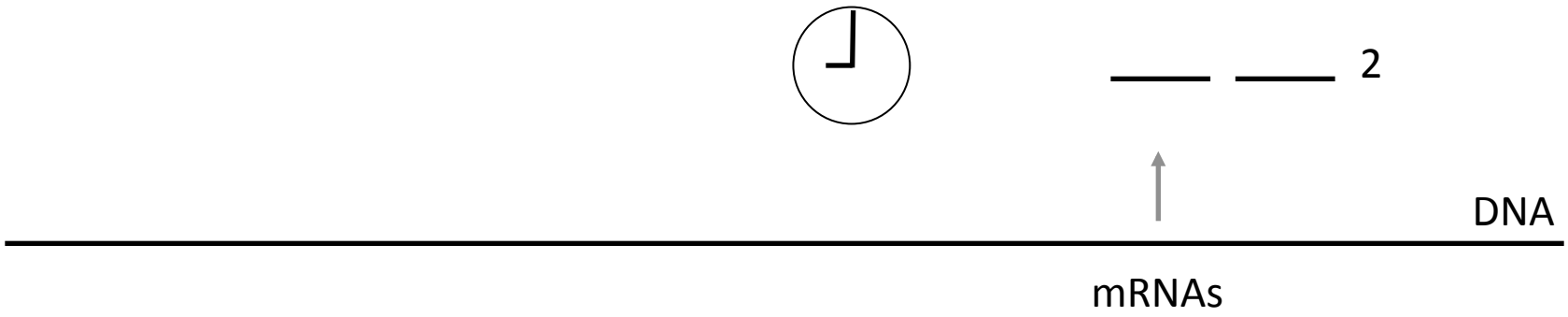
Why do we do RNA-seq?



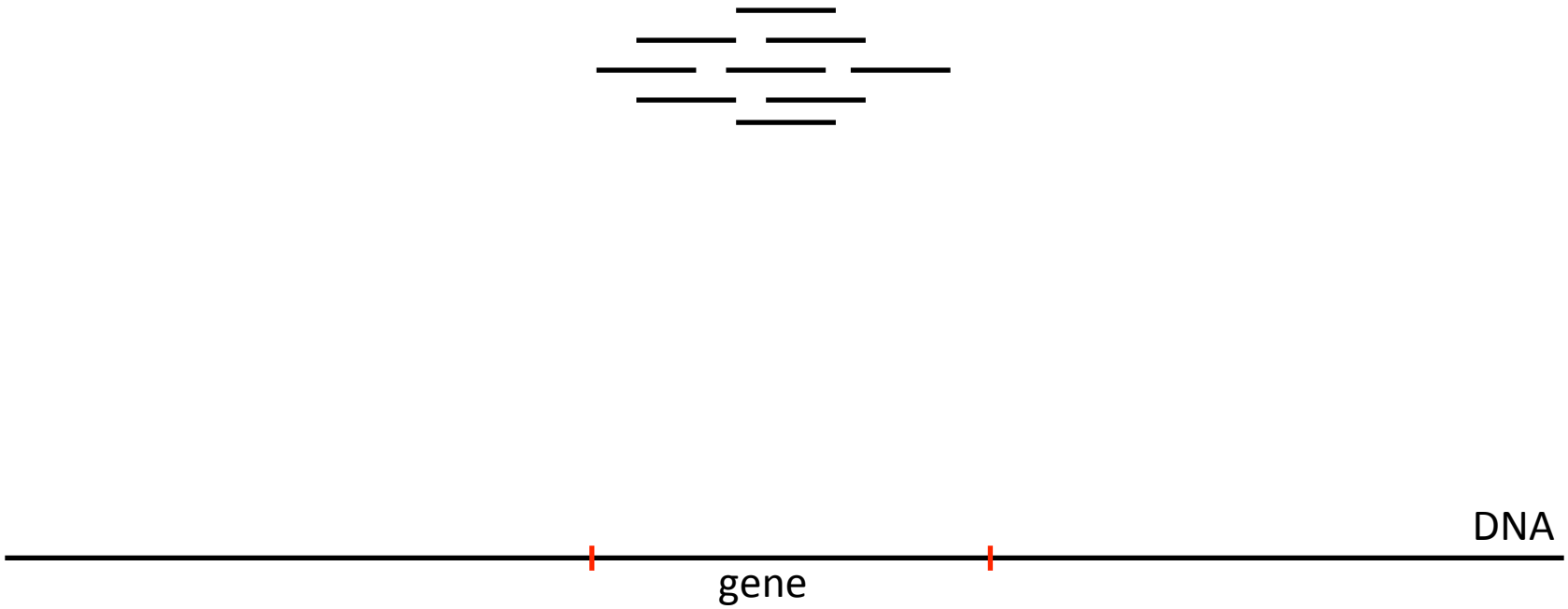
Why do we do RNA-seq?



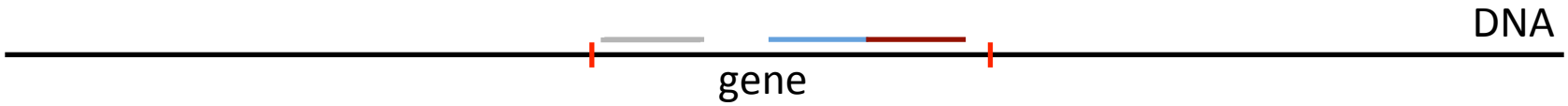
Why do we do RNA-seq?



Why do we do RNA-seq?

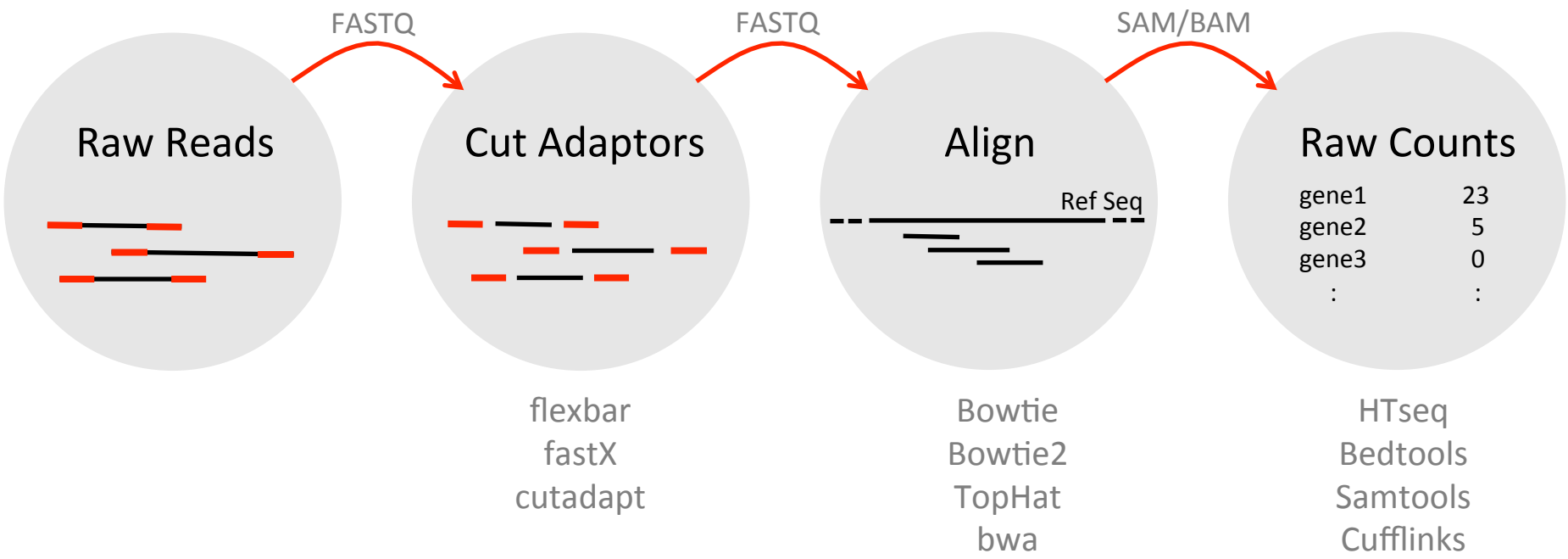


Why do we do RNA-seq?

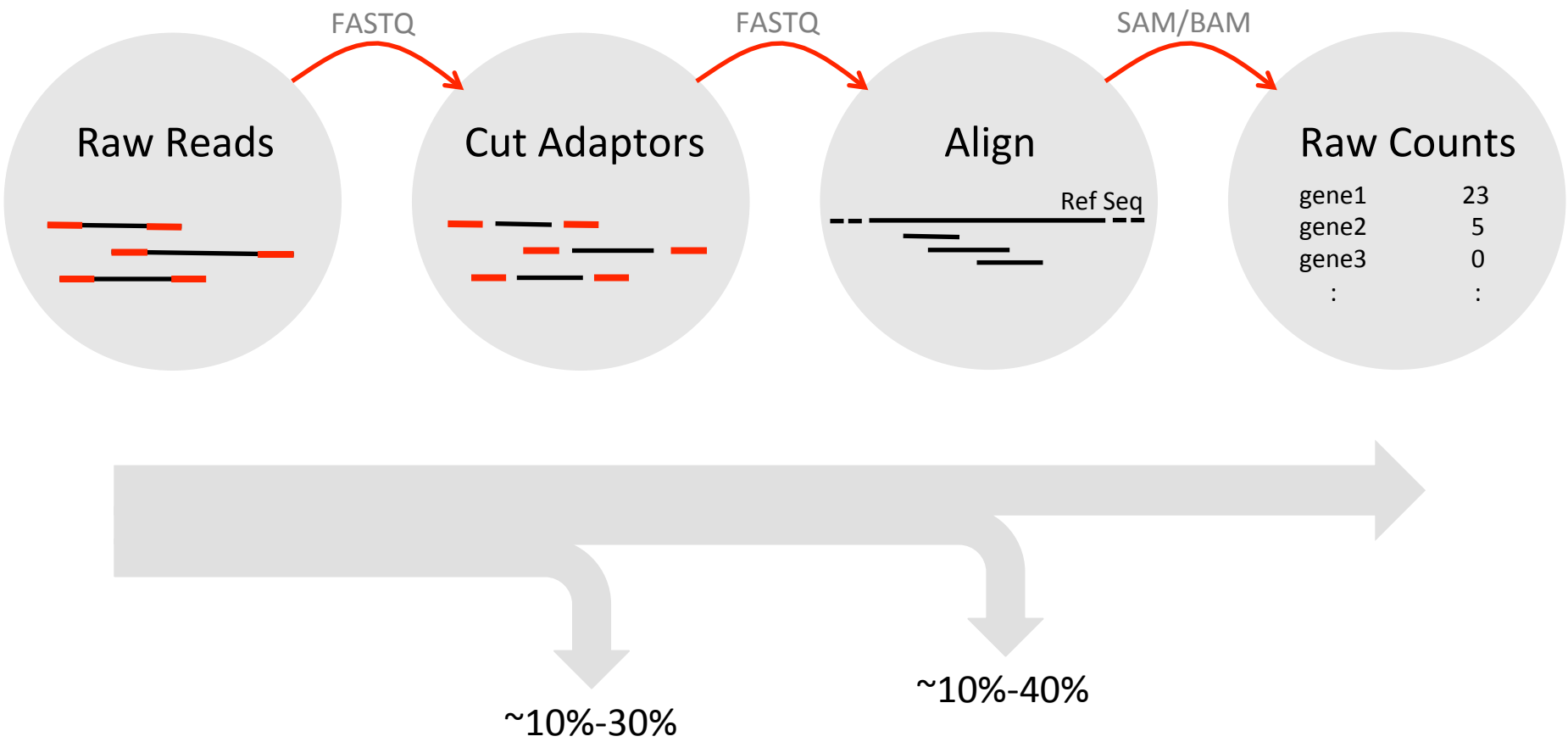


Next Gen Sequencing

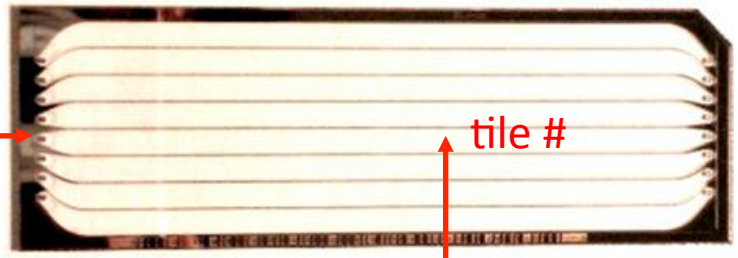
RNA-seq pipeline



RNA-seq pipeline



FASTQ file



lane #

tile #

Read identifier:

Machine name run ID flowcell ID x y 1/2:Y/N:0:index seq

```
@HWI-ST1097:104:D13TNACXX:4:1101:1715:2142 1:N:0:CGATGT
GCGTTGGTGGCATAGTGGTGAGCATAGCTGCCTTCCAAGCAGTTATGGGAG
+
=<@BDDD=A;+2C9F<CB?;CGGA<<ACEE*1?C:D>DE=FC*0BAG?DB6
```

seq

GSAF/sequence description

Ascii-encoded base quality score

How to get the number of reads?

```
grep ^@ <file.fastq> | wc -l

wc -l <file.fastq>
```

```
Daria-Sydykovas-MacBook-Pro:Downloads DariaSydykova$ grep ^@ sample_1.fq | wc -l
  750000
Daria-Sydykovas-MacBook-Pro:Downloads DariaSydykova$ wc -l sample_1.fq
 3000000 sample_1.fq
```

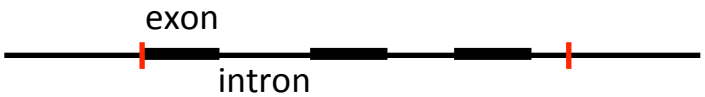
How to cut off the adaptors?

GCGTTGGTGGCATAGTGGTGAGCATAGCTGCCTTCCAAGCAGTTATGGGAG

```
flexbar –n 1 –t <new_file_name> –r <reads1.fastq>
–p <reads2.fastq> –a adaptors.fasta –f <fastq/fasta> > flexbar.out
```

Mapping

TopHat/Bowtie2

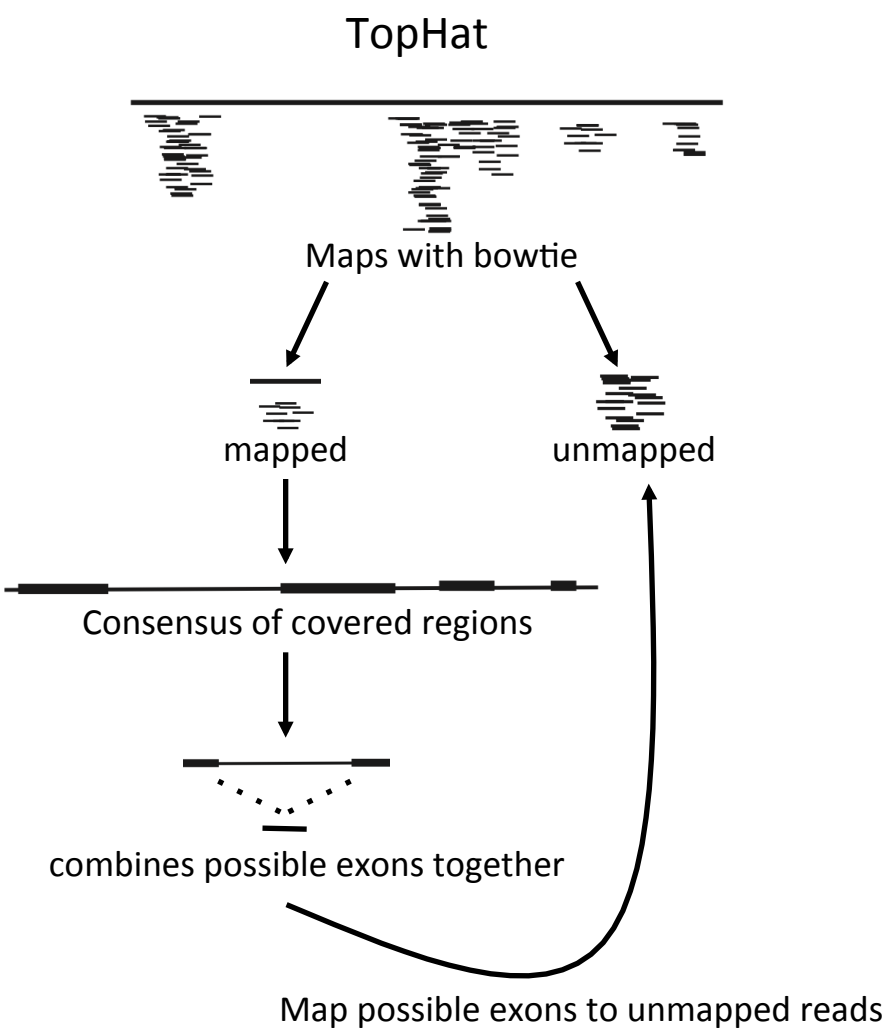


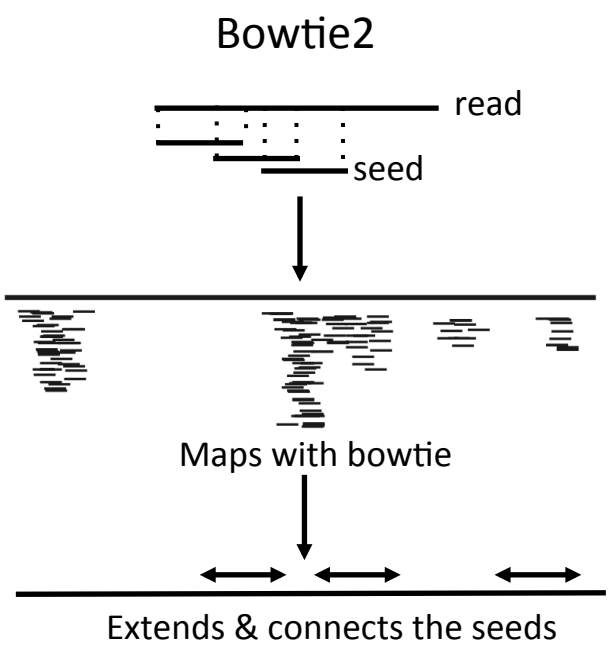
gene with introns

Bowtie/Bowtie2

gene without introns

Mapping


TopHat
Maps with bowtie
mapped
unmapped
Consensus of covered regions
combines possible exons together
Map possible exons to unmapped reads


Bowtie2
read
seed
Maps with bowtie
Extends & connects the seeds

Mapping

```
tophat –p 4 –G <gtf_ref_seq_file> -o <output.sam>
--no-novel-juncs <reference_index> <reads1.fastq>
<reads2.fastq>

bowtie2 –x <reference_index> -1 <reads1.fastq> -2
<reads2.fastq> -S <output.sam>
```

GTF file

```
REL606 .	exon	190	255	.	+	.	transcript_id "ECB_00001";
REL606 .	gene	190	255	.	+	.	transcript_id "ECB_00001"; gene_id "ECB_00001"; gene_name "thrL";
REL606 .	exon	336	2798	.	+	.	transcript_id "ECB_00002";
REL606 .	gene	336	2798	.	+	.	transcript_id "ECB_00002"; gene_id "ECB_00002"; gene_name "thrA";
REL606 .	exon	2800	3732	.	+	.	transcript_id "ECB_00003";
REL606 .	gene	2800	3732	.	+	.	transcript_id "ECB_00003"; gene_id "ECB_00003"; gene_name "thrB";
REL606 .	exon	3733	5019	.	+	.	transcript_id "ECB_00004";
REL606 .	gene	3733	5019	.	+	.	transcript_id "ECB_00004"; gene_id "ECB_00004"; gene_name "thrC";
REL606 .	exon	5232	5528	.	+	.	transcript_id "ECB_00005";
REL606 .	gene	5232	5528	.	+	.	transcript_id "ECB_00005"; gene_id "ECB_00005"; gene_name "yaaX";
REL606 .	exon	5681	6457	.	-	.	transcript_id "ECB_00006";
REL606 .	gene	5681	6457	.	-	.	transcript_id "ECB_00006"; gene_id "ECB_00006"; gene_name "yaaA";
REL606 .	exon	6527	7957	.	-	.	transcript_id "ECB_00007";
REL606 .	gene	6527	7957	.	-	.	transcript_id "ECB_00007"; gene_id "ECB_00007"; gene_name "yaaJ";
REL606 .	exon	8236	9189	.	+	.	transcript_id "ECB_00008";
REL606 .	gene	8236	9189	.	+	.	transcript_id "ECB_00008"; gene_id "ECB_00008"; gene_name "talB";
REL606 .	exon	9304	9891	.	+	.	transcript_id "ECB_00009";
REL606 .	gene	9304	9891	.	+	.	transcript_id "ECB_00009"; gene_id "ECB_00009"; gene_name "mogA";
REL606 .	exon	9926	10492	.	-	.	transcript_id "ECB_00010";
REL606 .	gene	9926	10492	.	-	.	transcript_id "ECB_00010"; gene_id "ECB_00010"; gene_name "yaaH";
REL606 .	exon	10641	11354	.	-	.	transcript_id "ECB_00011";
REL606 .	gene	10641	11354	.	-	.	transcript_id "ECB_00011"; gene_id "ECB_00011"; gene_name "yaaW";
REL606 .	exon	10828	11313	.	+	.	transcript_id "ECB_00012";
REL606 .	gene	10828	11313	.	+	.	transcript_id "ECB_00012"; gene_id "ECB_00012"; gene_name "htgA";
```

Mapping Output: SAM/BAM file

samtools view

Mate name | Mate position | Template length

Read identifier | FLAG | Ref Seq | position | Mapping quality: $-10 \log_{10}$P(mapping pos is wrong)

HWI-ST1001:137:C12FPACXX:7:1115:14131:66670 0 chr1 12805 1

CIGAR

42M4I5M * 0 0

TTGGATGCCCCTCCACACCCTCTTGATCTTCCCTGTGATGTCACCAATATG seq

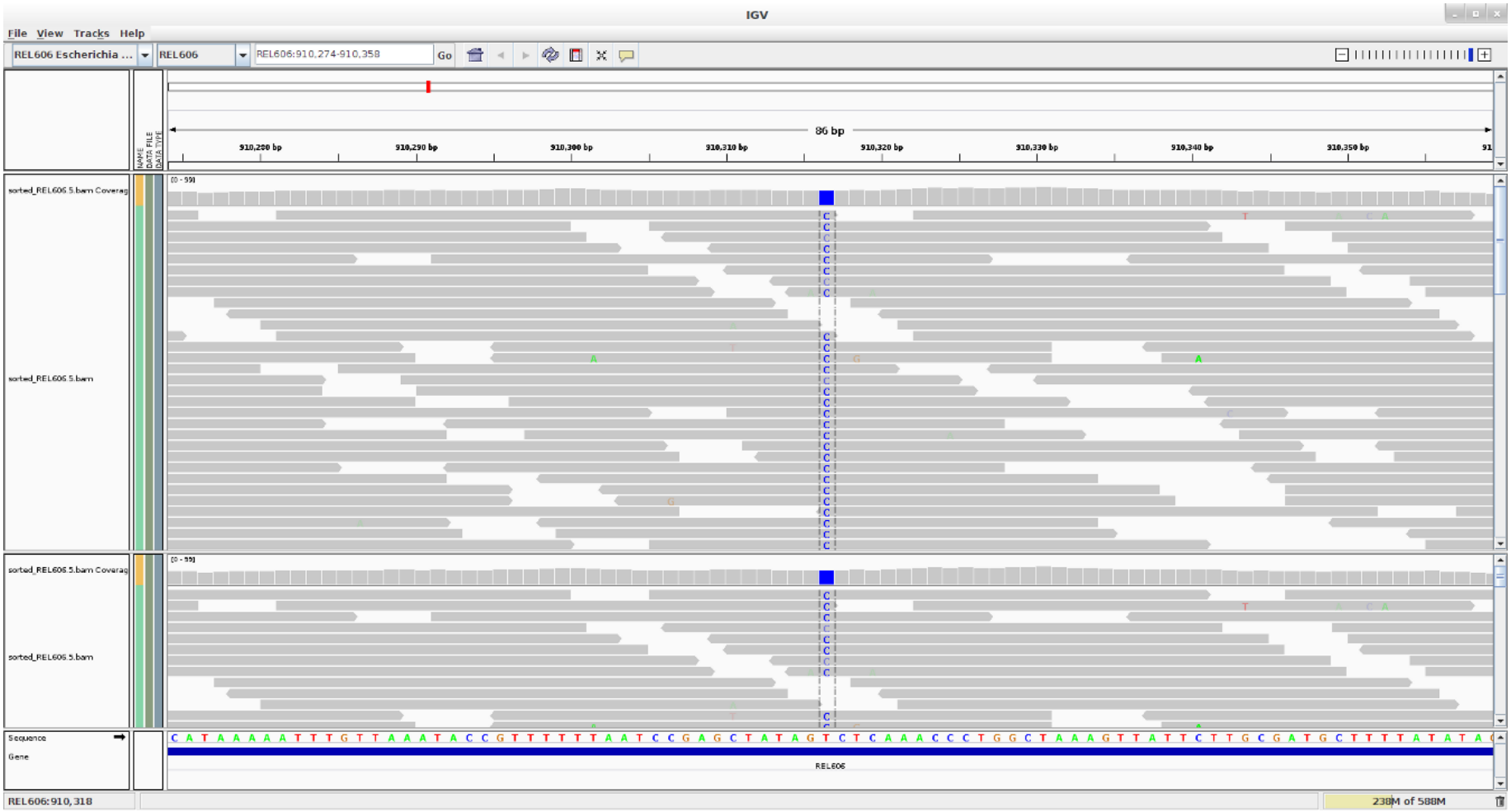
quality score

CCCFFFFFHHGHHJJJJJHJJJJJJJJJJJJJJJJIJJJJJJJJJJJJJIJJ AS:i:-28 XN:i:0 XM:i:2 XO:i:1XG:i:4 NM:i:6 MD:Z:2C41C2 YT:Z:UU NH:i:3 CC:Z:chr15 CP:i:102518319 XS:A:+ HI:i:0

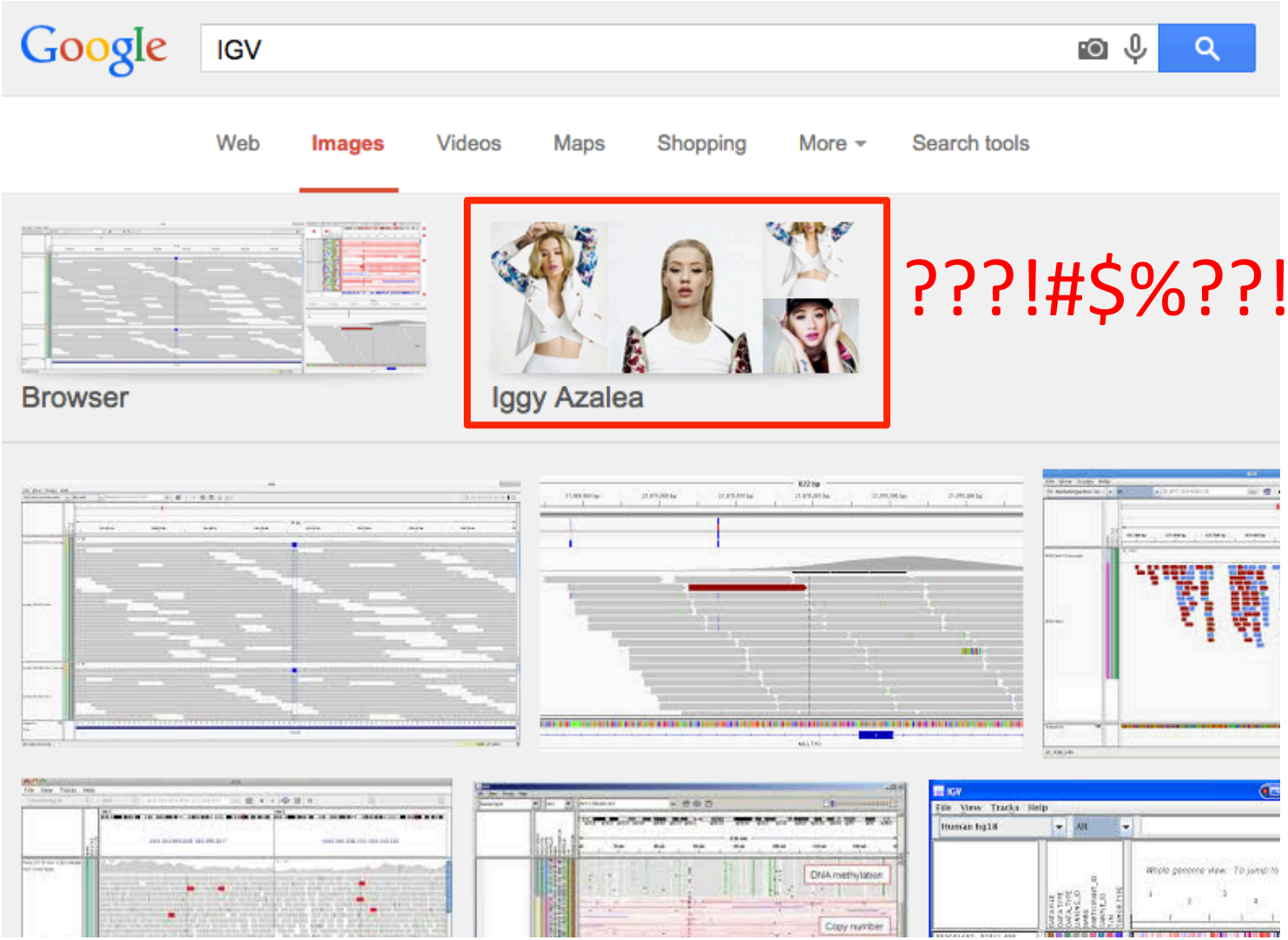
Optional fields: alignment details/program specific

Visualizing SAM files

Integrative genomics viewer (IGV):



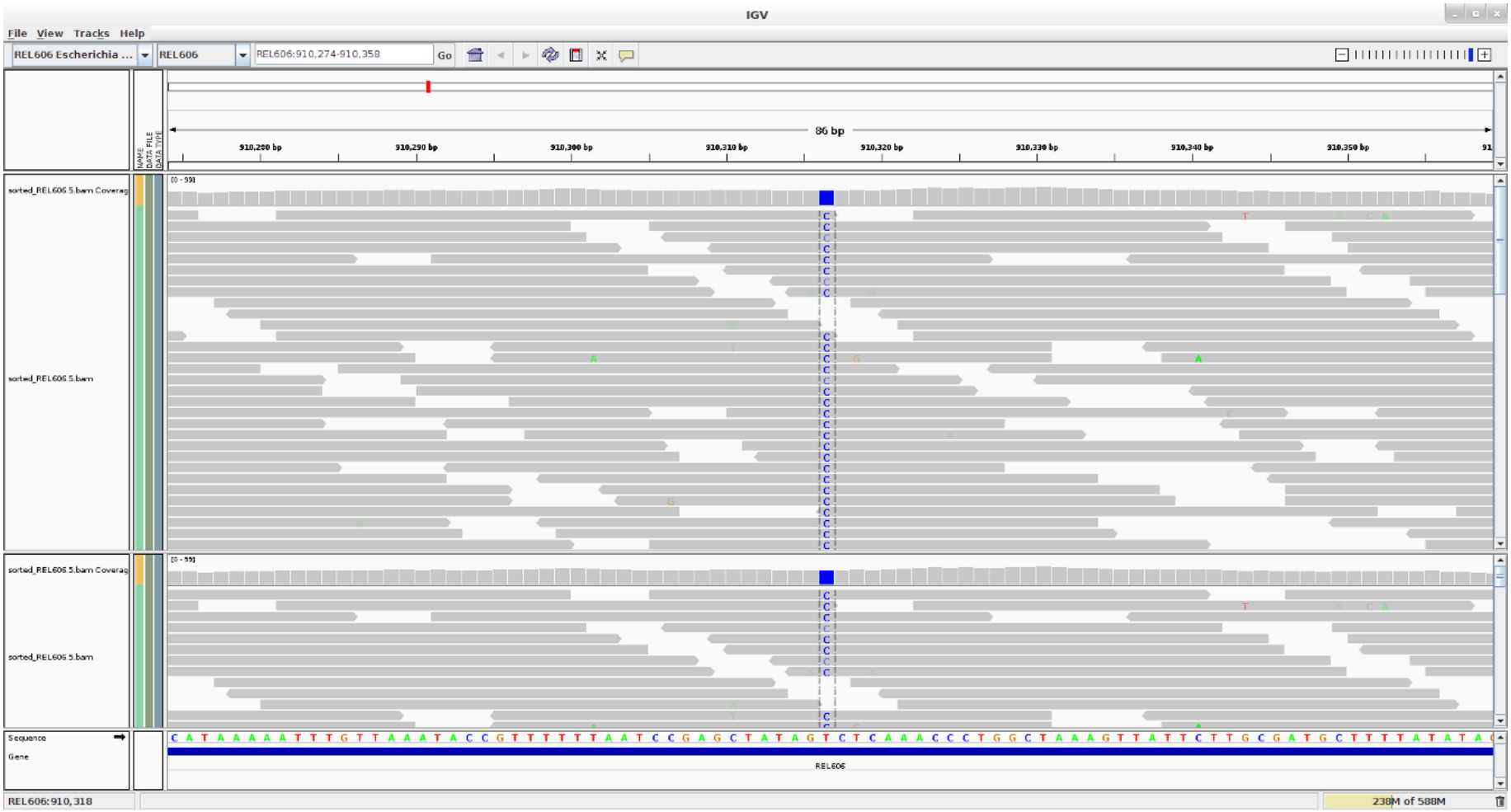
Visualizing SAM files



???!#$%??!?

Visualizing SAM files

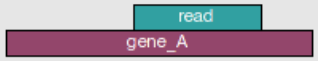
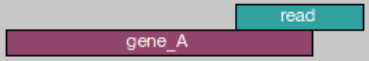
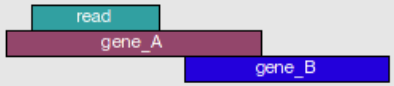
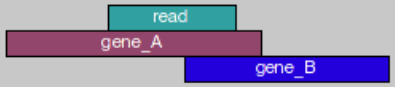
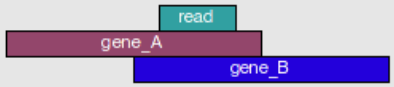
Integrative genomics viewer (IGV):



Quantifying Gene Counts

htseq-count –m <mode> -t <feature_type> -i <id_attribute> <output.sam> <gtf_ref_seq_file>

sorted
tools sort

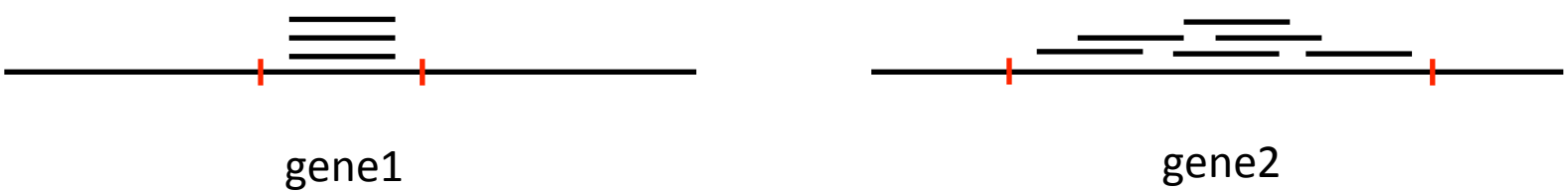
	union	intersection _strict	intersection _nonempty
	gene_A	gene_A	gene_A
	gene_A	no_feature	gene_A
	gene_A	no_feature	gene_A
	gene_A	gene_A	gene_A
	gene_A	gene_A	gene_A
	ambiguous	gene_A	gene_A
	ambiguous	ambiguous	ambiguous

Normalizing transcript counts

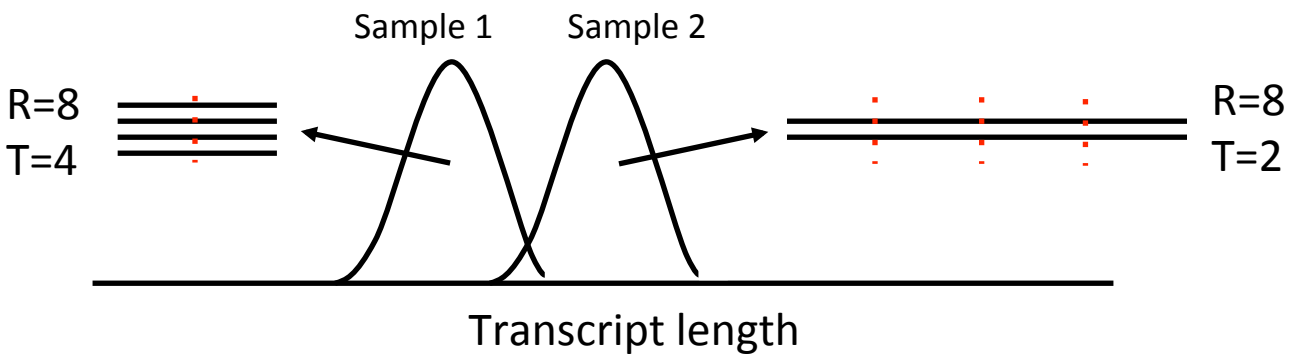
Reads (fragments) per kilobase of exon per million fragments mapped (RPKM or FPKM) normalizes for:

- Total number of reads
- Gene lengths

$$\text{RPKM} = \frac{\text{\# reads mapped to a gene}}{\text{exon length}/10^3 * \text{total \# reads mapped}/10^6}$$



Downside of RPKM



DESeq2

	sample1	sample2
gene1	4	2
gene2	10	5
Size Factor:	20	10

↓ Divide by size factor

	sample1	sample2
gene1	.2	.2
gene2	.5	.5

Questions?

TPM

Transcripts per million (TPM) normalizes:

- Total number of transcripts
- Gene lengths
- Read lengths/gene coverage

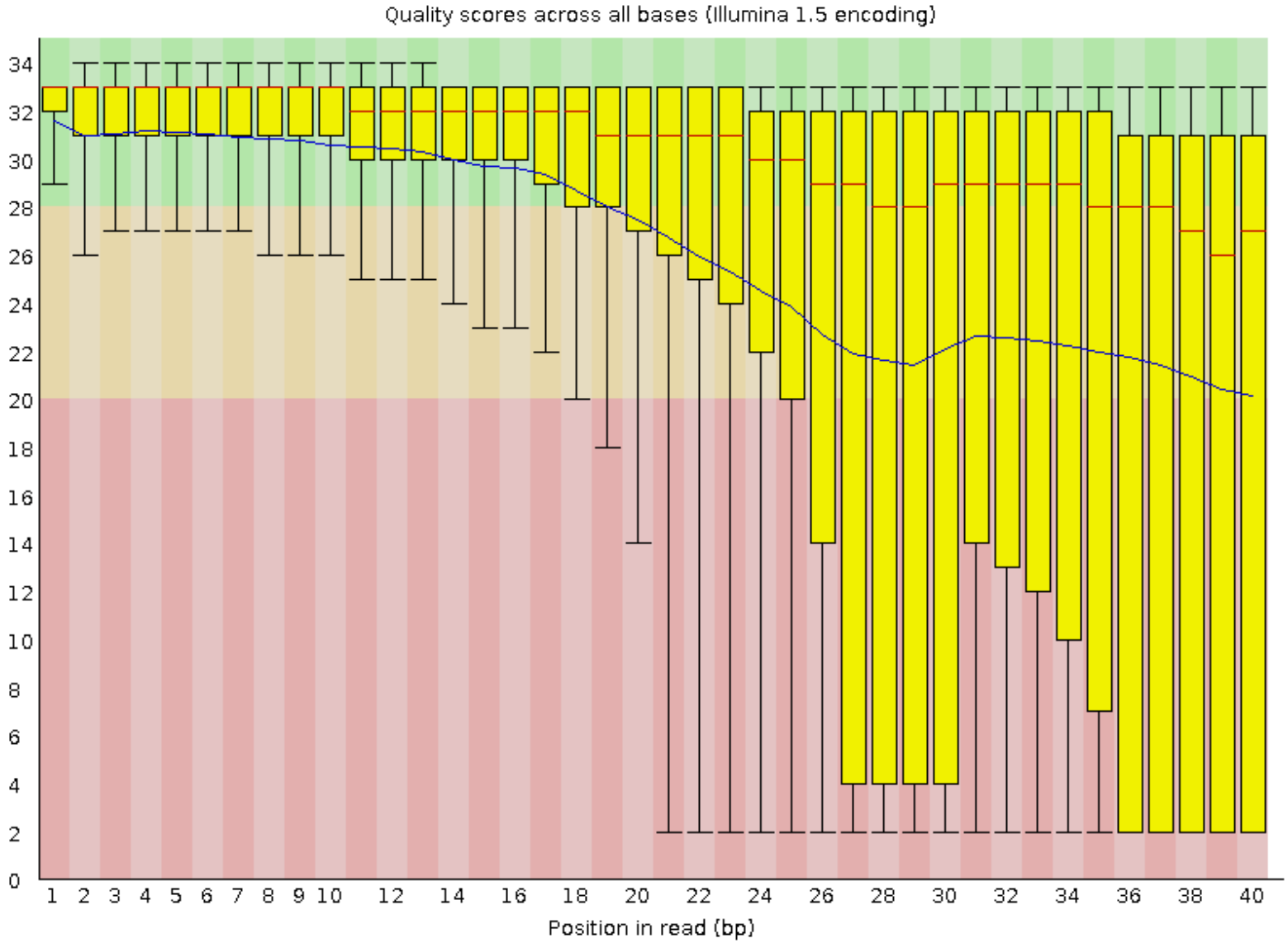
$$\text{TPM} = \frac{\text{\# reads mapped to a gene * mean read length}}{\text{exon length} * T/10^6}$$

$$T = \sum \frac{\text{\# reads mapped to a gene*mean read length}}{\text{exon length}}$$

What is the sequence quality of your reads?

```
fastqc <file.fastq>
```

Bad sequence quality



Good sequence quality