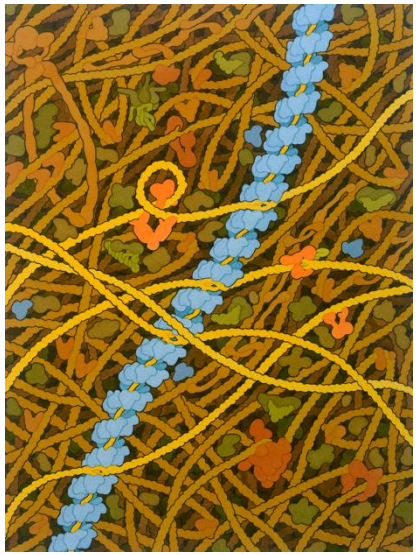
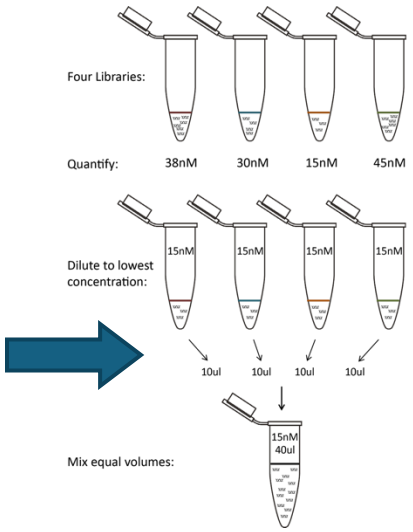
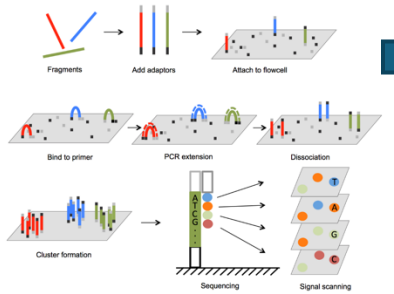
Week 4 Intro

NICHD RNAseq Course, Spring 2025

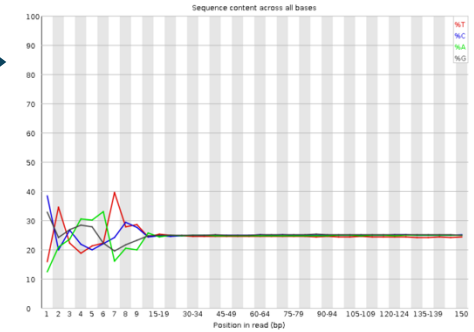
February 21st, 2025








```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

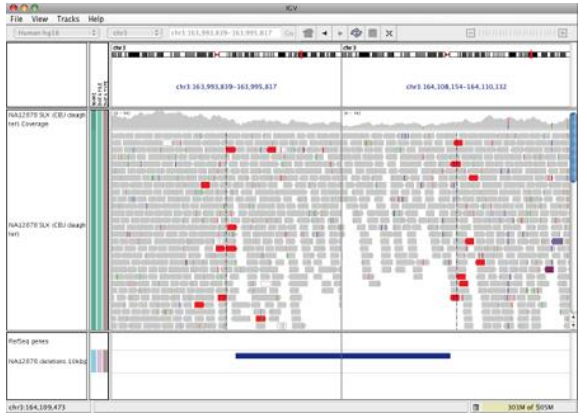


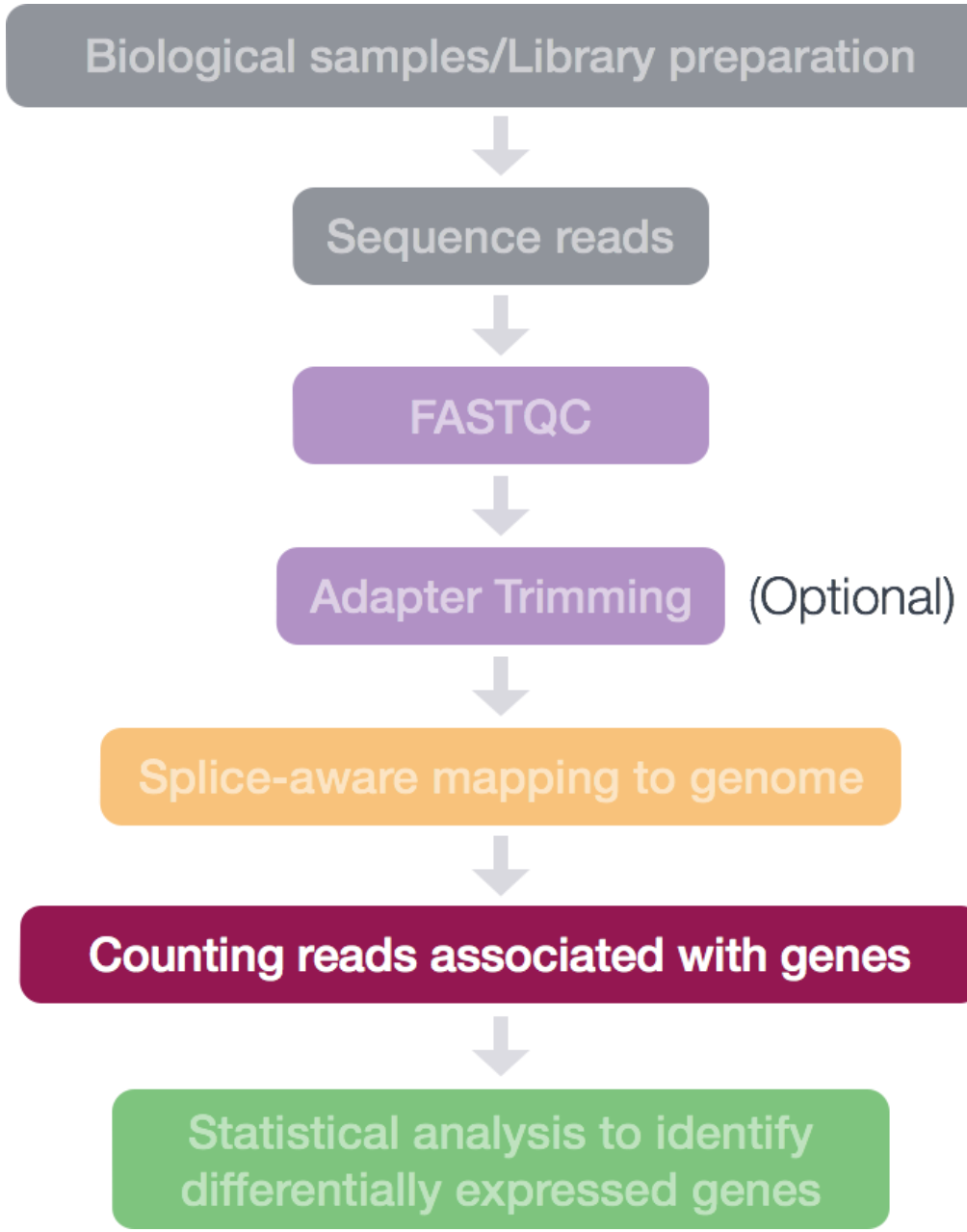
Aligner + index

```
SEQ_1  16  chr1  105  60  10M  *  0  0  GATTTGGGGT   !''*((((**   XG:i:0 NH:i:1 NM:i:1
```

For each sequence in each library, we now know where it came from in the genome. This is powerful information!

This week – quantifying!




Biological samples/Library preparation
Sequence reads
FASTQC
Adapter Trimming
(Optional)
Splice-aware mapping to genome
Counting reads associated with genes
Statistical analysis to identify differentially expressed genes

Counting reads in genes

- featureCounts is a popular command line tool for counting reads in features.
  - Input: BAMs (one for each sample) and GTF
  - Output: TSV with number of reads in each gene and each sample
- Only known genes can be quantified. We may have reads in novel genes, those reads will be ignored
- Typically, BAM reads falling in introns are also ignored
- There are alternative quantification methods (e.g. pseudoalignment with Salmon or Kallisto) that will not be discussed today

Simply speaking, the genomic coordinates of where the read is mapped (BAM) are cross-referenced with the genomic coordinates of whichever feature you are interested in counting expression of (GTF), it can be exons, genes or transcripts.

GTF

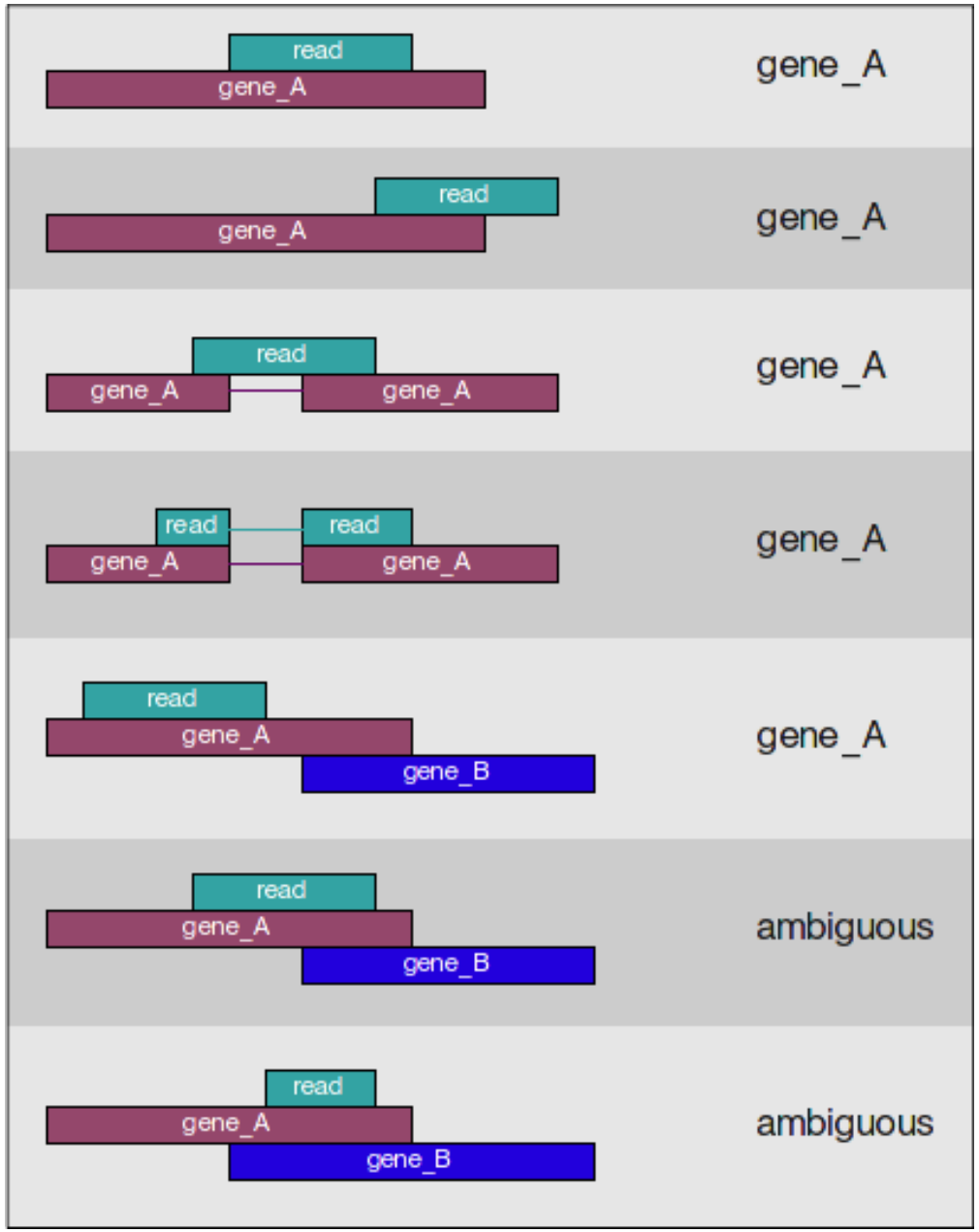
```
chr1 unknown  exon         113217048  113217252   .   +   .   gene_id "MOV10";p_id "P5535";transcript_id "NM_001130079"
chr1 unknown  exon         113217048  113217351   .   +   .   gene_id "MOV10";p_id "P5535";transcript_id "NM_020963"
chr1 unknown  exon         113217470  113217671   .   +   .   gene_id "MOV10";p_id "P5535";transcript_id "NM_001130079"
chr1 unknown  CDS          113217535  113217671   .   +   0   gene_id "MOV10";p_id "P5535";transcript_id "NM_001130079"
chr1 unknown  start_codon  113217535  113217537   .   +       gene_id "MOV10";p_id "P5535";transcript_id "NM_001130079"
```

feature type

feature

We are specifically looking at *gene* counts

Schematic for featureCounts assigns reads to genes:



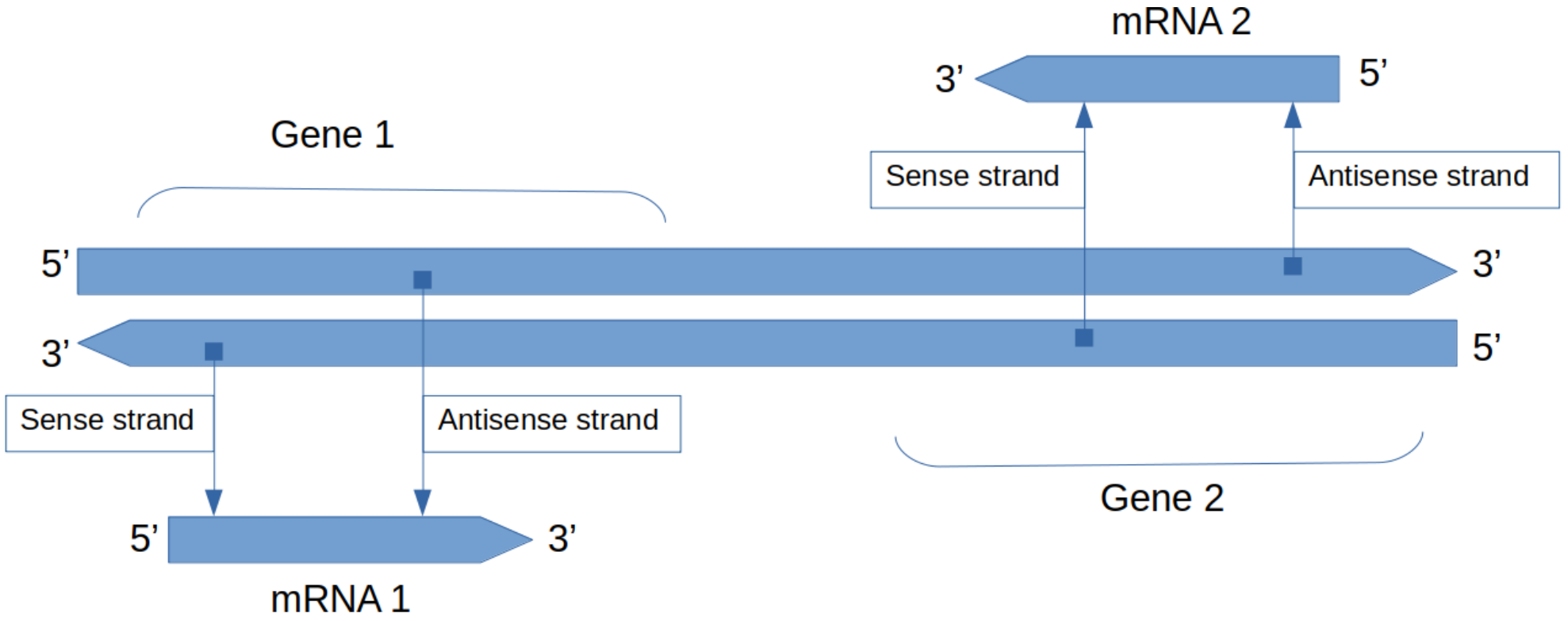
FeatureCounts output

Geneid	Irrel_kd_1	Irrel_kd_2	Irrel_kd_3	Mov10_oe_1	Mov10_oe_2	Mov10_oe_3
ENSG000002	0	0	0	0	0	1
ENSG000002	0	0	0	0	0	0
ENSG000003	82	86	65	119	106	52
ENSG000002	0	0	0	0	0	0
ENSG000002	0	0	0	0	0	0
ENSG000002	0	2	1	0	3	3
ENSG000002	0	0	0	0	0	0
ENSG000002	0	0	0	0	0	0
ENSG000003	0	0	0	0	0	0
ENSG000002	0	0	0	0	0	0
ENSG000002	0	0	0	0	0	0
ENSG000002	0	0	0	0	0	0
ENSG000001	0	0	0	0	0	0
ENSG000002	2	0	0	2	1	0
ENSG000002	0	0	0	0	0	0
ENSG000003	0	0	0	0	0	0
ENSG000002	0	0	0	0	0	0
ENSG000003	1	1	0	0	1	0

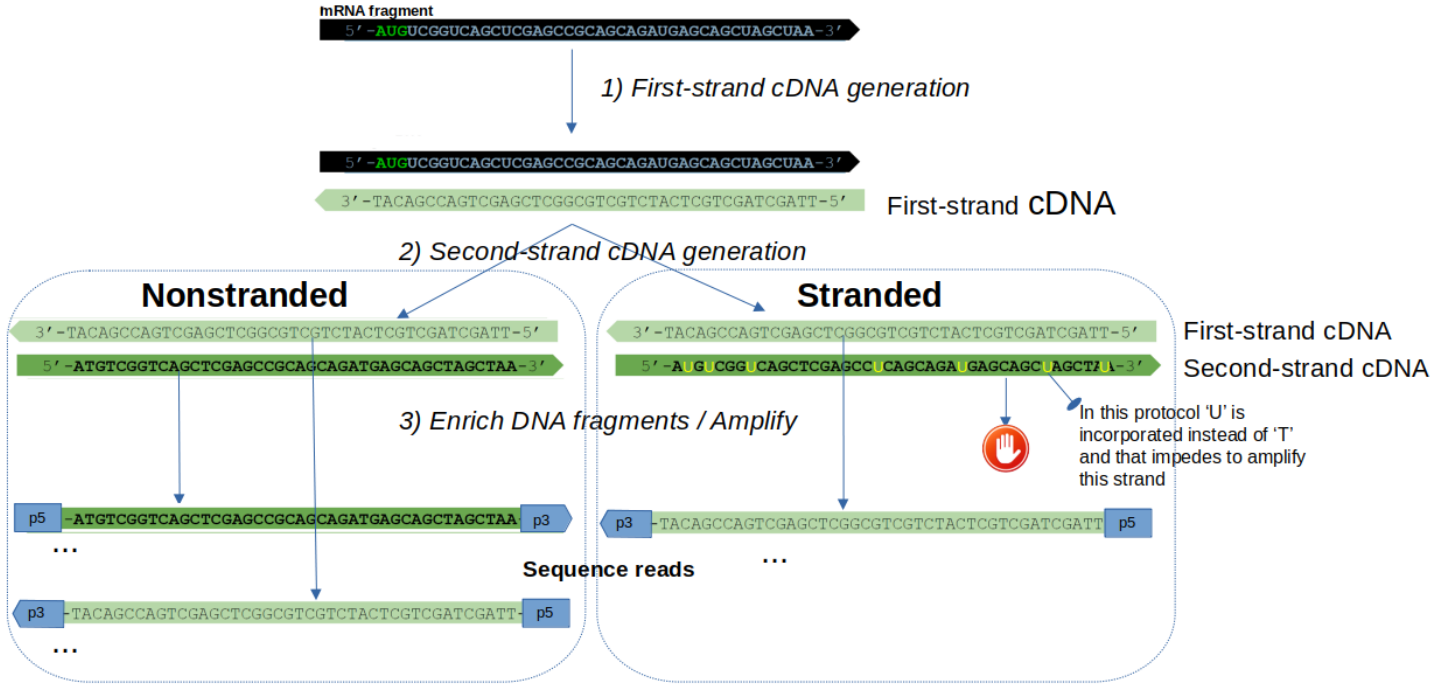
A *featurecounts* command

```
featureCounts -T 4 \  #four threads
 -s 2 \  #reverse-stranded
-a annotation.gtf \
-o your_output_featurecounts.txt
\ /STAR_results/*.bam
```

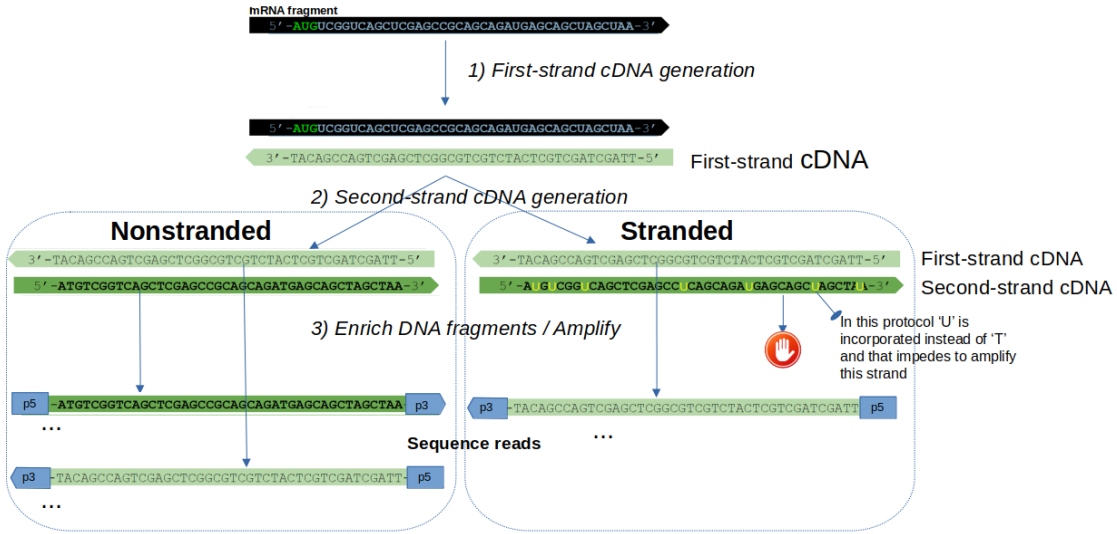
Strandedness of your data



https://bioinfogp.cnb.csic.es/tools/seqnjoy/help/#by-way-of-reminder-strandness-in-transcription



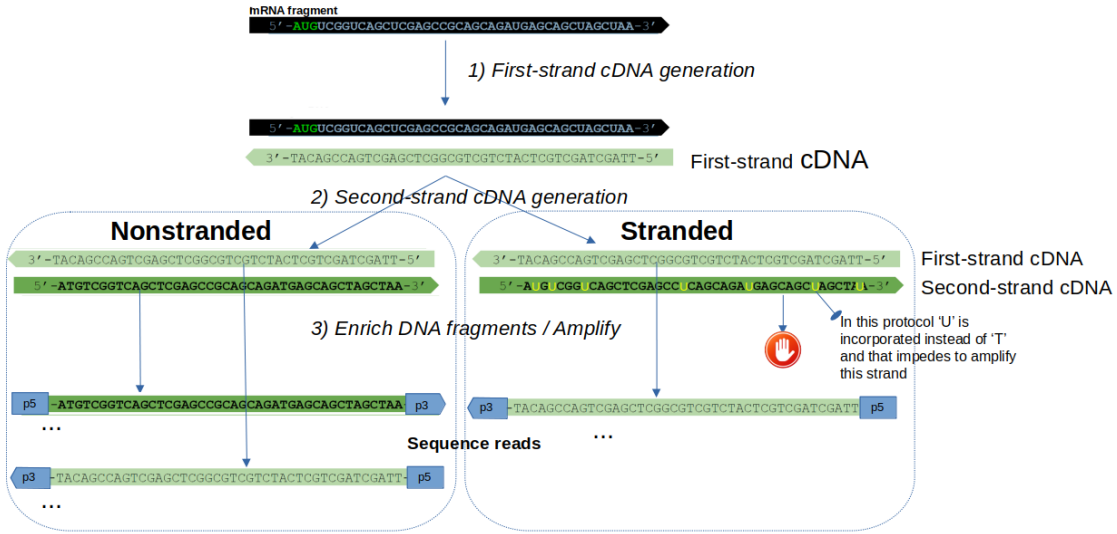
Different protocols preserve information about strandedness, or not



Protocol	Synonym	Examples/Kits[1]	Description
Unstranded	non-stranded	Standard Illumina, TruSeq RNA Sample Prep kit, NuGEN OvationV2, SMARTer universal low input (TaKara), GDC normalized TCGA data	Information regarding the strand is not conserved (lost)
Reverse	first_strand	dUTP, NSR, NNSR; TruSeq Stranded (Total RNA/mRNA), NEB Ultra Directional, Agilent SureSelect Strand-Specific	The second read (read 2) is from the original RNA strand/template, first read (read 1) is from the opposite (reverse) strand.
Direct	second_strand	Directional Illumina (Ligation), Standard SOLiD, ScriptSeq v2, SMARTer Stranded Total RNA, NuGEN Encore Complete, NuGEN SoLo, Illumina ScriptSeq	The first read (read 1) is from the original (direct sense) RNA strand/template, second read (read 2) is from the opposite strand.

Different protocols preserve information about strandedness, or not

https://bioinfogp.cnb.csic.es/tools/seqnjoy/help/#rna-seq-protocols-unstranded-vs.stranded



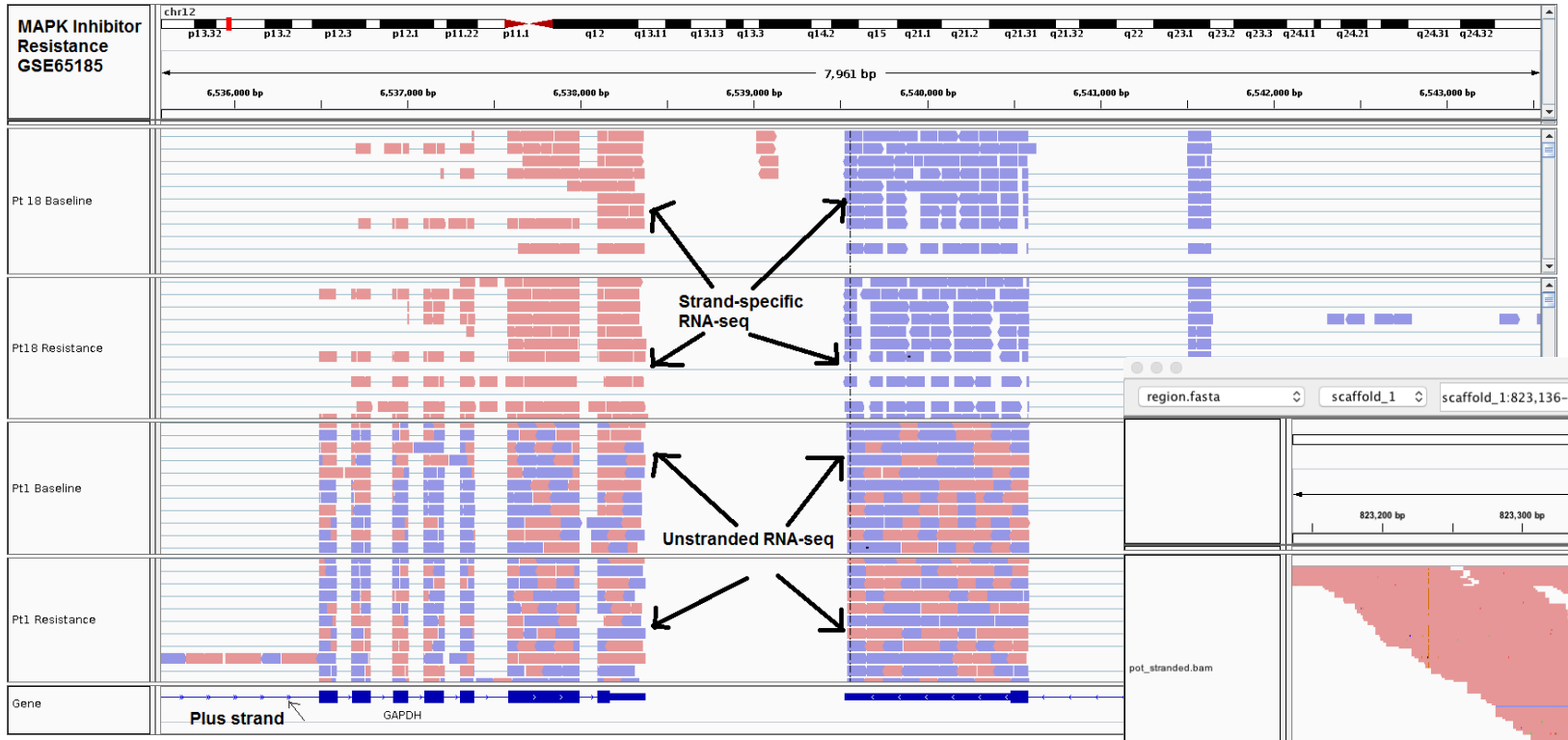
The example data is **single-end reverse-stranded**

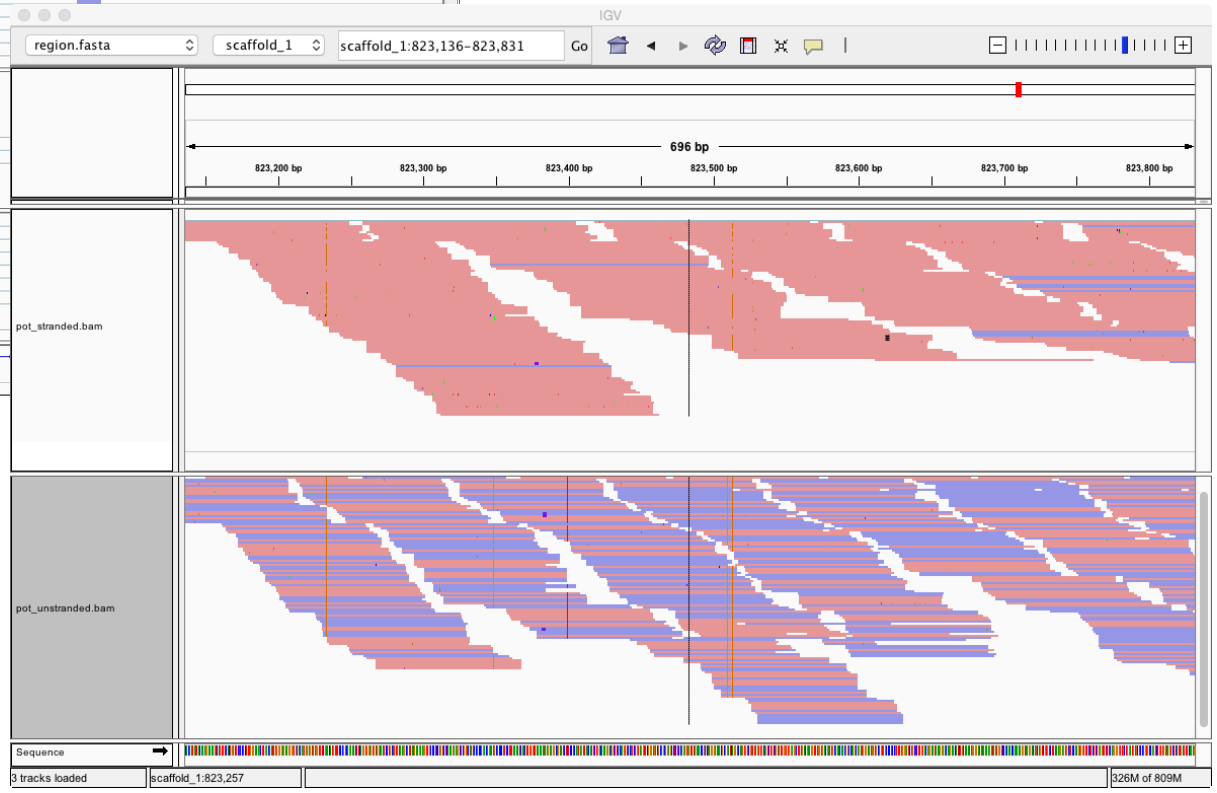
Protocol	Synonym	Examples/Kits[1]	Description
Unstranded	non-stranded	Standard Illumina, TruSeq RNA Sample Prep kit, NuGEN OvationV2, SMARTer universal low input (TaKara), GDC normalized TCGA data	Information regarding the strand is not conserved (lost)
Reverse	first_strand	dUTP, NSR, NNSR; TruSeq Stranded (Total RNA/mRNA), NEB Ultra Directional, Agilent SureSelect Strand-Specific	The second read (read 2) is from the original RNA strand/template, first read (read 1) is from the opposite (reverse) strand.
Direct	second_strand	Directional Illumina (Ligation), Standard SOLiD, ScriptSeq v2, SMARTer Stranded Total RNA, NuGEN Encore Complete, NuGEN SoLo, Illumina ScriptSeq	The first read (read 1) is from the original (direct sense) RNA strand/template, second read (read 2) is from the opposite strand.

Different protocols preserve information about strandedness, or not

https://bioinfogp.cnb.csic.es/tools/seqnjoy/help/#rna-seq-protocols-unstranded-vs.stranded

How can you tell strandedness?





visualized the mappings in IGV and turned on the option Color Alignments By -> First-of-Pair Strand

https://darencard.net/blog/2019-09-13-determining-stranded-rnaseq/

Strandedness: CLI tools

Infer_experiment.py from RSeQC package – also feeds into MultiQC!

Example 3: Single-end strand specific:

```
infer_experiment.py -r hg19.refseq.bed12 -i SingleEnd_StrandSpecific_36mer_Human_hg19.bam

#Output::

This is SingleEnd Data
Fraction of reads failed to determine: 0.0170
Fraction of reads explained by "++,--": 0.9669
Fraction of reads explained by "+-,-+": 0.0161
```

Interpretation: This is single-end, strand specific RNA-seq data. Strandness of reads are concordant with strandness of reference gene.

https://rseqc.sourceforge.net/#infer-experiment-py

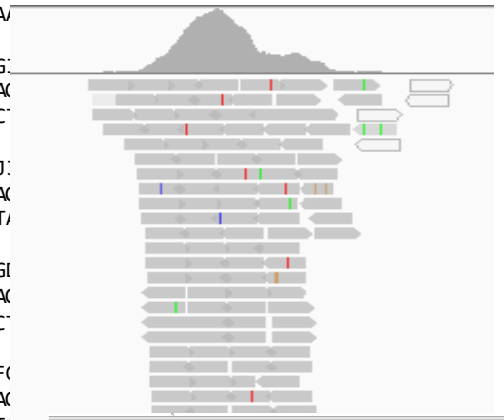
A *featurecounts* command

```
featureCounts -T 4 #four threads -s 2 \  #reverse-stranded
  -a annotation.gtf \
  -o your_output_featurecounts.txt \
   /STAR_results/*.bam
```

Output: TSV of counts from featureCounts, ready for downstream analysis in R

Geneid	Irrel_kd_1	Irrel_kd_2	Irrel_kd_3	Mov10_oe_1	Mov10_oe_2	Mov10_oe_3
ENSG000002	0	0	0	0	0	1
ENSG000002	0	0	0	0	0	0
ENSG000003	82	86	65	119	106	52
ENSG000002	0	0	0	0	0	0
ENSG000002	0	0	0	0	0	0
ENSG000002	0	2	1	0	3	3
ENSG000002	0	0	0	0	0	0
ENSG000002	0	0	0	0	0	0
ENSG000003	0	0	0	0	0	0
ENSG000002	0	0	0	0	0	0
ENSG000002	0	0	0	0	0	0
ENSG000002	0	0	0	0	0	0
ENSG000001	0	0	0	0	0	0
ENSG000002	2	0	0	2	1	0
ENSG000002	0	0	0	0	0	0
ENSG000003	0	0	0	0	0	0
ENSG000002	0	0	0	0	0	0
ENSG000003	1	1	0	0	1	0

```
@SRR948304.5141 UNC14-SN744:253:D135LACXX:5:1101:4054:4287
AGGACTTTGGCTGTCCCCAACCGTACAGGTGGGTCTCTCCCTCATGGA
+
CCCFFFFFHHHHHGHIIJIJJJIEHGIJJCGGJAFHEIJDHGIFGGE<
@SRR948304.5160 UNC14-SN744:253:D135LACXX:5:1101:4625:4251
CCACGAAGTCAAGATCGCCGACAAGGCCTTCCTGATGA
+
BCCFFFFFHGHHHJJJJJJJJIJGJJJJJJJJJJFIJJG
@SRR948304.5196 UNC14-SN744:253:D135LA
GGAACGCCTACTGGTACTATCTGAACATGGACTACTCC
+
CCCFFFFFFHHHHJGHIIJJJJJJJJJIJJJIJIIJJJJ
@SRR948304.5200 UNC14-SN744:253:D135LA
CAGGAAGAAGGAGTAGTCCATGTTCAGATAGTACCAGT
+
@@CFFBDDHHHBHIHIFGHHIGFFHHIJJJJFIIJIJG
@SRR948304.5215 UNC14-SN744:253:D135LA
CCTTCTTCAACGACTACTACACCAAGTACTTCAAGTTC
+
@@@FFFFFHGDHFIEHGIJIFEHIHH?EFGGIGGG:BF
@SRR948304.5247 UNC14-SN744:253:D135LA
CTACAGCTTCCGCAAGAACTACTACGAGATGGAGACCT
+
BB@FFFFFHHHHFIJIJJJIJIJIJJIGGGIIGGGIIJ
@SRR948304.5250 UNC14-SN744:253:D135LA
CCTGGTTGAACTCGTAGATATTCTCGCGCAGAATGAAC
+
CCCFFFFFHHHHHJHGHIJIJJJJJJJJJJJIIIJJIJ
```

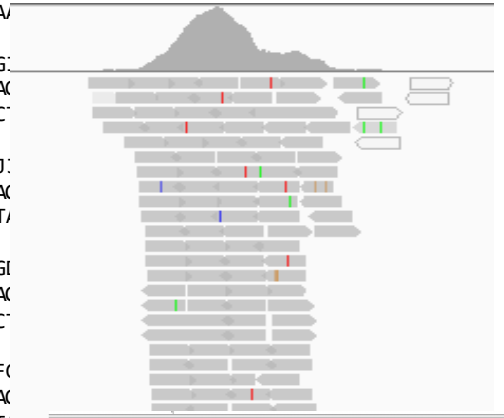


We are here!

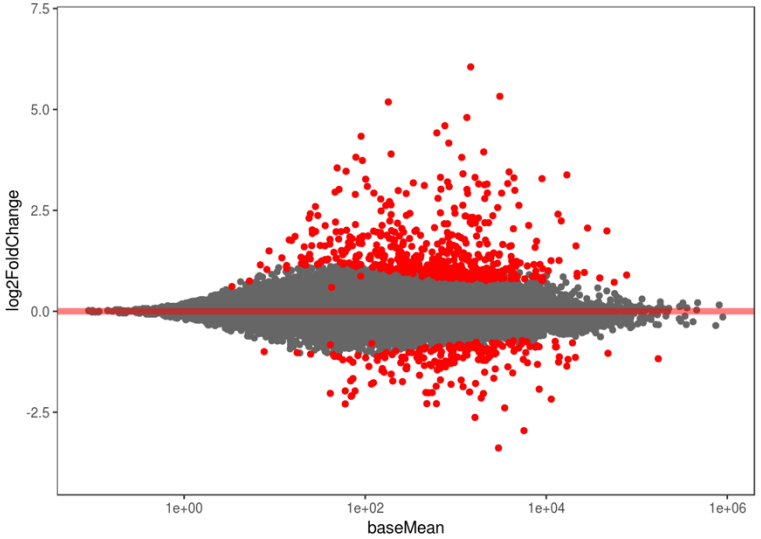
	A	B	C	D	E	F	G	H	I	J
1	Geneid	Chr	Start	End	Strand	Length	Ecoli_LB_04_1	Ecoli_LB_04_2	Ecoli_M63_04_1	Ecoli_M63_04_2
2	EG11277	chr	190	255	+	66	667	977	3906	4234
3	EG10998	chr	337	2799	+	2463	47	104	442	538
4	EG10999	chr	2801	3733	+	933	36	50	186	249
5	EG11000	chr	3734	5020	+	1287	60	86	386	451
6	G6081	chr	5234	5530	+	297	22	23	18	32
7	EG10011	chr	5683	6459	-	777	25	58	38	25
8	EG11555	chr	6529	7959	-	1431	17	37	35	32
9	EG11556	chr	8238	9191	+	954	609	998	454	573
10	EG11511	chr	9306	9893	+	588	91	149	41	54
11	EG11512	chr	9928	10494	-	567	22	30	29	20
12	G6082	chr	10643	11356	-	714	2	7	4	9
13	EG11509	chr	10830	11315	+	486	0	1	2	3
14	G8202	chr	11382	11786	-	405	1	2	2	0
15	EG10241	chr	12163	14079	+	1917	149	487	203	210
16	G0-8893	chr	14080	14168	+	89	17	41	13	21
17	EG10240	chr	14168	15298	+	1131	33	65	54	81
18	G6083	chr	15445	16557	+	1113	44	65	56	52
19	EG10373	chr	16751	16960	-	210	1	2	1	3
20	G0-9563	chr	16751	16903	-	153	0	0	0	0

Prepping this data for statistics

```
@SRR948304.5141 UNC14-SN744:253:D135LACXX:5:1101:4054:4287
AGGACTTTGGCTGTCCCCAACCGTACAGGTGGGTCTCTCCCTCATGGA
+
CCCFFFFFHHHHHGHIIJIJJJIEHGIJJCGGJAFHEIJDHGIFGGE<
@SRR948304.5160 UNC14-SN744:253:D135LACXX:5:1101:4625:4251
CCACGAAGTCAAGATCGCCGACAAGGCCTTCCTGATGA
+
BCCFFFFFHGHHHJJJJJJJJIJGJJJJJJJJJJFIJJG
@SRR948304.5196 UNC14-SN744:253:D135LA
GGAACGCCTACTGGTACTATCTGAACATGGACTACTCC
+
CCCFFFFFFHHHHJGHIIJJJJJJJJJIJJJIJIIJJJJ
@SRR948304.5200 UNC14-SN744:253:D135LA
CAGGAAGAAGGAGTAGTCCATGTTCAGATAGTACCAGT
+
@@CFFBDDHHHBHIHIFGHHIGFFHHIJJJJFIIJIJG
@SRR948304.5215 UNC14-SN744:253:D135LA
CCTTCTTCAACGACTACTACACCAAGTACTTCAAGTTC
+
@@@FFFFFHGDHFIEHGIJIFEHIHH?EFGGIGGG:BF
@SRR948304.5247 UNC14-SN744:253:D135LA
CTACAGCTTCCGCAAGAACTACTACGAGATGGAGACCT
+
BB@FFFFFHHHHFIJIJJJIJIJIJJIGGGIIGGGIIJ
@SRR948304.5250 UNC14-SN744:253:D135LA
CCTGGTTGAACTCGTAGATATTCTCGCGCAGAATGAAC
+
CCCFFFFFHHHHHJHGHIJIJJJJJJJJJJJIIIJJIJ
```



	A	B	C	D	E	F	G	H	I	J
1	Geneid	Chr	Start	End	Strand	Length	Ecoli_LB_04_1	Ecoli_LB_04_2	Ecoli_M63_04_1	Ecoli_M63_04_2
2	EG11277	chr	190	255	+	66	667	977	3906	4234
3	EG10998	chr	337	2799	+	2463	47	104	442	538
4	EG10999	chr	2801	3733	+	9				
5	EG11000	chr	3734	5020	+	12				
6	G6081	chr	5234	5530	+	2				
7	EG10011	chr	5683	6459	-	7				
8	EG11555	chr	6529	7959	-	14				
9	EG11556	chr	8238	9191	+	9				
10	EG11511	chr	9306	9893	+	5				
11	EG11512	chr	9928	10494	-	5				
12	G6082	chr	10643	11356	-	7				
13	EG11509	chr	10830	11315	+	4				
14	G8202	chr	11382	11786	-	4				
15	EG10241	chr	12163	14079	+	19				
16	G0-8893	chr	14080	14168	+					
17	EG10240	chr	14168	15298	+	11				
18	G6083	chr	15445	16557	+	11				
19	EG10373	chr	16751	16960	-	2				
20	G0-9563	chr	16751	16903	-	1				

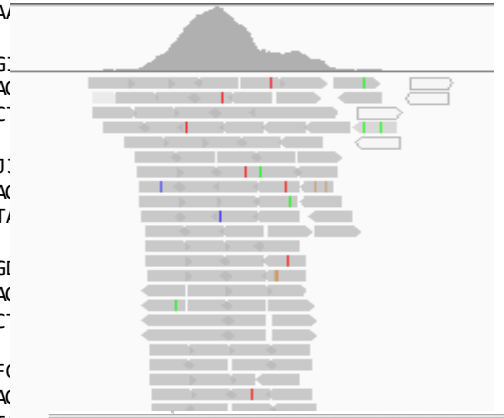


Still to come:

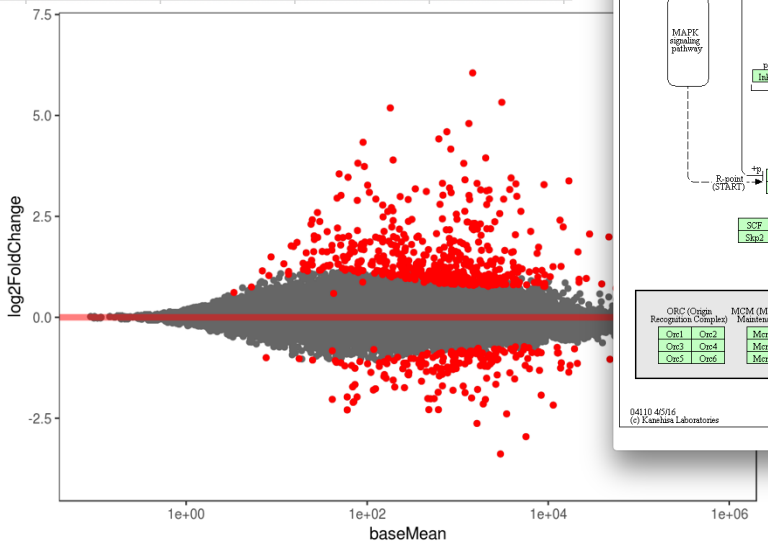
Statistical Testing

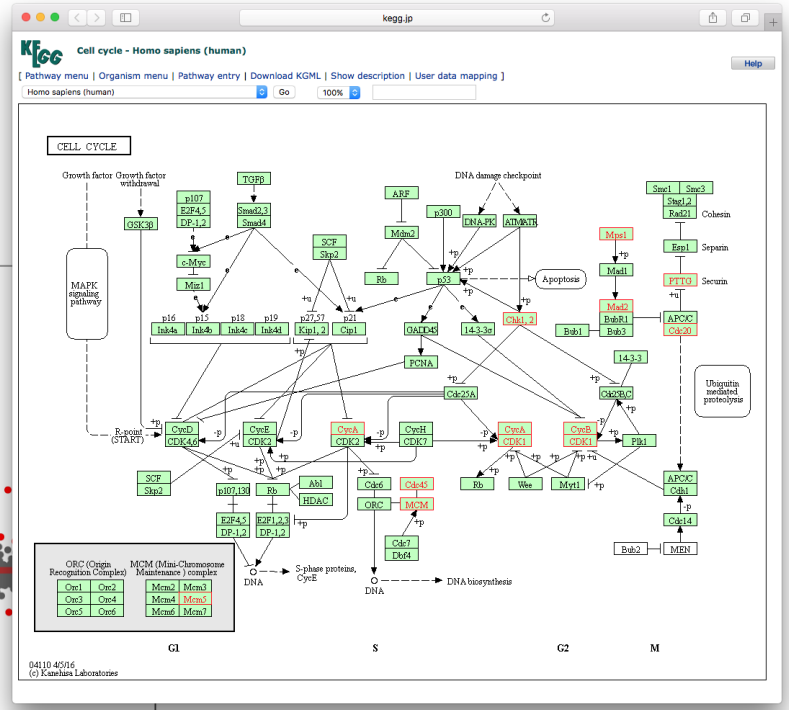
Still to come:

```
@SRR948304.5141 UNC14-SN744:253:D135LACXX:5:1101:4054:4287
AGGACTTTGGCTGTCCCCAACCGTACAGGTGGGTCTCTCCCTCATGGA
+
CCCFFFFFHHHHHGHIIJIJJJIEHGIJJCGGJAFHEIJDHGIFGGE<
@SRR948304.5160 UNC14-SN744:253:D135LACXX:5:1101:4625:4251
CCACGAAGTCAAGATCGCCGACAAGGCCTTCCTGATGA
+
BCCFFFFFHGHHHJJJJJJJJIJGJJJJJJJJJJFIJJG
@SRR948304.5196 UNC14-SN744:253:D135LA
GGAACGCCTACTGGTACTATCTGAACATGGACTACTCC
+
CCCFFFFFFHHHHJGHIIJJJJJJJJJIJJJJIIIJJJJ
@SRR948304.5200 UNC14-SN744:253:D135LA
CAGGAAGAAGGAGTAGTCCATGTTCAGATAGTACCAGT
+
@@CFFBDDHHHBHIHIFGHHIGFFHHIJJJJFIIJIJG
@SRR948304.5215 UNC14-SN744:253:D135LA
CCTTCTTCAACGACTACTACACCAAGTACTTCAAGTTC
+
@@@FFFFFHGDHFIEHGIJIFEHIHH?EFGGIGGG:BF
@SRR948304.5247 UNC14-SN744:253:D135LA
CTACAGCTTCCGCAAGAACTACTACGAGATGGAGACCT
+
BB@FFFFFHHHHFIJIJJJIJIJIJJIGGGIIGGGIIJ
@SRR948304.5250 UNC14-SN744:253:D135LA
CCTGGTTGAACTCGTAGATATTCTCGCGCAGAATGAAC
+
CCCFFFFFHHHHHJHGHIJIJJJJJJJJJJJJIIIJJIJ
```



	A	B	C	D	E	F	G	H	I	J
1	Geneid	Chr	Start	End	Strand	Length	Ecoli_LB_04_1	Ecoli_LB_04_2	Ecoli_M63_04_1	Ecoli_M63_04_2
2	EG11277	chr	190	255	+	66	667	977	3906	4234
3	EG10998	chr	337	2799	+	2463	47	104	442	538
4	EG10999	chr	2801	3733	+	9				
5	EG11000	chr	3734	5020	+	12				
6	G6081	chr	5234	5530	+	2				
7	EG10011	chr	5683	6459	-	7				
8	EG11555	chr	6529	7959	-	14				
9	EG11556	chr	8238	9191	+	9				
10	EG11511	chr	9306	9893	+	5				
11	EG11512	chr	9928	10494	-	5				
12	G6082	chr	10643	11356	-	7				
13	EG11509	chr	10830	11315	+	4				
14	G8202	chr	11382	11786	-	4				
15	EG10241	chr	12163	14079	+	19				
16	G0-8893	chr	14080	14168	+					
17	EG10240	chr	14168	15298	+	11				
18	G6083	chr	15445	16557	+	11				
19	EG10373	chr	16751	16960	-	2				
20	G0-9563	chr	16751	16903	-	1				





Biological insight

The few, the brave.

sed -i -e **'s/few/fun/g'** hello.txt

The fun, the brave.

The few, the brave.

sed -i -e **'** hello.txt**'s/few//g**
The , the brave.

/data/Bspc-training/shared/rnaseq_jan2025/results_for_counting/Mov10_oe_1Aligned.sortedByCoord.out.bam
/data/Bspc-training/shared/rnaseq_jan2025/results_for_counting/Mov10_oe_2Aligned.sortedByCoord.out.bam
/data/Bspc-training/shared/rnaseq_jan2025/results_for_counting/Mov10_oe_3Aligned.sortedByCoord.out.bam

Ideally we could do a search and replace like this:
:%s//**/data/Bspc-training/shared/rnaseq_jan2025/results_for_counting/**//g

But we need to “escape” the forward slashes that are part of the directory path by using backslashes for each one:
:%s/**/data\/Bspc-training\/shared\/rnaseq_jan2025\/results_for_counting\/**//g