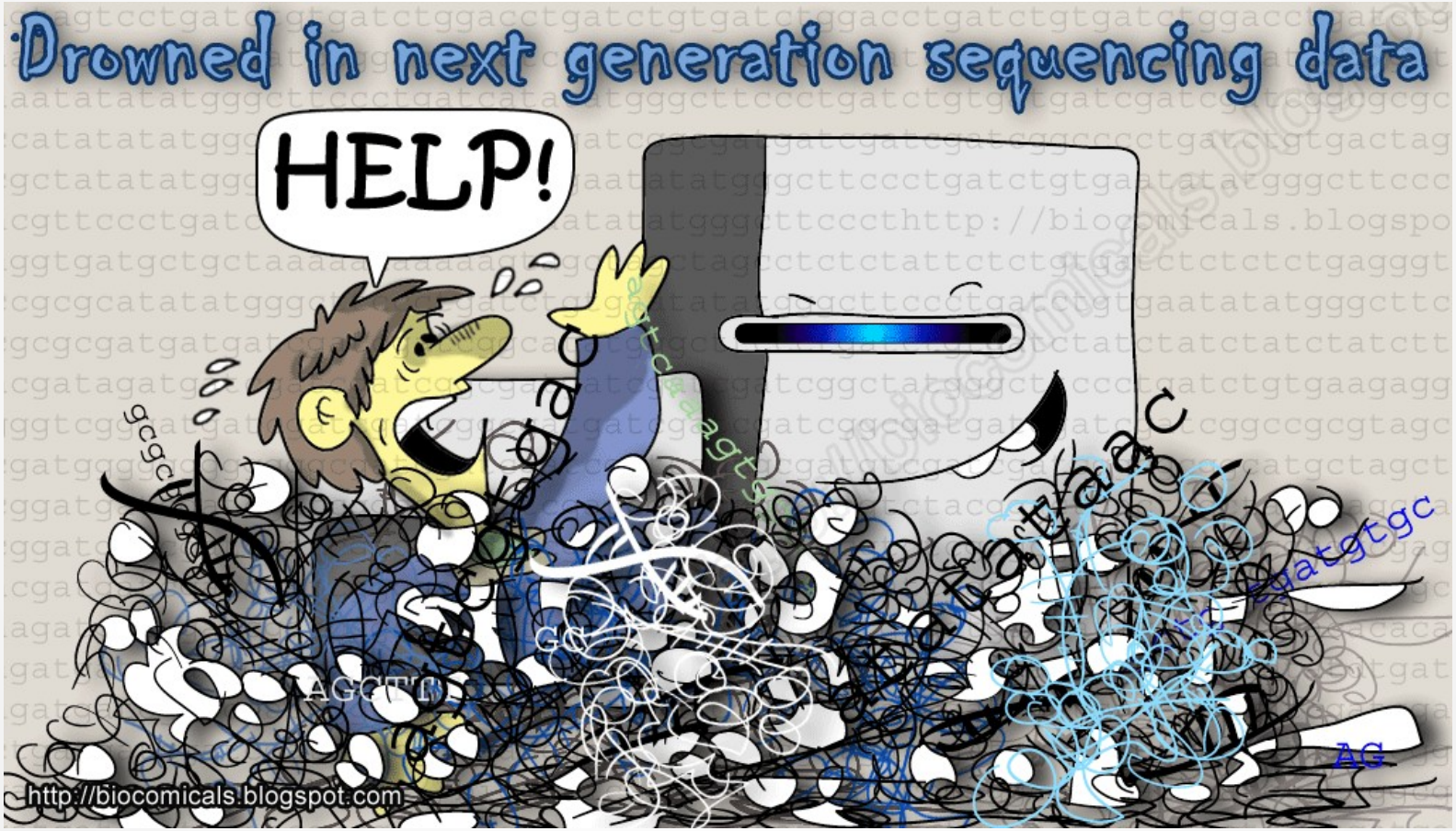
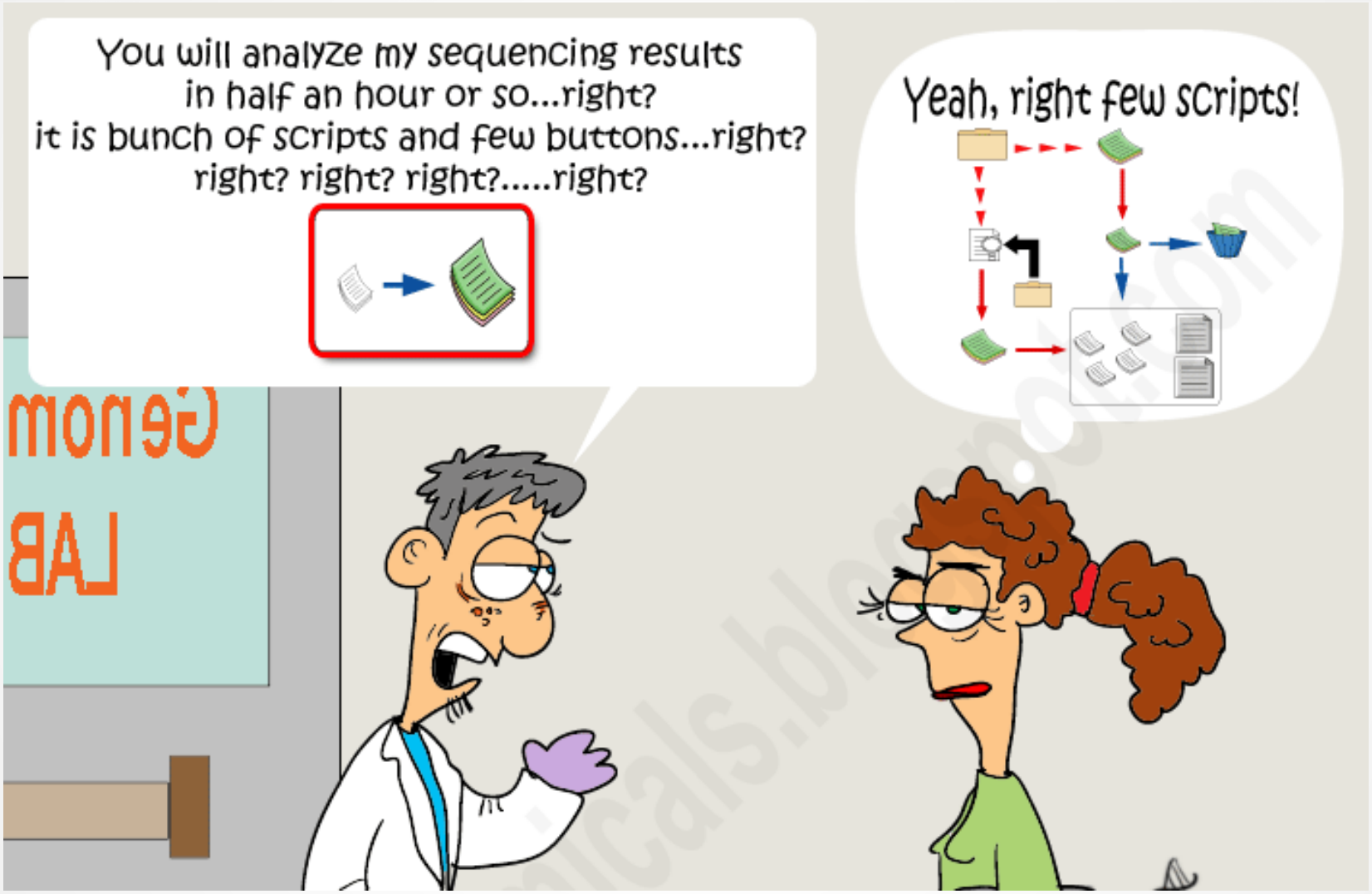
Basics of bioinformatics analysis

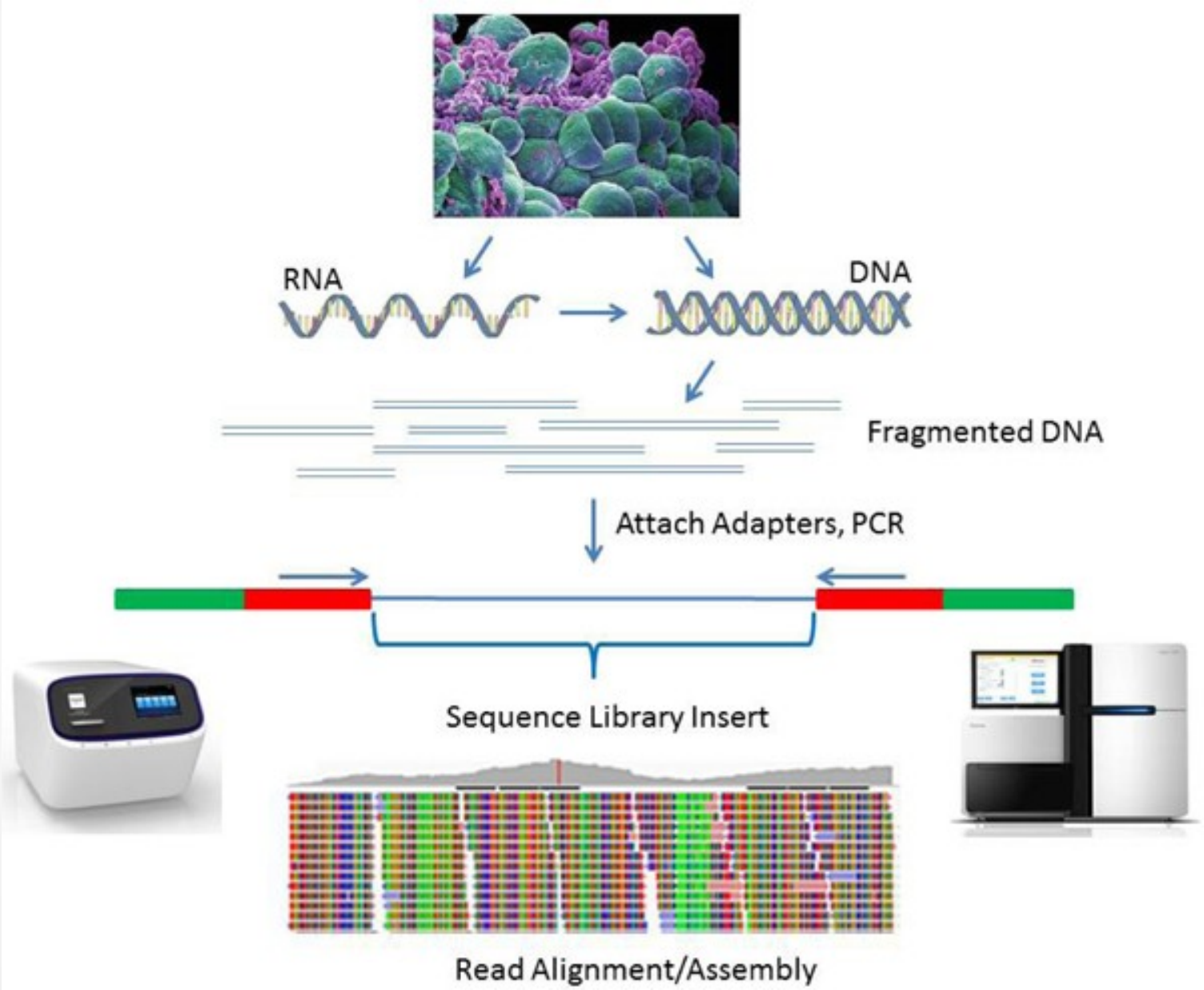


What is bioinfo for NGS analysis?



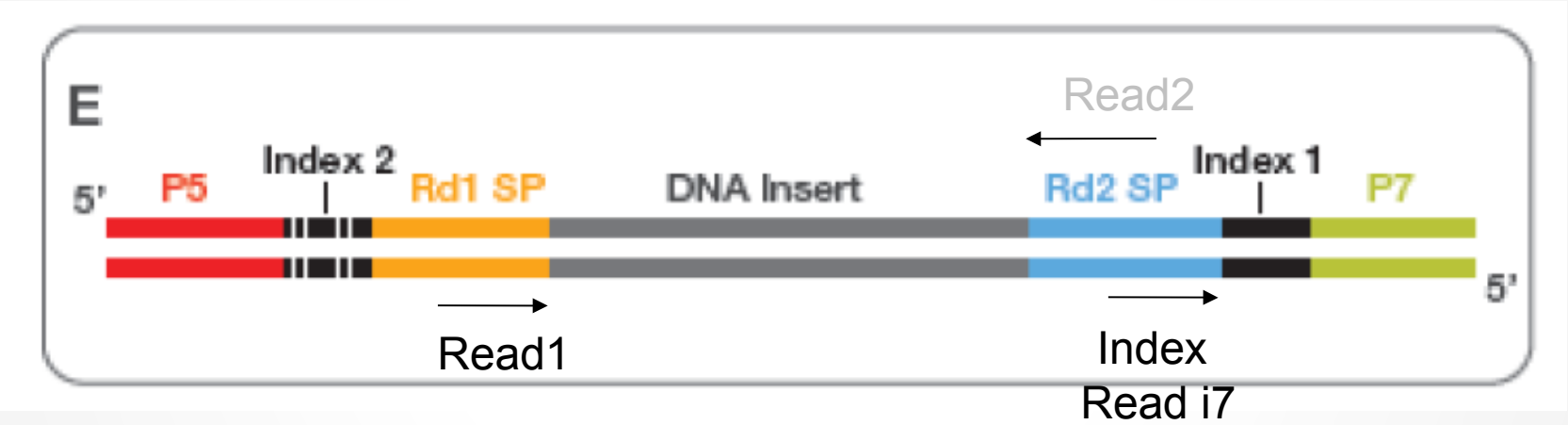
YES, Just few RIGHT(and PROPERLY WORKING) scripts !!

What to analyze?



Reminder on Illumina amplicon structure

(amplicon/library)

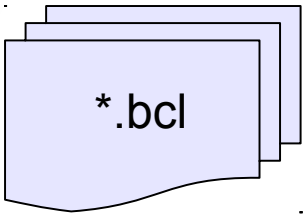


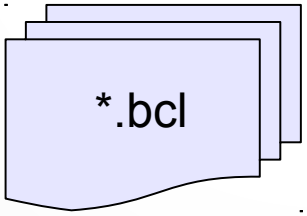
Raw sequencing data (*.bcl or intensity files)

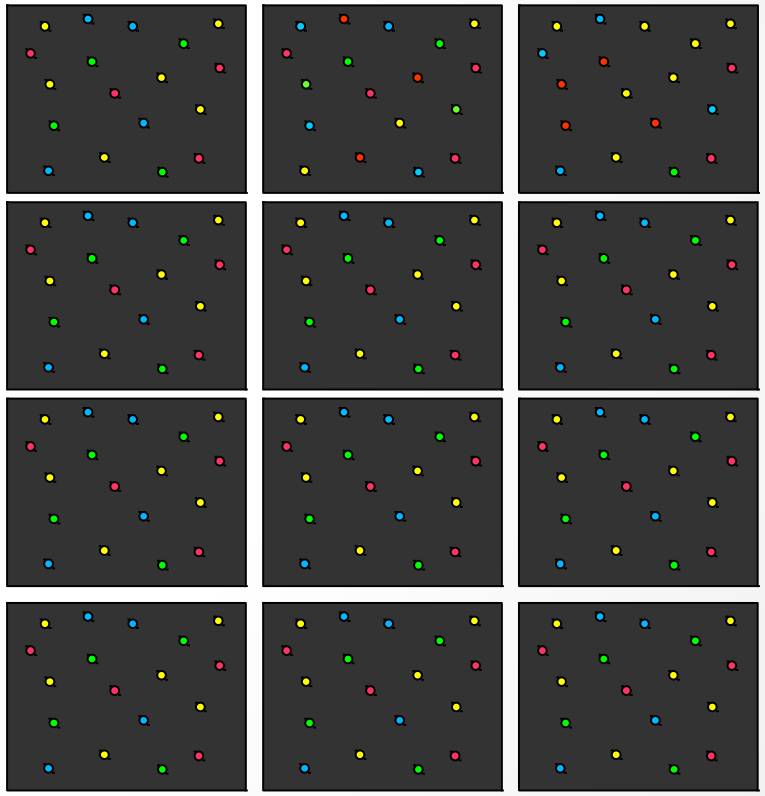
HiSeq

MiSeq

Intensity files

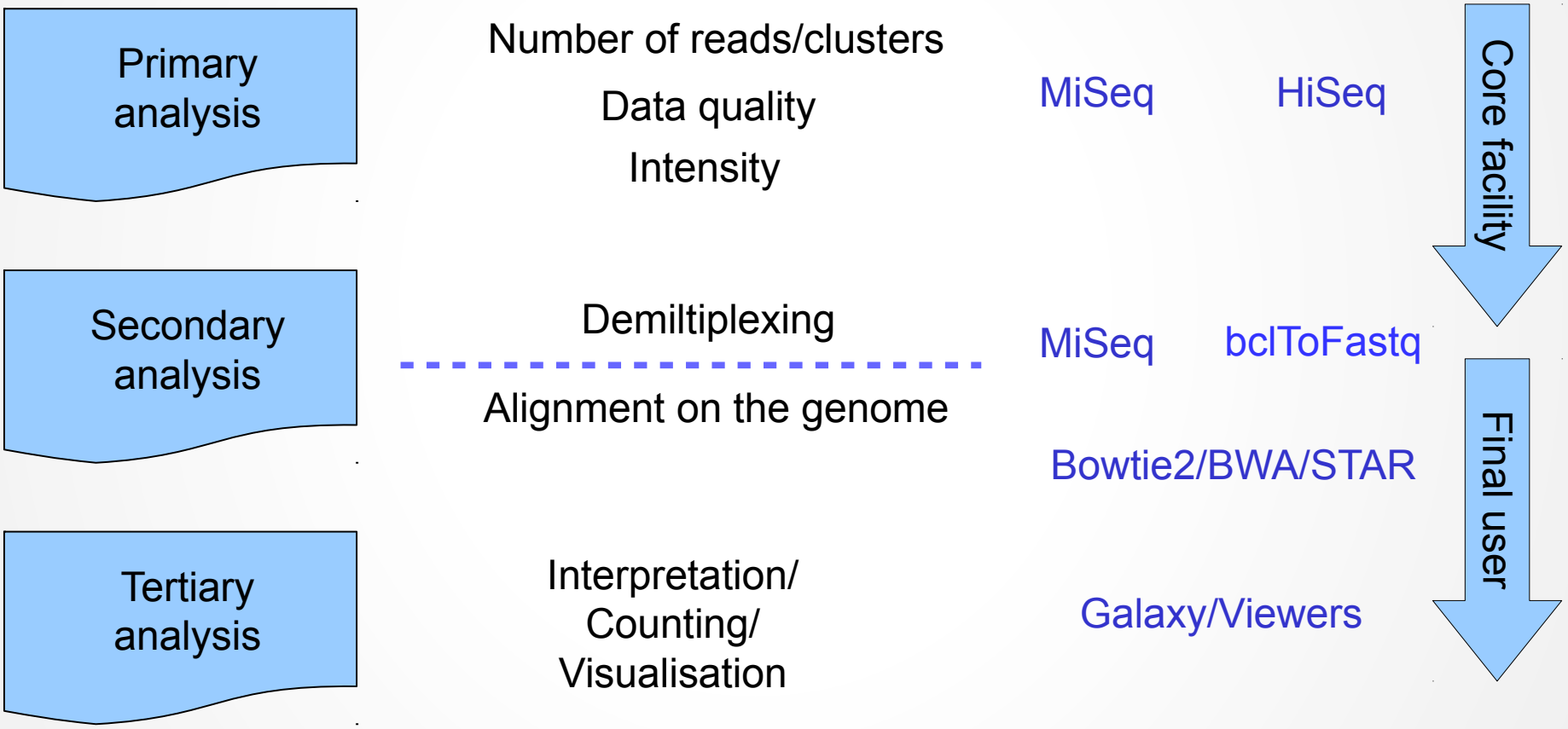






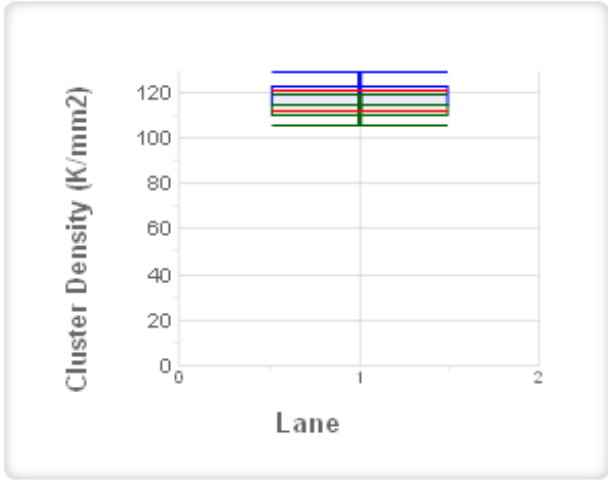
Coordinates X/Y of each cluster (spot) and intensity at a given cycle (binary file)

Notion of primary/secondary/tertiary data analysis

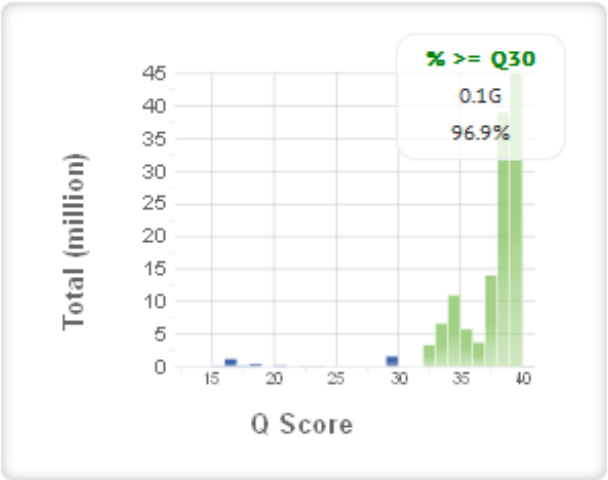


Primary analysis (raw data quality)

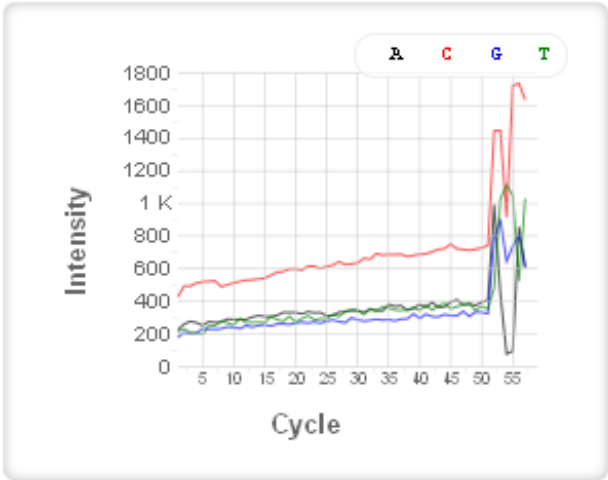
Number of reads/clusters

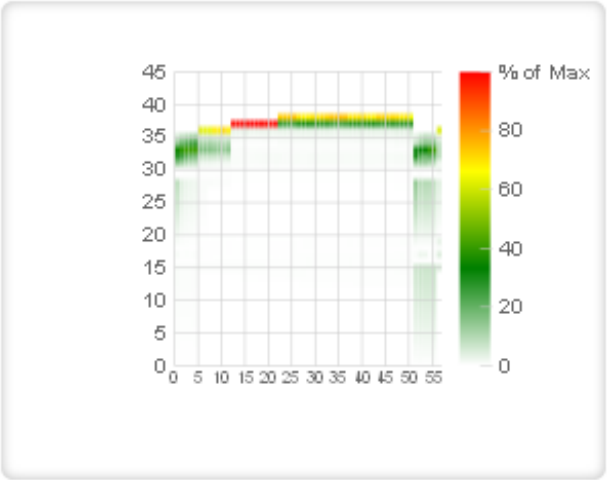


Data quality



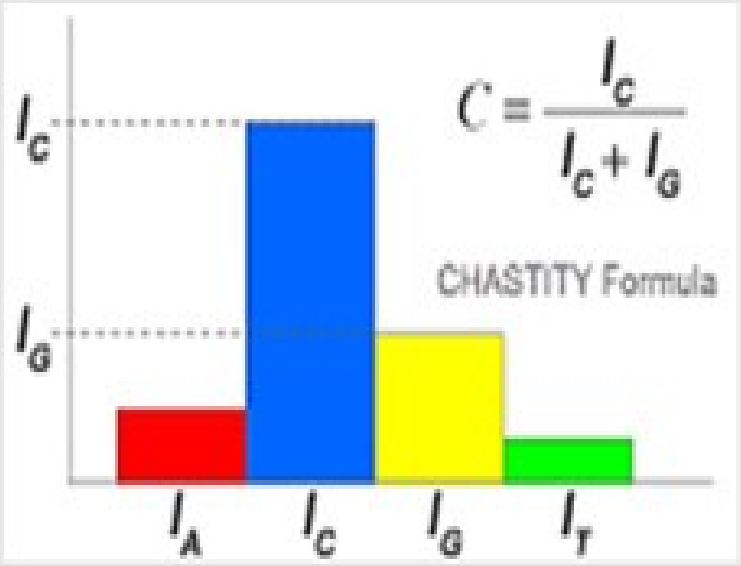
Intensity





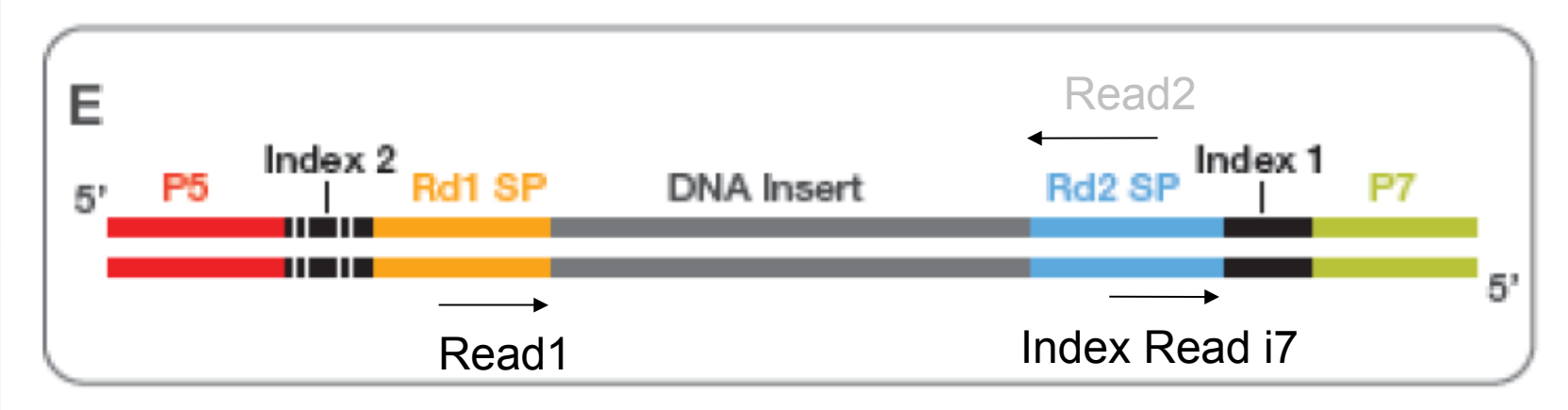
How to determine quality (CHASTITY filter) ?

Solexa CHASTITY filtering: Individual bases generated from original image files have quality scores which reflect the probability that a base-call is correct (or wrong), this is quantified by CHASTITY Formula (as shown in the figure below).



The chastity(C) of each base in the short reads is determined by the intensity of four colors (I_A, I_C, I_G, I_T here), the formula "the ratio of the highest (I_C here) of the four (base type) intensities to the sum of highest two(I_C and I_G here)." should be no less than 0.6 in the first 25 bases.

Reminder on indexing (barcoding)



These days the capacity of one sequencing lane is sufficient for many samples!

Multiplexing = mix several samples from one or different projects in one lane

Project 1

Sample 1	barcode 1
Sample 2	barcode 2
Sample 3	barcode 3
…..	…….

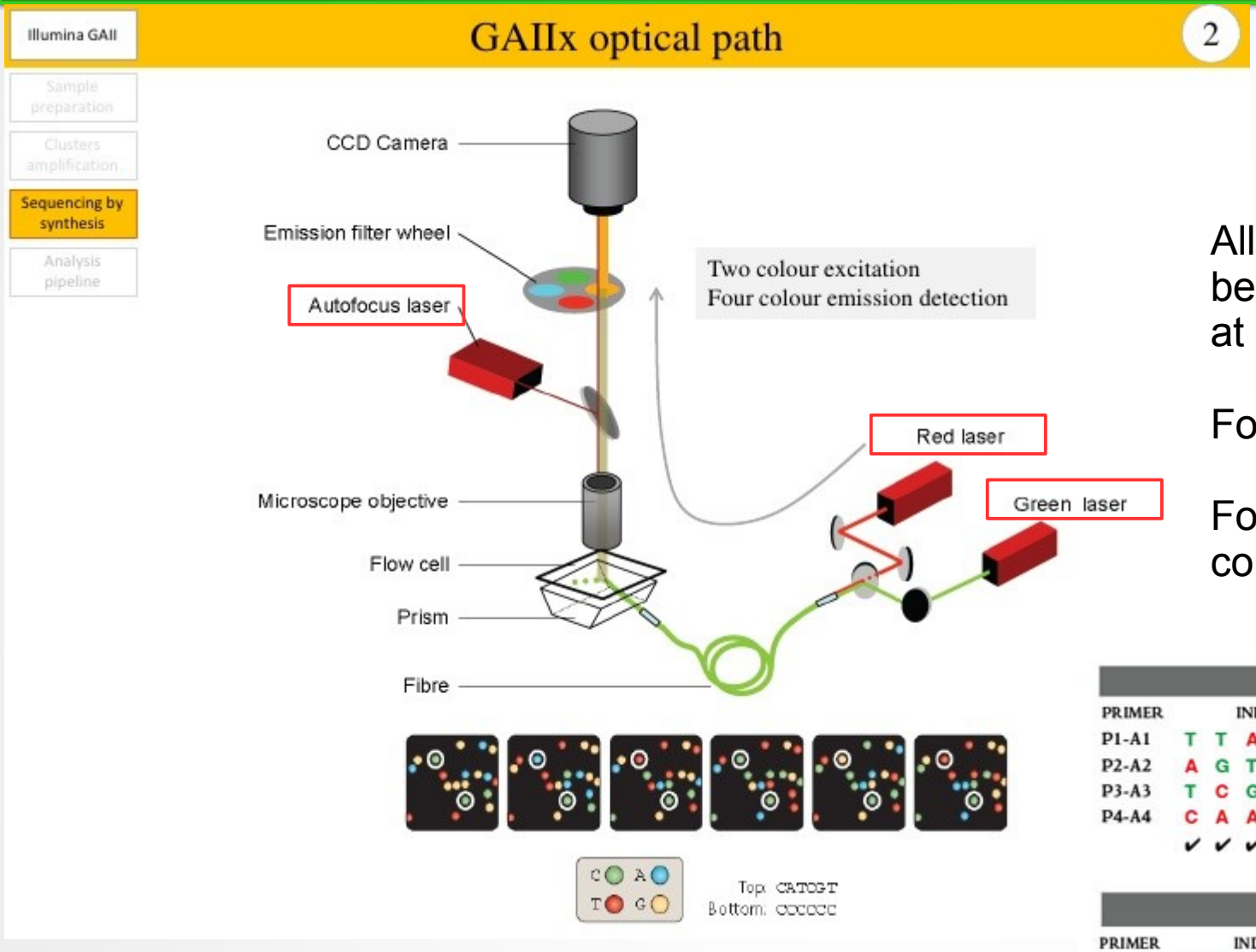
Project 2

Sample 1	barcode 11
Sample 2	barcode 12
Sample 3	barcode 13
…..	…….

Clusters

barcode 1
barcode 2
barcode 3
……..
barcode 11
barcode 12
barcode 13

Compatibility of indexes (barcodes) (Illumina rules)



Rule!

All four bases SHOULD be present in every cluster at every cycle!

For library diversity use PhiX

For indexes use correct combination of primers

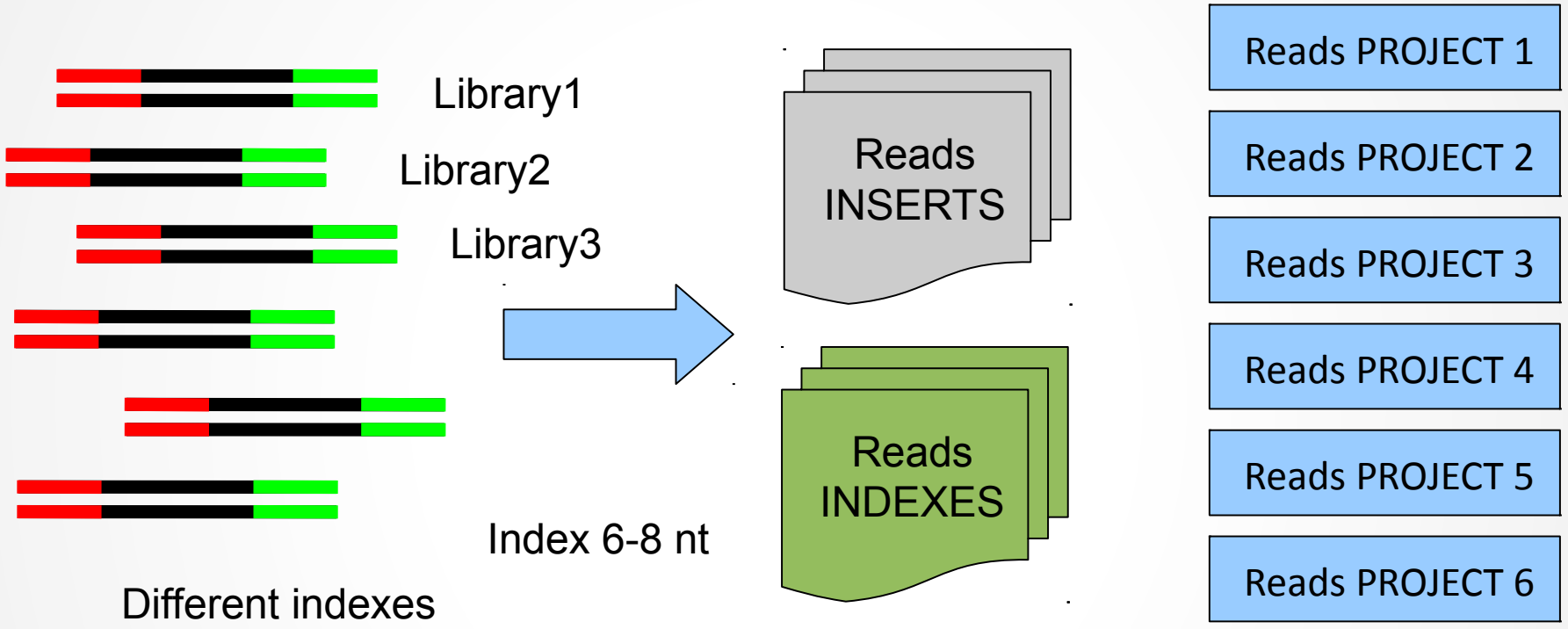
GOOD

PRIMER	INDEX SEQUENCE	PRIMER	INDEX SEQUENCE
P1-A1	T T A C C G A C	P41-D5	G A C G T C A T
P2-A2	A G T G A C C T	P42-D6	C T T A C A G C
P3-A3	T C G G A T T C	P43-D7	T C C A T T G C
P4-A4	C A A G G T A C	P44-D8	A G C G A G A T
	✔ ✔ ✔ ✔ ✔ ✔ ✔ ✔		✔ ✔ ✔ ✔ ✔ ✔ ✔ ✔

BAD

PRIMER	INDEX SEQUENCE	PRIMER	INDEX SEQUENCE
P9-A9	C G C A A C T A	P56-E8	T A T G G C A C
P10-A10	C G T A T C T C	P57-E9	C T C G A A C A
P11-A11	G T A C A C C T	P58-E10	C A A C T C C A
P12-A12	C G G C A T T A	P59-E11	G T C A T C G T
	✔ ✖ ✔ ✖ ✔ ✔ ✔ ✔		✔ ✔ ✔ ✔ ✔ ✖ ✔

Step 1 : Demultiplexing



Projects/samples and associated indexes (barcodes) are specified in the SampleSheet

And/or in Demiltiplexing sheet used by bclToFastq

Structure of FastQ file

Fastq file

unique instrument name
Run ID:flowcell ID

coordinates

N="passing filter"

Sequence

Q-score

```
@M01332:4:000000000-A76EU:1:1101:13298:1770 1:N:0:9
CCACCACGTTCCCGTGGTGGAATTCTCGGGTGCCAAGGAACTCCAGTCACG
+
BBBCCCCDEFFFGGGGGGGGGGHHHHHGGGHGGHHHHGHHHHHHHHHHHG
@M01332:4:000000000-A76EU:1:1101:16072:1846 1:N:0:9
CCGGCCGTGGAATTCTCGGGTGCCAAGGAACTCCAGTCACGGCTACATCTC
+
AAA3ADDAADAFDGGGFGGGGEFHHHHGFGHHHGFHFFGGGEEGGHBH5GH
@M01332:4:000000000-A76EU:1:1101:14553:2063 1:N:0:9
GGGCGGTGATGACCCCAACATGCCATCTGAGTGTCGGTGCTGAAATCC
+
>AAABCBBCFFFGGGGGGGGGGHHHHHHHHHFHGHGGGGGHHHHHHHH
@M01332:4:000000000-A76EU:1:1101:18239:2331 1:N:0:9
CTACGGGGATGATTTTACGAACTGAACTCTCTCTTTCTGATGGATTAGTGG
+
>BBBCCCDDDDFGGGGGGGGGGHHHHHHHHHHHHHHHHHGHHHHHHHHHH
@M01332:4:000000000-A76EU:1:1101:11665:2587 1:N:0:9
TATCTGTGATGATCTTATCCCGAACCTGAACTTCTGTTGAAAAAAAAAAAC
+
>AABBFFFFFFFGGGGGGGGGGGGGGHHHHHHHHHHHHHGHGHAFGGGGGG
```

Structure of FastQ file

Fastq file

unique instrument name
Run ID:flowcell ID

coordinates

N="passing filter"

Sequence

Q-score

```
@M01332:4:000000000-A76EU:1:1101:13298:1770 1:N:0:9
CCACCACGTTCCCGTGGTGGAATTCTCGGGTGCCAAGGAACTCCAGTCACG
+
BBBCCCCDEFFFGGGGGGGGGGHHHHHGGGHGGHHHHGHHHHHHHHHHHG
@M01332:4:000000000-A76EU:1:1101:16072:1846 1:N:0:9
CCGGCCGTGGAATTCTCGGGTGCCAAGGAACTCCAGTCACGGCTACATCTC
```

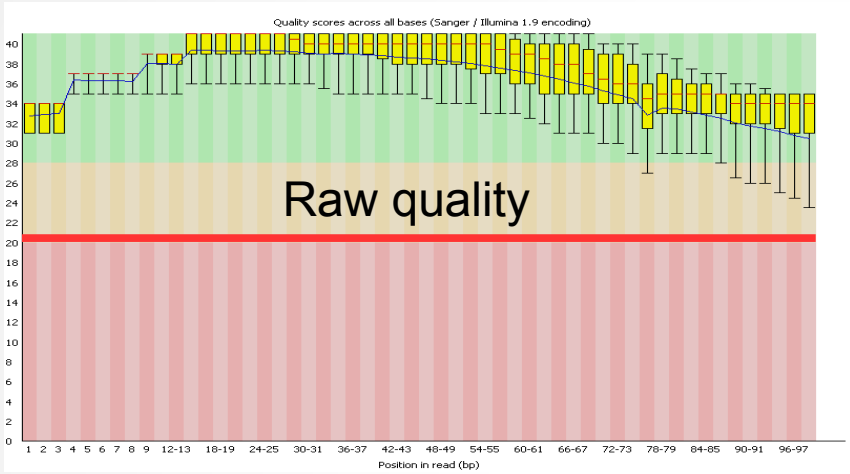
```
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJ
 0.2.....................26...31........41
```

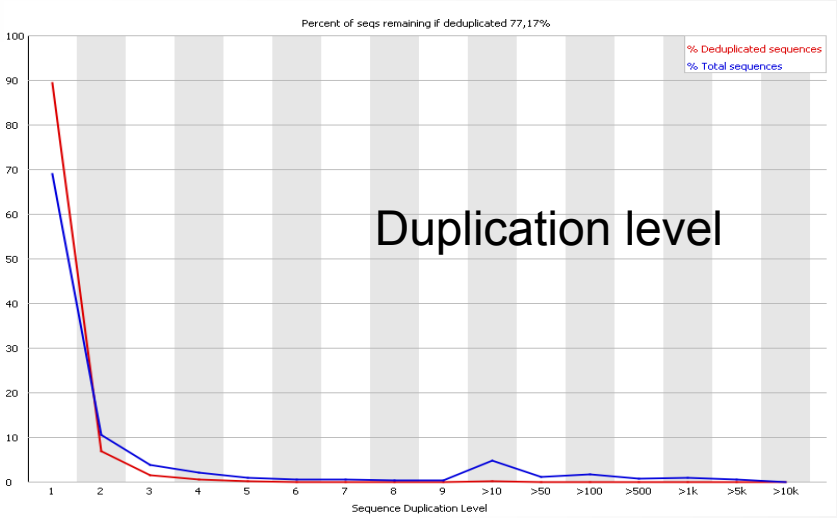
```
>AAABCBBCFFFGGGGGGGGGGHHHHHHHHHFHGHGGGGGHHHHHHHH
@M01332:4:000000000-A76EU:1:1101:18239:2331 1:N:0:9
CTACGGGGATGATTTTACGAACTGAACTCTCTCTTTCTGATGGATTAGTGG
+
>BBBCCCDDDDFGGGGGGGGGGHHHHHHHHHHHHHHHHHGHHHHHHHHHH
@M01332:4:000000000-A76EU:1:1101:11665:2587 1:N:0:9
TATCTGTGATGATCTTATCCCGAACCTGAACTTCTGTTGAAAAAAAAAAC
+
>AABBFFFFFFFGGGGGGGGGGGGGGHHHHHHHHHHHHHGHGHAFGGGGGG
```

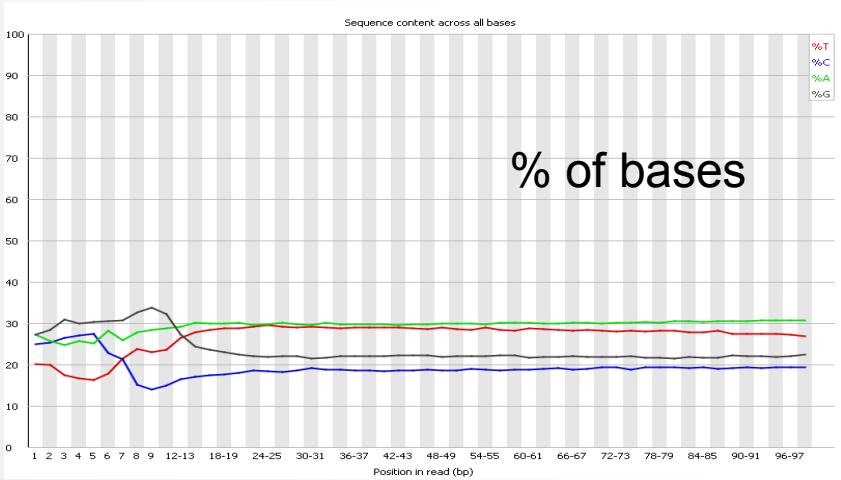
Significance of Q-score line

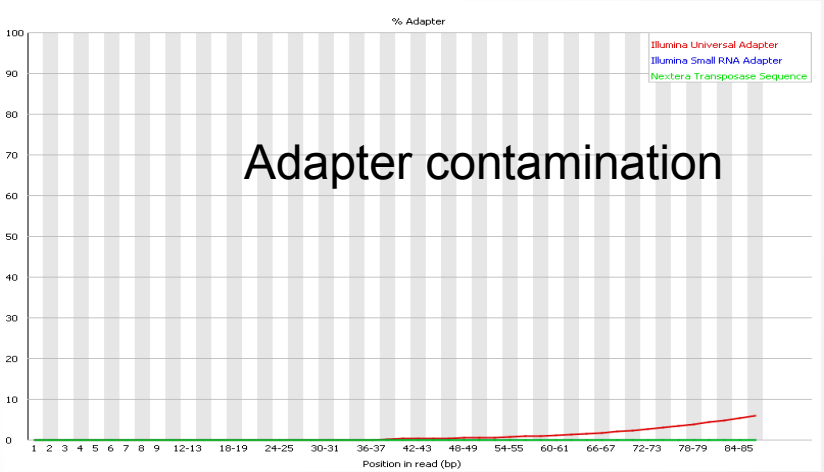
1 error/10 bases = Q-score 10
1 error/100 bases = Q-score 20
1 error/1000 bases = Q-score 30
1 error/10000 bases= Q-score 40

Utility FastQC to check quality and other parameters

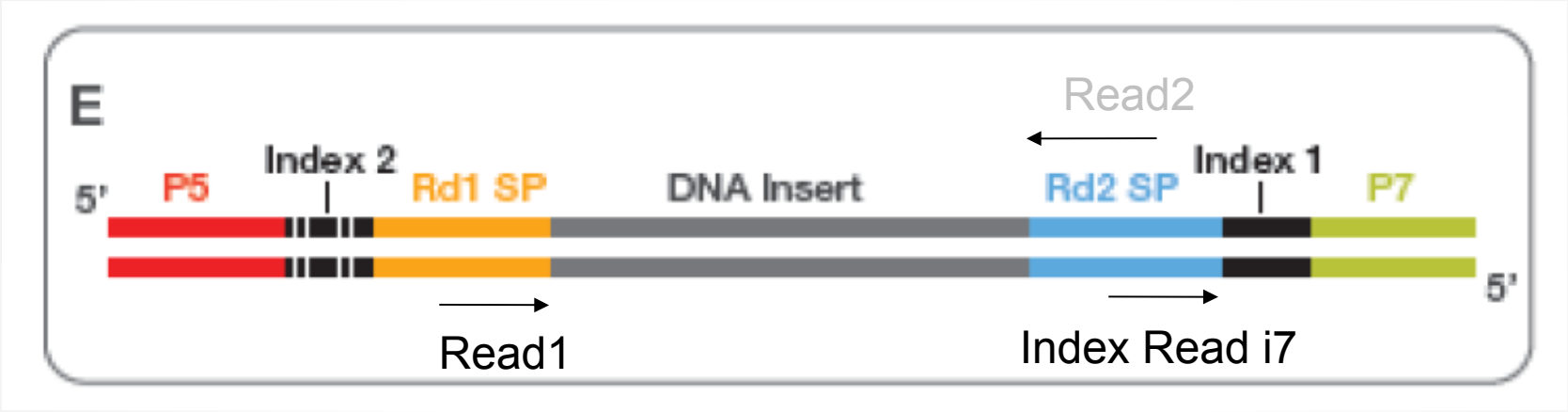


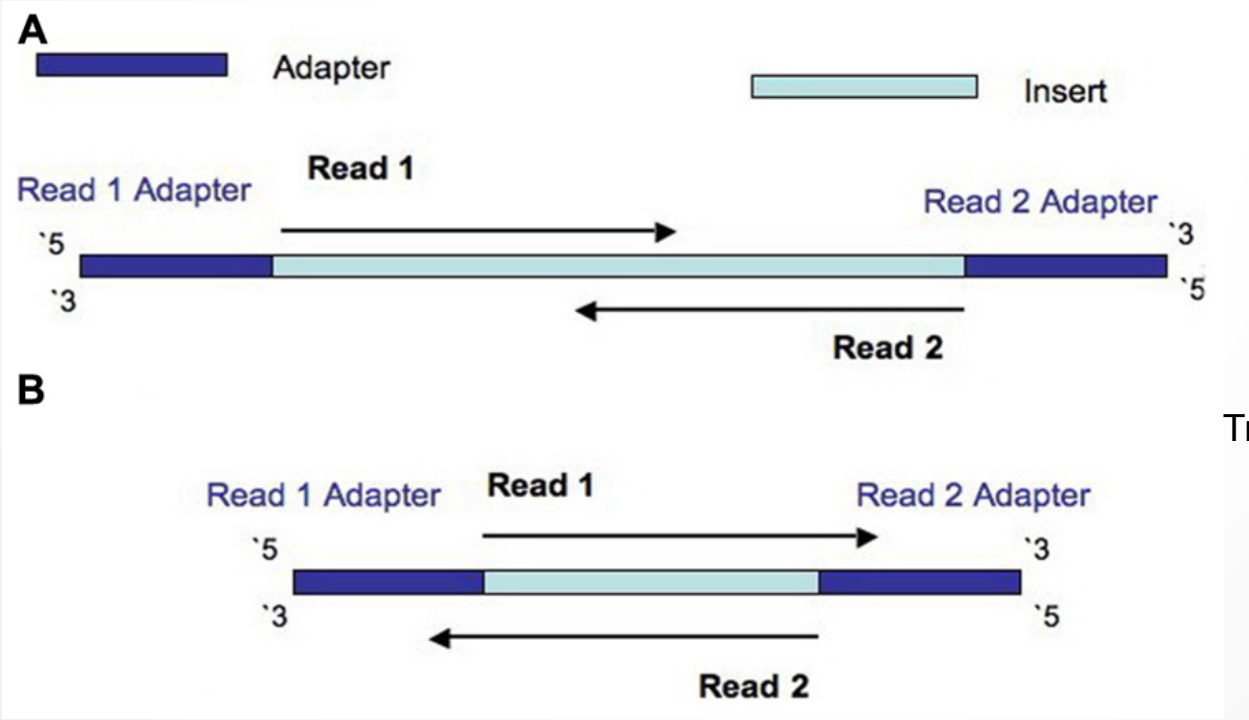






Trimming (adapter removal)

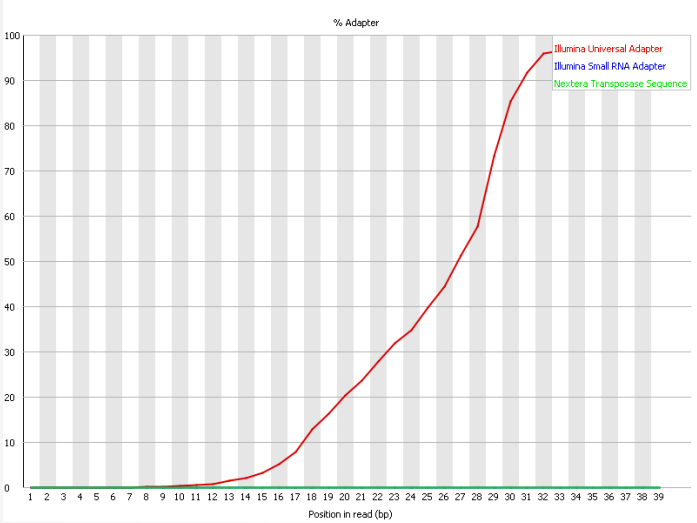




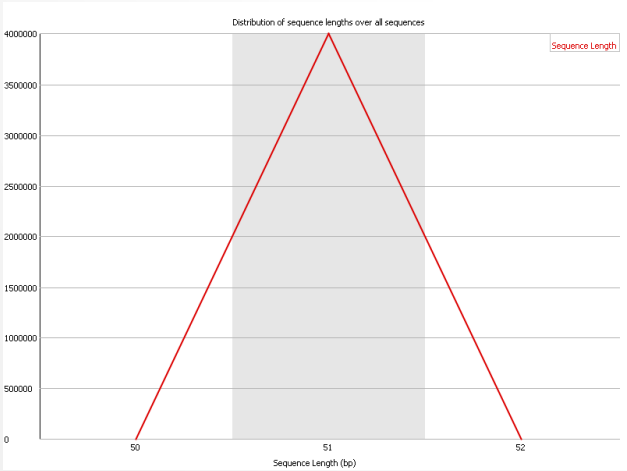
Trimming is essential for short Sequencing, like miRNA

Small RNA sequencing : trimming results

Adapter Content

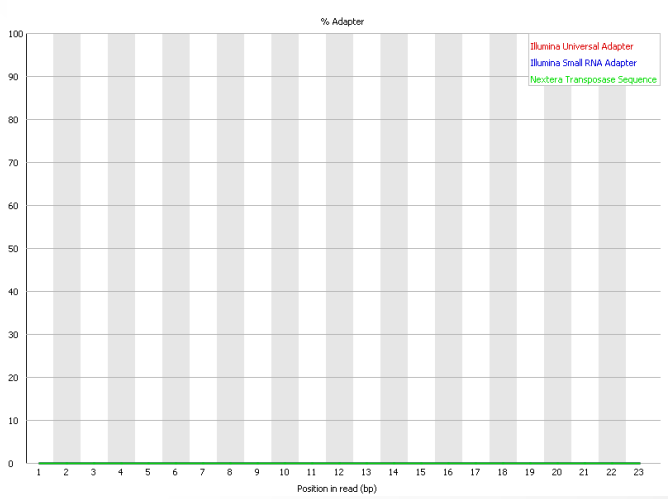


Sequence Length Distribution

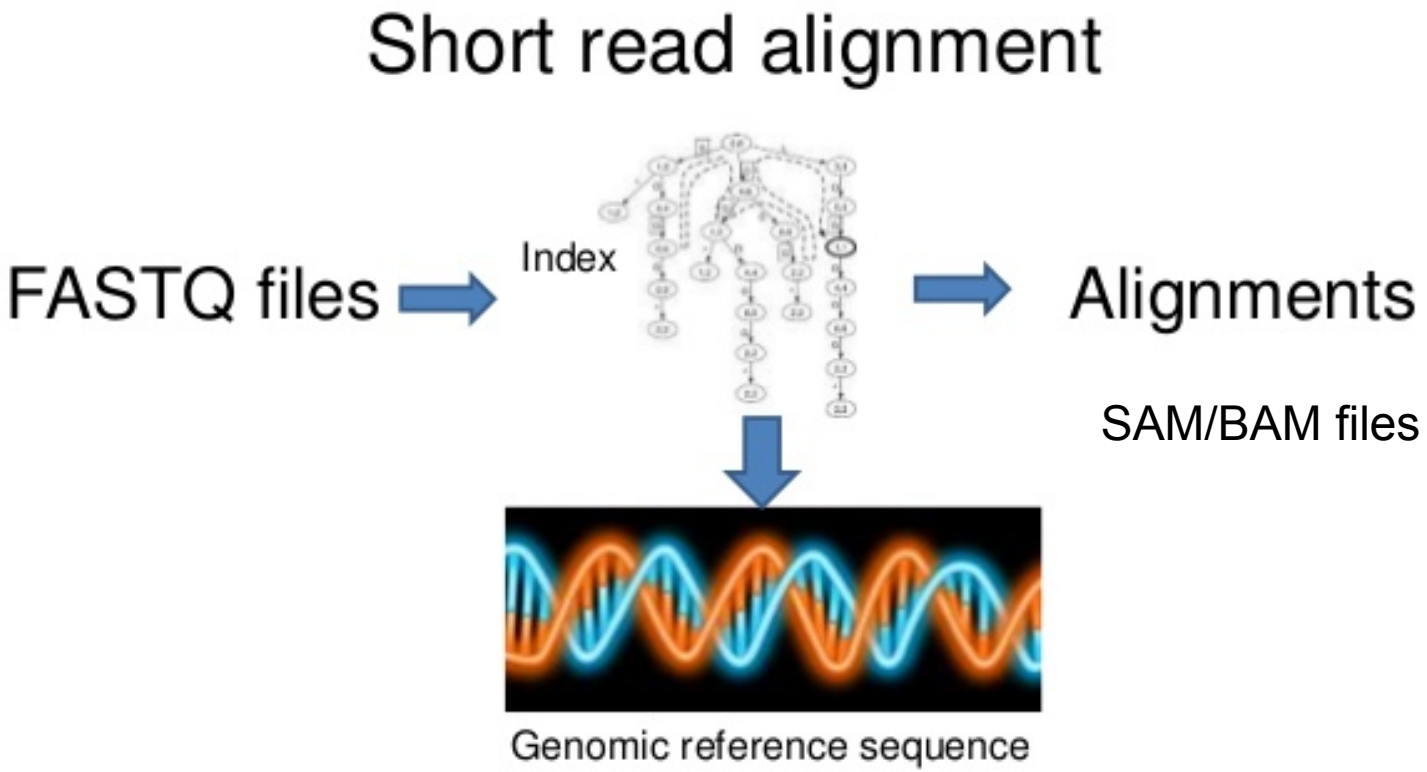


Trimming

Adapter Content

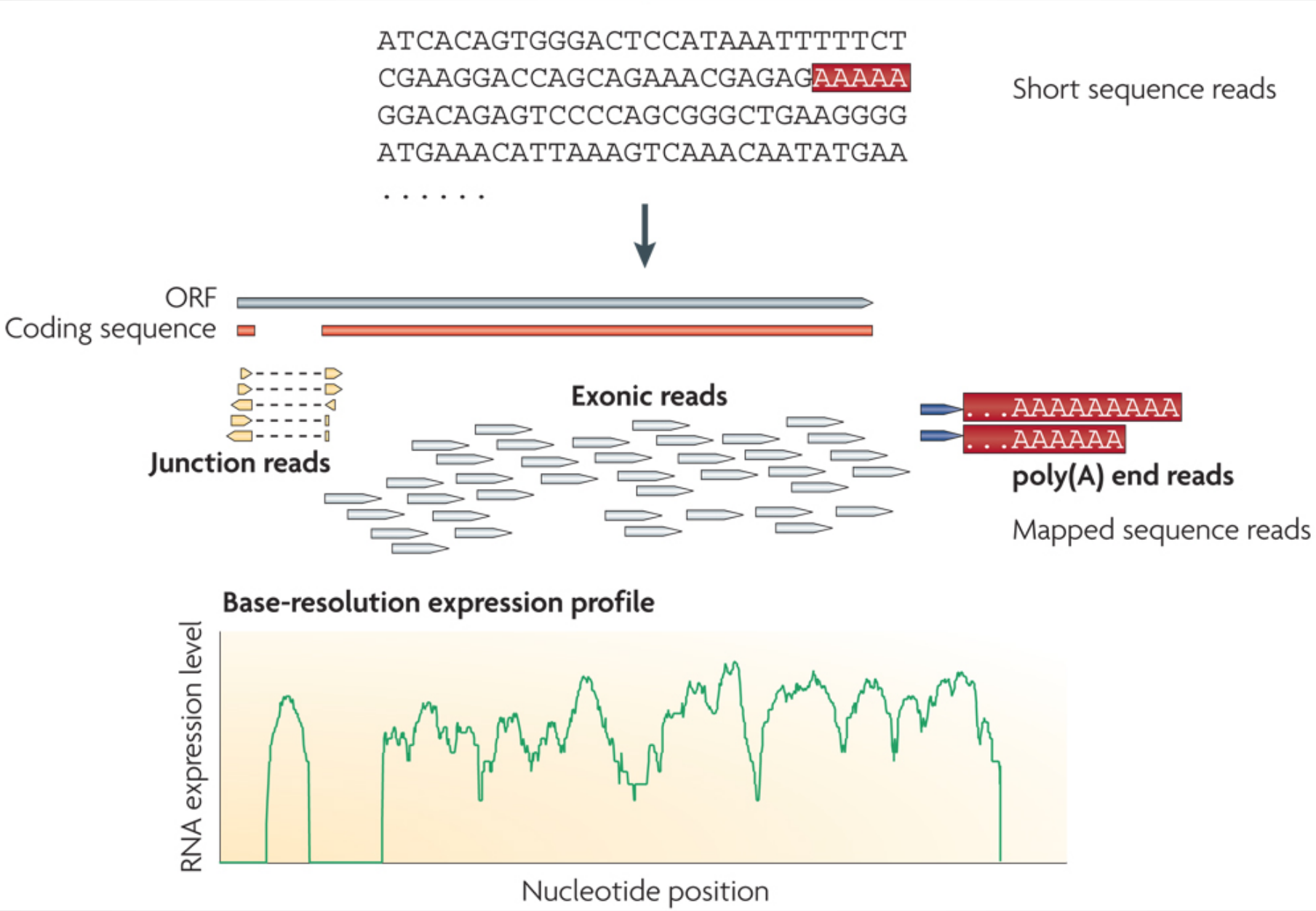


Alignment step



- Many choices: BWA, Bowtie, Maq, Soap, Star, Tophat, etc.

Alignment to the reference sequence... easy?



End-to-End vs Local alignment

Bowtie2 end-to-end Alignment

Global/end-to-end alignment (default)
- Align the whole read

Option: --end-to-end

Can control how many gaps, how long, etc.

```
Read:   ATGCAGCTAGCTAGCTAGCTAGCT
        |||||||||||||||||||||||
Genome: ATGCAGCTAGCTAGC---CTAGCT
```
3nt insertion

```
Read:   ATGCA--TAGCTAGCTAGCAAGCT
        |||||||||||||||||||||||
Genome: ATGCAGCTAGCTAGCTAGCTAGCT
```
2nt deletion
1 mismatch

Appropriate for exact mapping of Extremities
(small RNA sequencing, others

Bowtie2 Local Alignment

Local alignment
- Soft trimming/clipping
- Can include gaps

Option: --local
- Also has -local versions of preset options

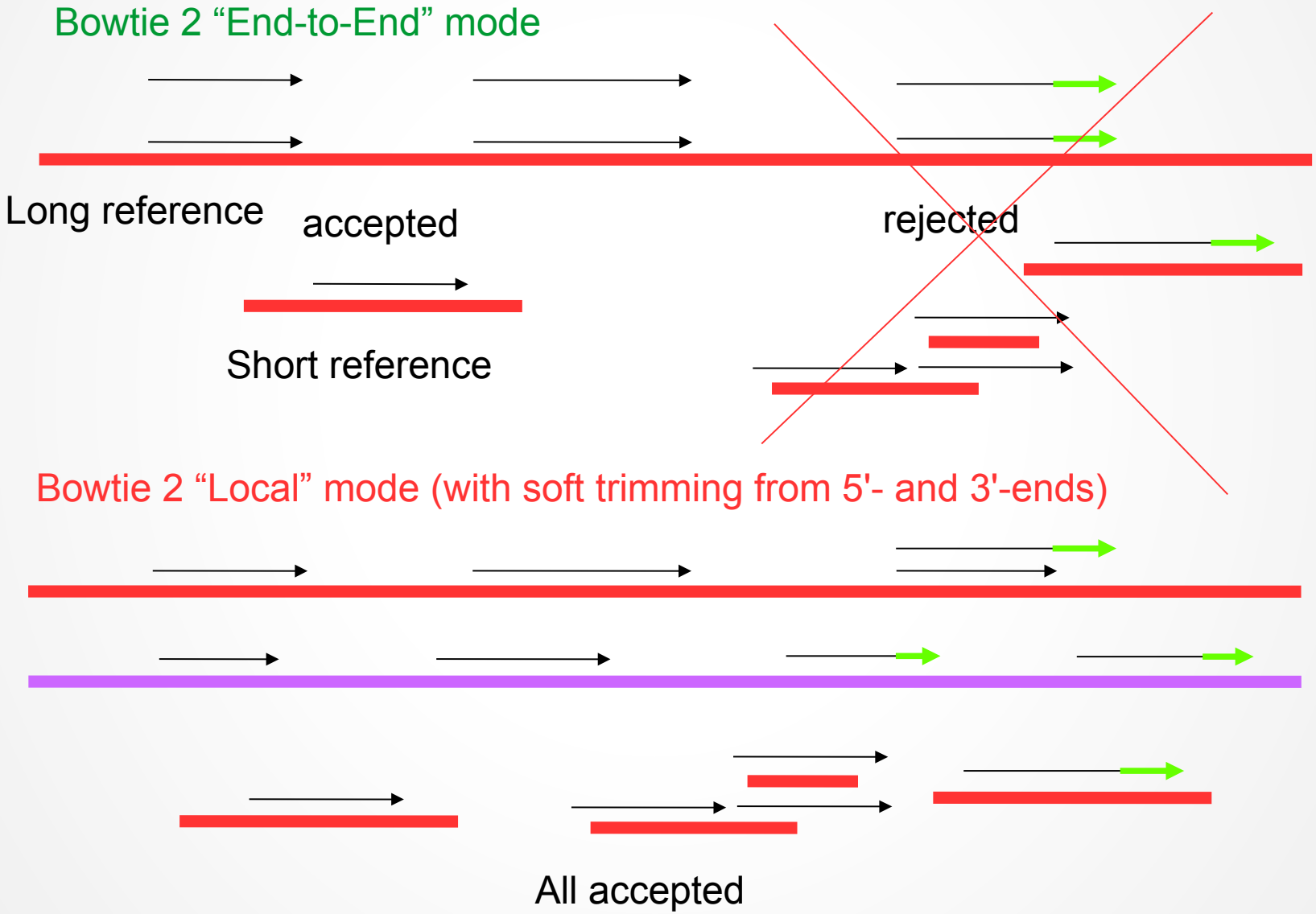
--sensitive-local (default) = -D 15 -R 2 -N 0 -L 20 -i S,1,0.75

```
Read:   ATGCAGCTAGCTAGCTAGCTAGCT
          ||||||||||||||||||
Genome: GCACAGCTAGCTAGCTAGCTAGAC
```
3' and 5' soft trimming

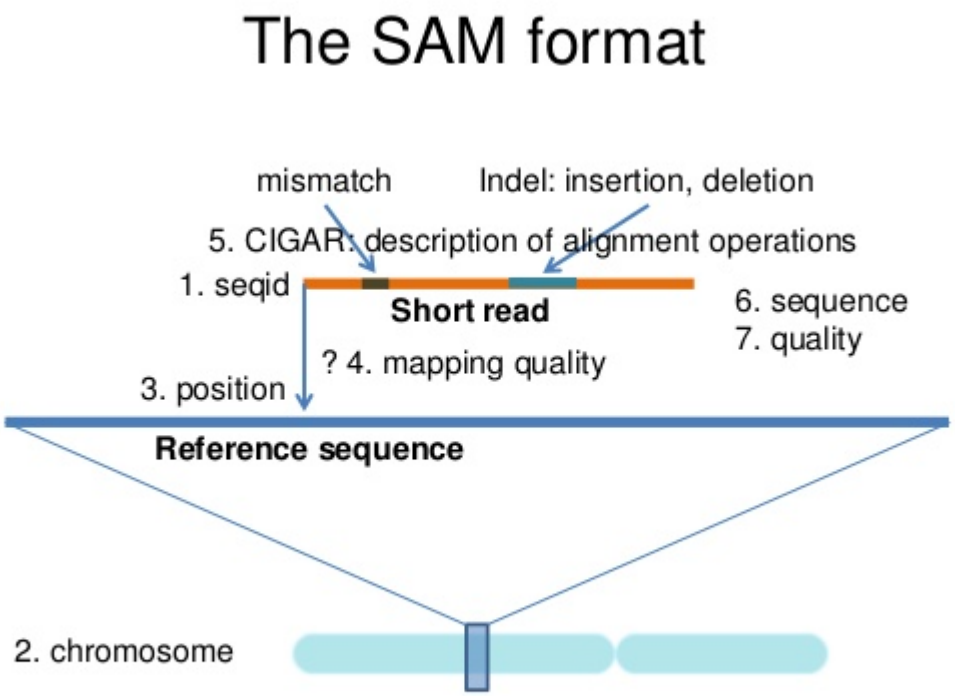
-N dictates mismatches allowed in seed region

Appropriate for counting of gene expression
(transcriptome analysis)

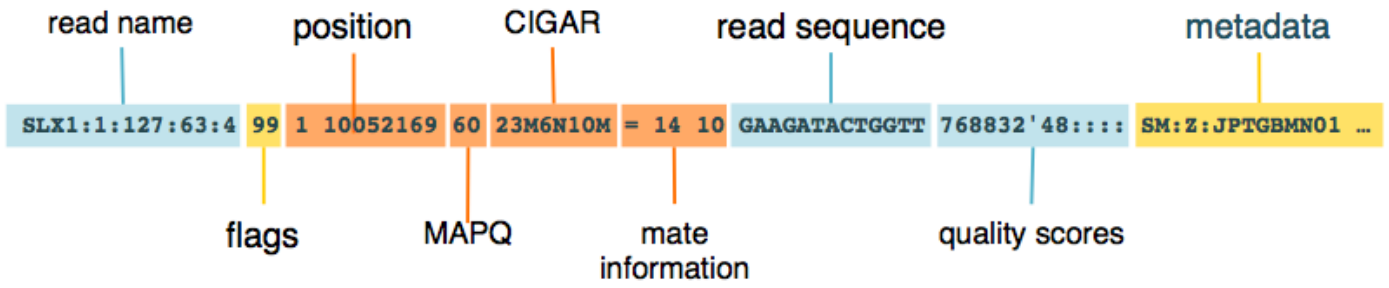
End-to-End vs Local alignment



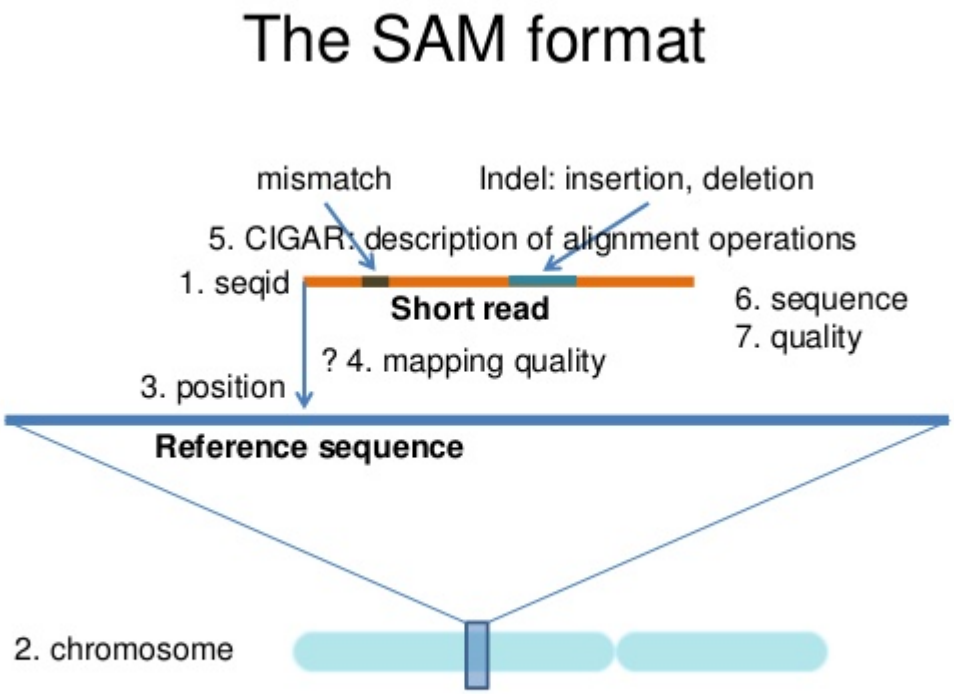
SAM file structure



HEADER containing metadata (sequence dictionary, read group definitions etc)
RECORDS containing structured read information (1 line per read record)



SAM file structure



Requires conversion to

- BAM format for most viewers

Conversion SAM/BAM in Galaxy (slow)
Or by SAMTools (to be installed on server)

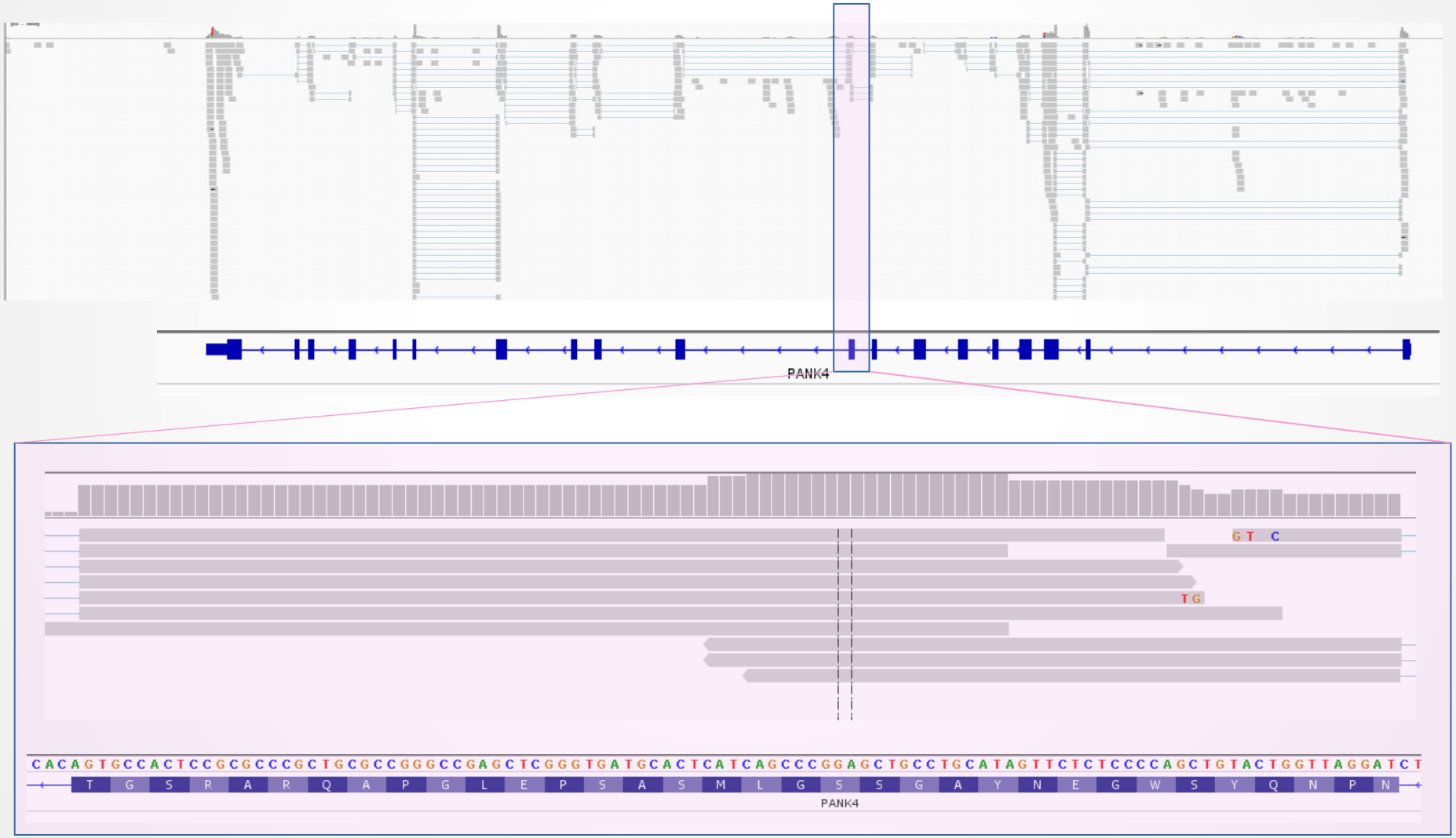
Read ID RefSeq ID pos CIGAR

```
HWI-ST1001:137:C12FPACXX:7:1115:14131:66670   0   chr1   12805   1   42M4I5M   *   0   0
TTGGATGCCCCTCCACACCCTCTTGATCTTCCCTGTGATGTCACCAATATG   CCCFFFFFHHGHHJJJJJHJJJJJJJJJJJJJJJIJJJJJJJJJJJIJJ
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
```

editing distance

```
HWI-ST1001:137:C12FPACXX:7:2313:17391:30032   272   chr1   13494   1   51M   *   0   0
ACTGCCTGGCGCTGTGCCCTTCCTTTGCTCTGCCCGCTGGAGACAGTGTTT   CFFFFHHJJJJIJJJJIJJJJJJJJJJJJJJJJJJJJJHHHHFA+FFFC@B
AS:i:-3 XN:i:0   XM:i:1   XO:i:0   XG:i:0NM:i:1   MD:Z:44G6   YT:Z:UU XS:A:+   NH:i:3   CC:Z:chr15
CP:i:102517626   HI:i:0
```

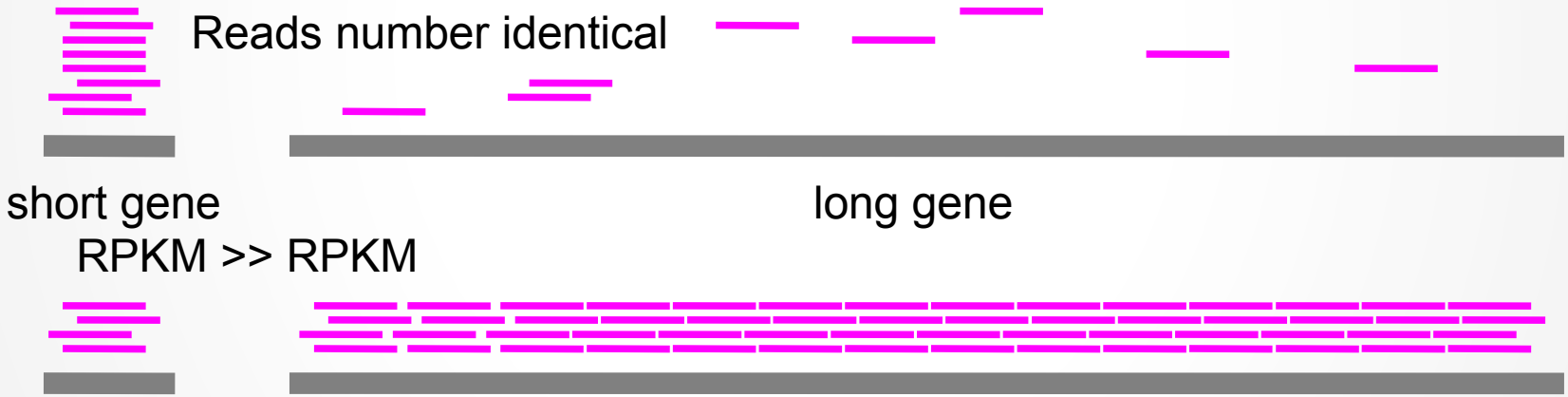
Visualization (IGV)



Counting gene expression

Reads Per Kilobase of exon per Million fragments mapped (RPKM)

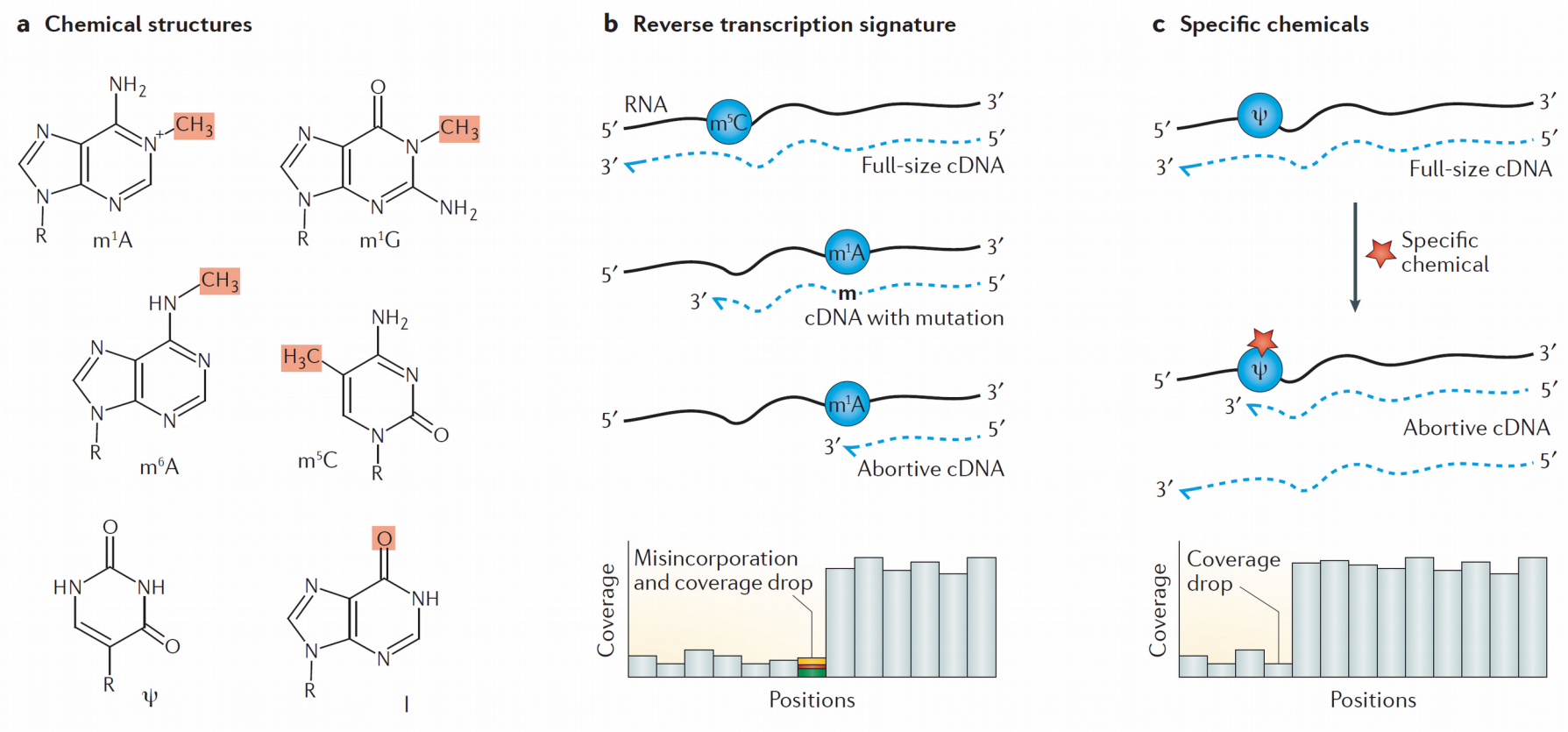
Normalization to the length of gene



$$RPKM(X) = \frac{\text{Reads per transcript}}{\text{million reads} \cdot \text{transcript length(kb)}}$$

Specific approaches for RNA modification studies

Reverse transcription (RT) signatures



Nucleotides with modified Watson-Crick edge (m^1A, m^3C, m^3U, $m^6{}_2A$, etc)

Detecting RNA modifications in the epitranscriptome: predict and validate

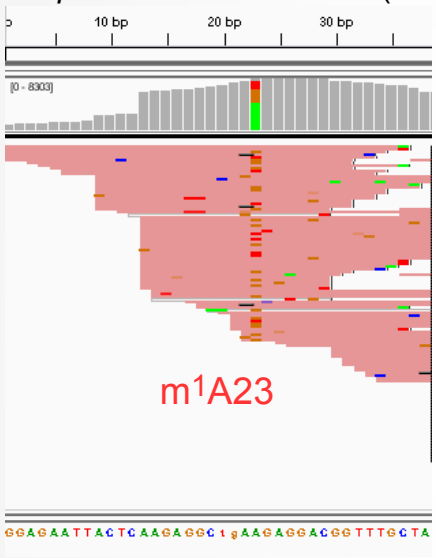
Mark Helm[1] and Yuri Motorin[2]

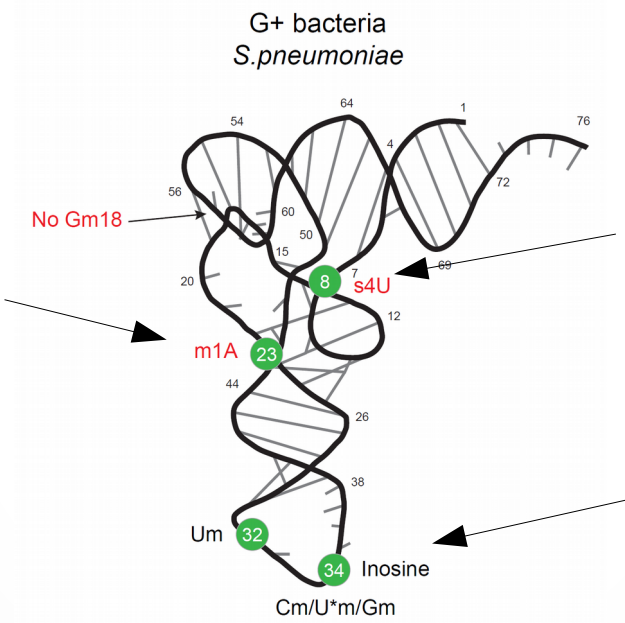
NATURE REVIEWS | GENETICS 2017

Are they visible in IGV viewer ?

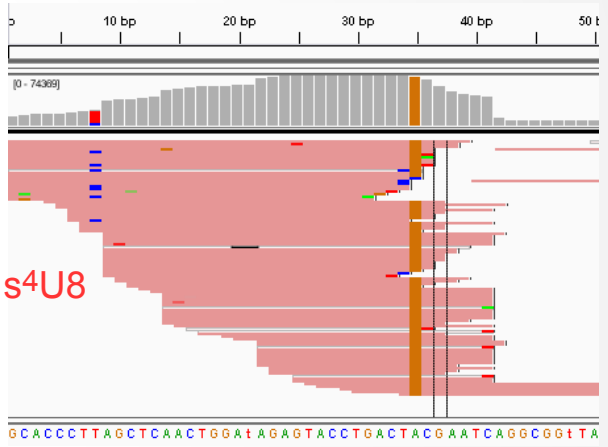
tRNA modifications

S.pneumoniae tRNA-Ser(GCT)





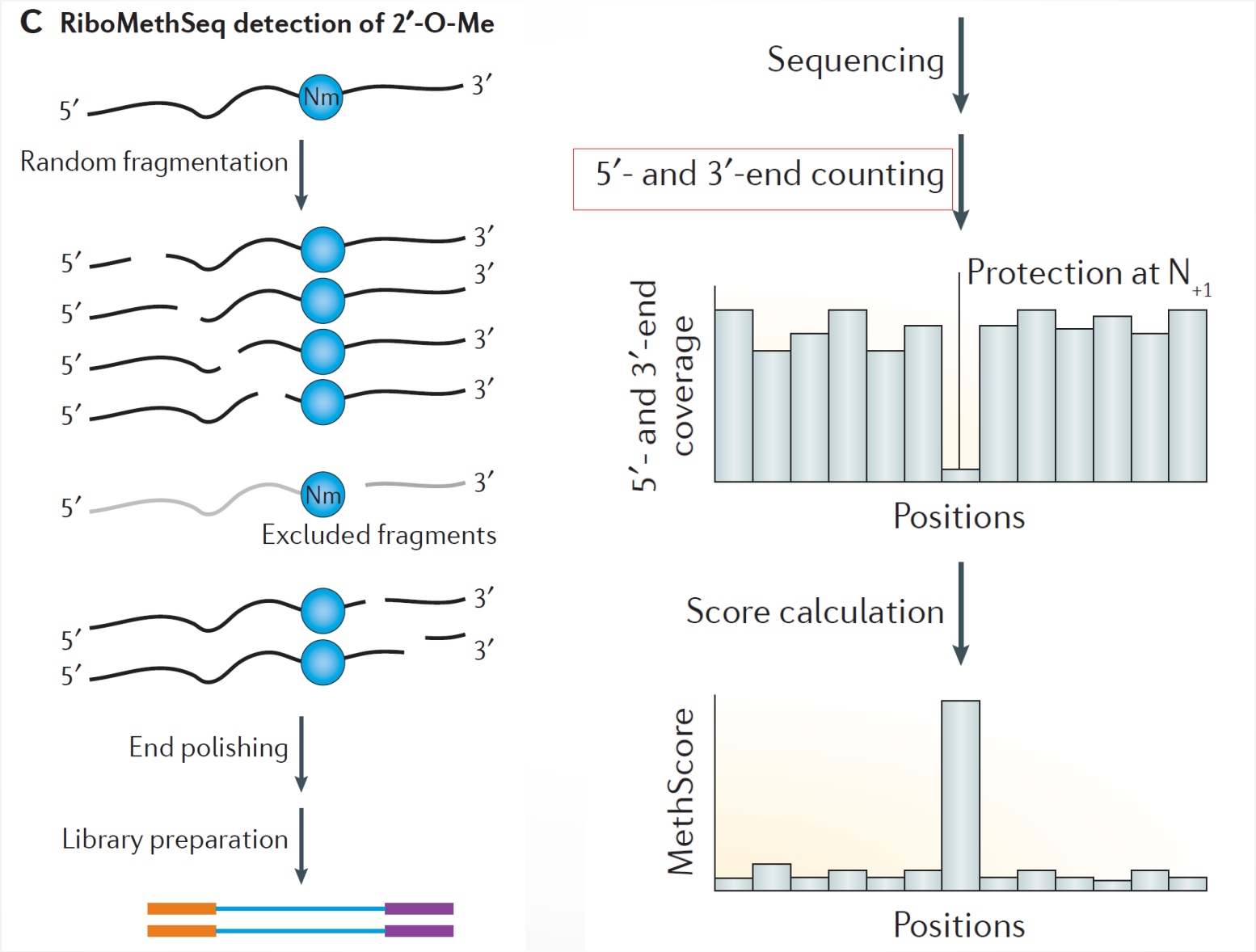
S.pneumoniae tRNA-Arg(ACG)



Sometimes difficult to distinguish from genomic tRNA variants !

If specific chemical product is used, check dependence of signature from treatment

RiboMethSeq and other methods : counting of 5’ and 3’ ends



Where to get information ?

Read ID RefSeq ID pos CIGAR

```
HWI-ST1001:137:C12FPACXX:7:1115:14131:66670   0   chr1   12805   1   42M4I5M   *   0   0
TTGGATGCCCCTCCACACCCTCTTGATCTTCCCTGTGATGTCACCAATATG   CCCFFFFFHHGHHJJJJJHJJJJJJJJJJJJJJJIJJJJJJJJJJJJIJJ
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
```

editing distance

```
HWI-ST1001:137:C12FPACXX:7:2313:17391:30032   272   chr1   13494   1   51M   *   0   0
ACTGCCTGGCGCTGTGCCCTTCCTTTGCTCTGCCCGCTGGAGACAGTGTTT   CFFFFHHJJJJIJJJJIJJJJJJJJJJJJJJJJJJJJJJHHHHFA+FFFC@B
AS:i:-3 XN:i:0   XM:i:1   XO:i:0   XG:i:0NM:i:1   MD:Z:44G6   YT:Z:UU XS:A:+   NH:i:3   CC:Z:chr15
CP:i:102517626   HI:i:0
```

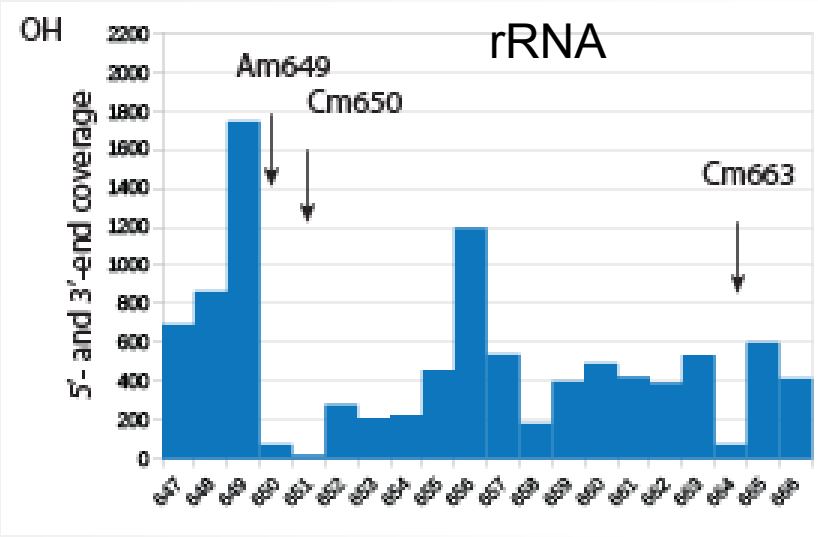
- Counting and quantification of mispaired nucleotides (RT-signatures)

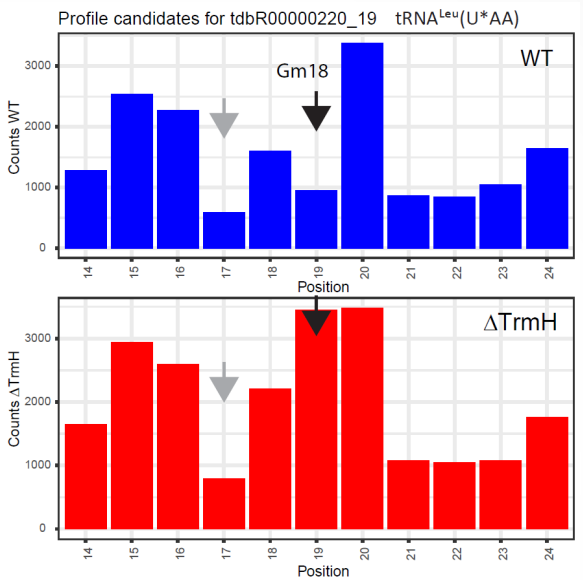
  Difficult to extract directly from SAM → conversion to mpileup file

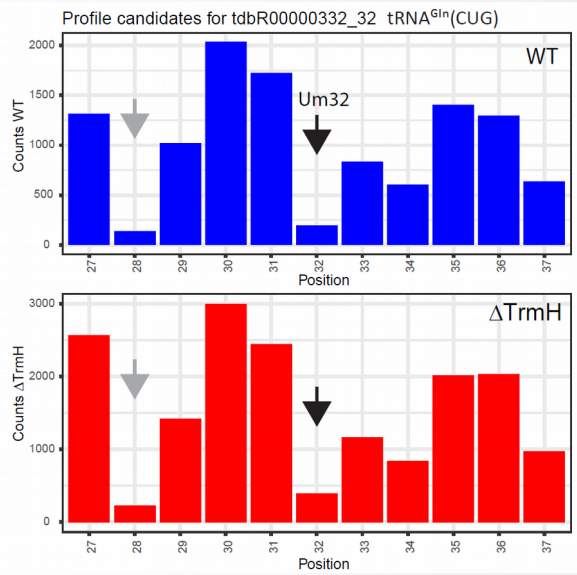
- Precise mapping of 5'- and 3'-ends and their counting

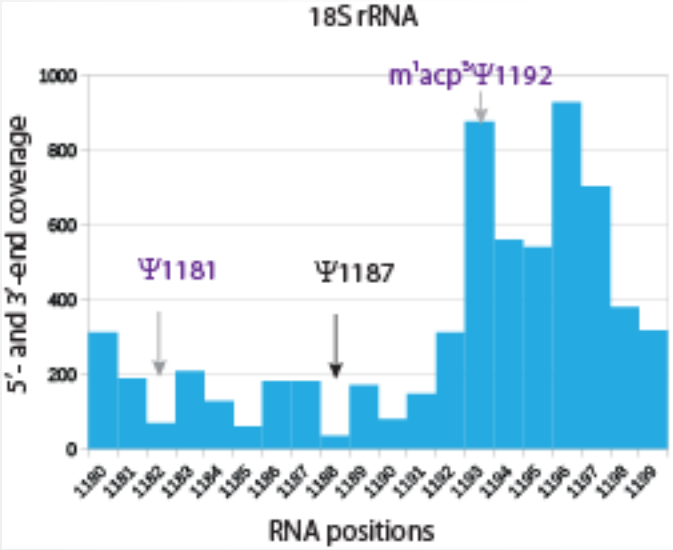
  Either directly from SAM for 5'-ends
  Or → conversion to *. bed format for both 5'- and 3'- counting

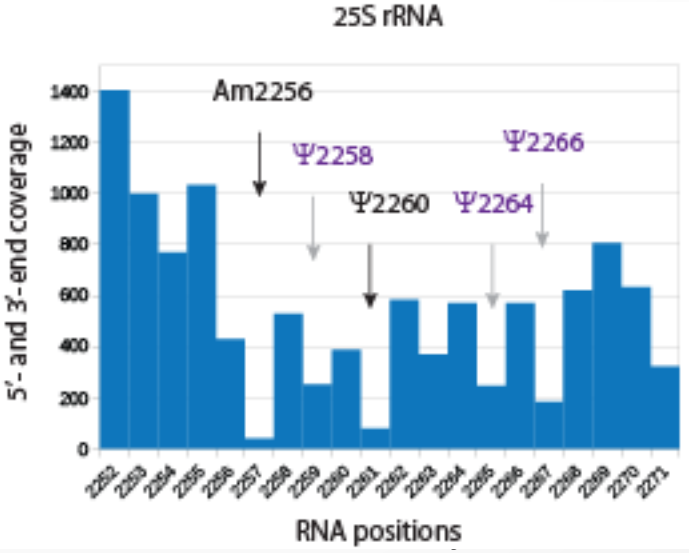
Interpretation RiboMethSeq profile











Bioinfo is FUN!