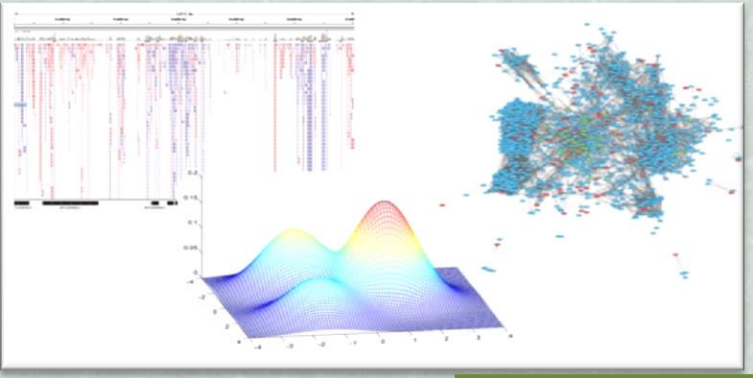


SPS
SCIENCES des PLANTES de SACLAY

Le 29 juin 2017, Module 2

- **Main approaches in Bioinformatic to explore RNA-Seq data**
- **From raw data to gene expression**

Claire Toffano-Nioche (I2BC)
Véronique Brunaud (IPS2)

High-Throughput Sequencing changes Bioinformatic approaches

Impacts in Computer sciences

- Run huge data : from 50 Gigas to Tera
- High-performing of computer and network
- Disk usage and backup

Impacts in Bioinformatics Tools

- Create new algorithms (more performing, sensibility/specificity)
- Use and evaluate many tools (known parameters, set of reference)

Impacts in Statistical methods

-Impact of technical methods (library preparation, sequencing)
-Change of data : type, quantity

RNA-Seq Applications

- **Applications on mRNA or non coding RNA**
- **Measure gene expression of annotated or *de novo* genome**
- **Differential expression (conditions, organs, genotypes…)**
- **Detect variants : allele specific expression, SNPs in genes**

→ **Goal**

1- assigned each read to a gene

2- obtain counts by gene

SPS Module2
Main approaches in Bioinformatic of RNA-Seq

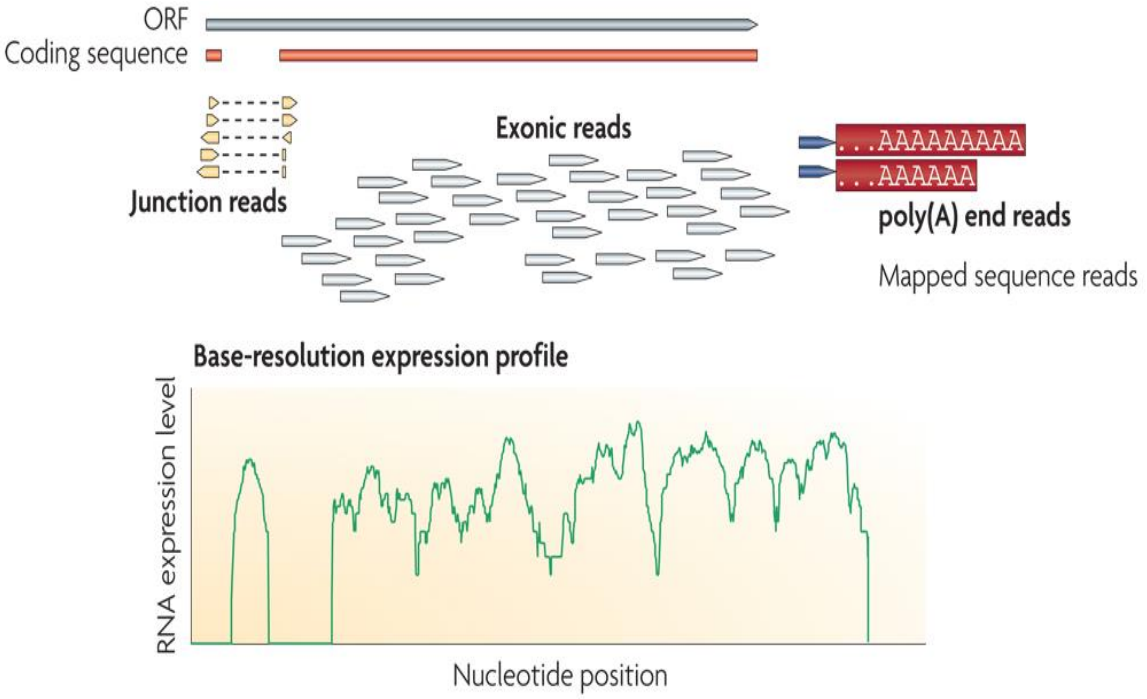
1/ Classical analyses of RNA-Seq (V. Brunaud)

- **Check quality , Trimming**
- **Mapping / counts**
- **Assembly**

2/ Specific applications of RNA-Seq (C. Toffano-Nioche)

- **Study smallRNA**
- **Gene expression by transcript (isoform)**

Bioinformatics from raw data to counts



RAW DATA

Pre-processing

Check Quality

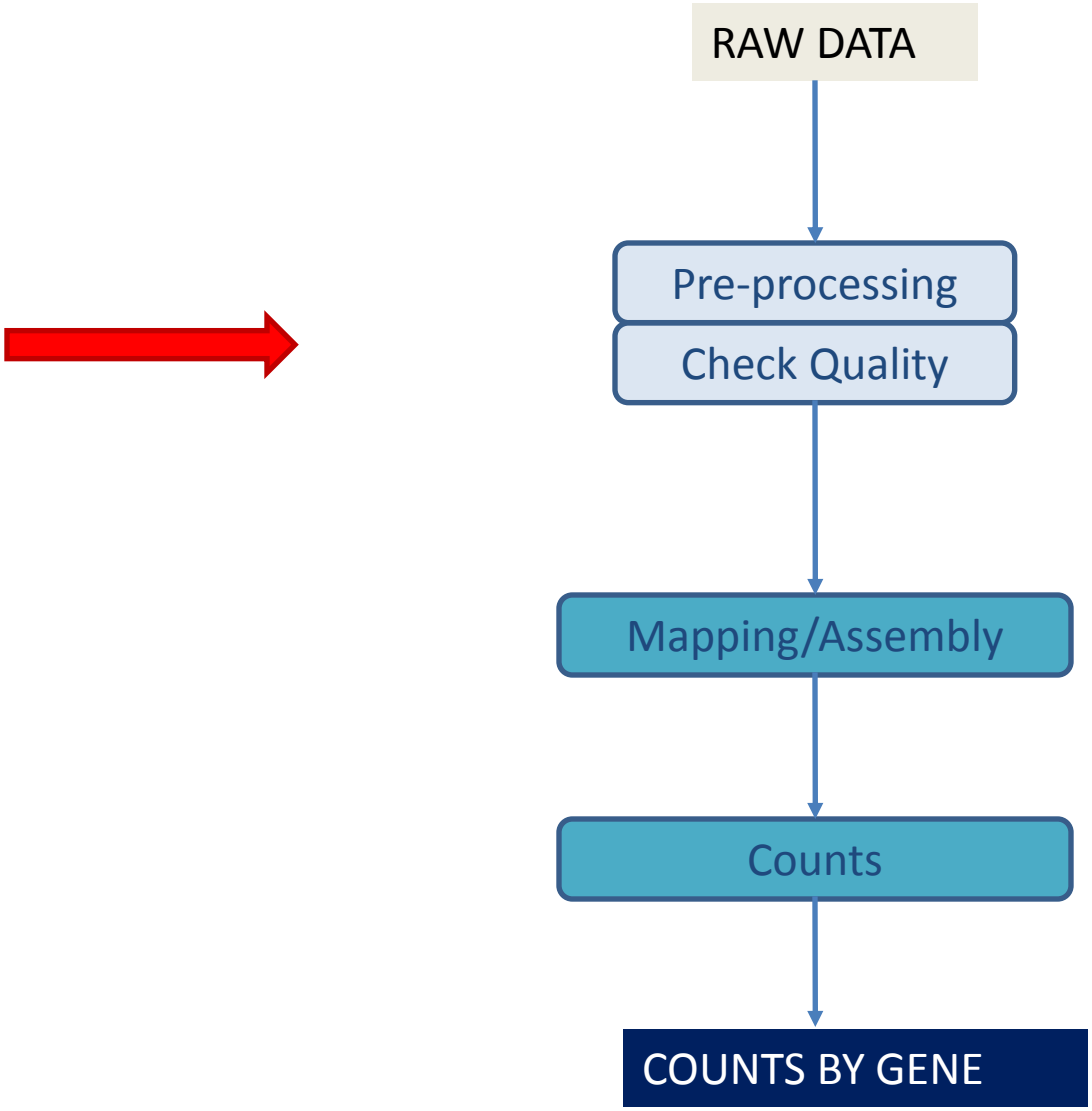
Mapping/Assembly

Counts

COUNTS BY GENE

Gene expression

Bioinformatics from raw data to count



Classic pre-processing

universal adapter Insert size Adapter with index

Classic pre-processing

universal adapter Insert size Adapter with index

BCL files
From sequencer

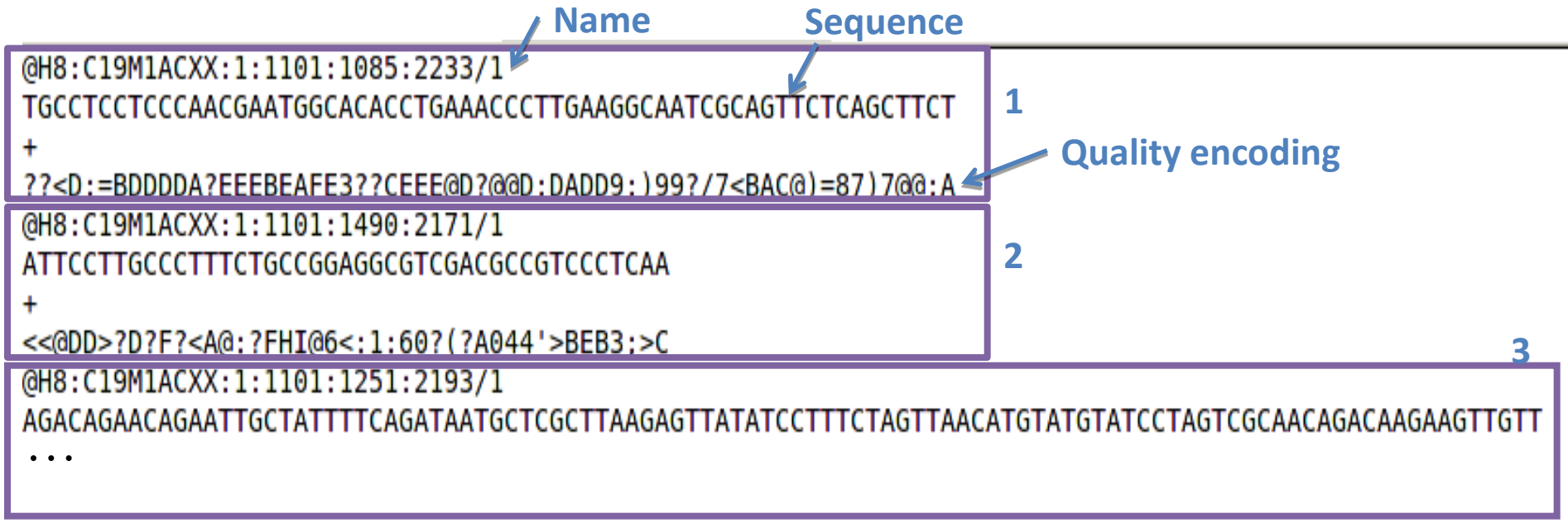
Bcl2fastq (Illumina) : convert base calling in fastq sequences, demultiplexing

Trimmomatic : remove adapters, reads with low quality, length too small (trimming)

SortMeRNA : remove rRNA sequences

Fastq results

Fastq format

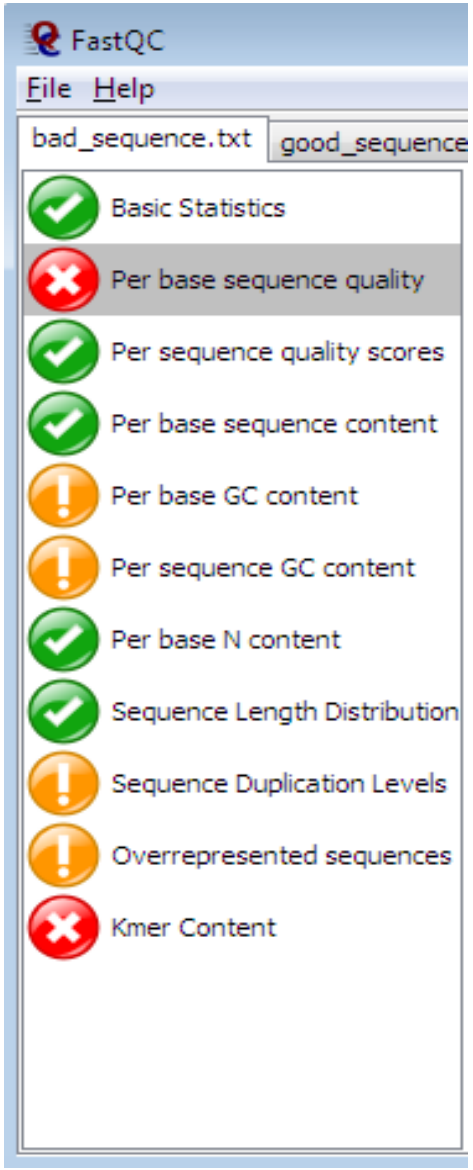


View the quality of reads with fastQC

FastQC → a report of quality on each sample

- Command line or interface viewer
- Generate a html report to check quality

http://www.bioinformatics.babraham.ac.uk/projects/fastqc/



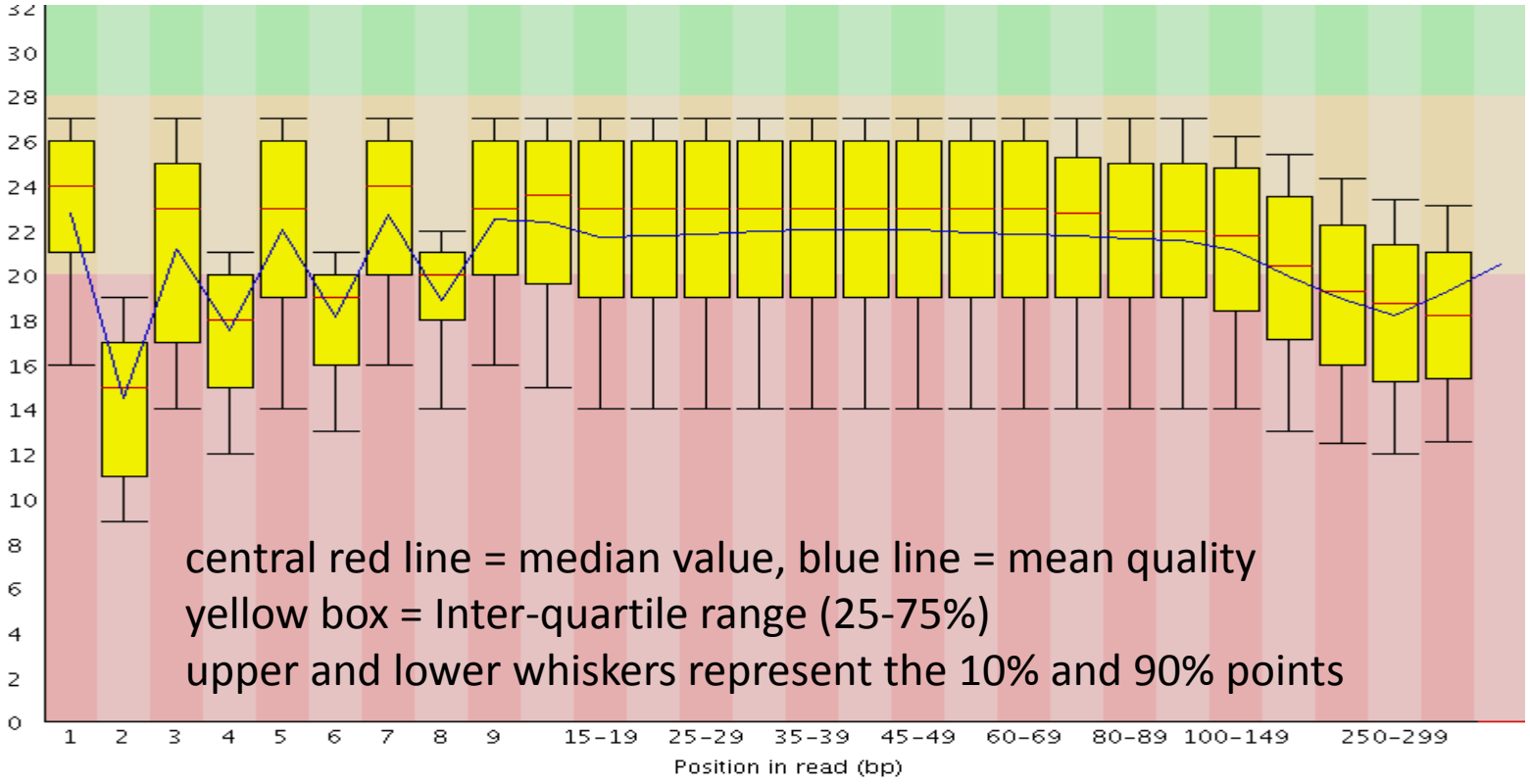
fastQC – Quality score

Quality score (Q-scores):

Q(B) =-10 log10(P(~B))

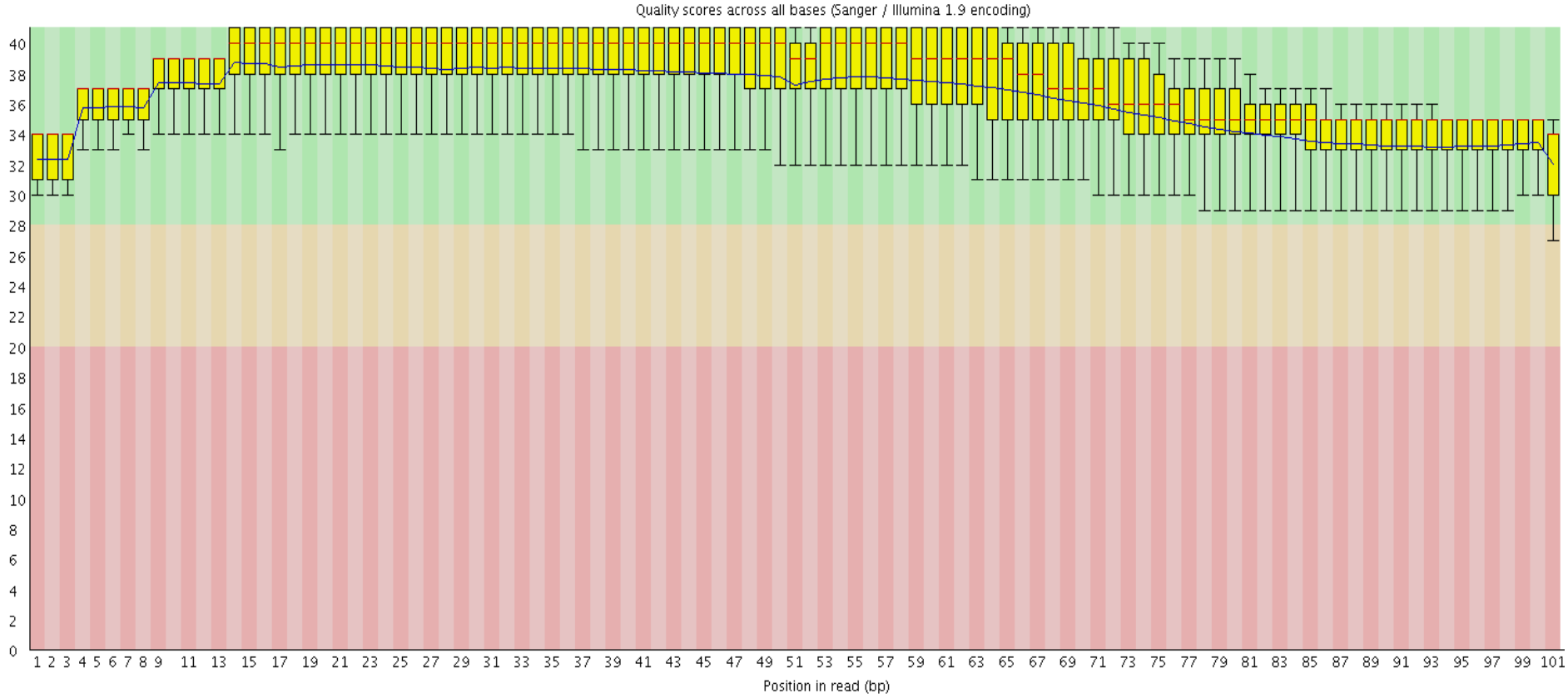
where *P(~B)* is the estimated probability of an assertion of Base being wrong.

Qscore=20 =1% error



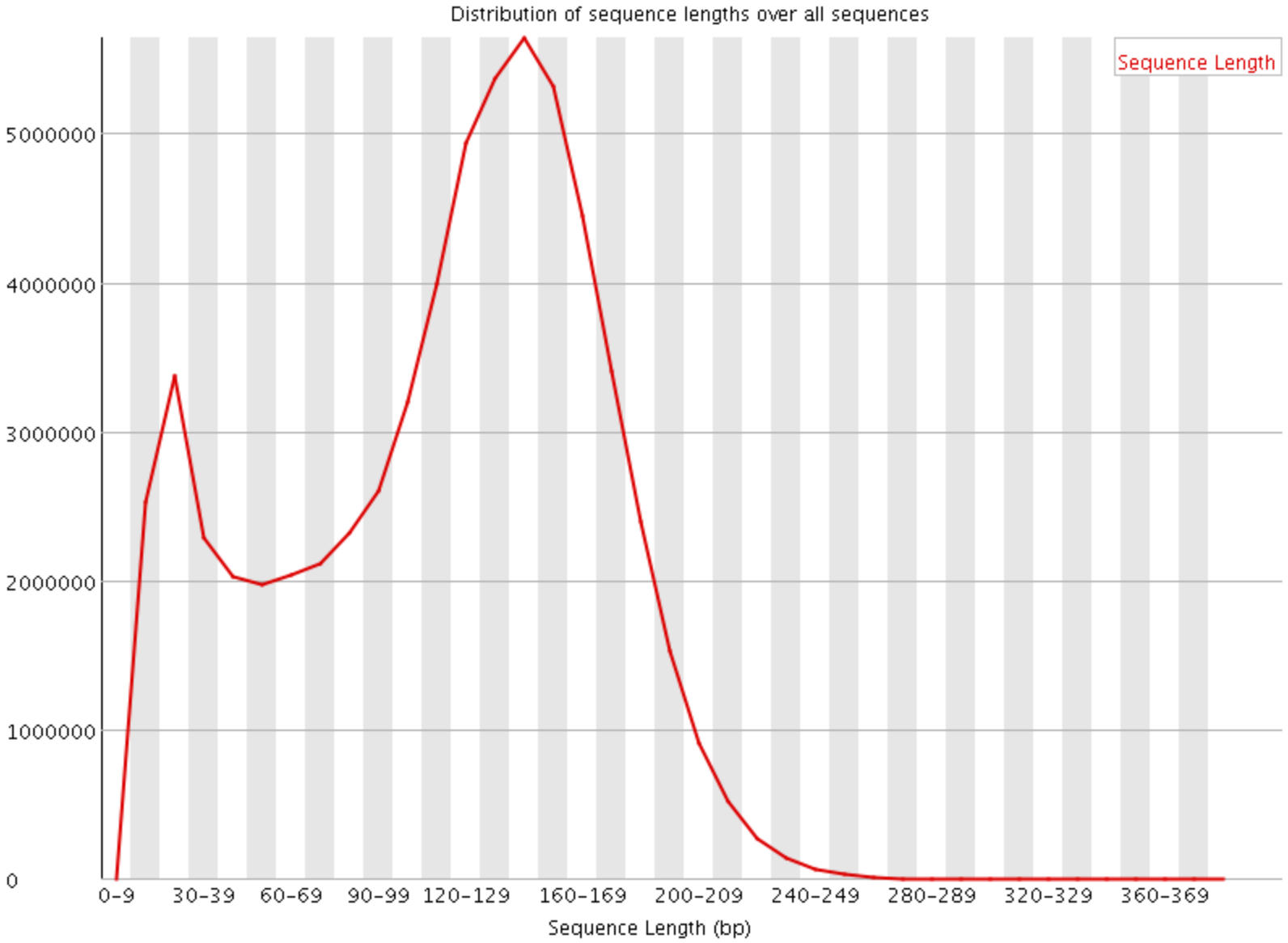
Generally trimming by 3'end as long as Qscore < 20

fastQC – Quality score

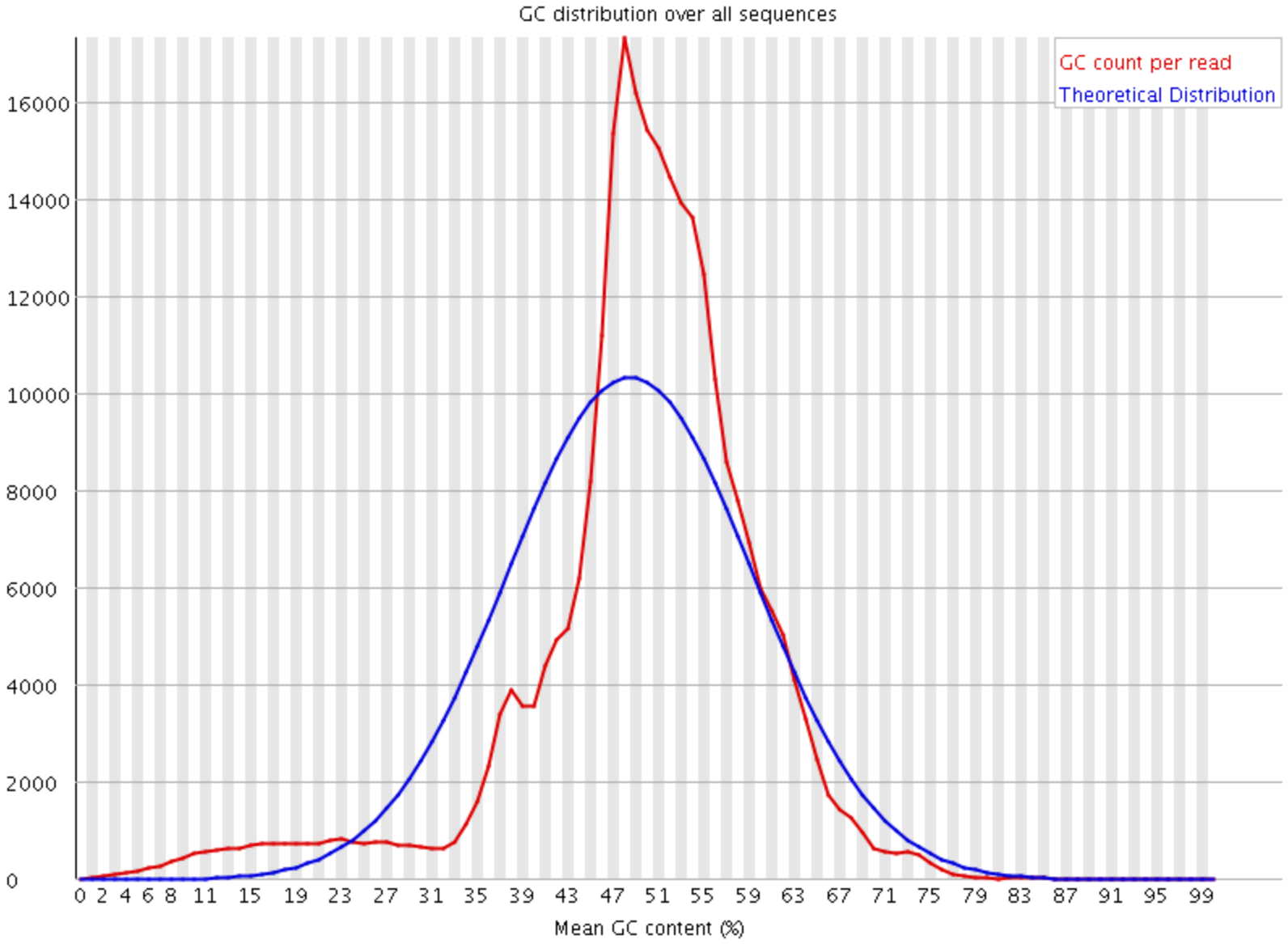


→ **Quality score very good Qscore > 30 (< 1‰)**

fastQC – length of reads



fastQC – GC %


GC distribution over all sequences
GC count per read
Theoretical Distribution
16000
14000
12000
10000
8000
6000
4000
2000
0
0 2 4 6 8 11 15 19 23 27 31 35 39 43 47 51 55 59 63 67 71 75 79 83 87 91 95 99
Mean GC content (%)

Duplicate reads = exactly same sequence for 2 reads
Is it a bias of PCR-duplication or a natural duplicate ?

- **Distinguish PCR- from natural duplicates :**

Natural duplicates are read duplicates that originated from different mRNA molecules.

→ Library using UMI=Unique Molecular Identifier method.

- **The impact of amplification on differential expression analyses by RNA-seq.** S. Parekh et al. (2016) in Scientific reports vol6

  - "We find that a large fraction of computationally identified read duplicates are not PCR duplicates and can be explained by sampling and fragmentation bias."

  - "Removal of duplicates does not improve the accuracy of quantification"

  - "Based on simulated differential expression…, we find that computational removal of duplicates has either a negligible or a negative impact on FDR and power"

→ **It's not necessary to remove duplicate reads**

Conclusion of pre-processing, check quality

Current trimming

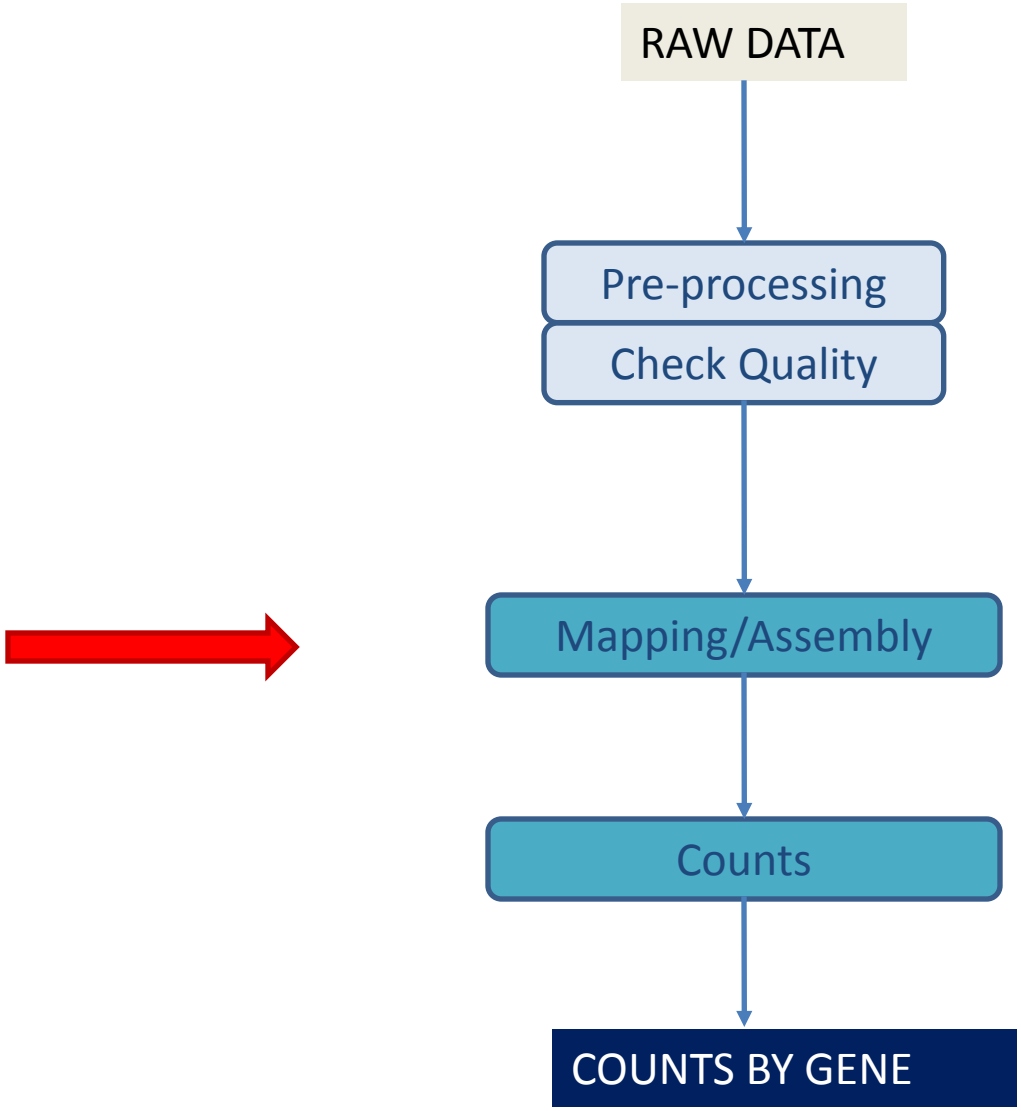
- mRNA : after quality trimming length > 30 bases
- smallRNA : no trimming quality, select by size length (see next talk)
- no undetermined base in read (for assembly)
- remove the both reads of one Paired-ends read (same fragment)
- remove ribosome

Trimmomatic
remove adapter,
Trimming low quality,
length control

SortMeRNA
remove rRNA sequences

Depends on the biological object:
mRNA, lncRNA, smallRNA…

Mapping

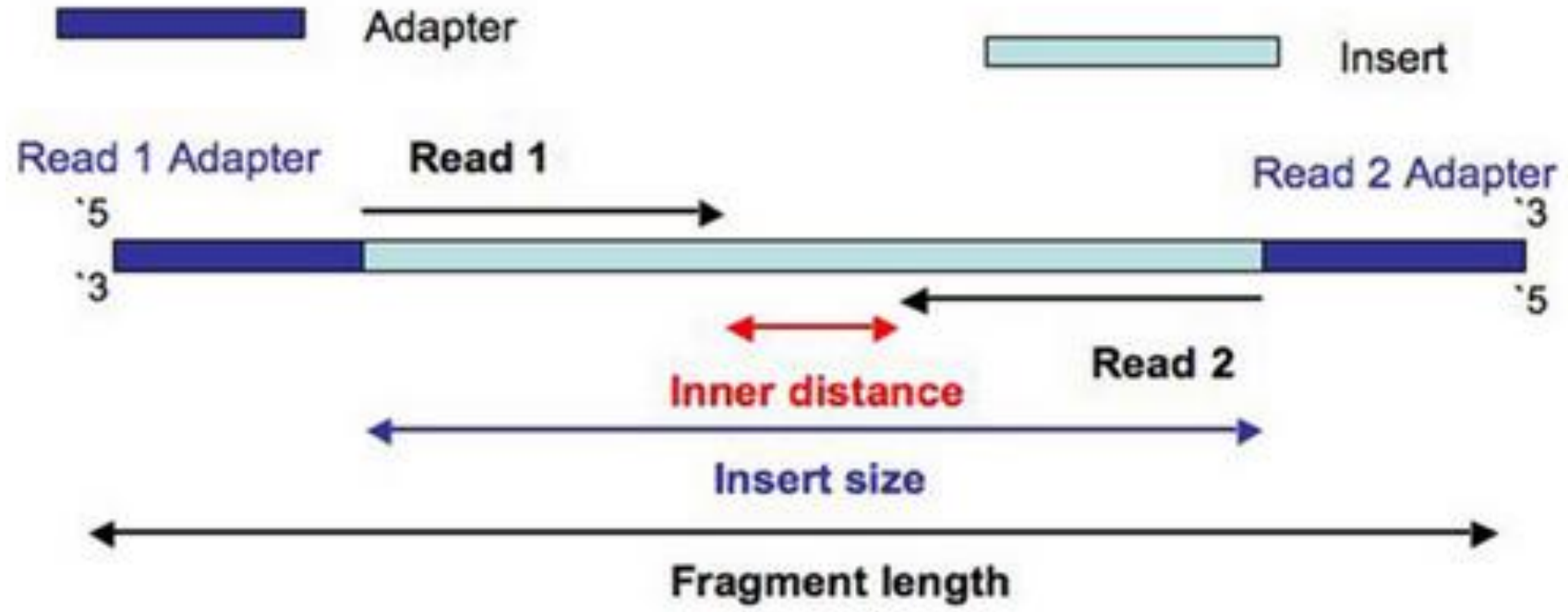

RAW DATA
Pre-processing
Check Quality
Mapping/Assembly
Counts
COUNTS BY GENE

Sample characteristics before mapping what kind of data ?

Depend on type of library

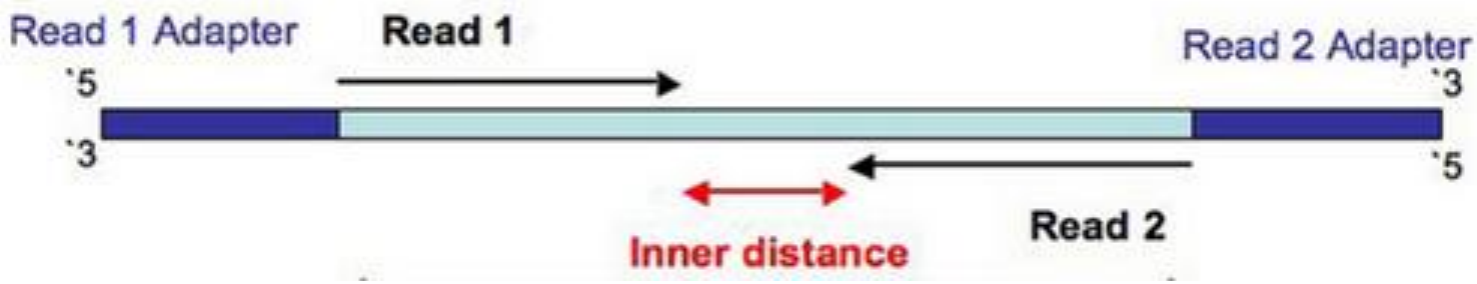
- Paired-end reads or single reads
- Stranded or not
- Sizing, Size of reads 75,100,150…
- Library depth

Definitions of fragment and insert size



See the above figure (from https://www.biostars.org/p/106291/)

Paired-end reads (PE) versus single reads (SR/SE)



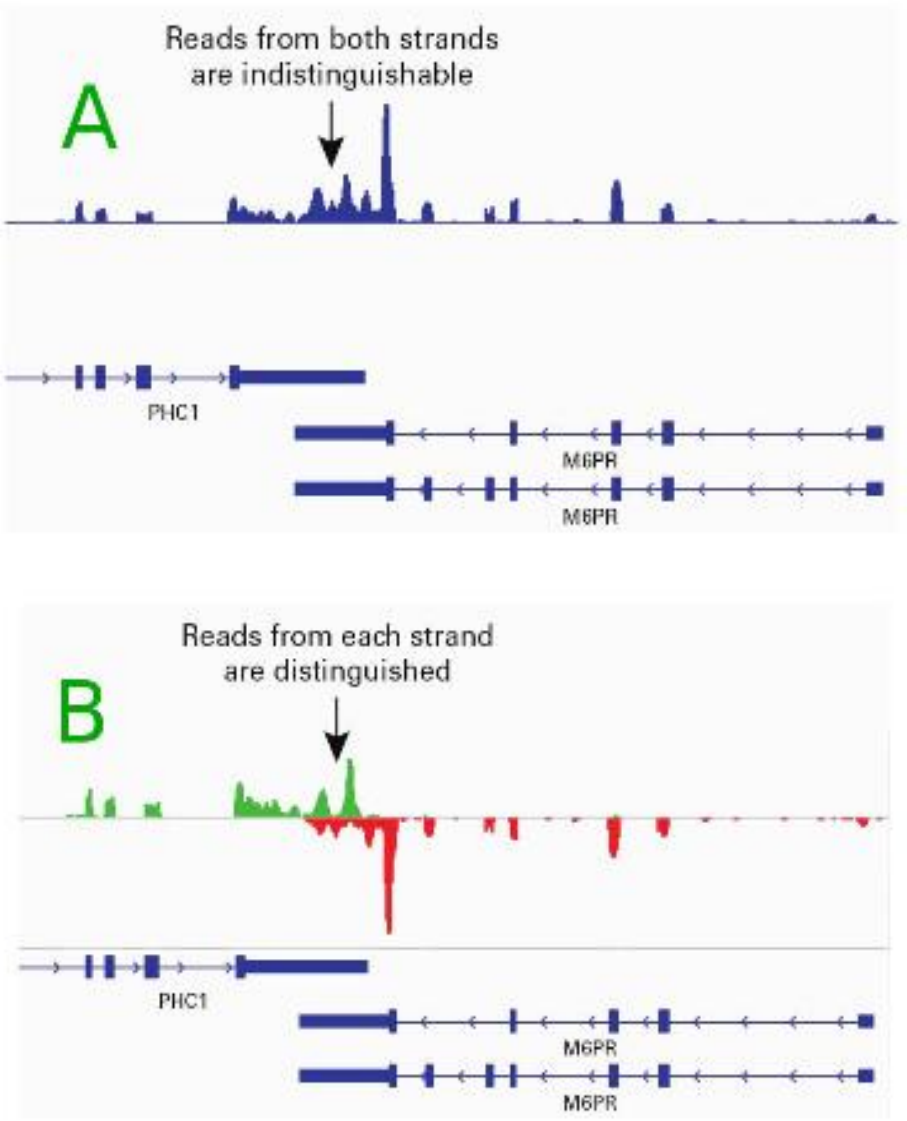
Paired-ends reads
- More strict, accurate on mapping and counts
- Less duplicate reads

But depend on
→ the transcriptome studied
→ the biological question

Chhangawala et al. Genome Biology (2015) 16:131
→ For DE analyses, same list of genes for 50bp in PE and 75bp in SE
→ For detect splicing junction, PE is better

Z. Chang et al. (2014) - PLOS one
→For Assembly (*de novo* genome) read length of 100 or more is better (organism dependent)

Sequencing Stranded or not ?



Sample characteristics before mapping what sort of data ?

Depend on type of library

- Paired-ends reads or Single reads
- Stranded or not
- Size of reads 75, 100,150
- Depth of library

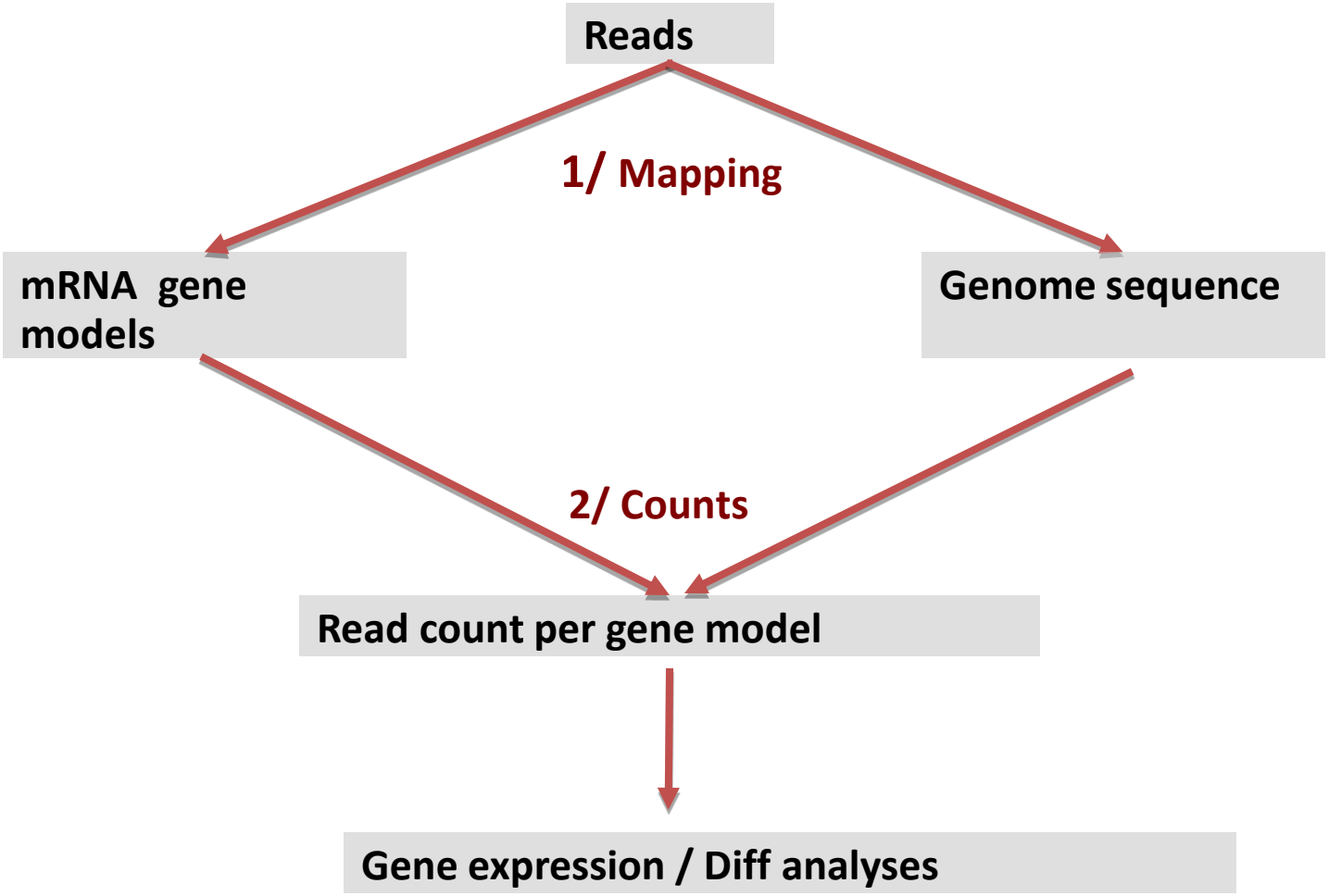
Depend on knowledge about the organism

Is there a genome sequence ?

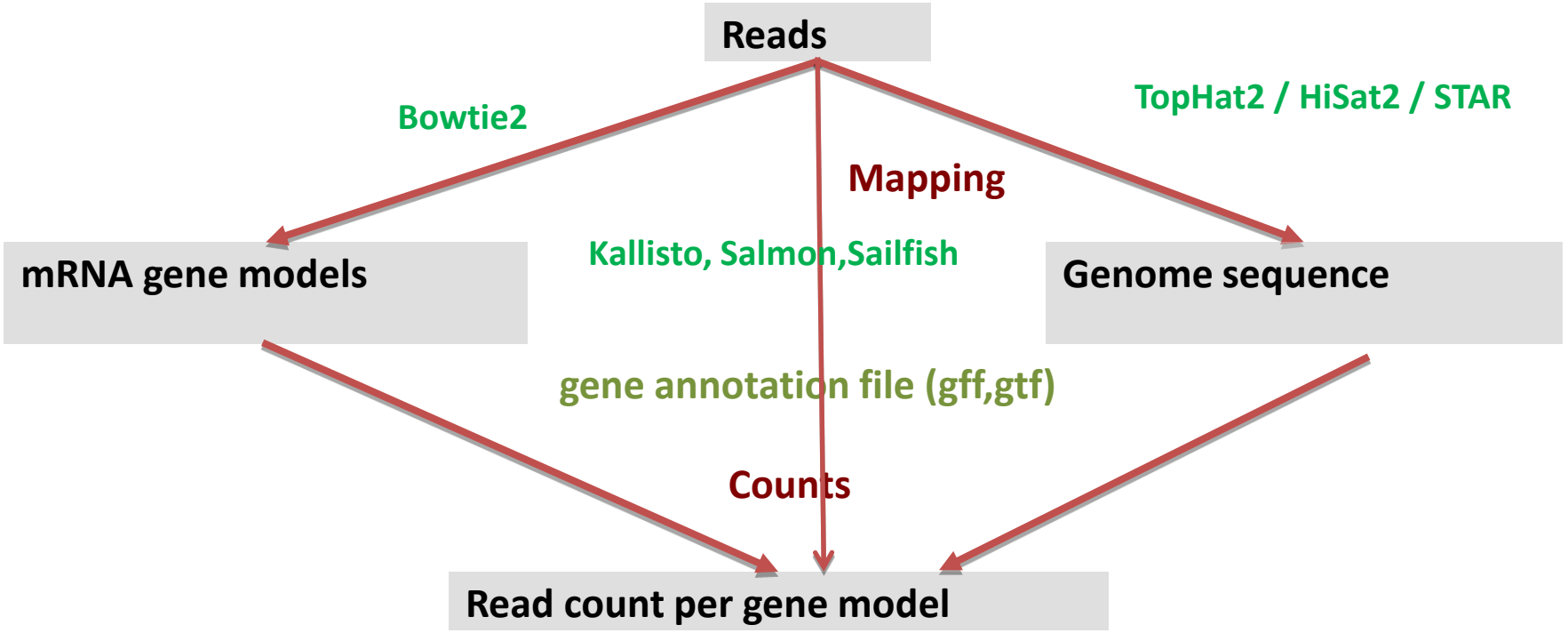
Is there a transcriptome reference ?

Is there a quality of these references ?

1st strategy : mapping RNA-Seq against a transcriptome or a genome



1st strategy : mapping RNA-Seq against a genome (transcripts or genome)



\+ classical trimming, time saving

\- confidence of gene annotation, no new genes detected

Mapper: different types of tools

Versus transcriptome: bowtie2 (one isoform / gene)

Versus genome with alignment: Tophat2(bowtie2)/HiSat2, STAR

→ Search for the best ‘exact’ alignment

→ Generate sam/bam files = describe alignments

Alignment: essential parameter

BOWTIE2 : search for the best seed alignment

Read: TAGCTACGCTCTACGCTATCATGCATAAAC

Seed 1 fw: TAGCTACGCT
Seed 2 fw: ……CGCTCTACGC
Seed 3 fw: ……ACGCTATCAT
…..
Seed n fw: ……ATGCATAAAC

Some parameters are essential like end-to-end (default)

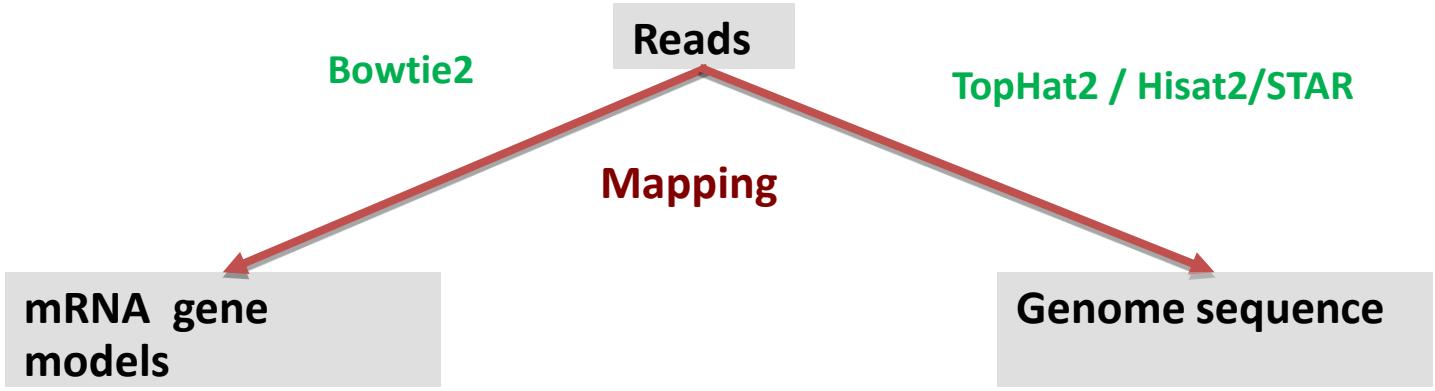
end-to-end (no trimming)

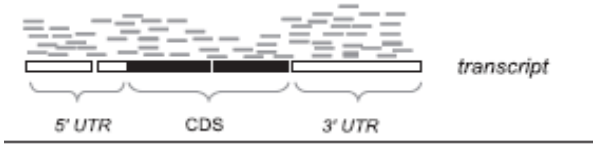
```
Alignment:
  Read:      GACTGGGCGATCTCGACTTCG
             |||||  |||||||||| |||
  Reference: GACTG--CGATCTCGACATCG
```

local (soft clipped)

```
Alignment:
  Read:      ACGGTTGCGTTAA-TCCGCCACG
                 ||||||||| ||||||
  Reference: TAACTTGCGTTAAATCCGCCTGG
```

Mapping





- No splicing
- A reference transcript
- No alternative transcript

- Information on genome (min/max of intron length
- All transcripts (gene models) for 1 gene

Mappers: different type of tools

Versus transcriptome: bowtie2 (one isoform /gene)

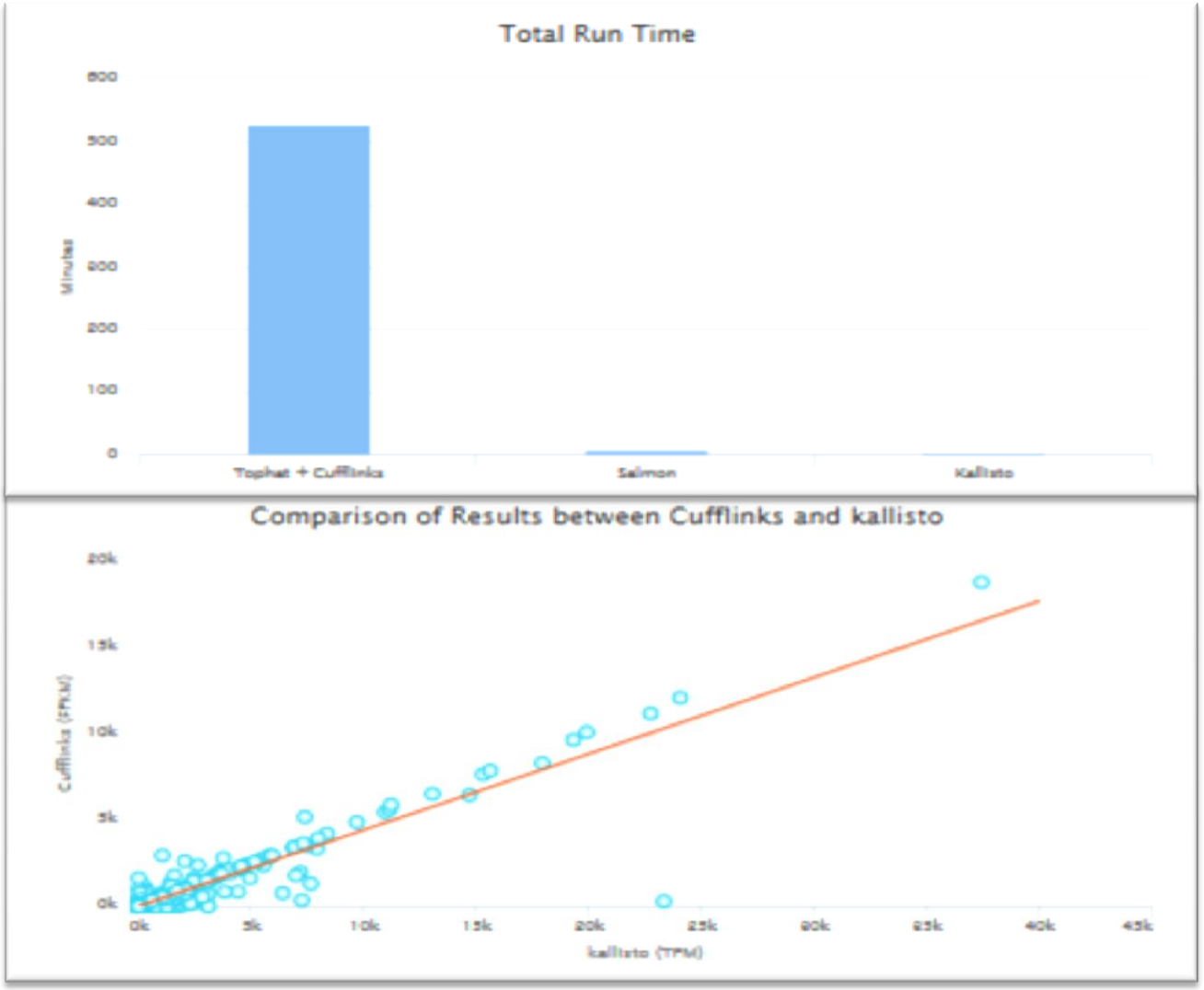
Versus genome with alignment: Tophat2(bowtie2)/HiSat2, STAR
→ Search for the best 'exact' alignment
→ Generate sam/bam files = describe alignments

Versus genome last new tools (Free-alignment) 2015-2017: Kallisto,Salmon,Sailfish
no real alignment : the information is not *where* a read aligns in a transcript , but only *which* transcripts could have generated the read.
→Estimation of k-mer assignment by Expectation-Maximization
→Generate expected counts /transcripts –gene (TPM)
→No sam/bam files, from fastq to TPM

Mappers: different type of tools

The Genomics Core Facility @ NYU CGSB

Web site, June 2016

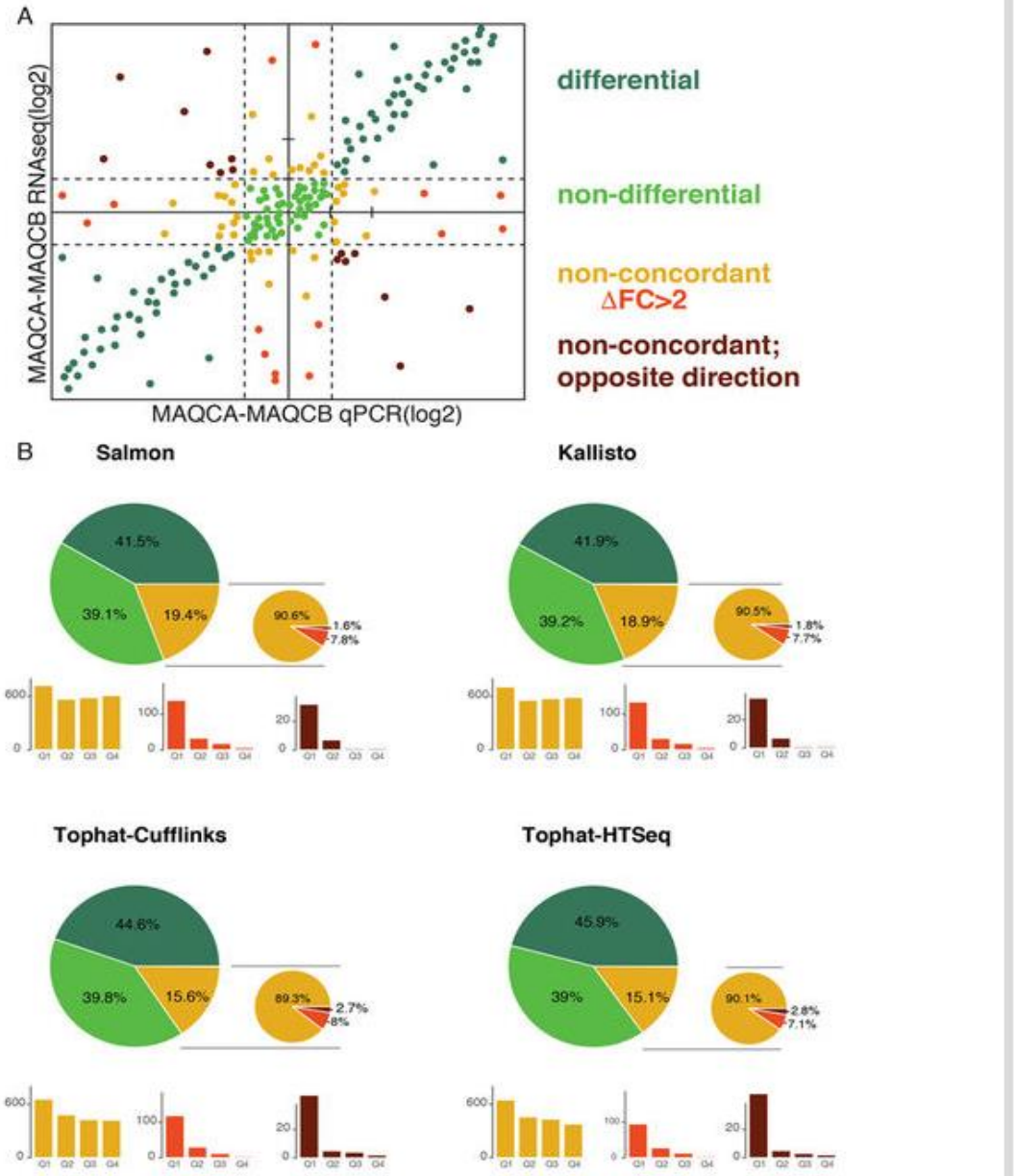


New tools are Faster →no finish photo:) & equivalent accuracy

C. Everaert et al.(2017)
Benchmarking of RNA-sequencing analysis workflows using whole-transcriptome RT-qPCR expression data. *Scientific Reports* **7**, 1559

Equivalent results of workflows

Figure 4

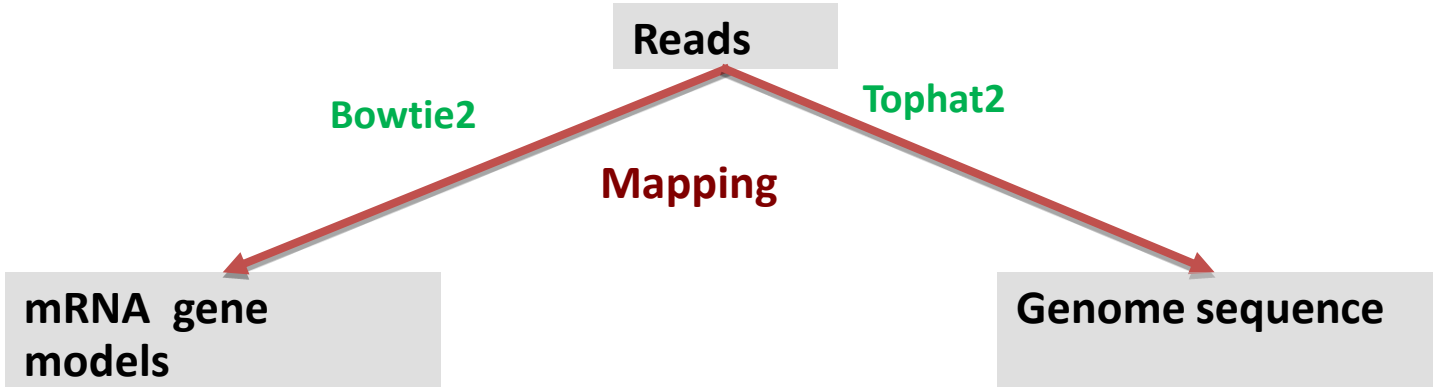


Quantification of non-concordant genes reveals that the numbers are low and similar between workflows. (**A**) A schematic overview of different classes of genes, used for further analysis, by means of a

Mapper parameters (example)

each tool = many parameters with default (75 for tophat2)

G. Baruzzo et al. dec 2016 - Nature methods



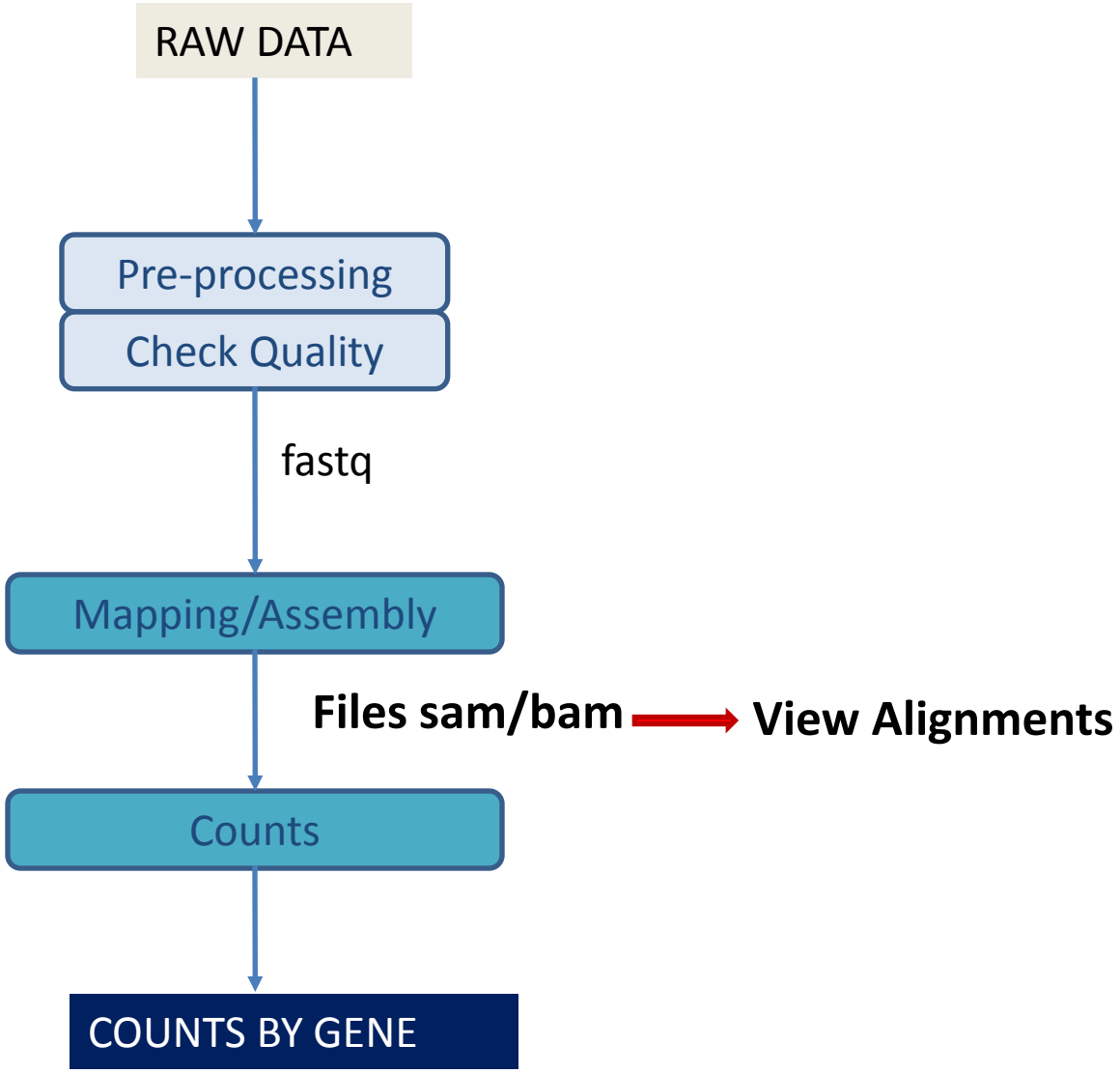
Bowtie2
–x Arabidopsis_transcripts_index
-1 read1.fastq -2 read2.fastq --local

Default is –end-to-end and not --local

Tophat2
--min-intron-length 10
--max-intron-length 70000
-G Arabidopsis_TAIR10.gff
Arabidopsis_genome_index
read1.fastq read2.fastq

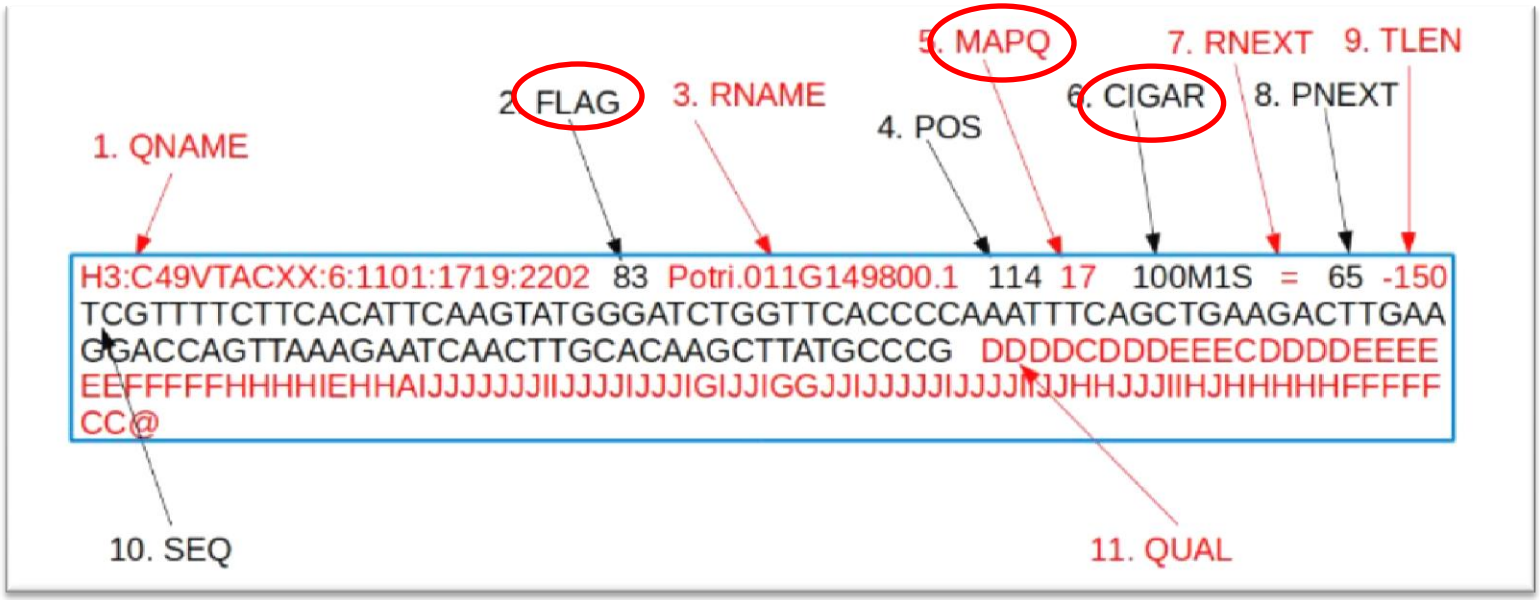
Default min-intron=70 and max-intron=500 000

Alignments


RAW DATA
Pre-processing
Check Quality
fastq
Mapping/Assembly
Files sam/bam
View Alignments
Counts
COUNTS BY GENE

Mapping : SAM/BAM file

https://samtools.github.io/hts-specs/SAMv1.pdf



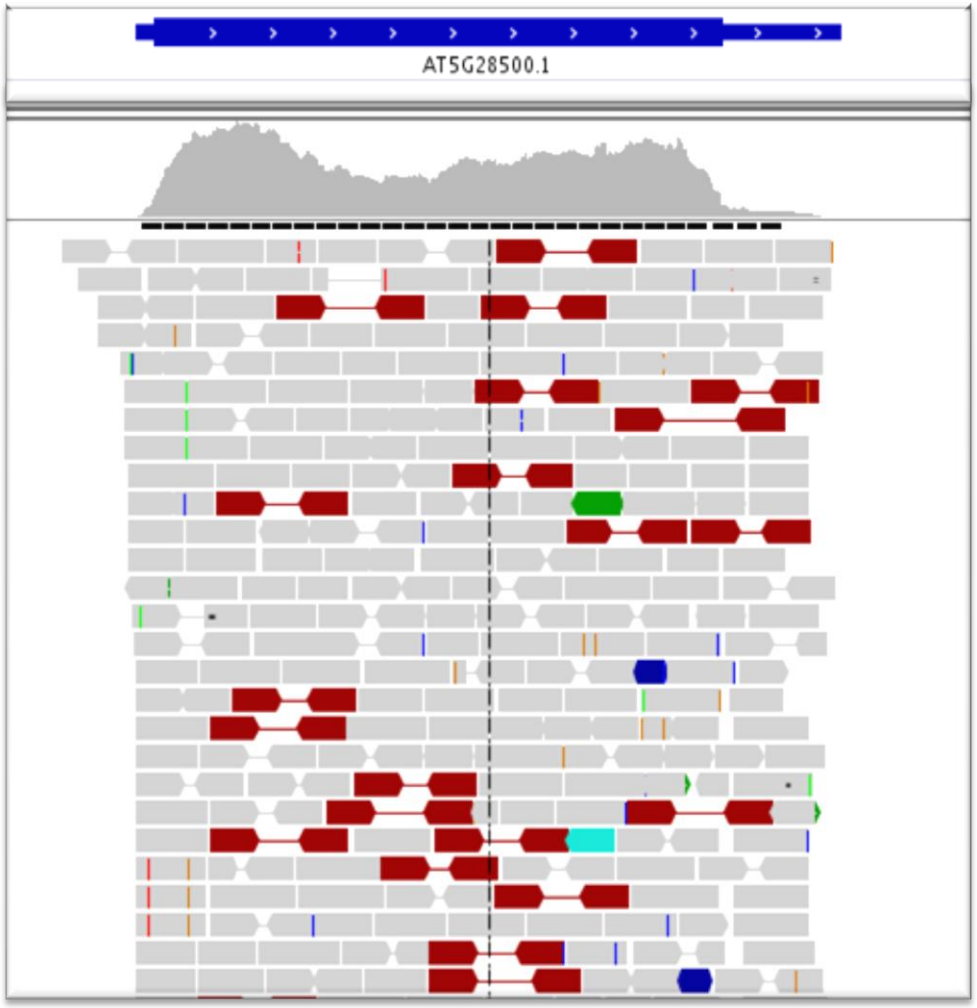
View read alignment via IGV (Integrative Genome Viewer)

J.T. Robinson Integrative Genomics Viewer. Nature Biotechnology 29, 24–26 (2011)

Gene

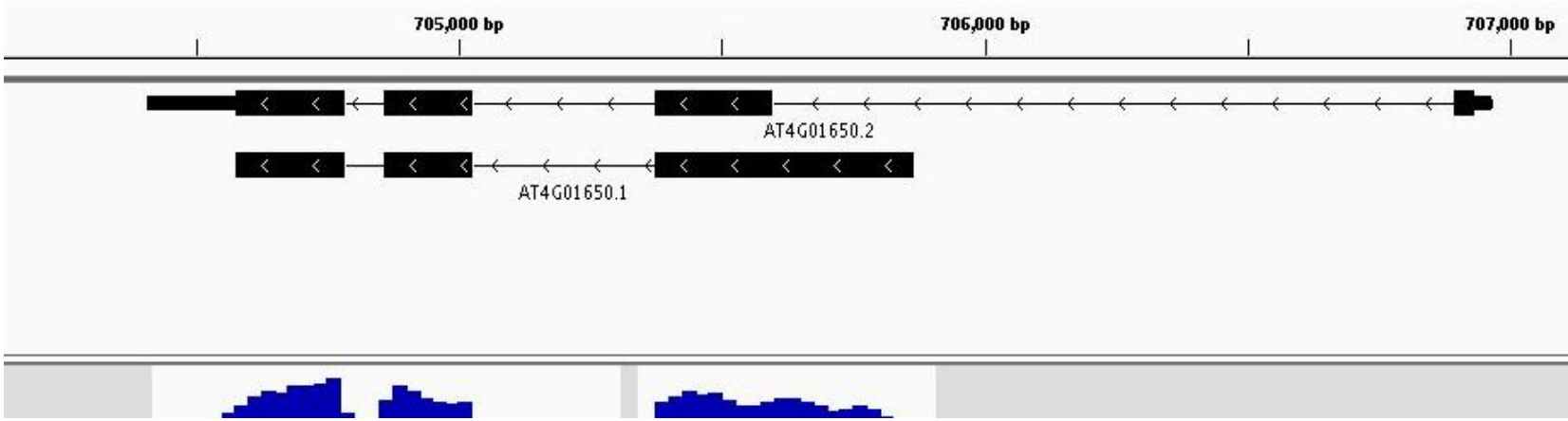
Read density

Read view



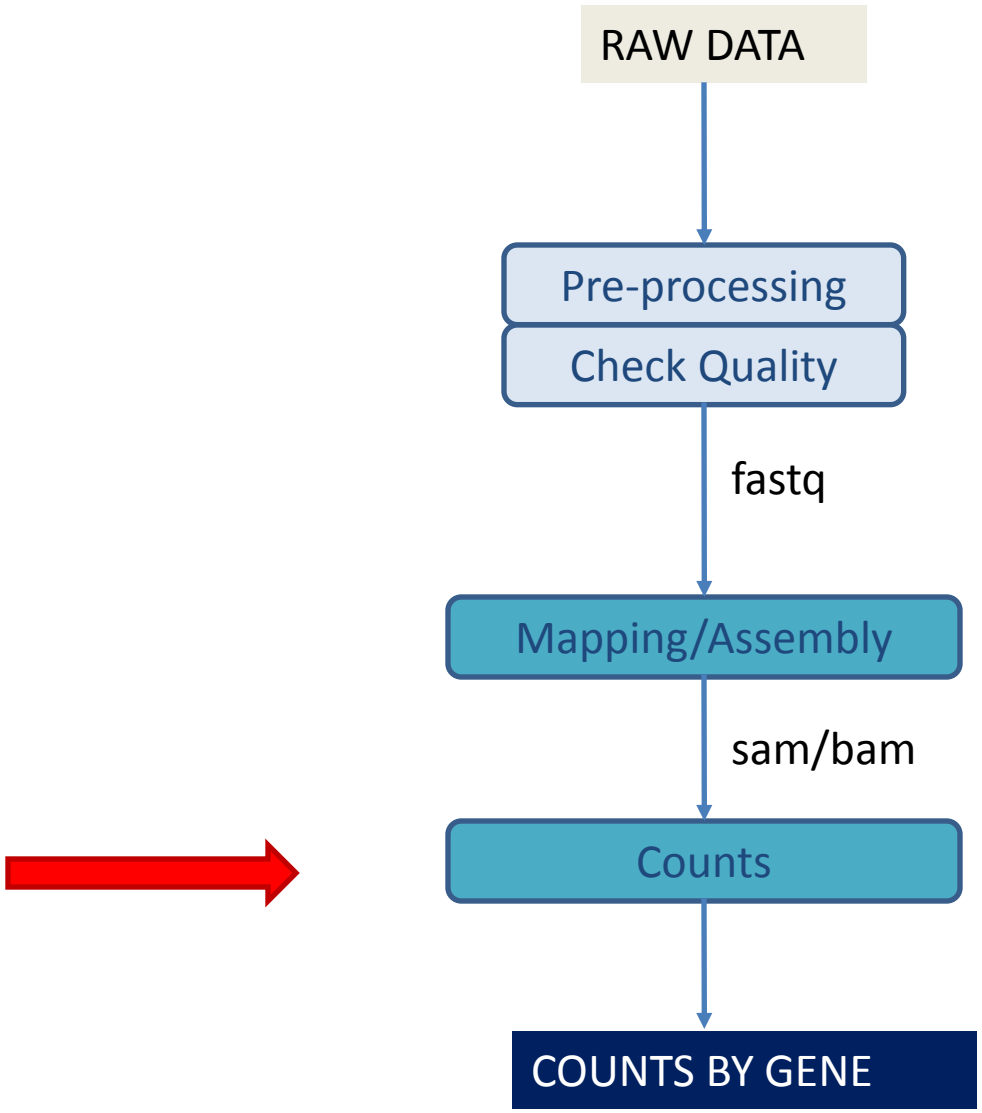
View read density and gene annotations via IGV

2 Isoforms
TAIR10

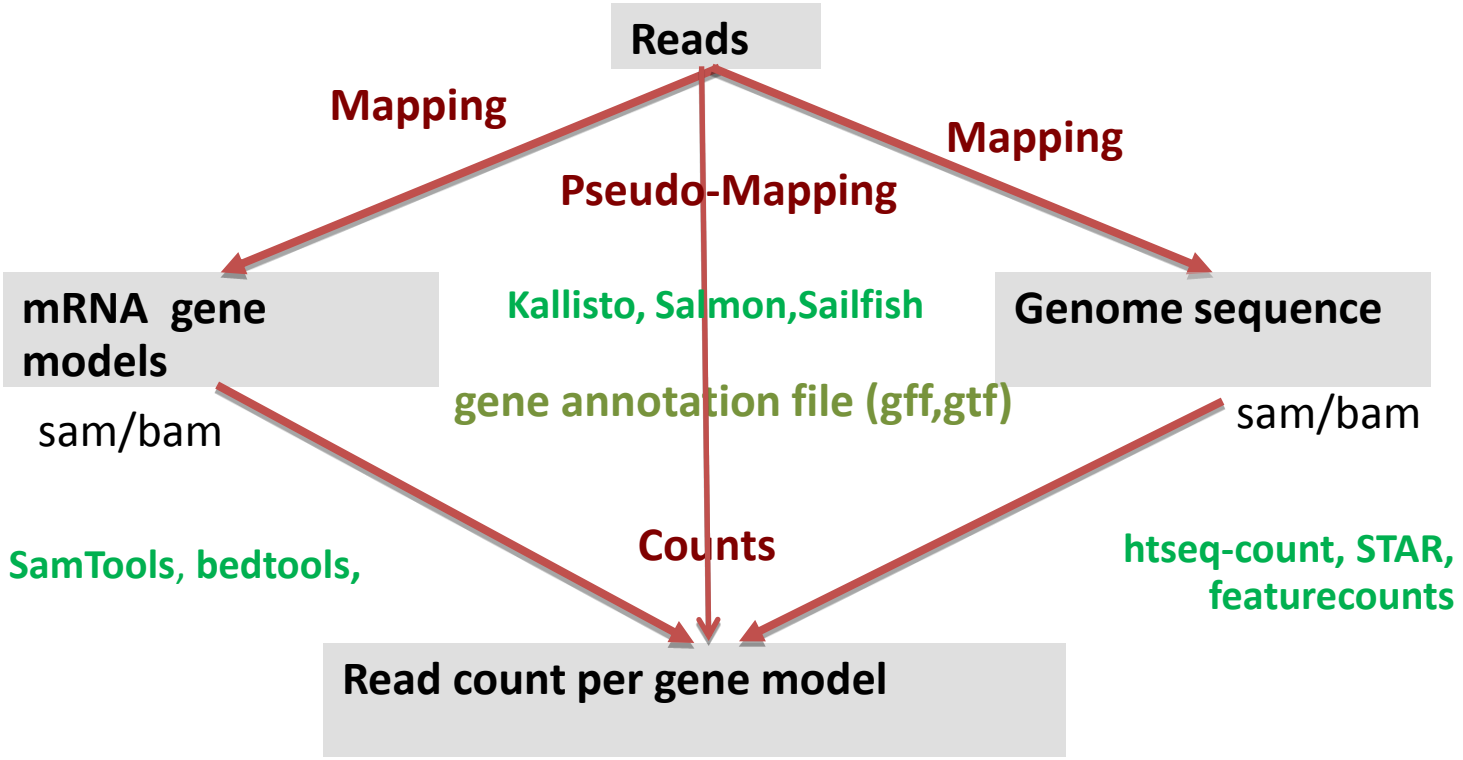


Read density

Counts


RAW DATA
Pre-processing
Check Quality
fastq
Mapping/Assembly
sam/bam
Counts
COUNTS BY GENE

Counts



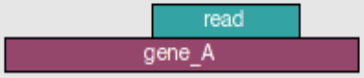
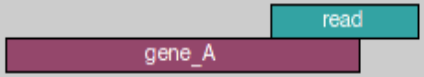
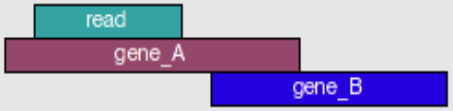
Type of counts : raw count (nb assigned reads), estimated counts, normalized counts RPKM/FPKM/TPM (size of library & gene/transcript)

How to count ?

Counts depend on the type of library

- Stranded or not
- Single reads or Paired-end reads

Counts by isoform or by gene

	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

From htseq-count website

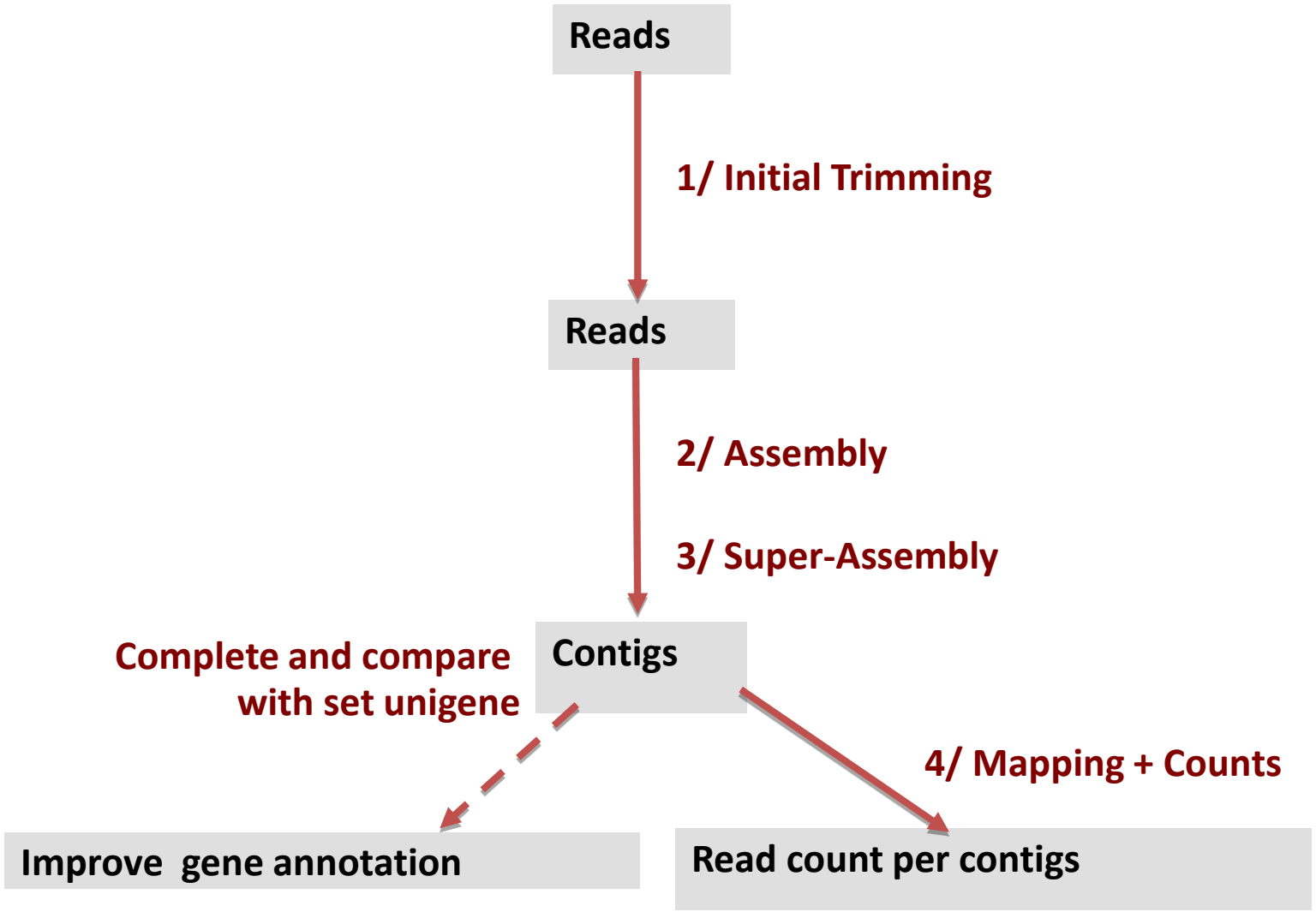
Conclusion on Mapping / counts

- Understand the main characteristics of tools (splicing or not)
  → Know the essential parameters and the default values

- Adjust parameters to your genome or question
  → Study coding or no coding RNA
  → Size of introns of organism
  → Repeated regions : multi-hits
  default 1 best hit randomly chosen

- Counts by gene / transcripts → see next talk

2nd strategy : de novo Assembly of RNA-Seq (without genome)



\+ defined new gene models

\- Assembly: not perfect (contig quality), time and memory consuming

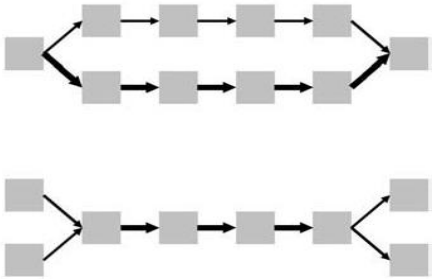
2 methods for *de novo* assembly of RNA-Seq

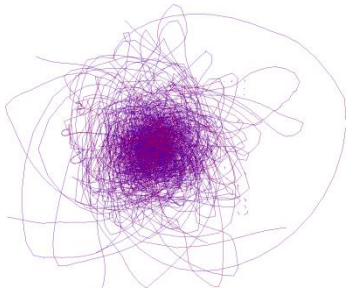
1- OLC : overlap layout consensus : newbler for 454
Research all overlap both reads to form a consensus=contig

→ Too expensive computer resources for million of reads treated
→ Adapted for seq. length > 300bases

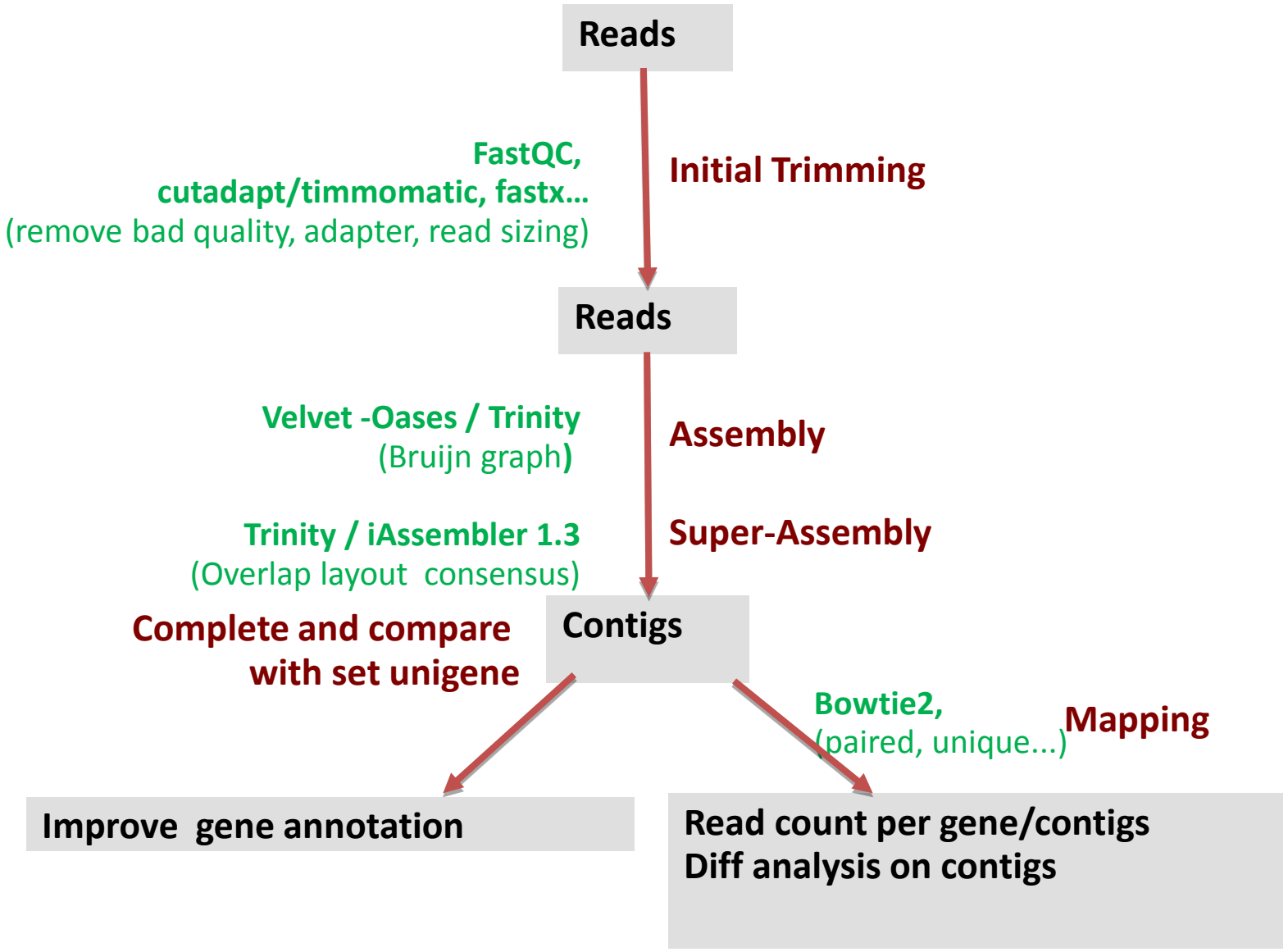
2- Bruijn Graph : velvet, trinity
Cut reads in kmer and overlap k-mer to form graph
Each graph path form a contig

→ If sequencing errors : many contigs
→ Need memory (150G – 500G)





2nd strategy : de novo Assembly of RNAseq (without genome)

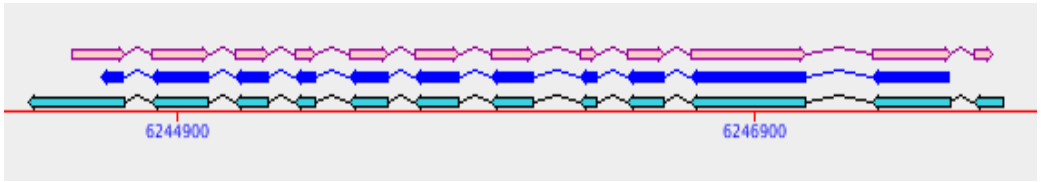


Assembly Results

F1_Mplex	
Nb of PE reads	43 030 388 PE
Nb of contigs	33 736 (length mean 1360)
Nb of mapped contigs Genome TAIR10	33 072 98%

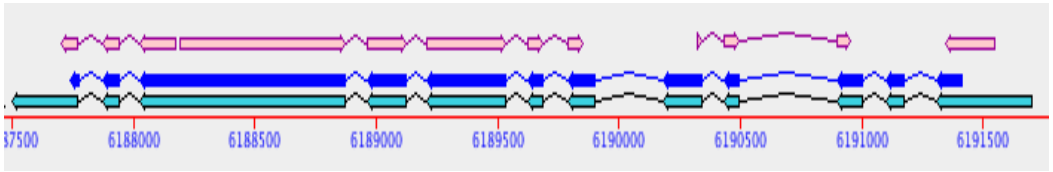
Data from Illumina HiSeq2000
- **Velvet/oases (kmer 61,71)**
- **iAssembler**

1 gene – 1 contig
same gene model



88% of genes confirmed by at least one contig

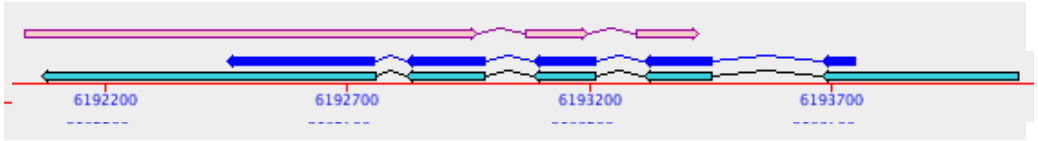
1 gene – 2 or more contigs
same gene model



Quality of Assembly : contig versus gene annotation

35% of contigs with other gene models (isoforms)

1 gene – 1 or n contigs
with other gene models



→3% of contigs with no annotated genes

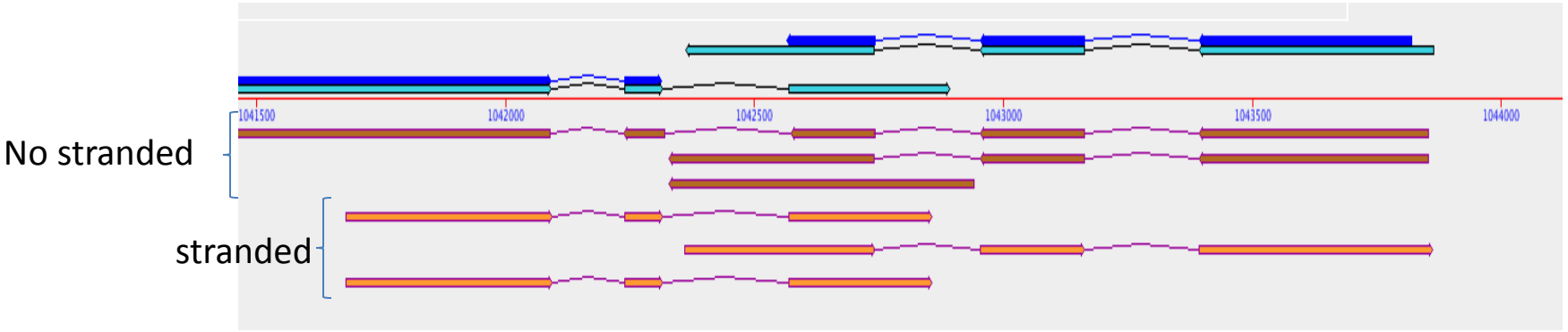
Check contig quality

- Number of contigs near number of expected genes
- Median length of contigs (N50)
- % of reads that maps contigs, redundancy of the contigs “multi-hits”
- % contigs that encode proteins ‘known in bank’

→ Trinotate (Trinity suite)

Assembly

Chimeric contigs → remove in part with library stranded



Conclusion for assembly

→ A good quality of contigs, efficient to detect new gene models
→ Problems: distinct false/good gene models, chimera that increase with read number
→Improving Assembly tools with PE, oriented, tuning parameters (coverage)

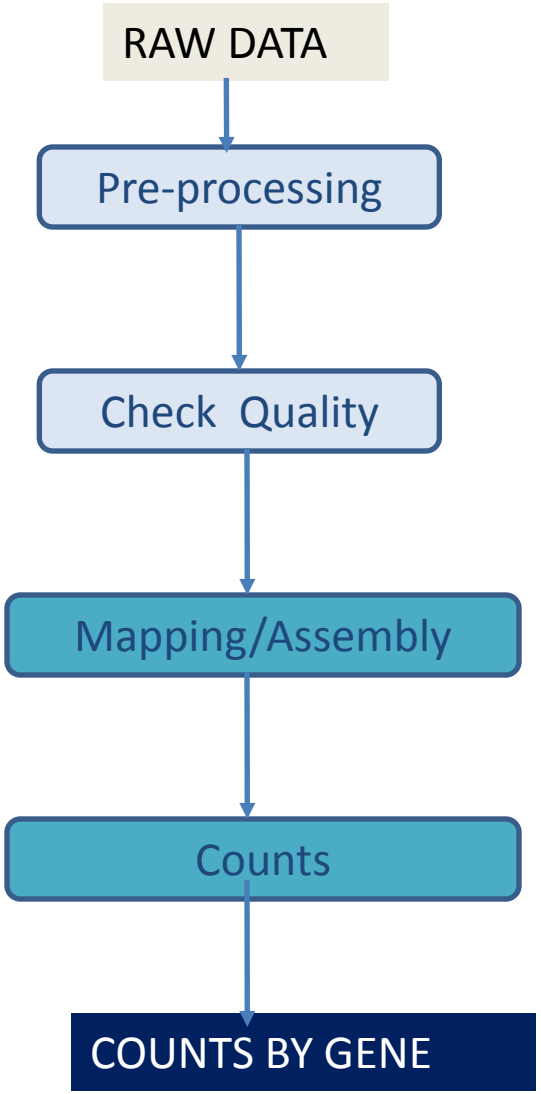
Be careful, assembly can be difficult if genome
contains many repeats, heterozygous regions, polyploidy …
→ A great number of contigs (ex > 500.000 without change parmeters)

Conclusion on Bioinformatic usage

- Don’t forget the biological question
- If you work with results done by other group
  - ask information on the tools used
  - just take time to check essential parameters

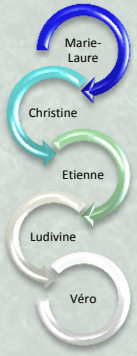
- All seem easy when all is working well !

→ **RUN/TEST and ANALYSE results is the best usage :!)**



THANKS

Gnet Team



OGE Team

RAW DATA

Pre-processing

Check Quality

Mapping/Assembly

Counts

COUNTS BY GENE

Module 3: Diff Analysis