RNA-seq

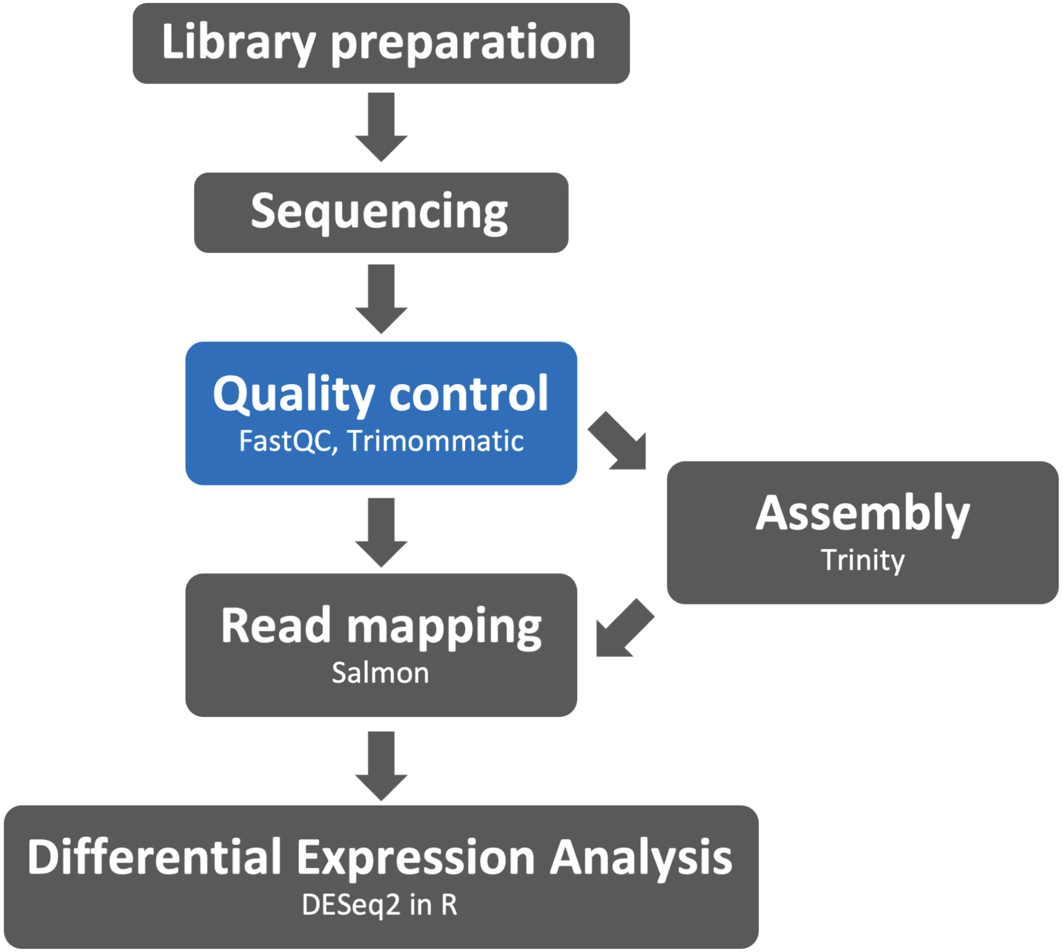
Quality Control

Aula 02

https://tttorres.github.io/transcriptomics/

RNA-seq workflow

Quality Control



Library preparation

Sequencing

Quality control
FastQC, Trimommatic

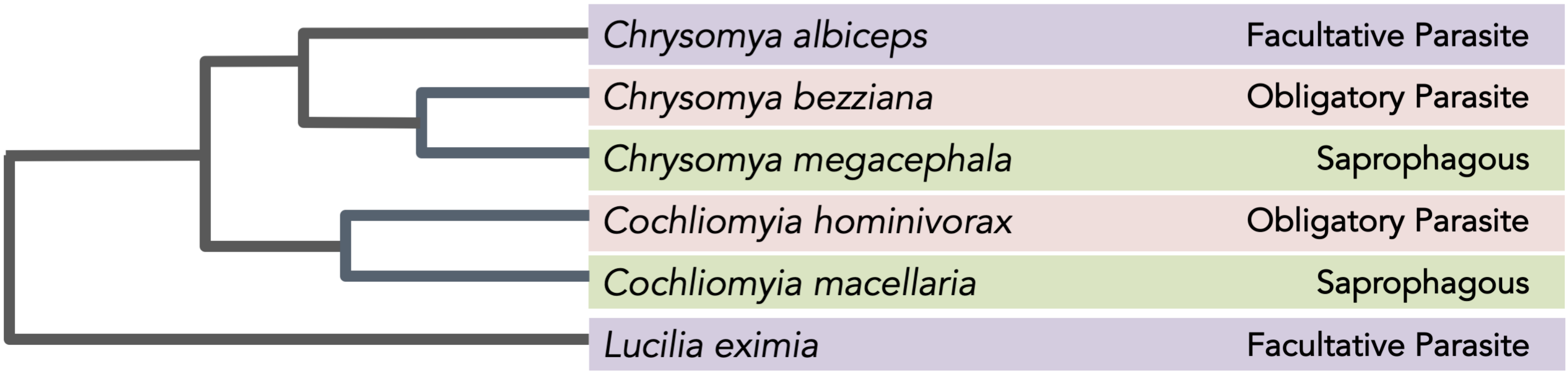
Assembly
Trinity

Read mapping
Salmon

Differential Expression Analysis
DESeq2 in R

RNA-seq workflow

Dataset



Chrysomya albiceps	Facultative Parasite
Chrysomya bezziana	Obligatory Parasite
Chrysomya megacephala	Saprophagous
Cochliomyia hominivorax	Obligatory Parasite
Cochliomyia macellaria	Saprophagous
Lucilia eximia	Facultative Parasite

RNA-seq workflow

Dataset

- Calb Female & Larvae
- Cbez Female & Larvae
- Chom Female & Larvae
- Cmac Female & Larvae
- Cmeg Female & Larvae
- Lexi Female & Larvae

RNA-Seq: initial processing

Quality check

1. Go to home folder and create a new folder called `rnaseq`

```
cd ~
mkdir rnaseq
```

2. Go to the `rnaseq` folder and create a new folder called `01-RawReads`

```
cd rnaseq
mkdir 01-RawReads
```

3. Go to the `01-RawReads` folder

```
cd rnaseq
mkdir 01-RawReads
```

4. Download the files (R1 and R2) for your sample from the link indicated on the course page ("save link as")

RNA-Seq: initial processing

Quality check

5. Check the md5sum of each file

6. View the first 10 lines of each file

RNA-Seq: initial processing

Quality check

5. Check the md5sum of each file

```
md5sum Chom_R1.fastq.gz
md5sum Chom_R2.fastq.gz
```

6. View the first 10 lines of each file

```
```

RNA-Seq: initial processing

Quality check

5. Check the md5sum of each file

```
md5sum Chom_R1.fastq.gz
md5sum Chom_R2.fastq.gz
```

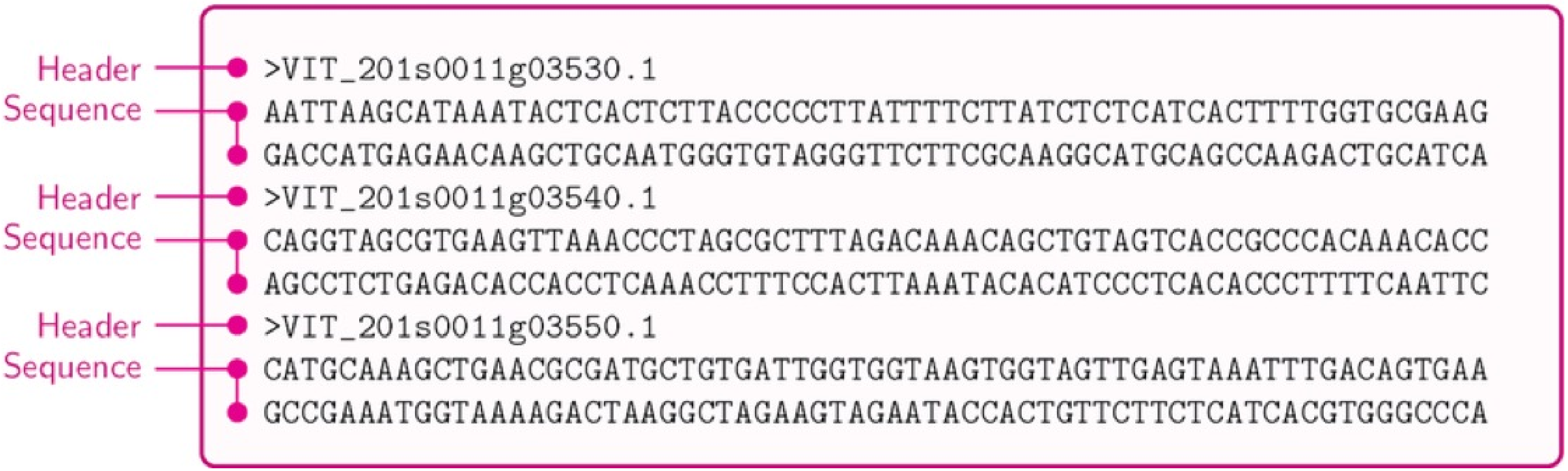
6. View the first 10 lines of each file

```
head Chom_R1.fastq.gz
head Chom_R2.fastq.gz
```

RNA-Seq analyses

Sequence formats

Fasta

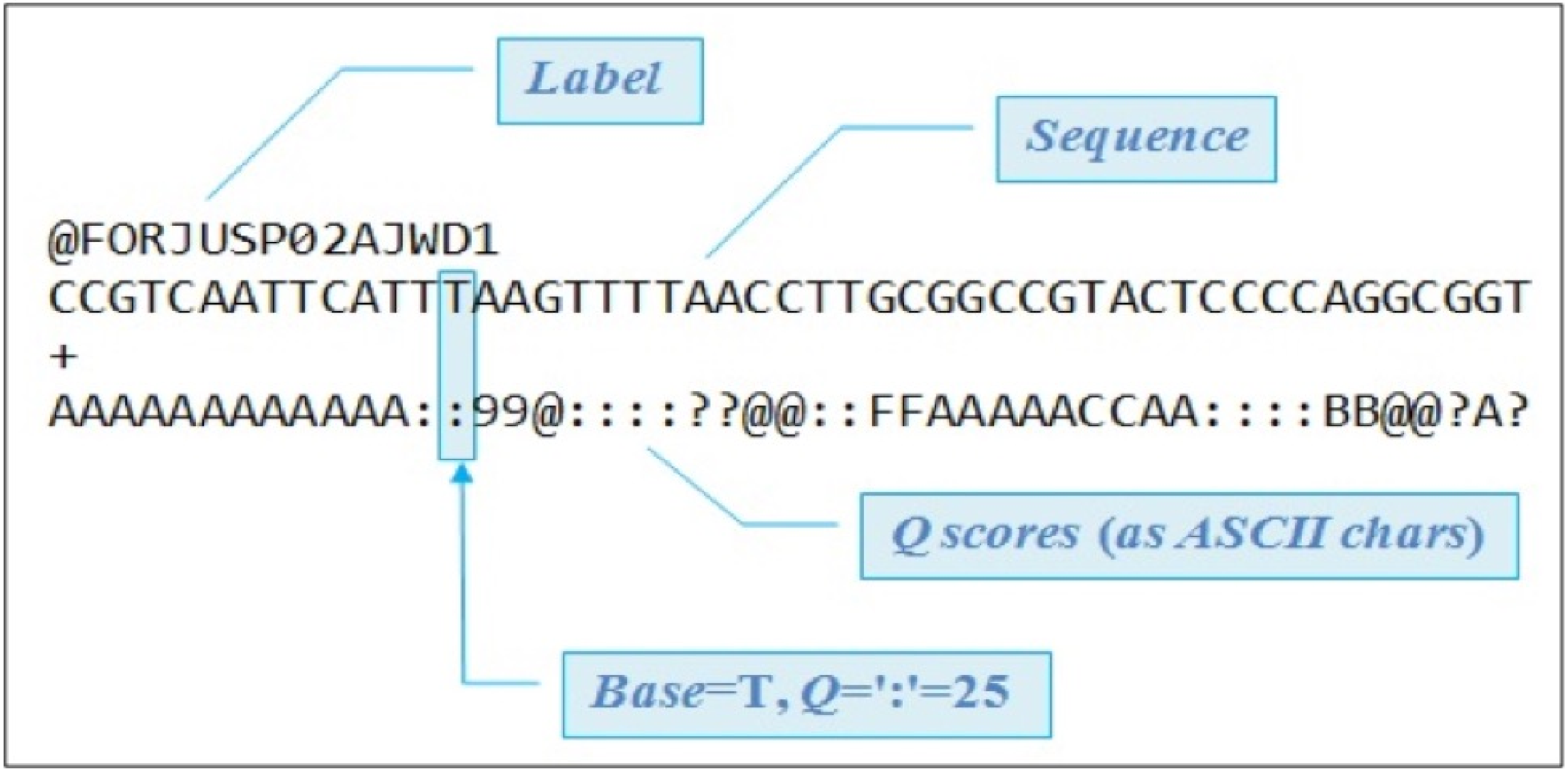


```
>VIT_201s0011g03530.1
AATTAAGCATAAATACTCACTCTTACCCCCTTATTTTCTTATCTCTCATCACTTTTGGTGCGAAG
GACCATGAGAACAAGCTGCAATGGGTGTAGGGTTCTTCGCAAGGCATGCAGCCAAGACTGCATCA
>VIT_201s0011g03540.1
CAGGTAGCGTGAAGTTAAACCCTAGCGCTTTAGACAAACAGCTGTAGTCACCGCCCACAAACACC
AGCCTCTGAGACACCACCTCAAACCTTTCCACTTAAATACACATCCCTCACACCCTTTTCAATTC
>VIT_201s0011g03550.1
CATGCAAAGCTGAACGCGATGCTGTGATTGGTGGTAAGTGGTAGTTGAGTAAATTTGACAGTGAA
GCCGAAATGGTAAAAGACTAAGGCTAGAAGTAGAATACCACTGTTCTTCTCATCACGTGGGCCCA
```

RNA-Seq analyses

Sequence formats

Fastq



```
@FORJUSP02AJWD1
CCGTCAATTCATTTAAGTTTTAACCTTGCGGCCGTACTCCCCAGGCGGT
+
AAAAAAAAAAAA::99@::::??@@::FFAAAAACCAA::::BB@@?A?
```

RNA-Seq analyses

Usual steps

- Initial Processing
- Assembly (model *vs* non-model)
- Annotation of transcripts
- Alignment/pseudoalignment reads to a reference
- Differential gene expression

RNA-Seq analyses

Initial Processing

- Demultiplexing
- Removing adapters
- Trimming
- Kmer Normalization

RNA-Seq: initial processing

Demultiplexing

Multiplexing is a common strategy due to the sheer amount of data a sequencer is able to generate.

> After sequencing, each read may be traced back to its original sample using the index sequence and binned accordingly.

- FASTX-Toolkit
- Stacks
- biopieces

RNA-Seq: initial processing

Removing adapters

> If cDNA insert sizes are sufficiently small and sequencing read lengths sufficiently long, it is possible to generate sequencing reads that contain a portion of adapter sequence at the 3′-end.

- FASTX-Toolkit
- Stacks
- biopieces
- trimmomatic

RNA-Seq: initial processing

Trimming

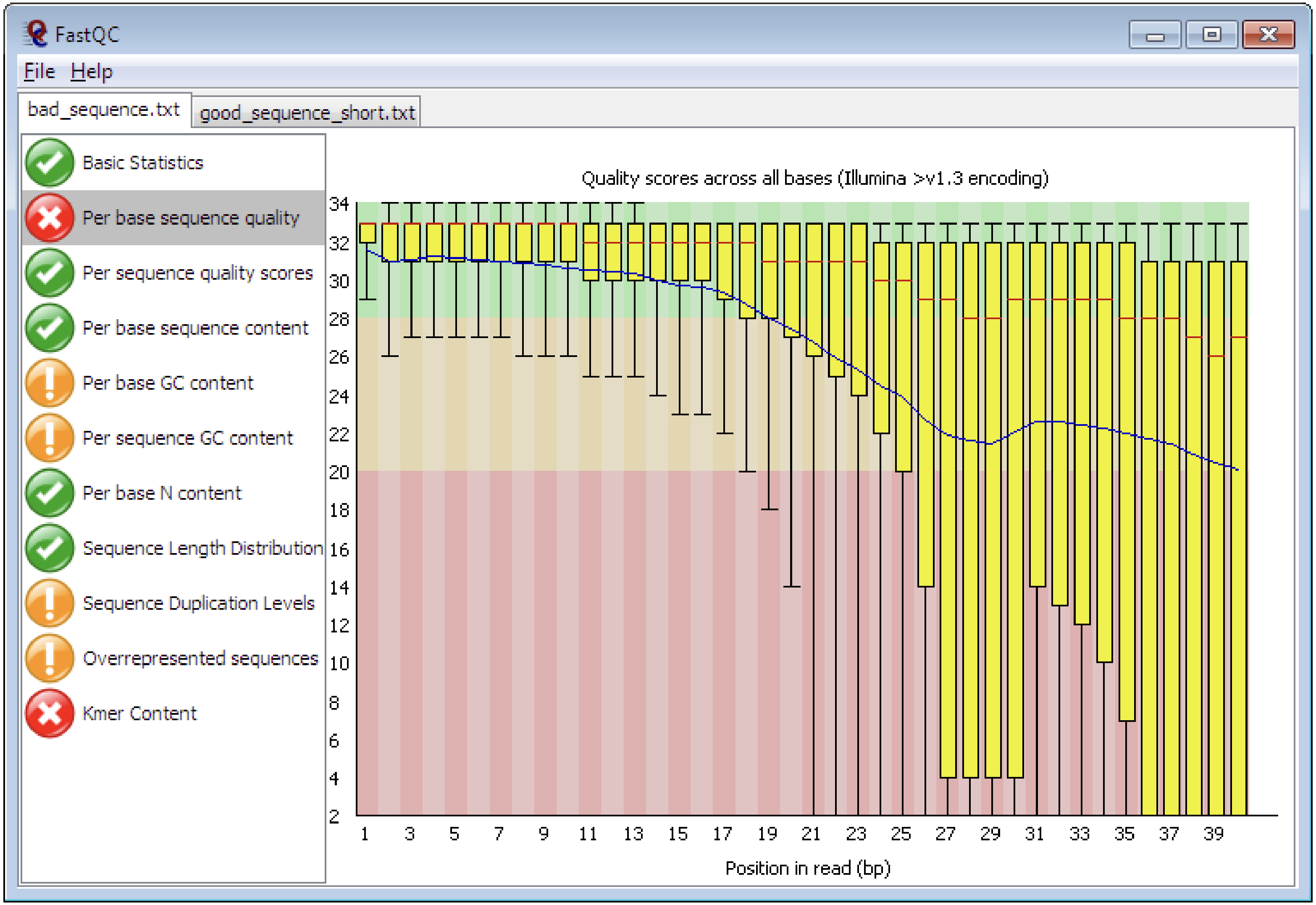
> Reads likely to contain multiple sequencing errors provide less biological information and are expected to hinder assembly and alignment.
> Some analyses are more tolerant of error than others. For example, de novo assembly requires much cleaner reads than alignment to a reference genome.

- FASTX-Toolkit
- Stacks
- biopieces
- trimmomatic

How do we know our sequences need trimming?

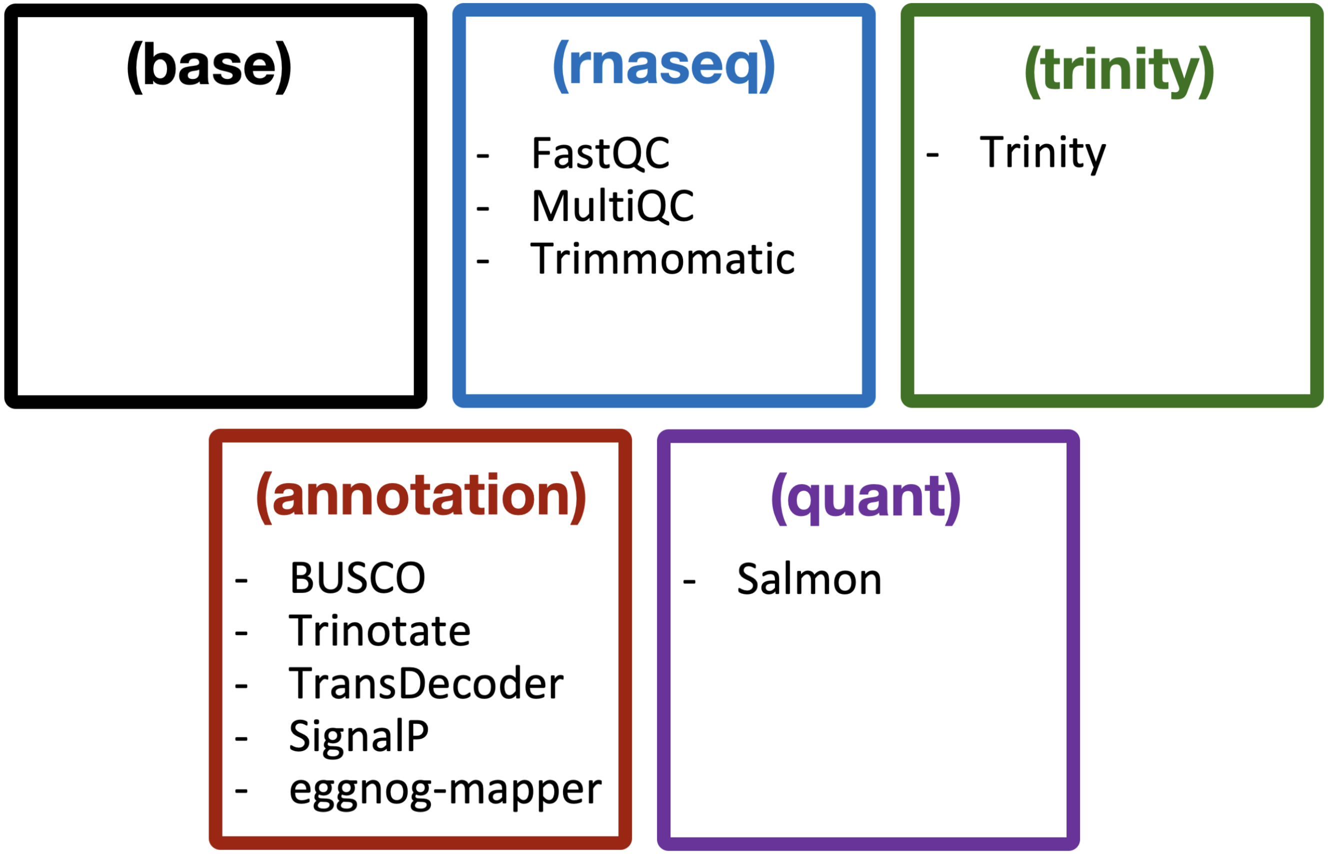
RNA-Seq: initial processing

Quality check


FastQC
File Help
bad_sequence.txt
good_sequence_short.txt
✓ Basic Statistics
✗ Per base sequence quality
✓ Per sequence quality scores
✓ Per base sequence content
! Per base GC content
! Per sequence GC content
✓ Per base N content
✓ Sequence Length Distribution
! Sequence Duplication Levels
! Overrepresented sequences
✗ Kmer Content
Quality scores across all bases (Illumina >v1.3 encoding)
34
32
30
28
26
24
22
20
18
16
14
12
10
8
6
4
2
1
3
5
7
9
11
13
15
17
19
21
23
25
27
29
31
33
35
37
39
Position in read (bp)

RNA-Seq: initial processing

Conda environments



(base)

(rnaseq)

- FastQC
- MultiQC
- Trimmomatic

(trinity)

- Trinity

(annotation)

- BUSCO
- Trinotate
- TransDecoder
- SignalP
- eggnog-mapper

(quant)

- Salmon

RNA-Seq: initial processing

Quality check

1. In the folder with the sequences, activate `rnaseq` environment

```
conda activate rnaseq
```

2. Call fastqc

```
fastqc
```

RNA-Seq: initial processing

Quality check: FASTQC

The main functions of FASTQC are:

- Import of data from BAM, SAM or FastQ files
- Providing a quick overview to tell you in which areas there may be problems
- Summary graphs and tables to quickly assess your data
- Export of results to an HTML based permanent report
- Offline operation to allow automated generation of reports without running the interactive application

RNA-Seq: initial processing

Running FASTQC

1. GUI: Call fastqc

```
fastqc
```

2. Individual sequence

```
fastqc SpeciesF_R1.fastq.gz
```

3. Bulk (using wildcard `*`)

```
fastqc *.gz
```

4. Using multiple threads

```
fastqc *.gz -t 8
```

RNA-Seq: initial processing

Running FASTQC

Using multiple threads

```
fastqc *.gz -t 8
```

In my computer:

- it took ~5s for two files:
  - `ChomF_R1.fastq.gz` with 6.1MB
  - `ChomF_R2.fastq.gz` with 6.0MB
- for 2 files with ~2GB each (20million reads), it took ~5min

RNA-Seq: initial processing

Running FASTQC

In your computer:

Use fastqc in a single file:

```
fastqc Species_R1.fastq.gz
```

RNA-Seq: initial processing

Running FASTQC

In your computer:

Use fastqc in a single file:

```
fastqc Species_R1.fastq.gz
```

FASTQ will generate two files:

- Species_R1_fastqc.html
- Species_R1_fastqc.zip

RNA-Seq: initial processing

FASTQC report

- Open `Species_R1_fastqc.html` (double-click):

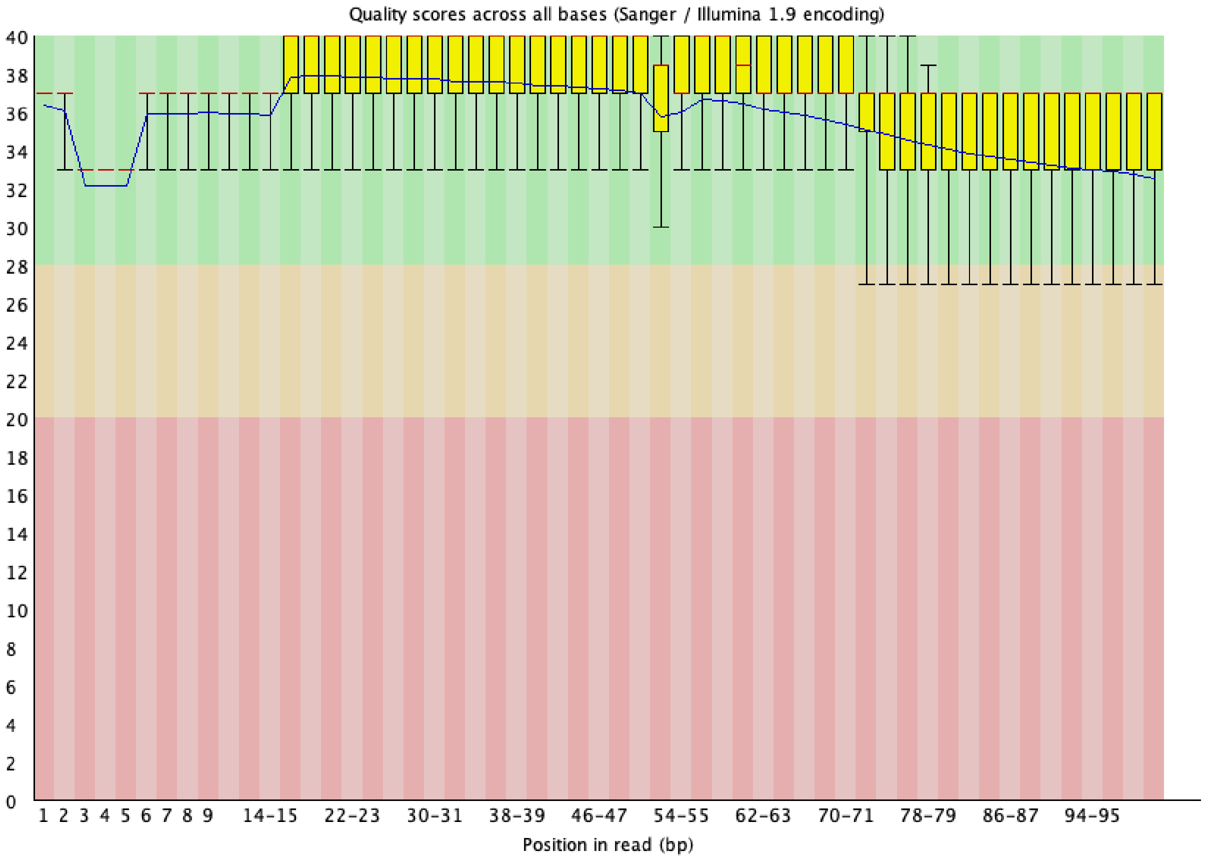
RNA-Seq: initial processing

FASTQC report

- Basic Statistics
- Per base sequence quality
- Per tile sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content

RNA-Seq: initial processing

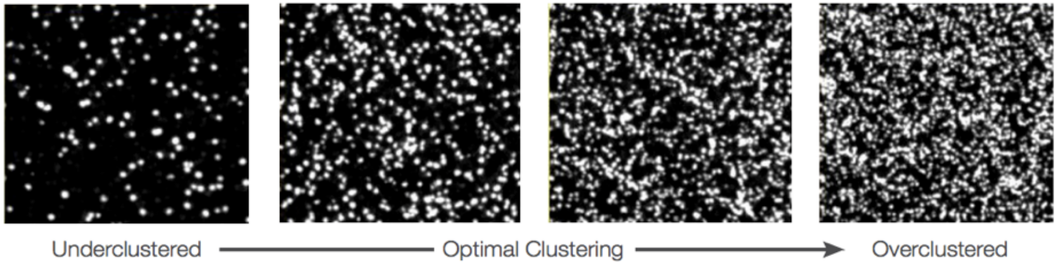
Per base sequence quality

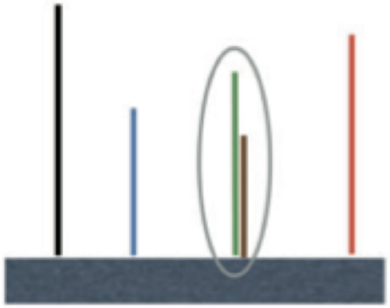


RNA-Seq: initial processing

Per base sequence quality

Overclustering

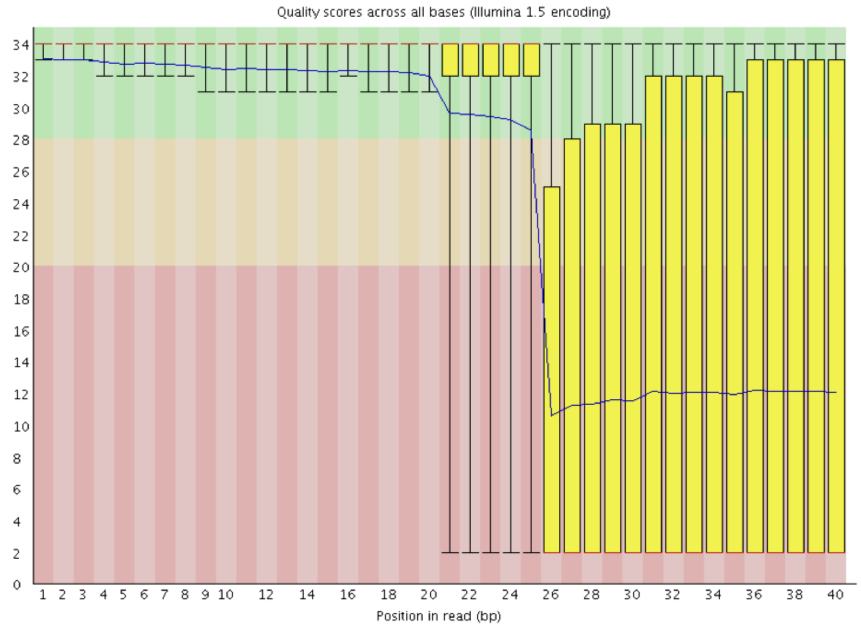


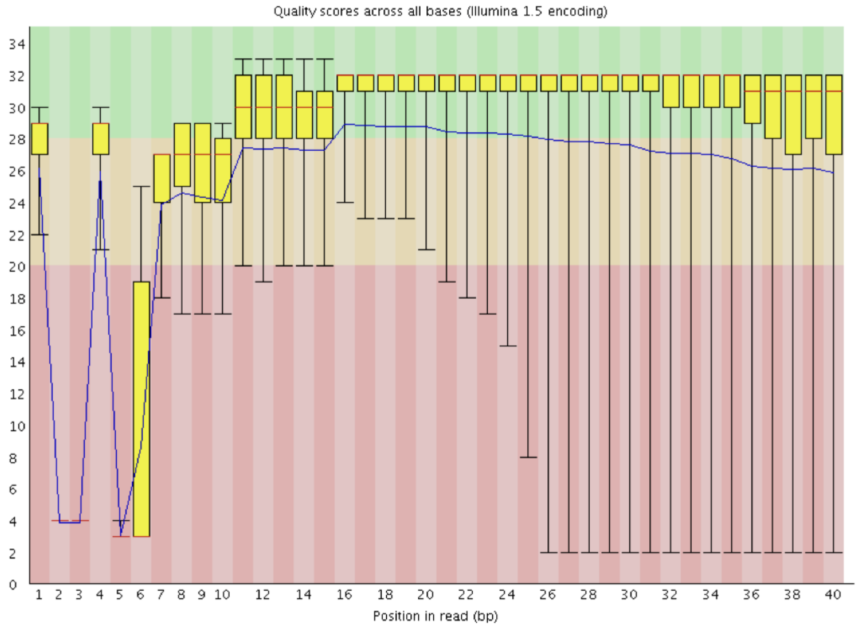


RNA-Seq: initial processing

Per base sequence quality

Issues with the sequencing instruments





RNA-Seq: initial processing

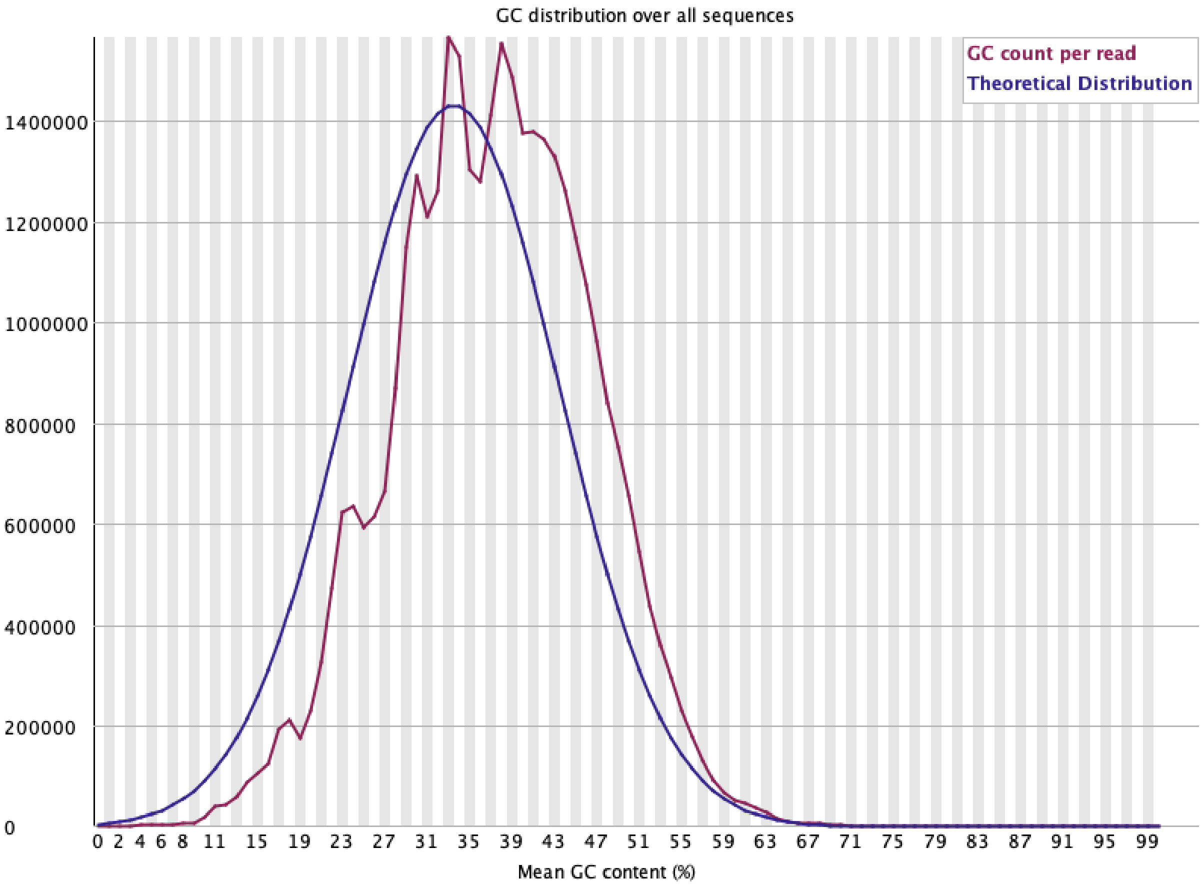
Per base sequence content

RNA-seq

> Always gives a FAIL for RNA-seq data. This is because the first 10-12 bases result from the ‘random’ hexamer priming that occurs during RNA-seq library preparation. This priming is not as random as we might hope giving an enrichment in particular bases for these intial nucleotides.

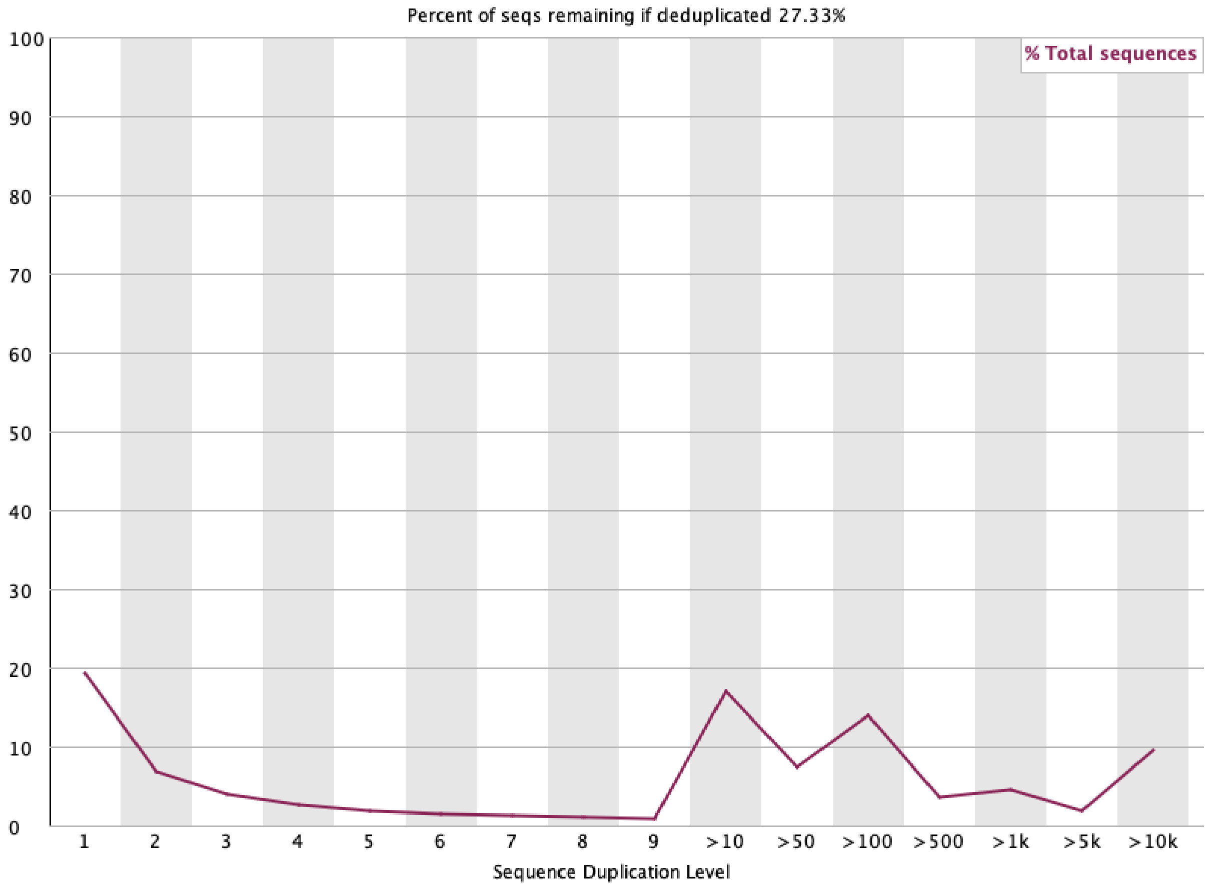
RNA-Seq: initial processing

Per sequence GC content



RNA-Seq: initial processing

Sequence Duplication Levels



RNA-Seq: initial processing

Trimming and filtering: Trimmomatic

- Cut adapter / other illumina-specific sequences from the reads;
- Cut bases off the start of a read, if below a threshold quality
- Cut bases off the end of a read, if below a threshold quality
- Drop the read if it is too short (below 25 bases)
- Drop low quality reads

RNA-Seq: initial processing

Trimming and filtering: Trimmomatic

1. In `rnaseq`, create folder `02-FilteredReads`

```
mkdir 02-FilteredSeqs
cd 02-FilteredReads
```

2. Save the file `TruSeq3-PE.fa` within `02-FilteredReads`

3. Run trimmomatic in the 'rnaseq' environment

```
trimmomatic
Usage: PE [-version] [-threads <threads>] [-phred33|
       -phred64] [-trimlog <trimLogFile>] [-summary
       <statsSummaryFile>] [-quiet] [-validatePairs]
       [-basein <inputBase> | <inputFile1> <inputFile2>]
       [-baseout <outputBase> | <outputFile1P>
       <outputFile1U> <outputFile2P> <outputFile2U>]
       <trimmer1>...
```

RNA-Seq: initial processing

Running Trimmomatic

Option	Function
SE or PE	Reads are single end or paired end
ILLUMINACLIP	Perform adapter removal
SLIDINGWINDOW	Perform sliding window trimming
LEADING	Cut bases off the start of a read, if below a threshold
TRAILING	Cut bases off the end of a read, if below a threshold
CROP	Cut the read to a specified length.
HEADCROP	Cut a specified number of bases from the start of the read
MINLEN	Drop an entire read if it is below a specified length
TOPHRED33	Convert quality scores to Phred-33
TOPHRED64	Convert quality scores to Phred-64

RNA-Seq: initial processing

Running Trimmomatic

3. Copy the command to run trimmomatic

4. Change the name of the `SPECIES` variable

```
trimmomatic PE -threads 8 -phred33 \
~/rnaseq/01-RawReads/Chom_R1.fastq.gz \
~/rnaseq/01-RawReads/Chom_R2.fastq.gz \
Chom_R1_paired.fastq Chom_R1_unpaired.fastq \
Chom_R2_paired.fastq Chom_R2_unpaired.fastq \
ILLUMINACLIP:../02-FilteredReads/TruSeq3-PE.fa:2:30:10:2:T
SLIDINGWINDOW:4:20 \
LEADING:3 TRAILING:3 MINLEN:36 >Chom_R1-report.txt
```

RNA-Seq: initial processing

Running Trimmomatic

3. Copy the command to run trimmomatic

4. Change the name of the `SPECIES` variable

```
trimmomatic PE \
            -threads 8 \
            -phred33 \
            -summary ${SPECIES}-report.txt \
            ~/rnaseq/01-RawReads/${SPECIES}_R1.fastq.gz \
            ~/rnaseq/01-RawReads/${SPECIES}_R2.fastq.gz \
            ${SPECIES}_R1_paired.fastq ${SPECIES}_R1_unpaired.fastq \
            ${SPECIES}_R2_paired.fastq ${SPECIES}_R2_unpaired.fastq \
            ILLUMINACLIP:../TruSeq3-PE.fa:2:30:10:2:True \
            SLIDINGWINDOW:4:20 \
            LEADING:3 \
            TRAILING:3 \
            MINLEN:50
```

RNA-Seq: initial processing

Running Trimmomatic

```
Input Read Pairs: 100000
Both Surviving: 91421 (91.42%)
Forward Only Surviving: 4513 (4.51%)
Reverse Only Surviving: 2034 (2.03%)
Dropped: 2032 (2.03%)
TrimmomaticPE: Completed successfull
```

RNA-Seq: initial processing

Running FASTQC on trimmed data

In your computer:

Use fastqc in a single file:

```
fastqc Species_R1_paired.fastq
```

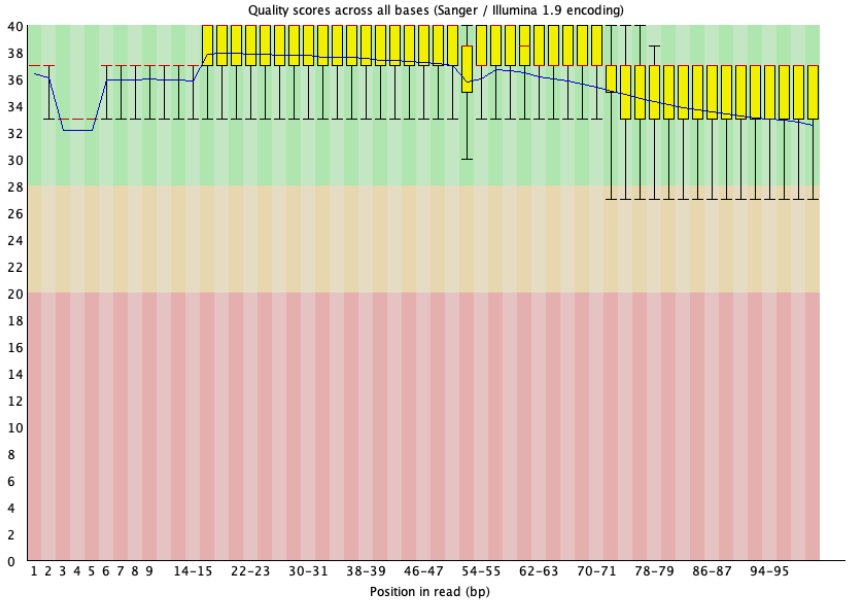
FASTQ will generate two files:

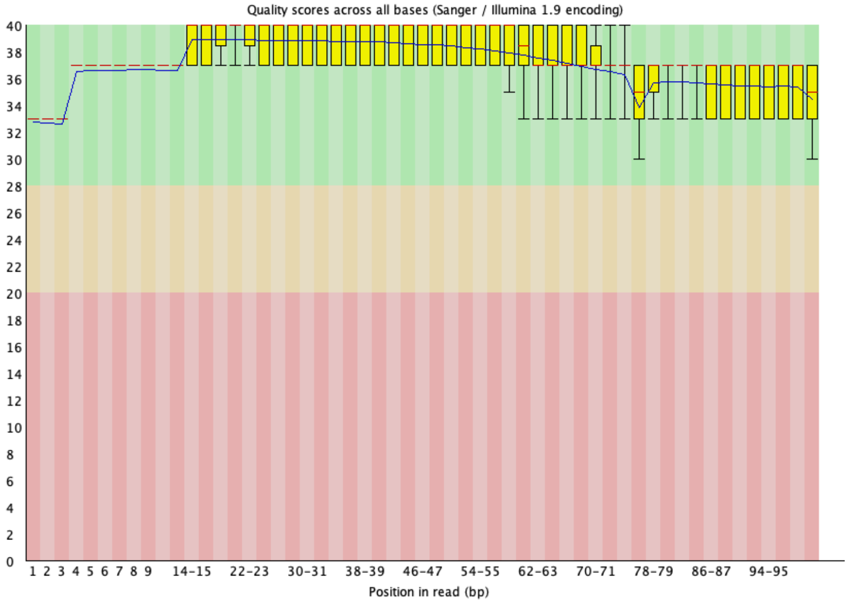
- Species_R1_paired_fastqc.html
- Species_R1_paired_fastqc.zip

RNA-Seq: initial processing

Running FASTQC on trimmed data

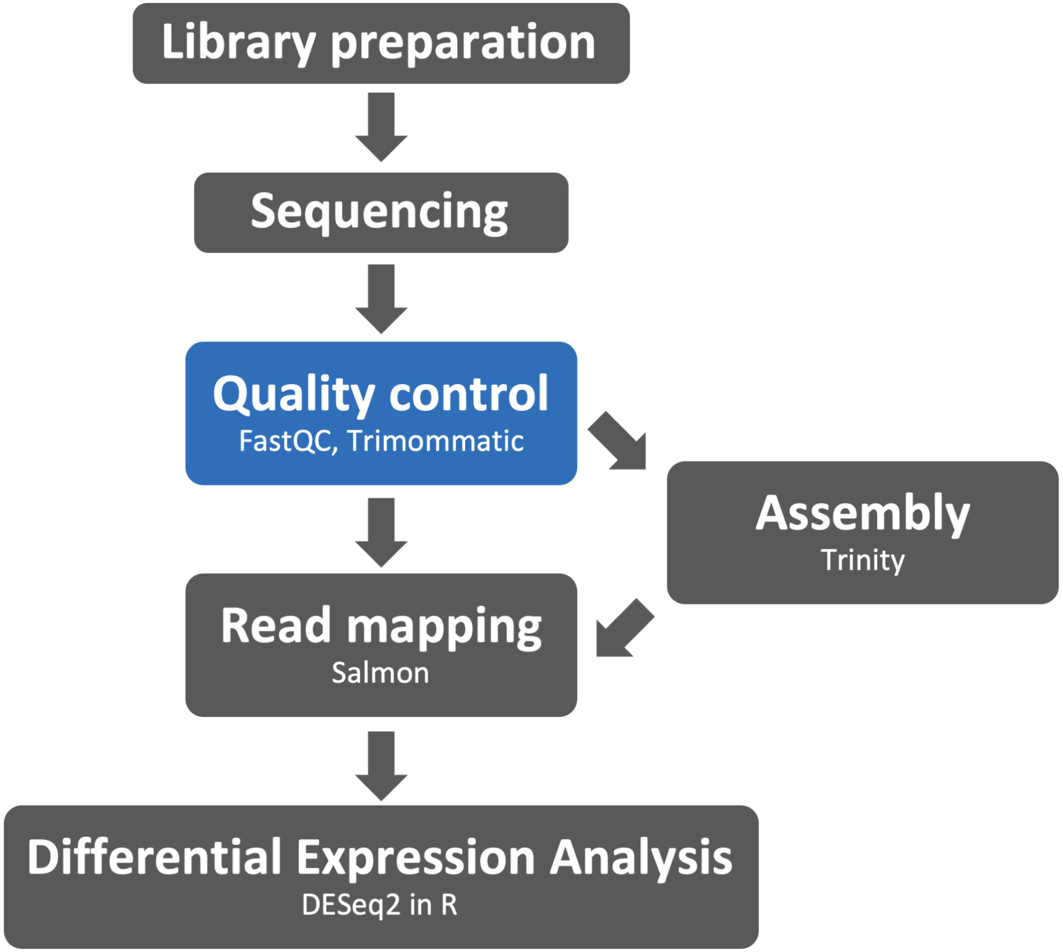
Compare before and after reports





RNA-seq workflow

Quality Control



Library preparation

Sequencing

Quality control
FastQC, Trimommatic

Assembly
Trinity

Read mapping
Salmon

Differential Expression Analysis
DESeq2 in R