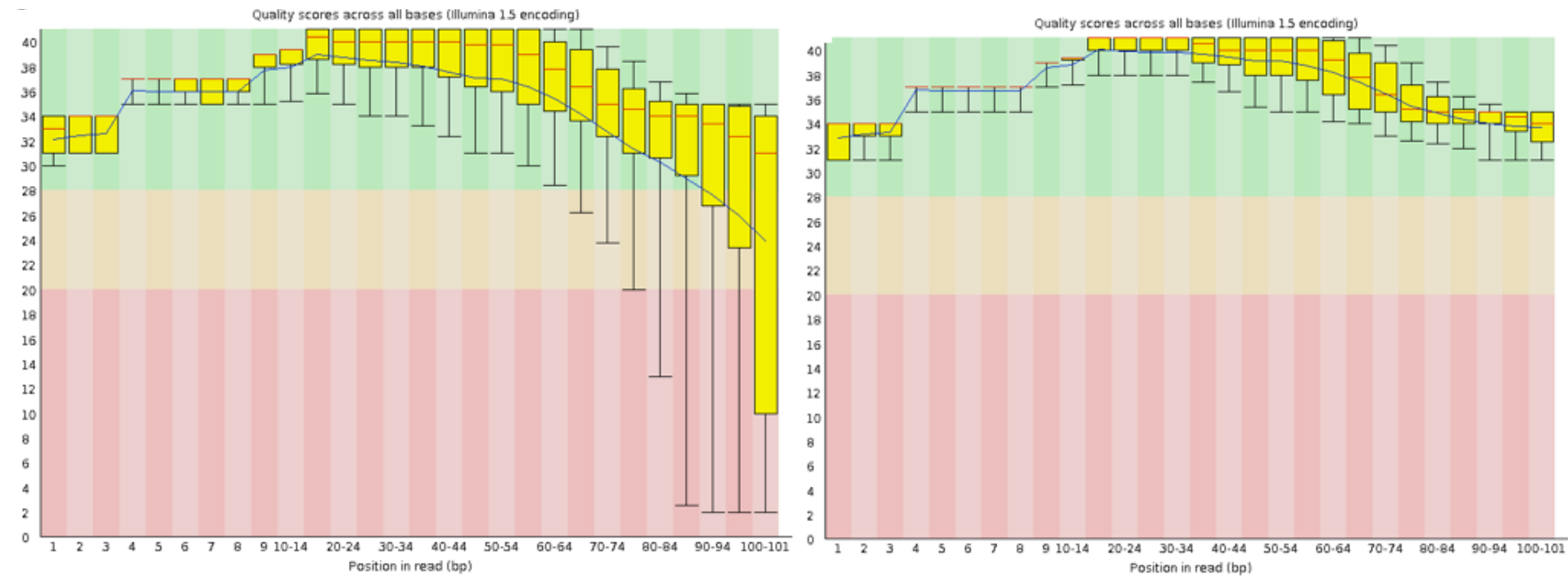


Swiss Institute of
Bioinformatics

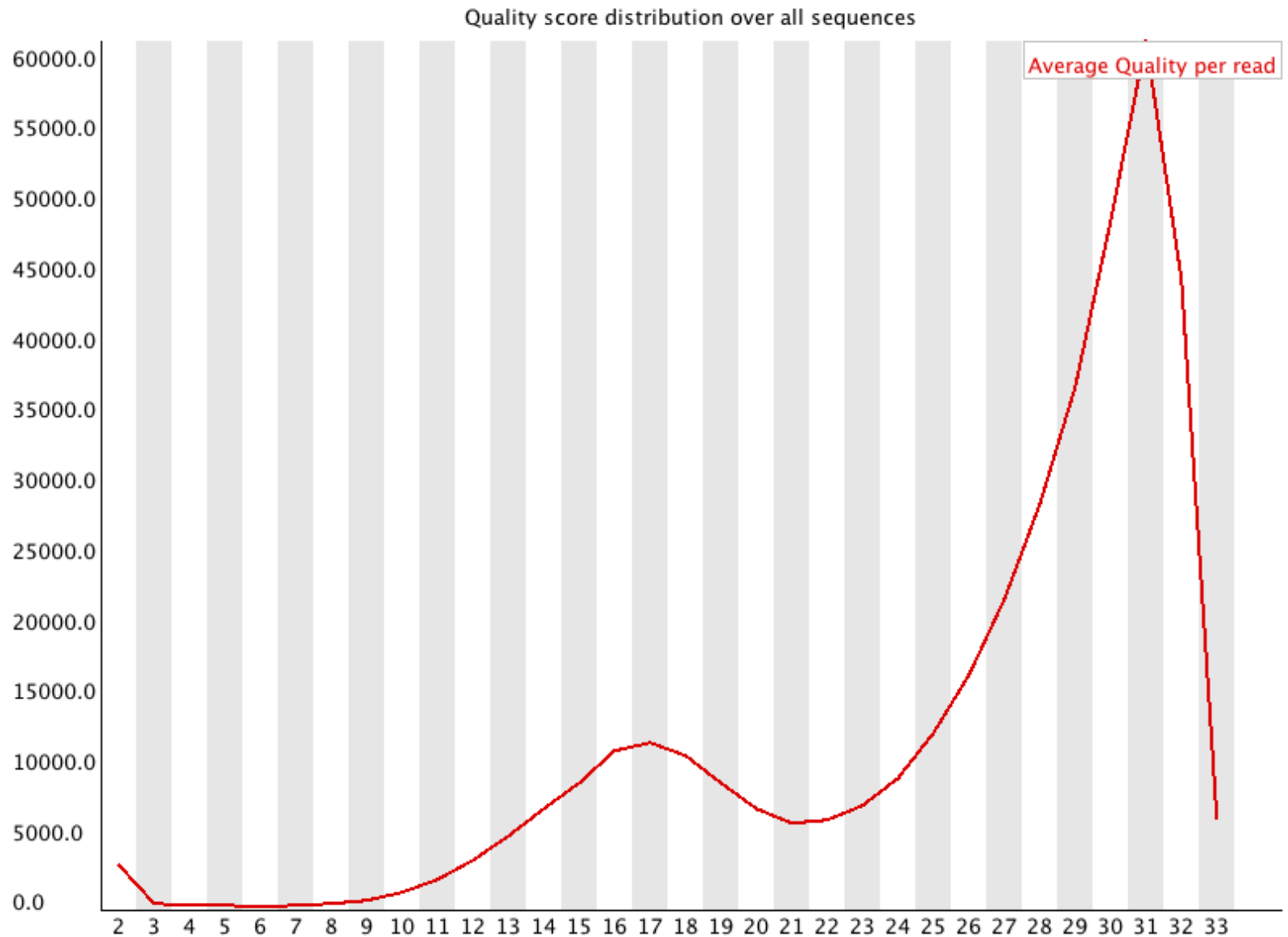
Introduction to RNA-Seq – Sequence trimming

Wandrille Duchemin

Per Base Sequence Quality



Per Sequence Quality Scores



Overrepresented Sequences

FastQC Report

Mon 24 Feb 2014
SRR576933.fastq

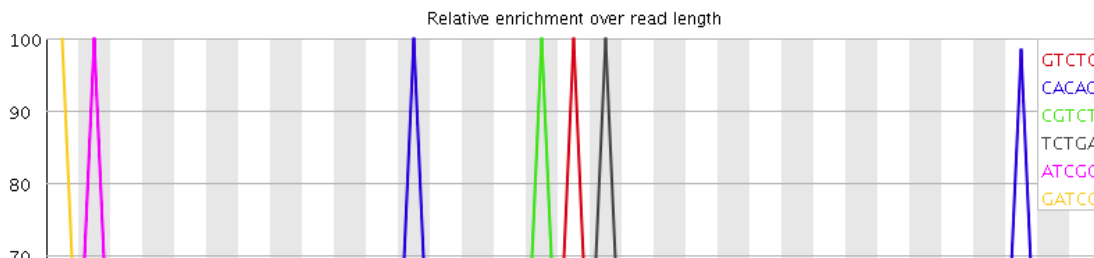
Summary

- ✓ [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](#)

✗ Overrepresented sequences

Sequence	Count	Percentage	Possible Source
GATCGGAAGAGCACACGTCTGAACTCCAGTCACACA	1060621	29.432719567181643	TruSeq Adapter, Index 5 (100% over 36bp)
GCTAACAAATACCCGACTAAATCAGTCAAGTAAATA	13630	0.37823875606902535	No Hit
NATCGGAAGAGCACACGTCTGAACTCCAGTCACACA	11728	0.3254573830651159	TruSeq Adapter, Index 5 (97% over 36bp)
GTTAGCTATTTACTTGACTGATTTAGTCGGGTATTT	10983	0.304783291115635	No Hit
GATCGGAAGAGCACACGTCTGAACTCCAGTCACACC	3658	0.10151117899490057	TruSeq Adapter, Index 1 (97% over 36bp)

✗ Kmer Content



Trimming – doing it or not?

Depends on your use case

Trimming – doing it or not?



Raw reads

trimming/
filtering

Trimmed/filtered
reads

assembly

Assembled
transcriptome

expression
quantification

variant
calling

isoform
description

...

Transcriptome assembly:

DO trimming

Trimming – doing it or not?



Sequencing
facility

Raw reads

trimming/
filtering

Trimmed/filtered
reads

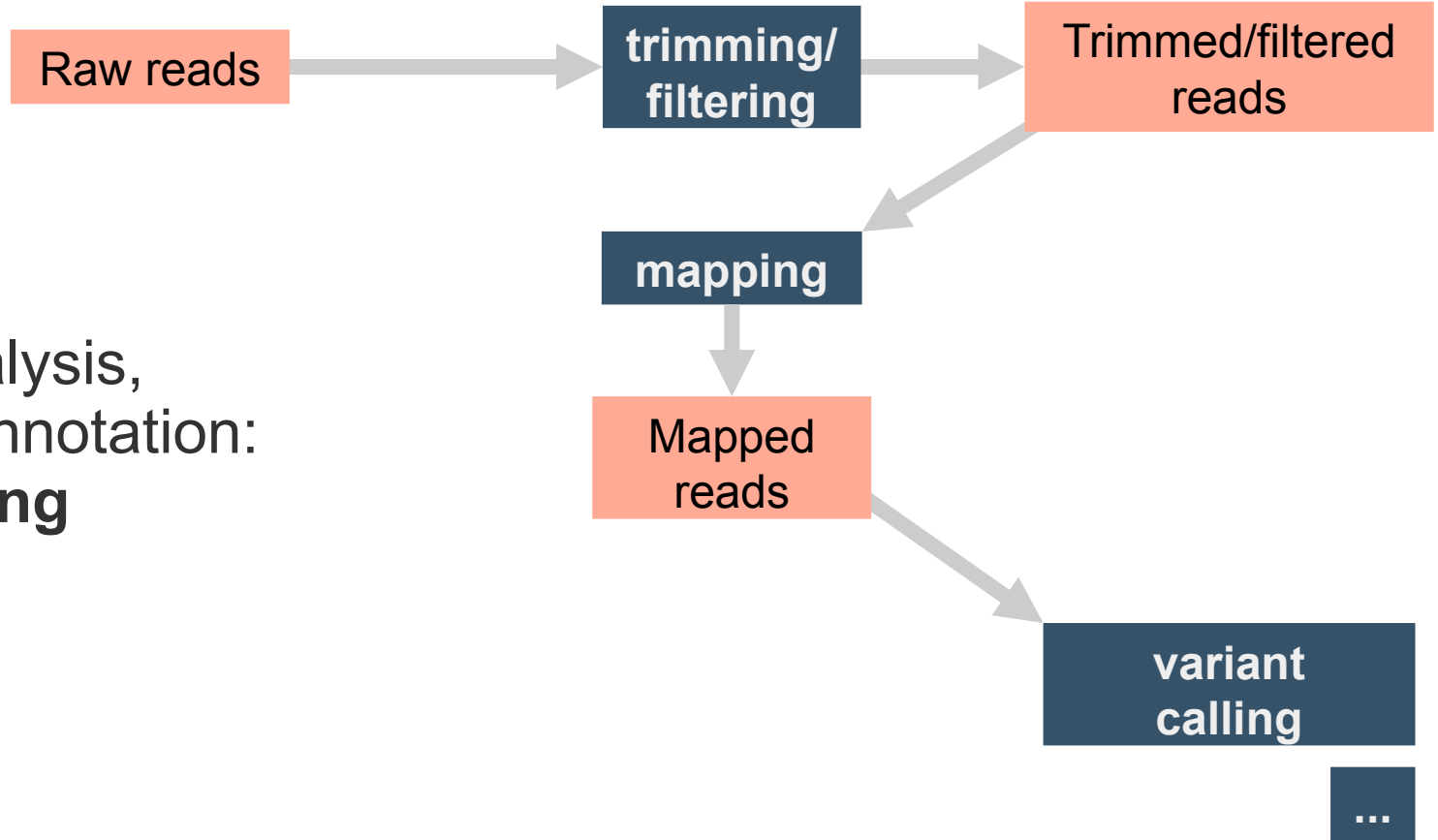
mapping

Mapped
reads

variant
calling

...

Variant analysis,
genome annotation:
DO trimming



Trimming – doing it or not?



Sequencing
facility

Raw reads

trimming/
filtering

Trimmed/filtered
reads

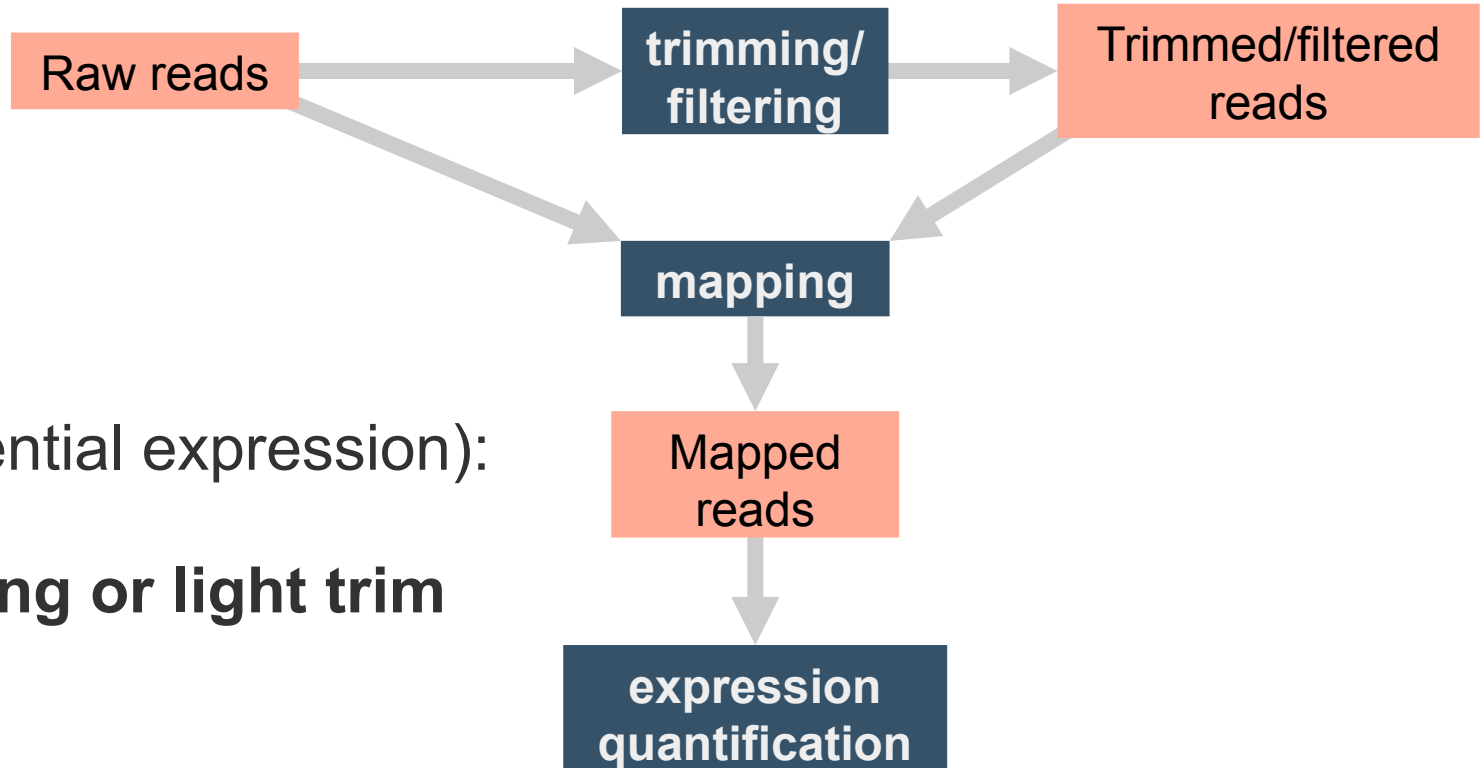
mapping

Mapped
reads

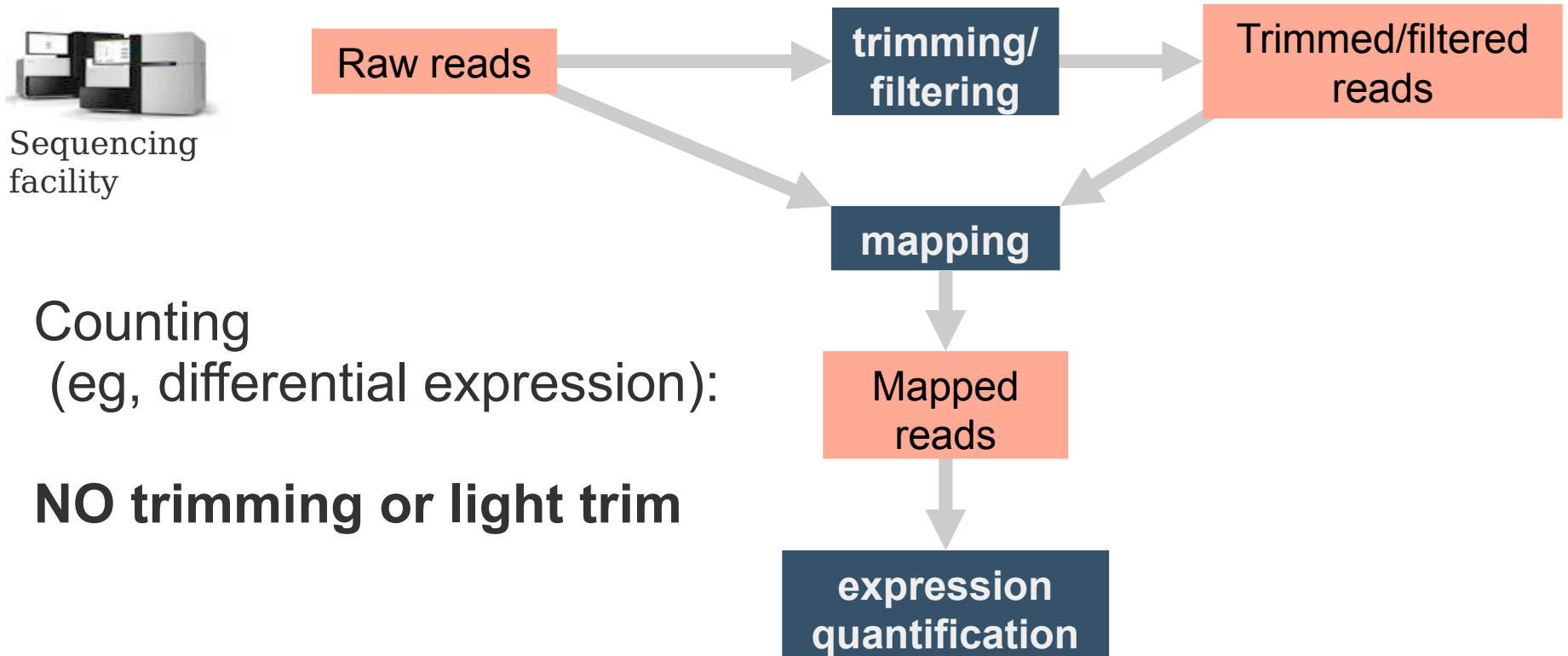
expression
quantification

Counting
(eg, differential expression):

NO trimming or light trim



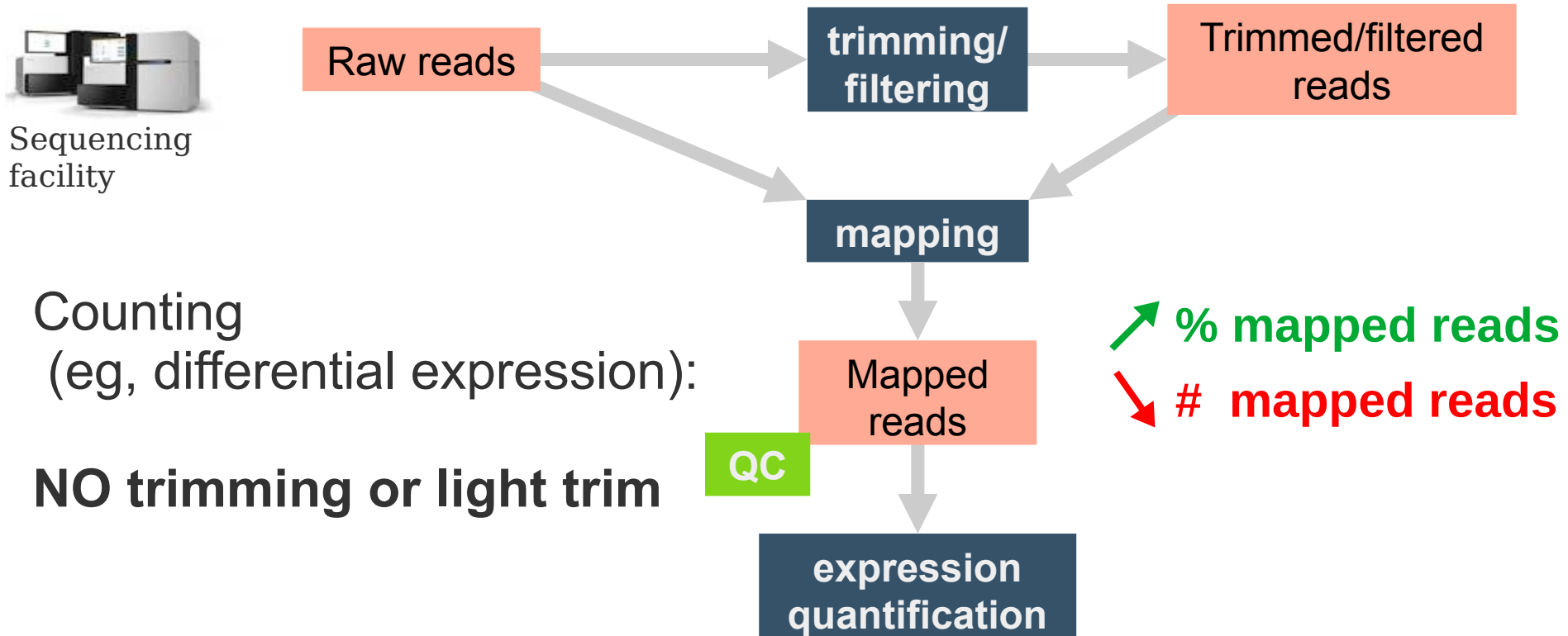
Trimming – doing it or not?



Aggressive trimming (high quality threshold, low length filtering) can have negative impacts on expression quantification

see : William et al. (2016) Trimming of sequence reads alters RNA-Seq gene expression estimates. BMC bioinfo

Trimming – doing it or not?



Aggressive trimming (high quality threshold, low length filtering) can have negative impacts on expression quantification

see : William et al. (2016) Trimming of sequence reads alters RNA-Seq gene expression estimates. BMC bioinfo

Trimmomatic – main options

ILLUMINACLIP: Cut adapter and other illumina-specific sequences from the read.

SLIDINGWINDOW: Performs a sliding window trimming approach. It starts scanning at the 5' end and clips the read once the average quality within the window falls below a threshold.

MAXINFO: An adaptive quality trimmer which balances read length and error rate to maximise the value of each read

LEADING: Cut bases off the start of a read, if below a threshold quality

TRAILING: Cut bases off the end of a read, if below a threshold quality

CROP: Cut the read to a specified length by removing bases from the end

HEADCROP: Cut the specified number of bases from the start of the read

MINLEN: Drop the read if it is below a specified length

AVGQUAL: Drop the read if the average quality is below the specified level

Trimmomatic – main options

ILLUMINACLIP: Cut adapter and other illumina-specific sequences from the read.

SLIDINGWINDOW: Performs a sliding window trimming approach. It starts scanning at the 5" end and clips the read once the average quality within the window falls below a threshold.

LEADING: Cut bases off the start of a read, if below a threshold quality

TRAILING: Cut bases off the end of a read, if below a threshold quality

CROP: Cut the read to a specified length by removing bases from the end

HEADCROP: Cut the specified number of bases from the start of the read

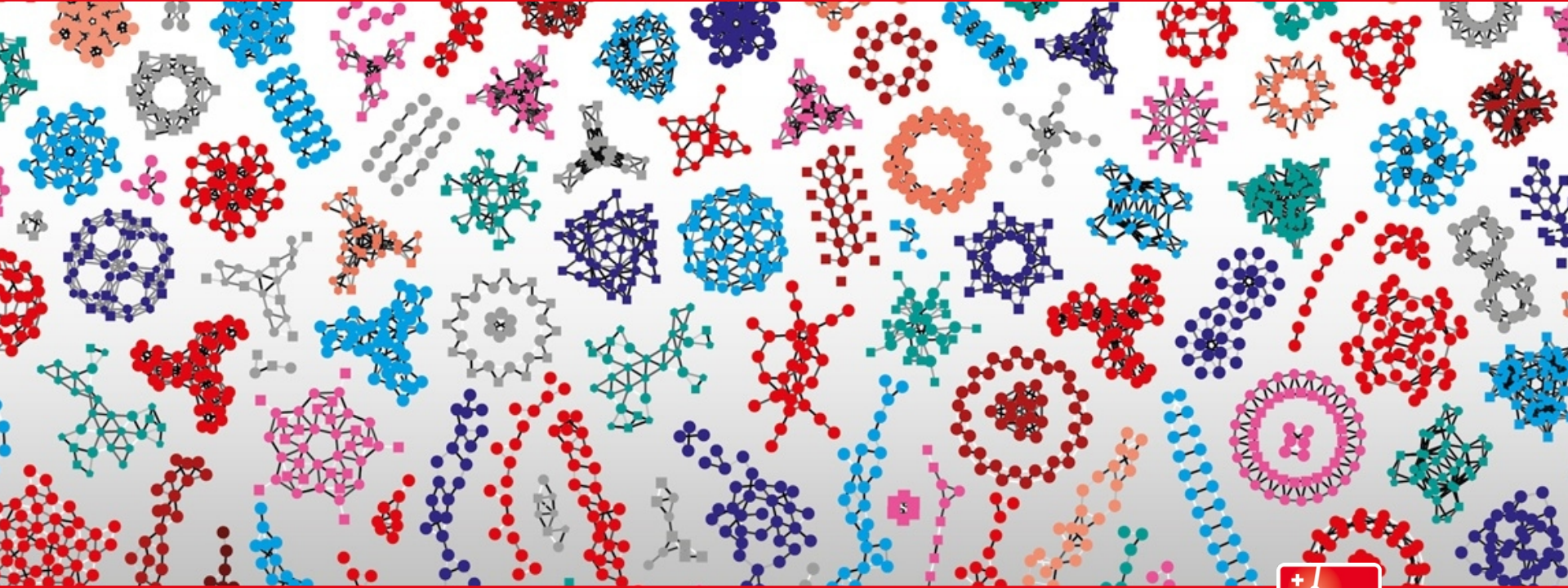
MINLEN: Drop the read if it is below a specified length

AVGQUAL: Drop the read if the average quality is below the specified level

<http://www.usadellab.org/cms/?page=trimmomatic>

Practical

- **Go to the website and do the trimming practical**



Swiss Institute of
Bioinformatics

Contributors:

Wandrille Duchemin

Geoffrey Fucile

Walid Gharib

Pablo Escobar Lopez

Mihaela Zavolan



www.sib.swiss