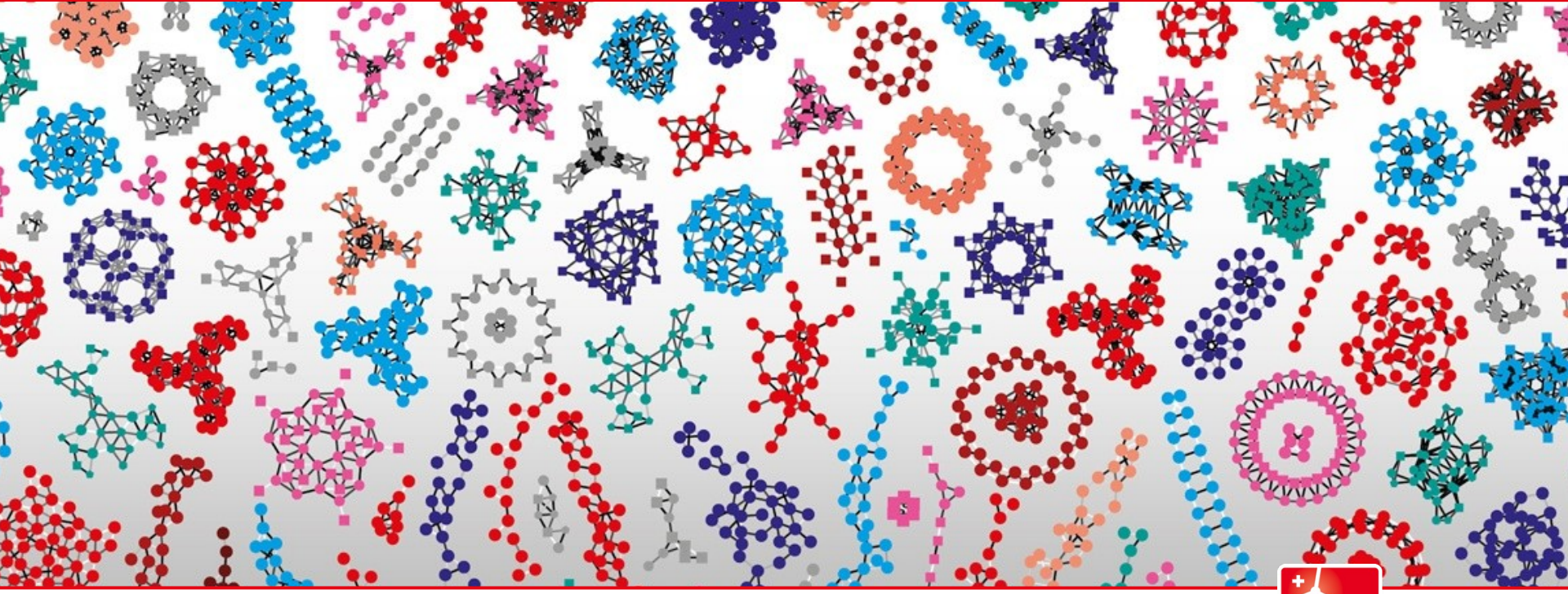
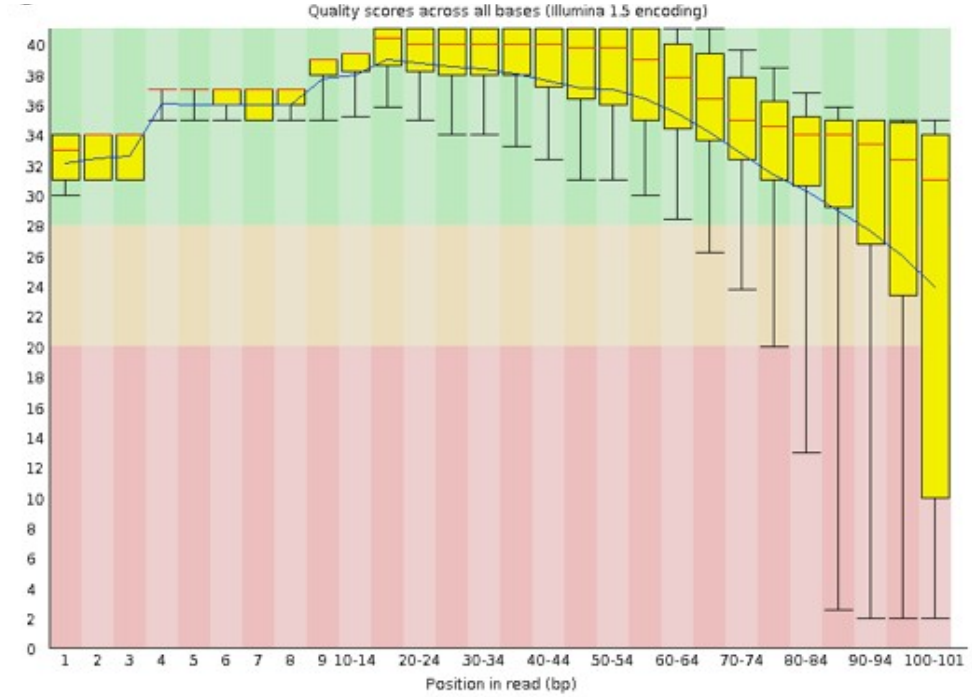


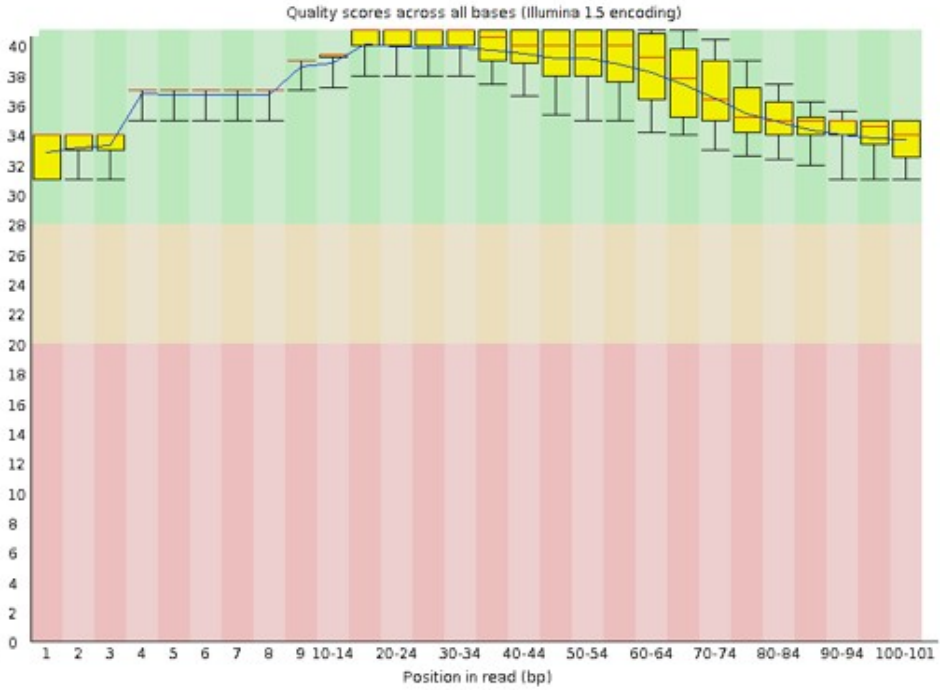
Swiss Institute of Bioinformatics

Introduction to RNA-Seq – Sequence trimming

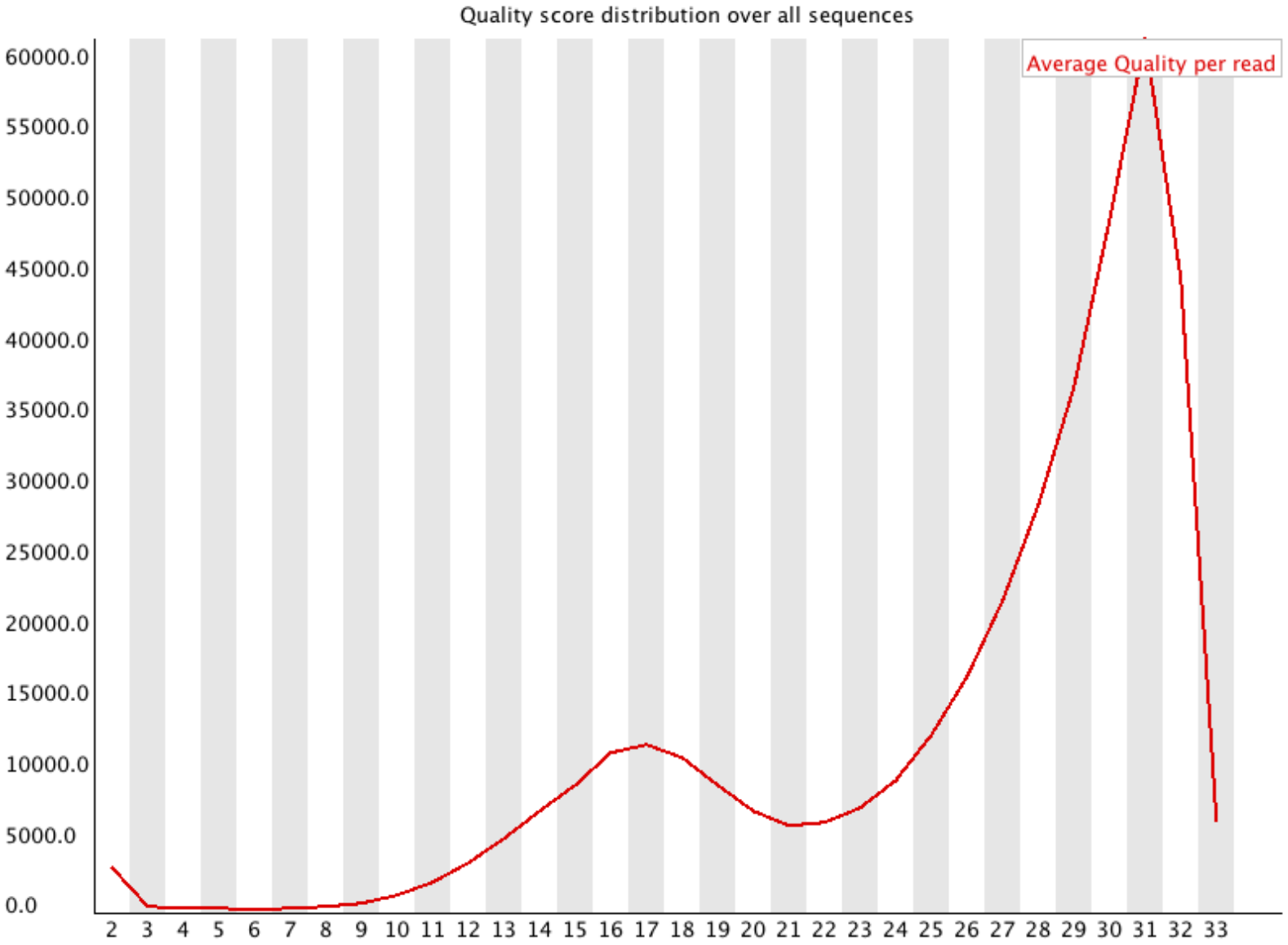
Wandrille Duchemin

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

Produced by FastQC (version 0.10.1)

Trimming – Modifying with Trimmomatic

- **Some dataset require modification prior to alignment, *eg.* trimming low quality base calls**
- **The decision to modify the dataset depends on the nature of the dataset and the question(s) asked**
  - Transcriptome assembly, variant analysis, genome annotation → **DO TRIMMING**
  - Counting (eg. differential expression) : most software can handle no trimming but still recommend some light trim. (continued)

Trimming – Modifying with Trimmomatic

- **Some dataset require modification prior to alignment, *eg.* trimming low quality base calls**
- **The decision to modify the dataset depends on the nature of the dataset and the question(s) asked**
  - Trimming increases the rate of mapped reads BUT diminish the absolute numbers of reads
  - 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 – Clipping Adapter Sequences

- **Generally, your sequencing facility will send you FASTQ files where all sequencing adapter sequences have been removed**
- **In practice, there can sometimes be unclipped adapter sequences. Why ?**
- **Trimmomatic option:**

ILLUMINACLIP:<fastaWithAdaptersEtc>:<seed mismatches>:<palindrome clip threshold>:<simple clip threshold>

ILLUMINACLIP:fastafilePATH:2:30:10

Trimmomatic 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

TOPHRED33: Convert quality scores to Phred-33

TOPHRED64: Convert quality scores to Phred-64

Trimmomatic

- ***eg* Trim a paired-end dataset**

```
trimmomatic PE \
reads_1.fq reads_2.fq \
./trimmed_data/paired_trimmed_1.fq. /trimmed_data/unpaired_trimmed_1.fq \
./trimmed_data/paired_trimmed_2.fq. /trimmed_data/unpaired_trimmed_2.fq \
SLIDINGWINDOW:4:10
```

Read the manual for more information. This is good practice, as software is constantly evolving and you can't always rely on recipes provided by courses ;-)

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

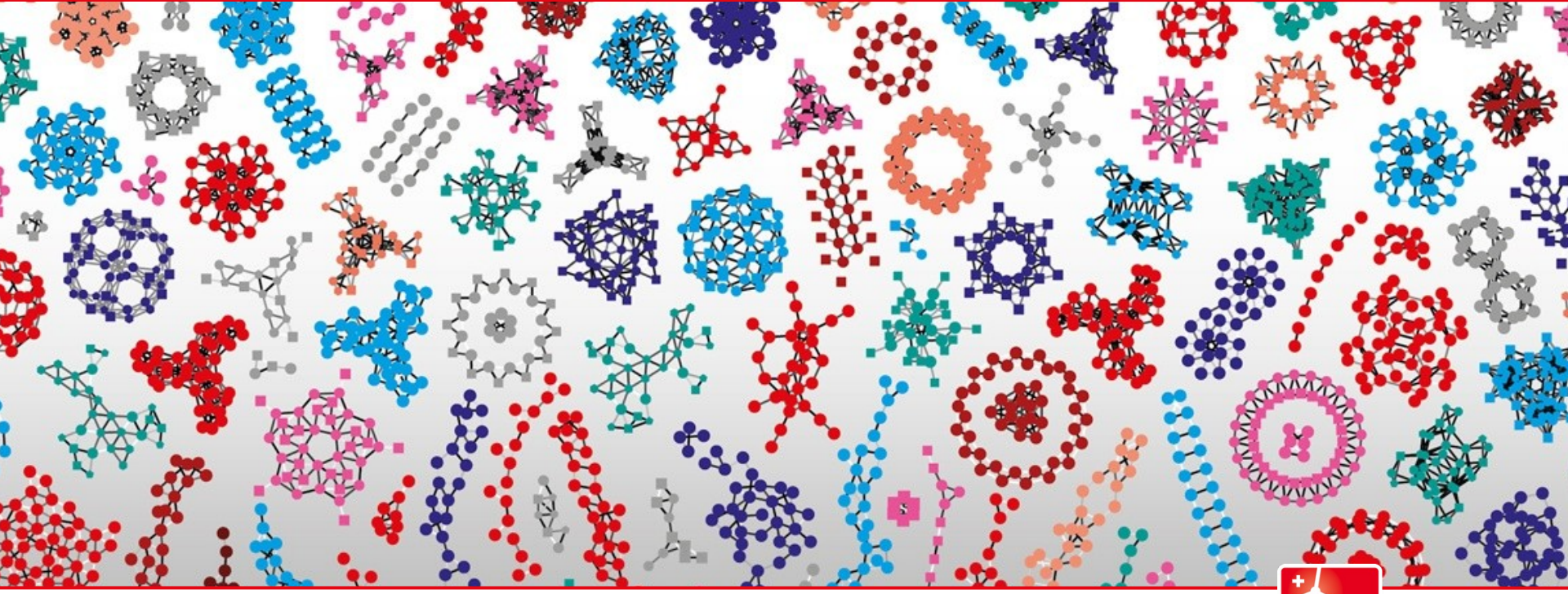
REFERENCES

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

Bolger *et al* (2014) “Trimmomatic: A flexible trimmer for Illumina sequence data” Bioinformatics 30(15): 2114-2120.

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