DAY 2

File formats and data QC

24.11.2015

Pille Hallast pille.hallast@ut.ee

UNIVERSITY OF TARTU
1632

Overview

- Common NGS data file formats
  - Fasta, fastq, SAM/BAM, VCF, GTF/GFF
- Terminology
  - Single end, paired end, mate pair
- NGS-Seq data Quality Control
  - FastQC, Trimmomatic

File Formats

- Fasta (sequences)
- Fastq (read data)
- SAM/BAM (short read alignment)
- VCF (variant information)
- GFF/GTF (annotation data)

Fasta

- Text file for nucleotide or peptide sequences
  - Line 1 begins with a '>' character, followed by an optional sequence identifier
  - Line 2 is the raw sequence letters

```
>OVAX_CHICK - GENE X PROTEIN (OVALBUMIN-RELATED)
MKDLLVSSSTDLDTTLVLVNAIYFKGMWKTAFNAEDTREMPFHVTKQESKPVQMMCMNNSFN
MKILELPFASGDLSMLVLLPDEVSDLERIEKTINFEKLTEWTNPNTMEKRRVKVYLPQMKIE
LMALGMTDLFIPSANLTGISSAESLKISQAVHGAFMELSEDGIEMAGSTGVIEDIKHSPESE
LFLIKHNPTNTIVYFGRYWSP
```

  - Typical file extensions (.fasta, .fa, .fna, .ffn, .faa)

Fastq

- Text file for read sequences & quality scores
  - Line 1 = '@' character, followed by information about the sequencing run (also a sequence identifier)
  - Line 2 = raw sequence letters
  - Line 3 = '+' character and is *optionally* followed by the same run info
  - Line 4 = encodes the quality values for the sequence in Line 2, and must contain the same number of symbols as letters in the sequence

```
@HS21_07614:7:1202:2173:113319#20/1
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*(((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

Phred Quality Score

- Indicates the probability that a given base is called correctly by the sequencer
- Derived from Sanger sequencing
- Parameters relevant to a particular sequencing chemistry are analyzed for a large empirical data set of known accuracy
- The resulting quality score lookup tables are used to calculate a quality score for de novo next generation sequencing data
- Q = -10 log 10 P
  - Q10 = incorrect base 1 in 10 (90% accuracy)
  - Q20 = incorrect base 1 in 100 (99% accuracy)
  - Q30 = incorrect base 1 in 1000 (99.9% accuracy)
  - Q40 = incorrect base 1 in 10000 (99.99% accuracy)
- Low Q scores can lead to increased false positive variant calls, inaccurate conclusions & high experimental validation costs

Fastq

- Text file for read sequences & quality scores
  - Line 1 = '@' character, followed by a sequence identifier and an optional description.
  - Line 2 = raw sequence letters
  - Line 3 = '+' character and is *optionally* followed by the same seq identifier
  - Line 4 = encodes the quality values for the sequence in Line 2, and must contain the same number of symbols as letters in the sequence

```
@HS21_07614:7:1202:2173:113319#20/1
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*(((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

SAM/BAM

- Sequence Alignment/Map format (tab delim text)
  - Header section (optional)

```
@HD VN:1.3 SO:coordinate
@SQ SN:ref LN:45
```

  - Alignment section (11 mandatory fields + optional)

```
R001 163 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002   0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA *
r003   0 ref  9 30 5H6M       *  0   0 AGCTAA * NM:i:1
r004   0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC *
r003  16 ref 29 30 6H5M       *  0   0 TAGGC * NM:i:0
r001  83 ref 37 30 9M         =  7 -39 CAGCGCCAT *
```

- BAM = binary SAM

Variant Call Format (VCF)

- Text file – SNP, ins/del, CNV
  - Meta information

```
##fileformat=VCFv4.1
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,species="Homo sapiens"
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of samples">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
```

Variant Call Format (VCF)

- Text file – SNP, ins/del, CNV
  - Header Line (8 fixed mandatory columns)
  - Data lines (corresponding to header columns)

```
#CHROM POS     ID          REF ALT QUAL FILTER INFO
20     14370   rs6054257   G   A   29   PASS   NS=3;DP=14;AF=0.5;DB;H2
20     17330   .           T   A   3    q10    NS=3;DP=11;AF=0.017
20     1110696 rs6040355   A   G,T 67   PASS   DP=10;AF=0.333,0.667;AA=T
20     1230237 .           T   .   47   PASS   NS=3;DP=13;AA=T
20     1234567 microsat1   GTC G   50   PASS   NS=3;DP=9;AA=G GT:GQ:DP
```

GFF/GTF

- General Feature Format annotation file (tab delim text)
  - http://www.ensembl.org/info/website/upload/gff.html
- 1 line per feature, 9 columns of data, plus optional track definition lines

```
X  Ensembl      Repeat      2419108  2419128   42      .   .   hid=trf; hstart=1; hend=21
X  Ensembl      Repeat      2419108  2419410   2502    -   .   hid=AluSx; hstart=1; hend=303
X  Ensembl      Repeat      2419108  2419128   0       .   .   hid=dust; hstart=2419108; hend=2419128
X  Ensembl      Pred.trans. 2416676  2418760   450.19  -   2   genscan=GENSCAN00000019335
X  Ensembl      Variation   2413425  2413425   .       +   .
X  Ensembl      Variation   2413805  2413805   .       +   .|
```

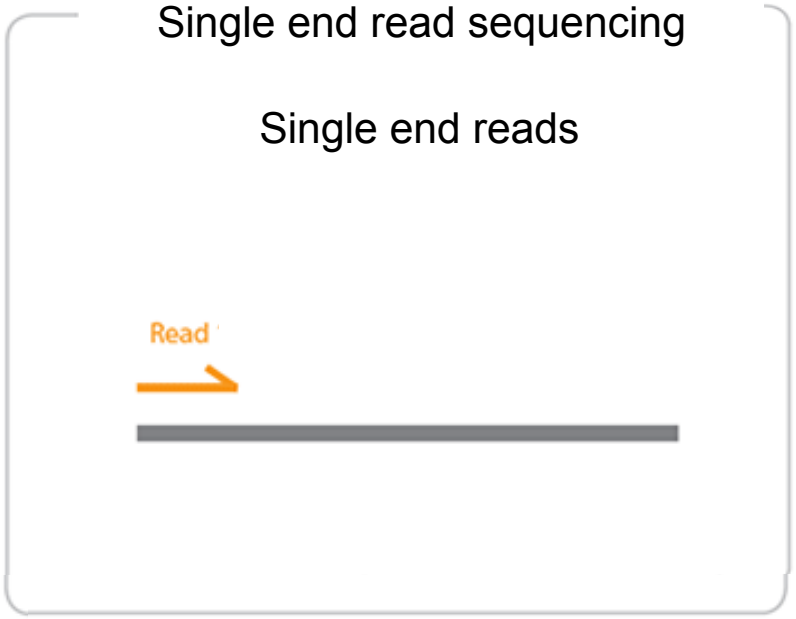
Single, Paired End & Mate Pair

- **Single-end**
  Each read is a single sequence from one end of a DNA fragment (single fastq file). The fragment is usually 200-800bp long, with the amount being read can be chosen between 50 and 300 bp

- **Paired-end**
  Each read is two sequences (a pair) from each end of the same DNA fragment. The distance between the reads on the original genome sequence is equal to the length of the DNA fragment that was sequenced, usually 200-800 bp

- **Mate-pair**
  Each read is two sequences from each end of the same DNA fragment, but the distance between the reads on the original genome sequence is much longer, e.g. 3000-10000 bp

Single, Paired End & Mate Pair



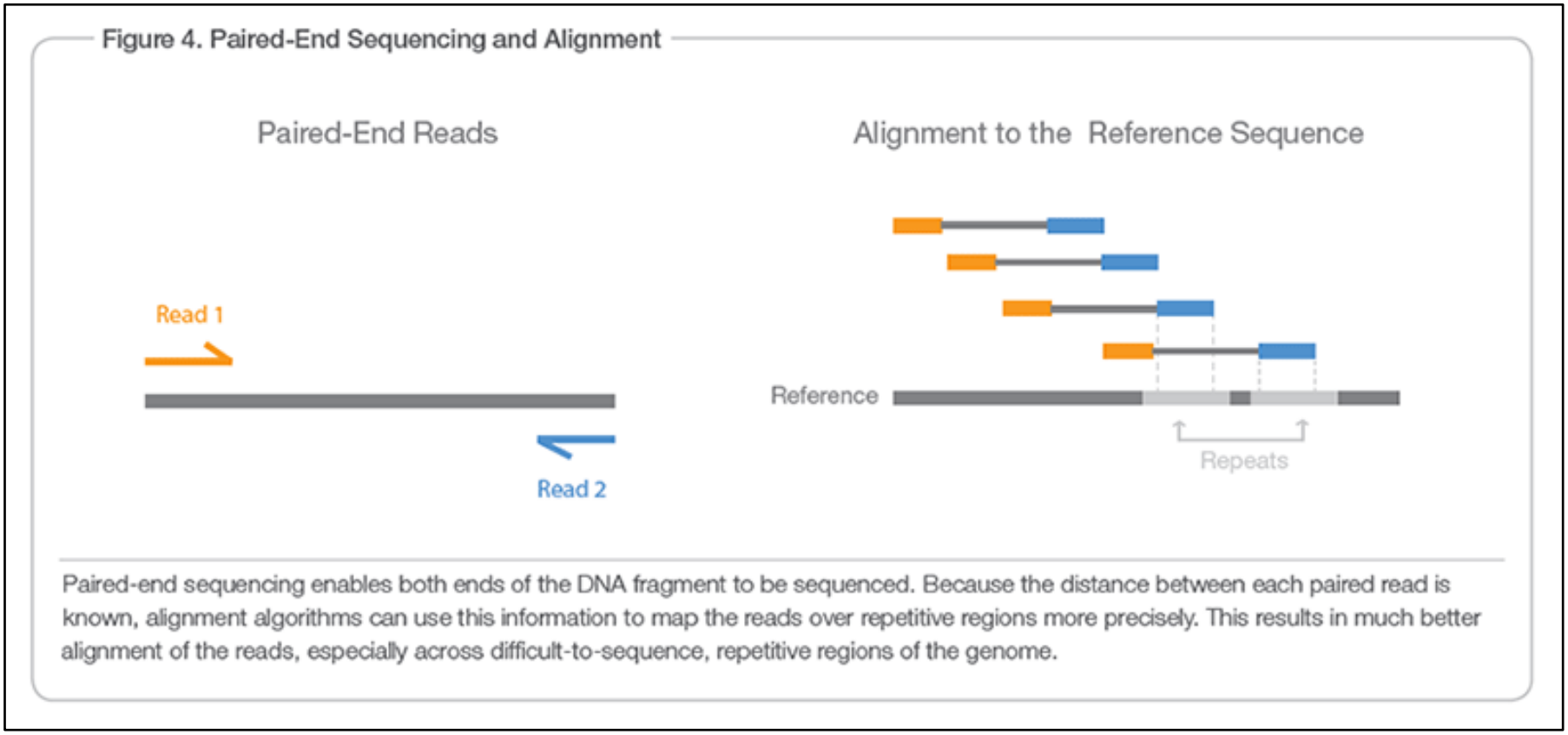
Single, Paired End & Mate Pair

- **Single-end**
  Each read is a single sequence from one end of a DNA fragment (single fastq file). The fragment is usually 200-800bp long, with the amount being read can be chosen between 50 and 300 bp

- **Paired-end**
  Each read is two sequences (a pair) from each end of the same DNA fragment (2 fastq files). The distance between the reads on the original genome sequence is equal to the length of the DNA fragment that was sequenced, usually 200-800 bp

- **Mate-pair**
  Each read is two sequences from each end of the same DNA fragment (2 fastq files), but the distance between the reads on the original genome sequence is much longer, e.g. 3000-10000 bp

Single, Paired End & Mate Pair



Single, Paired End & Mate Pair

- **Single-end**
  Each read is a single sequence from one end of a DNA fragment (single fastq file). The fragment is usually 200-800bp long, with the amount being read can be chosen between 50 and 300 bp

- **Paired-end**
  Each read is two sequences (a pair) from each end of the same DNA fragment (2 fastq files). The distance between the reads on the original genome sequence is equal to the length of the DNA fragment that was sequenced, usually 200-800 bp (2 fastq files)

- **Mate-pair**
  Each read is two sequences from each end of the same DNA fragment (2 fastq files), but the distance between the reads on the original genome sequence is much longer, e.g. 3000-10000 bp

Single, Paired End & Mate Pair

Figure 5. *De Novo* Assembly with Mate Pairs

Using a combination of short and long insert sizes with paired-end sequencing results in maximal coverage of the genome for de novo assembly. Because larger inserts can pair reads across greater distances, they provide a better ability to read through highly repetitive sequences and regions where large structural rearrangements have occurred. Shorter inserts sequenced at higher depths can fill in gaps missed by larger inserts sequenced at lower depths. Thus a diverse library of short and long inserts results in better de novo assembly, leading to fewer gaps, larger contigs, and greater accuracy of the final consensus sequence.

Read Quality Control/Pre-Processing

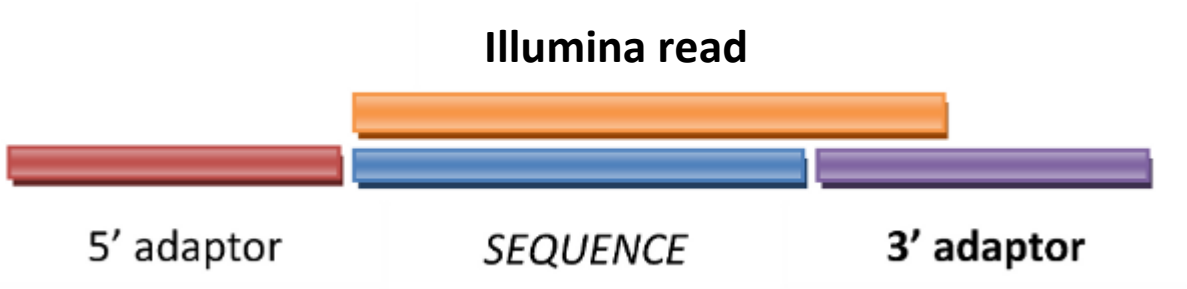
- Assess data quality, remove/trim poor quality reads and adapter sequences
- Improves quality of downstream analyses (particularly de-novo assembly)
- Reduces CPU and storage overhead
- FastQC, Trimmomatic, Fastx Toolkit & Cutadapt
  - http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
  - http://hannonlab.cshl.edu/fastx_toolkit/
  - http://code.google.com/p/cutadapt/

Read Quality Control/Pre-Processing

- Assess data quality
  - Trim adapter sequences
  - Remove/trim poor quality reads
  - Sequencing error correction
  - PCR De-duplication
  - Digital normalization

Read Quality Control/Pre-Processing

- Assess data quality
  - Trim adapter sequences

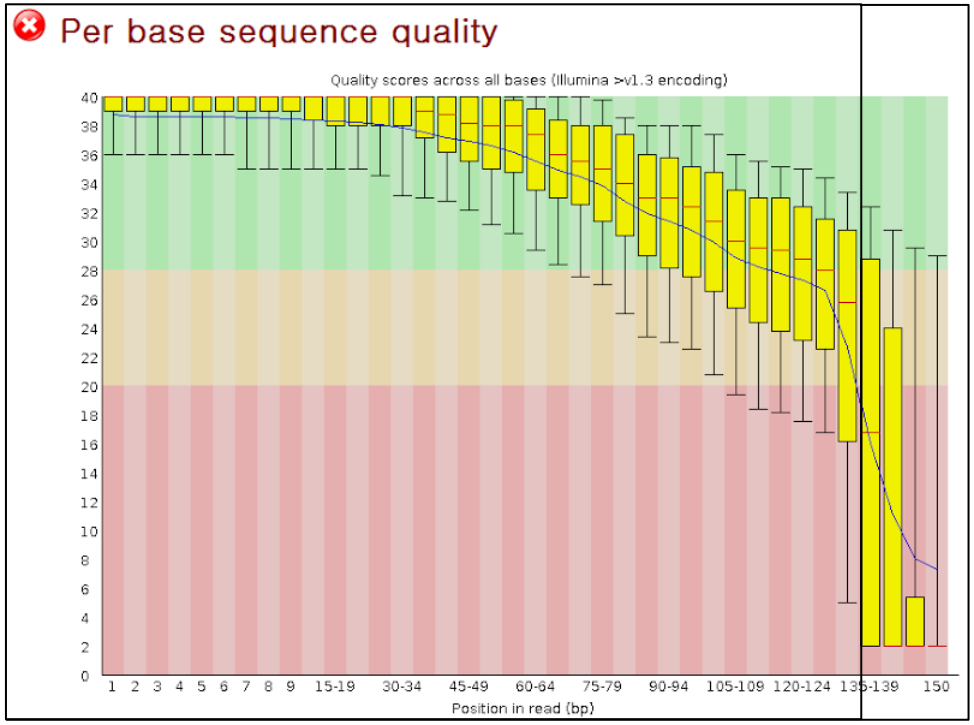


AGATCGGAAGAGCACAAACGATCTCGTATGCCGTCCTTG**ATCTCGTATGCCGTCTTCTGCTT G**

AGATCGGAAGAGCACAAACGATCTCGTATGCCGTCCTTG

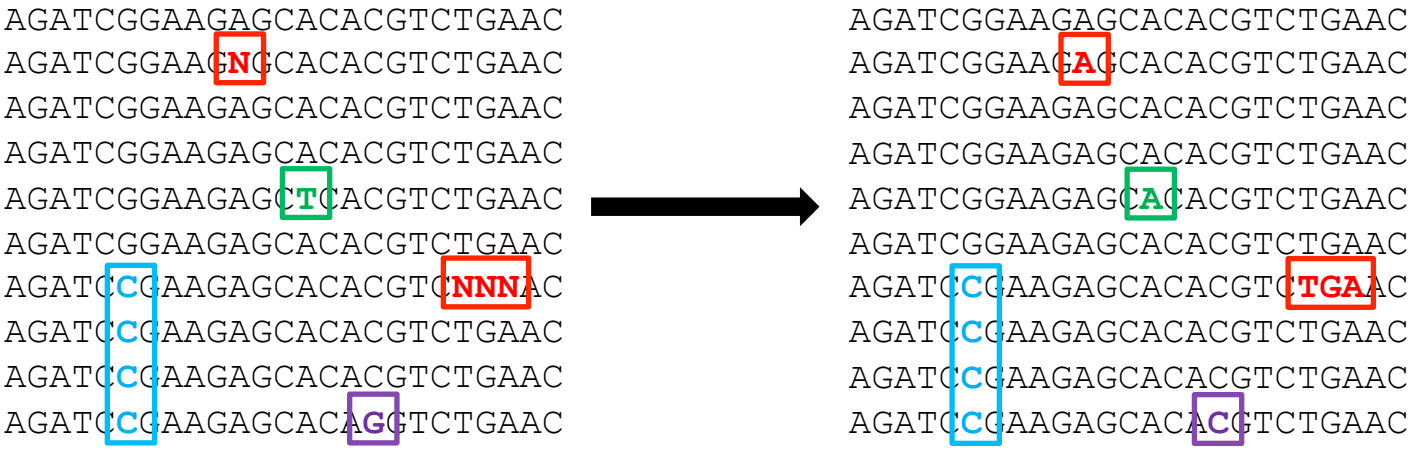
Read Quality Control/Pre-Processing

- Assess data quality
  - Remove/trim poor quality reads
  - Quality of data begins to fade towards 3’ end



Read Quality Control/Pre-Processing

- Assess data quality
  - Sequencing error correction
  - Illumina data predominantly substitution errors
  - Split reads in to kmers (e.g. 25bp)

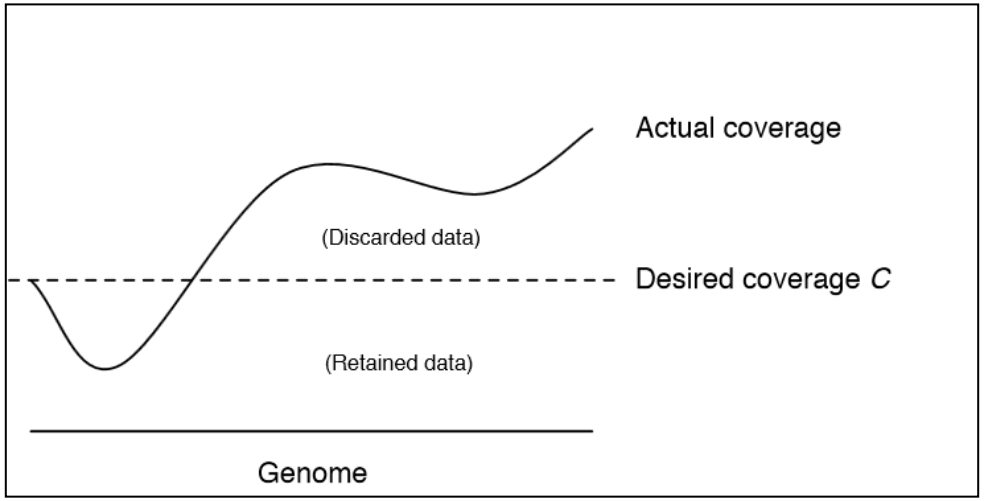


Read Quality Control/Pre-Processing

- Assess data quality
  - PCR De-duplication
  - Step in library prep may produce many identical reads
  - Problematic for DNA-Seq
  - Affects RAM and CPU requirements
  - In RNA-Seq could also be highly expressed transcript

Read Quality Control/Pre-Processing

- Assess data quality
  - Digital normalization (Brown, *et, al.* 2012)
  - Removes redundant (high coverage) reads
  - Normalizes average coverage
  - Reduces computational overhead

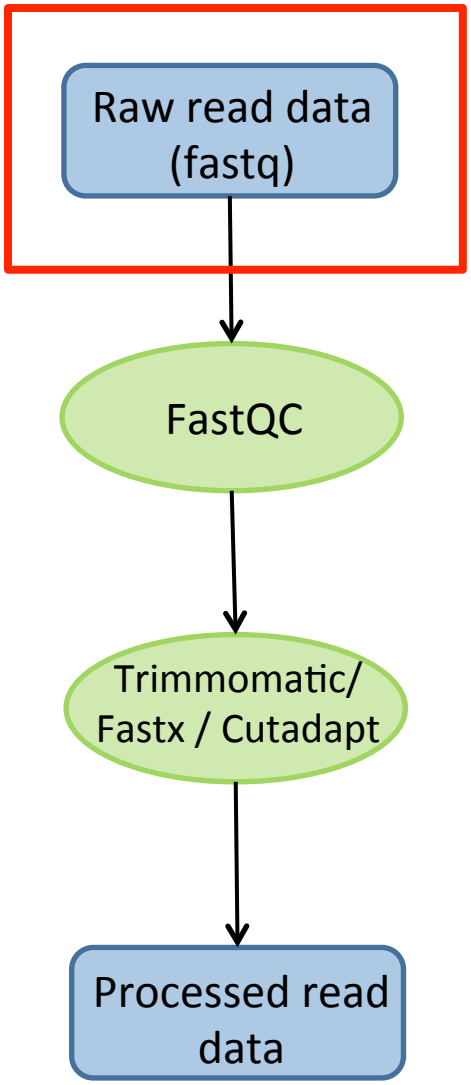


Read Quality Control/Pre-Processing

- Quality Assessment
  - FastQC
- Trim adapter sequences
  - Trimmomatic, Cutadapt
- Remove/trim poor quality reads
  - Trimmomatic, FastX Toolkit
- Sequencing error correction
  - Reptile, SOAPec
- PCR De-duplication
  - FastUniq, Picard
- Digital normalization
  - Diginorm

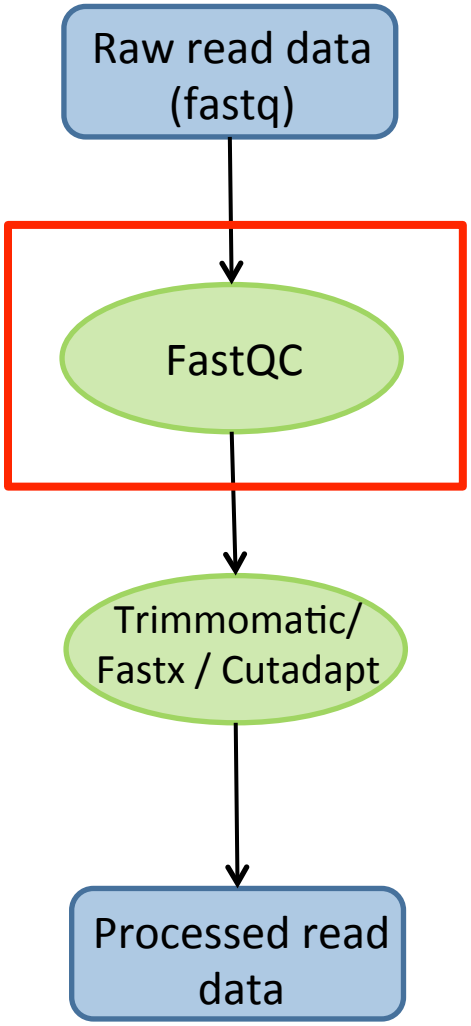
Read Quality Control/Pre-Processing

- QC Workflow



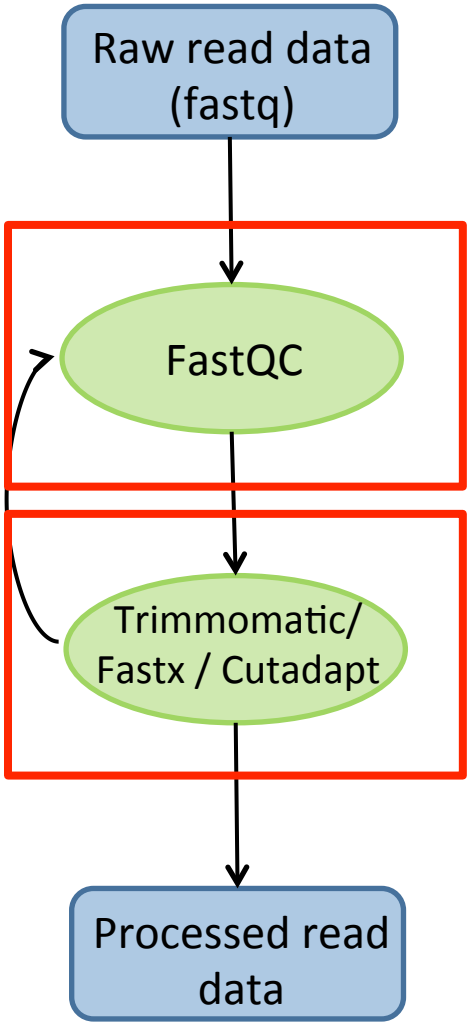
Read Quality Control/Pre-Processing

- QC Workflow
  - Number of reads
  - Quality per base (Phred score)
  - GC content, identify over-represented sequences



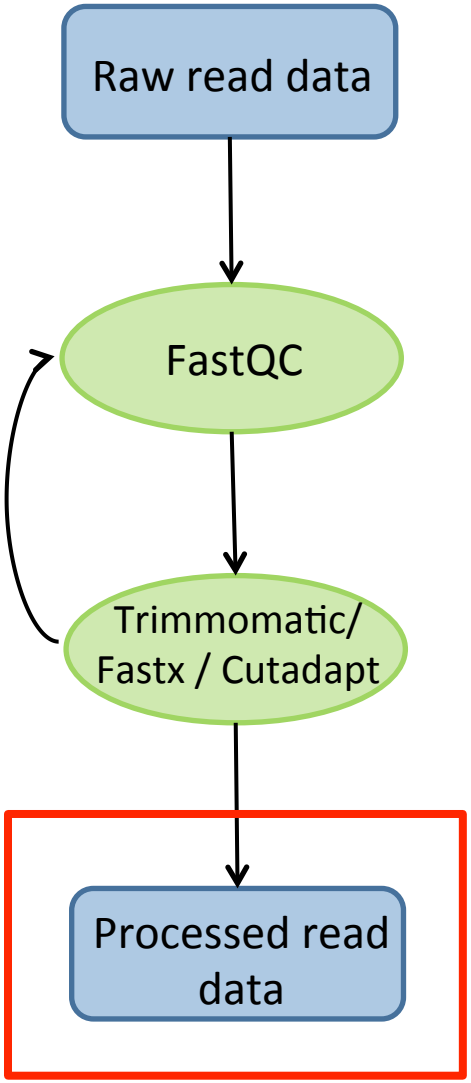
Read Quality Control/Pre-Processing

- QC Workflow
  - Number of reads
  - Quality per base (Phred score)
  - GC content, identify over-represented sequences
  - Remove or trim low quality reads and sequence adapters
  - Error correction
  - Repeat FastQC analysis to assess improvement



Read Quality Control/Pre-Processing

- QC Workflow
  - Number of reads
  - Quality per base (Phred score)
  - GC content, identify over-represented sequences
  - Remove or trim low quality reads and sequence adapters
  - Error correction
  - Repeat FastQC analysis to assess improvement
  - Remove singletons
  - Read data ready for analysis

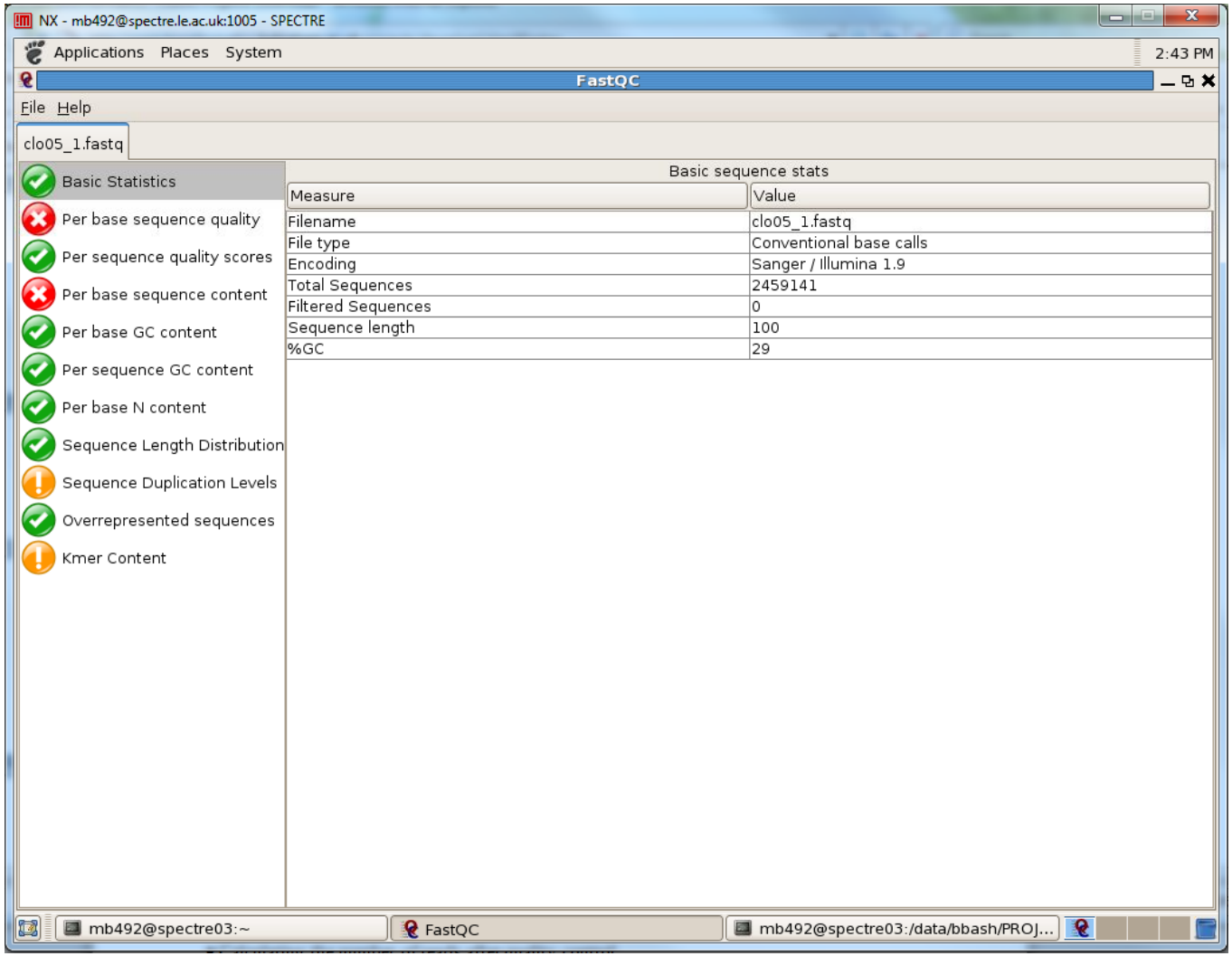


FastQC

- Load fastq file (accepts Casava, Colorspace and Gzip FastQ, SAM & BAM)
- Analysis modules

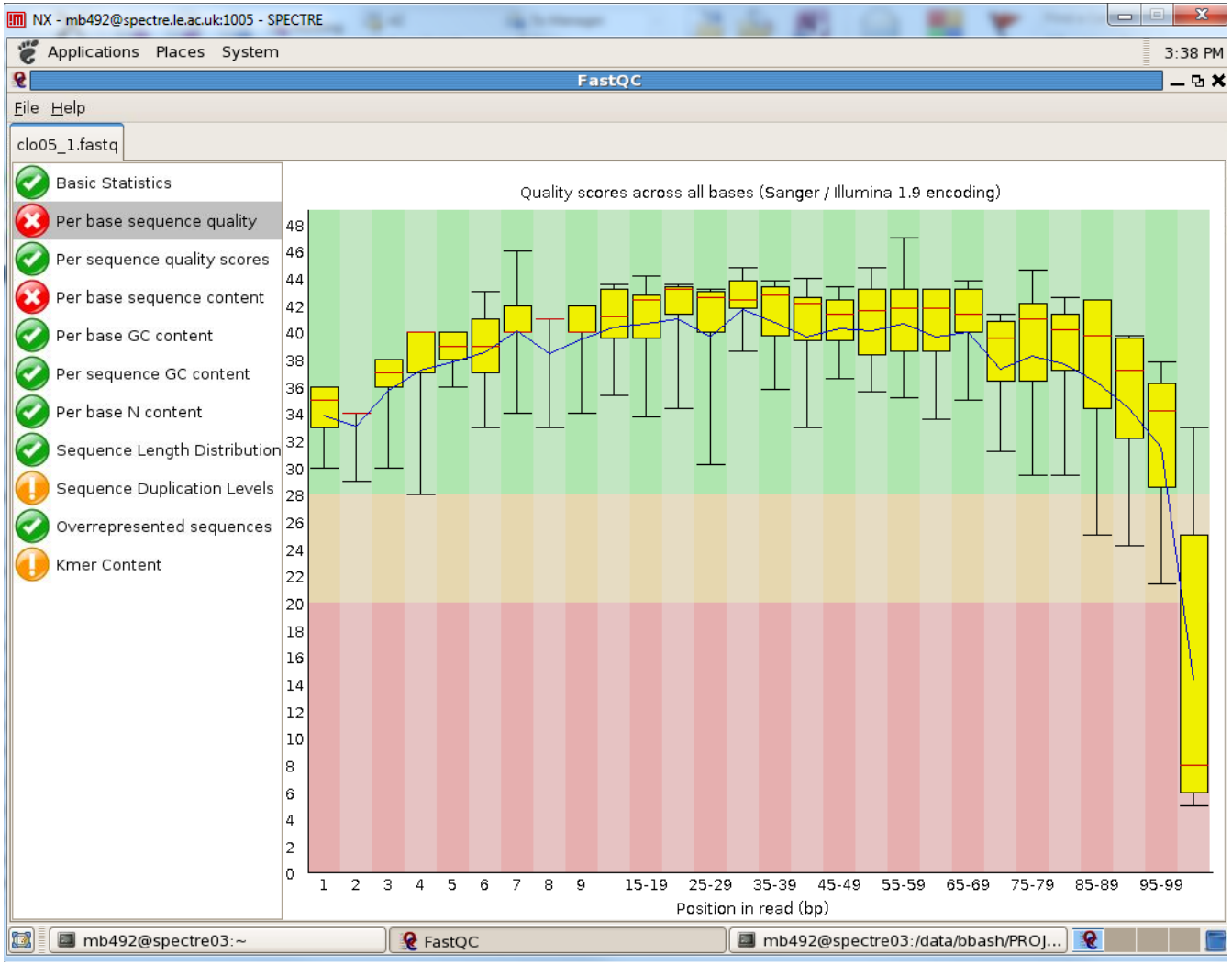
University of Leicester

FastQC Analysis Modules



University of Leicester

Per Base Sequence Quality



Yellow box:
inter-quartile range (25-75%)

Upper and lower whiskers:
10% and 90% points

median

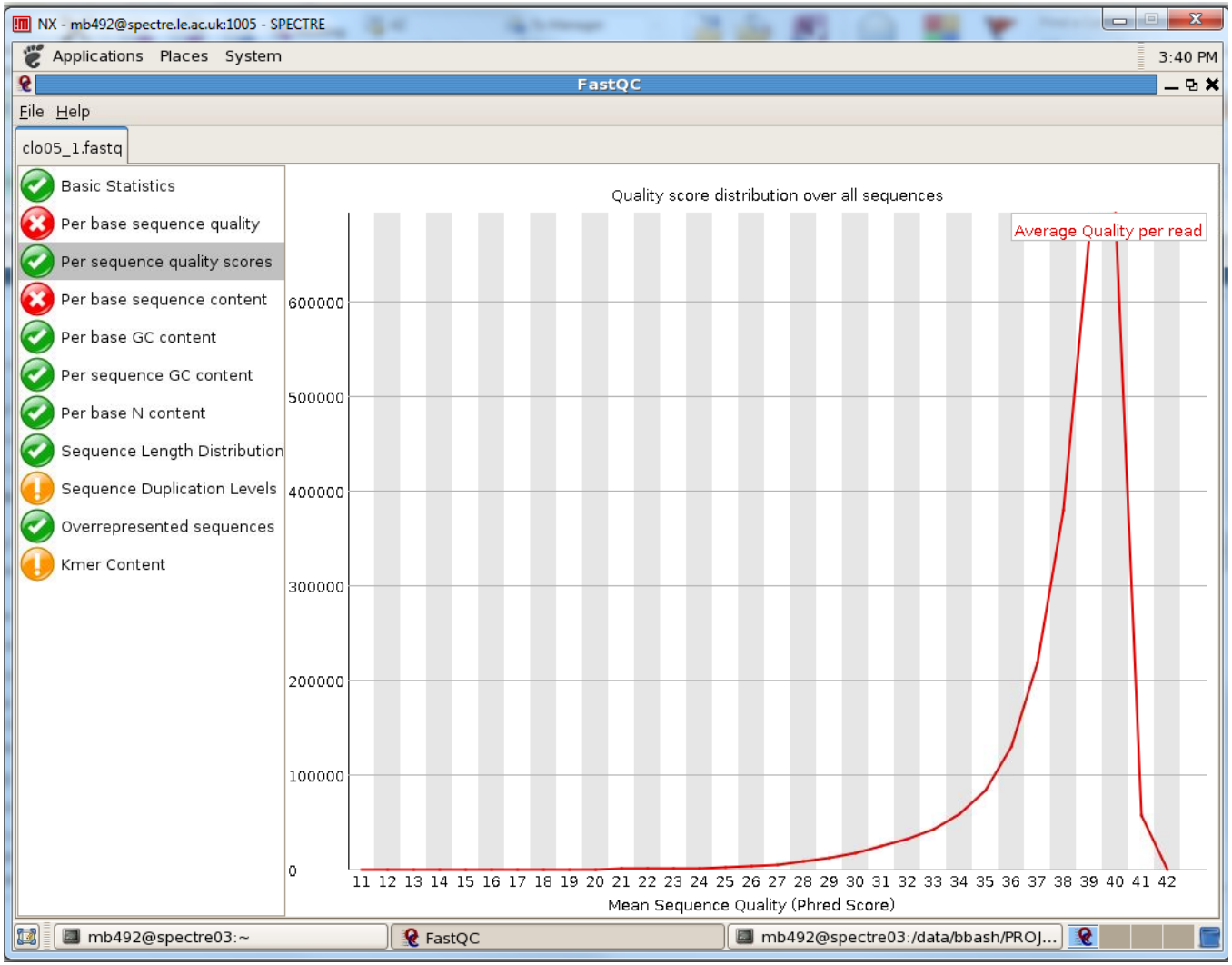
mean

Warning:
lower quartile for any base is < 10, or the median for any base <25

Fail:
lower quartile for any base <5 or the median for any base <20

University of Leicester

Per Sequence Quality Score

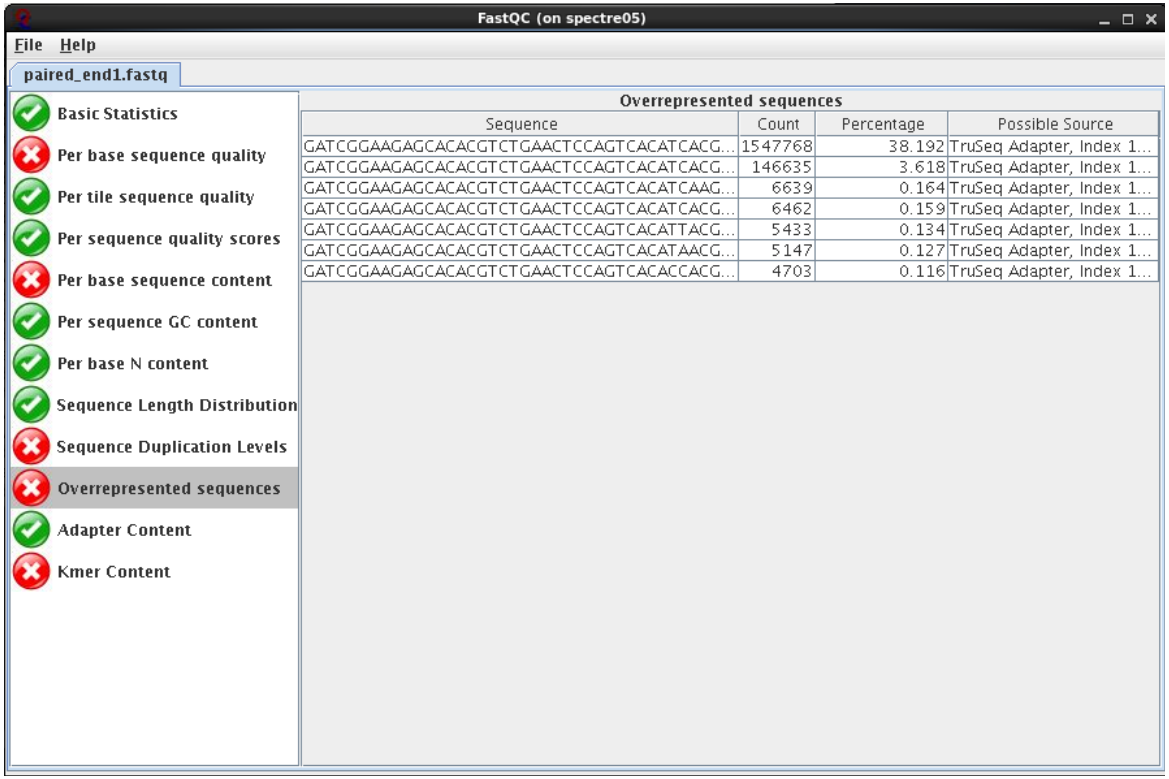


Warning:
mean quality <27

Fail:
mean quality <20

University of Leicester

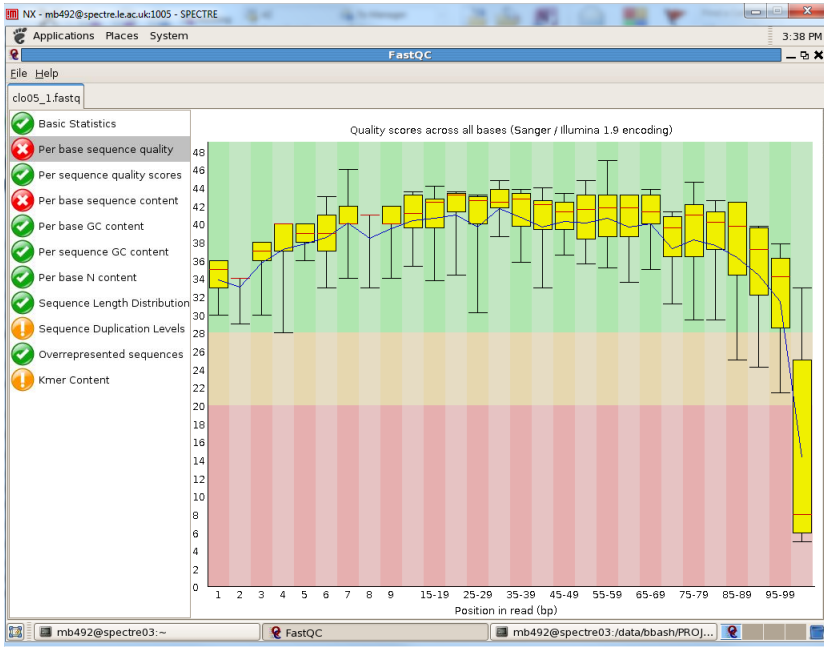
Overrepresented Sequences



University of Leicester

FastQC Summary

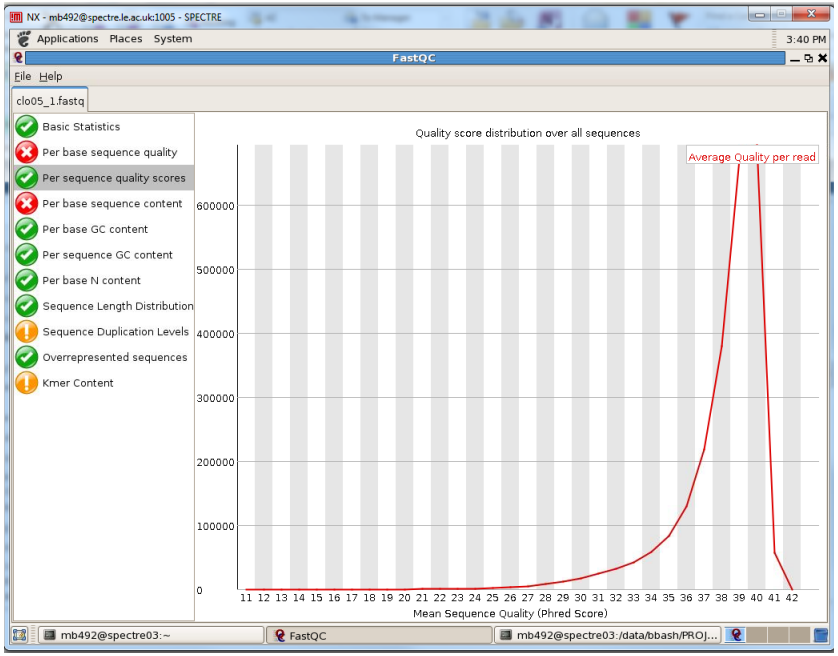
- What has FastQC told us?
  - Read trimming



University of Leicester

FastQC Summary

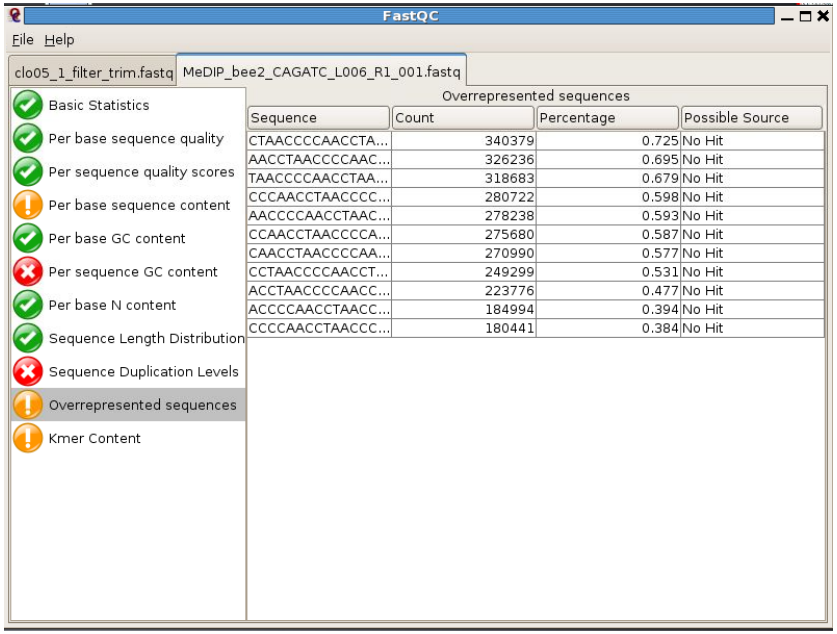
- What has FastQC told us?
  - Read trimming
  - Remove poor quality reads



University of Leicester

FastQC Summary

- What has FastQC told us?
  - Read trimming
  - Remove poor quality reads
  - Adapter removal



University of Leicester

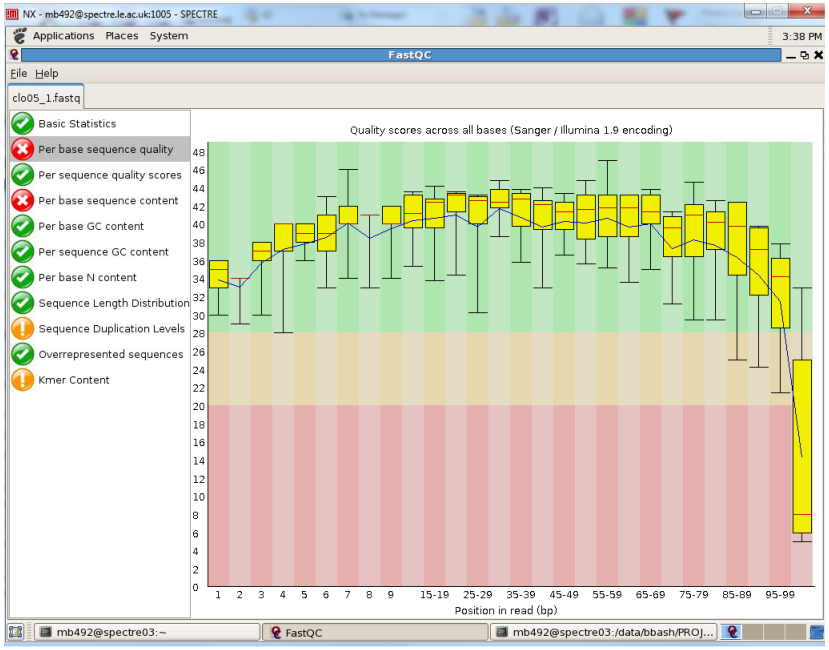
Trimmomatic

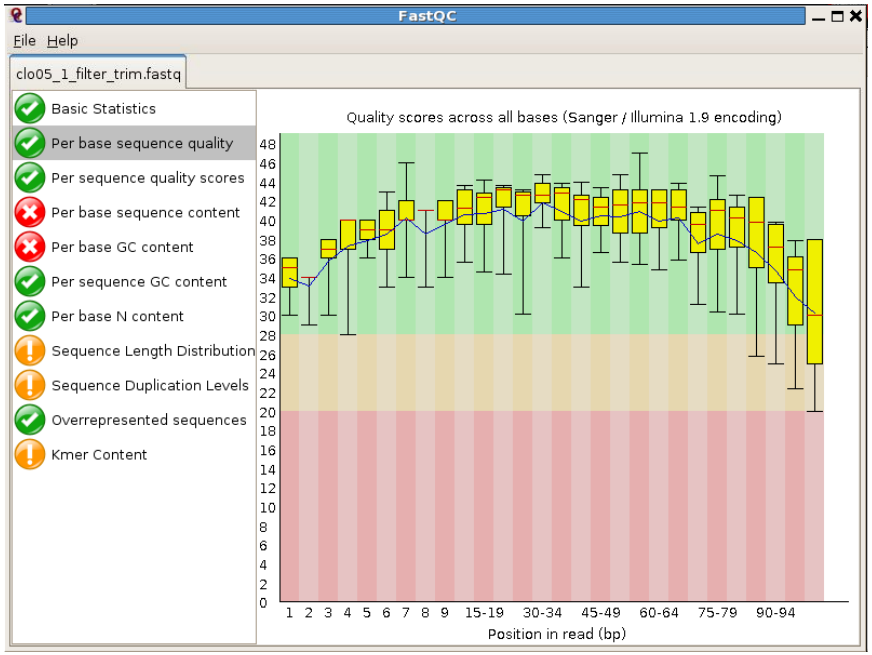
- A flexible trimmer for Illumina sequence reads
- Command line tools

University of Leicester

FastQC

- Re-run FastQC to assess improvement

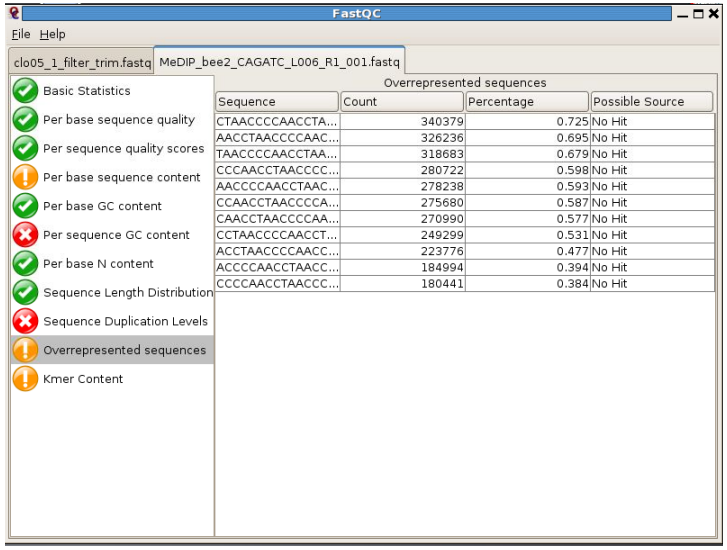


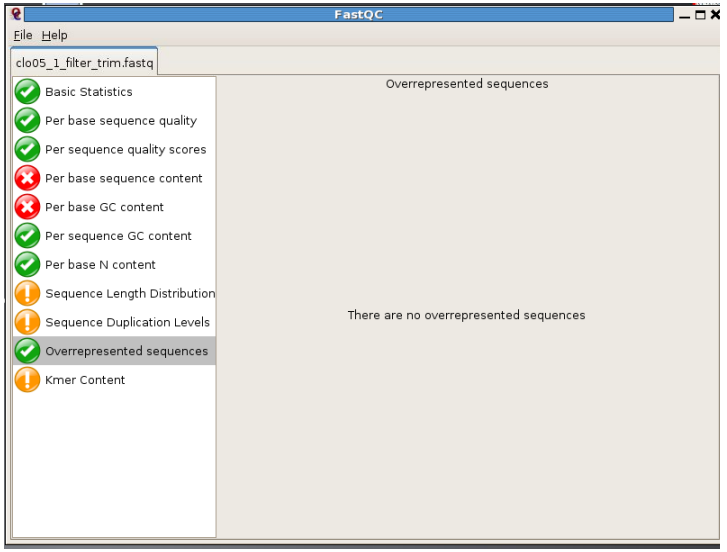


University of Leicester

FastQC

- Overrepresented sequences

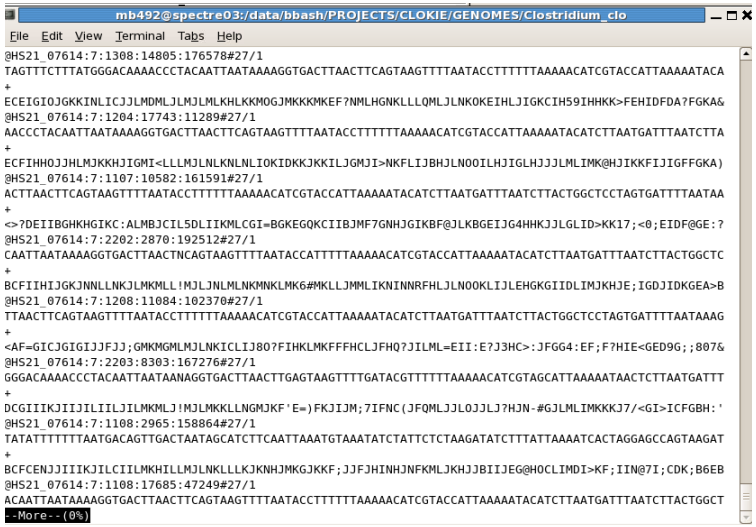




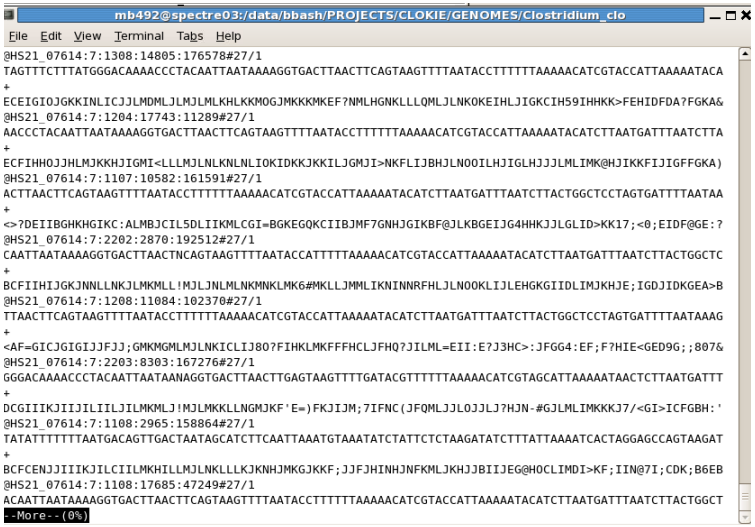
University of Leicester

Final data checks

Paired end 1



Paired end 2



Raw data	2459141 reads	2459141 reads
QC	2450224 reads	2448972 reads

Acknowledgements:

Kate Lee and Dr Matt Blades (B/BASH, University of Leicester, UK)

University of Leicester

Practical session

- Paired end illumina data (*Vibrio cholera* data)
- Run FastQC
- Decide what filtering/trimming is needed
- Run Trimmomatic to remove sequence adapters and low quality bases
- Check effects of filtering/trimming

Log into pico:

```
ssh -X your_user@login.pico.cineca.it  <- USE THE -X option!!!
```

Make a folder for the data and copy the data:

```
cd /pico/scratch/usertrain/your_user/
pwd
mkdir Data_QC
cd Data_QC
Pwd
```

Copy the files:

```
cp /pico/scratch/userexternal/phallast/DataQC/paired_end1.fastq .
cp /pico/scratch/userexternal/phallast/DataQC/paired_end2.fastq .
cp /pico/scratch/userexternal/phallast/DataQC/TruSeq3-PE-2.fa .
```

The folder contains:

Paired end 1 data file	= paired_end1.fastq
Paired end 2 data file	= paired_end2.fastq
Text file of adapter sequences	= TruSeq3-PE-2.fa

Look at how the files look:

```
more paired_end1.fastq
head paired_end1.fastq
```

Assess the quality of the data using FastQC

```
module load autoload fastqc/0.11.2
fastqc &
```

Click Open and find both .fastq files

Look at the fastqc modules on the left side, especially:
'Basic Statistics'
'Per Base Sequence Quality'
'Per Sequence Quality Scores'
'Overrepresented Sequences'

Look at the manual page of Fastqc and try to understand what these modules are telling you?

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

Use Trimmomatic to remove adapter sequences:

```
module load profile/advanced
module load autoload trimmomatic/0.33

java -Xmx1G -jar /cineca/prod/applications/trimmomatic/0.33/
binary/bin/trimmomatic-0.33.jar PE -phred33 -threads 8 -trimlog
logfile paired_end1.fastq paired_end2.fastq Left_paired.fastq
Left_unpaired.fastq Right_paired.fastq Right_unpaired.fastq
ILLUMINACLIP:TruSeq3-PE-2.fa:2:40:15 MINLEN:36
```

The parameters used for **Trimmomatic** are defined as follows:

PE	(data is paired end)
phred33	(Quality scores are 33 offset)
threads 8	(number of threads to use)
trimlog logfile	(name of logfile for summary information)
Left_paired.fastq	(paired trimmed output fastq file for left reads)
Left_unpaired.fastq	(unpaired trimmed output fastq file for left reads)
Right_paired.fastq	(paired trimmed output fastq file for right reads)
Right_unpaired.fastq	(unpaired trimmed output fastq file for right reads)
ILLUMINACLIP	(parameters for the adapter clipping)
TruSeq3-PE-2.fa	(text file of adapter sequences to search for)
:2:40:15	(adapter-read alignment settings – see manual)
MINLEN:36	(delete reads trimmed below length MINLEN)

Use Trimmomatic to trim low quality bases:

```
java -Xmx1G -jar /cineca/prod/applications/trimmomatic/0.33/
binary/bin/trimmomatic-0.33.jar PE -phred33 -threads 8 -trimlog
logfile2 paired_end1.fastq paired_end2.fastq
Left_trim_paired.fastq Left_trim_unpaired.fastq
Right_trim_paired.fastq Right_trim_unpaired.fastq LEADING:3
TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36
```

The parameters used for **Trimmomatic** are defined as follows:

PE	(data is paired end)
-phred33	(Quality scores are 33 offset)
-threads 8	(number of threads to use)
-trimlog logfile2	(name of logfile for summary information)
Left_paired.fastq	(name of input adapter trimmed left fastq file)
Right_paired.fastq	(name of input adapter trimmed right fastq file)
Left_trim_paired.fastq	(paired trimmed output fastq file for left reads)
Left_unpaired.fastq	(unpaired trimmed output fastq file for left reads)
Right_paired.fastq	(paired trimmed output fastq file for right reads)
Right_unpaired.fastq	(unpaired trimmed output fastq file for right reads)
LEADING:3	(Trim 5’ bases with quality score < 3)
TRAILING:3	(Trim 3’ bases with quality score < 3)
SLIDINGWINDOW:4:15	(see manual for explanation)
MINLEN:36	(delete reads trimmed below length MINLEN)

Assess again the quality of the data using FastQC

What are the main changes you see in the read quality?