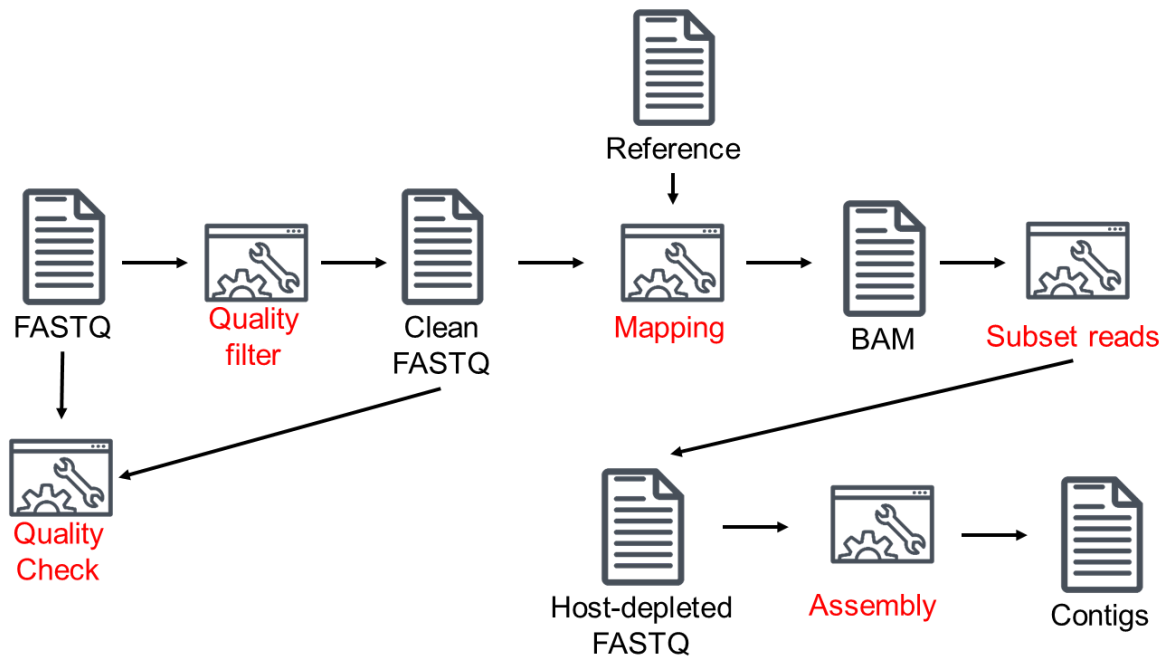


De novo assembly with short reads

Overview



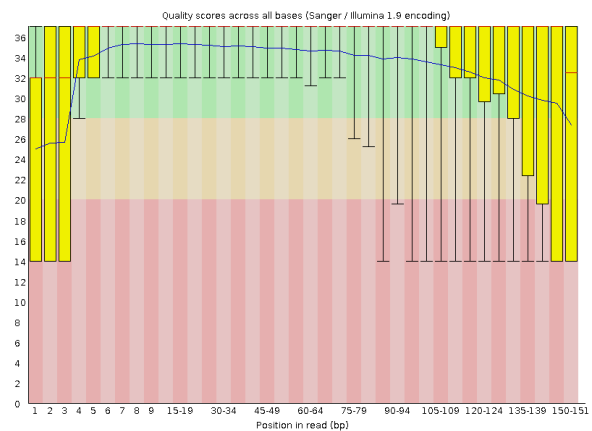
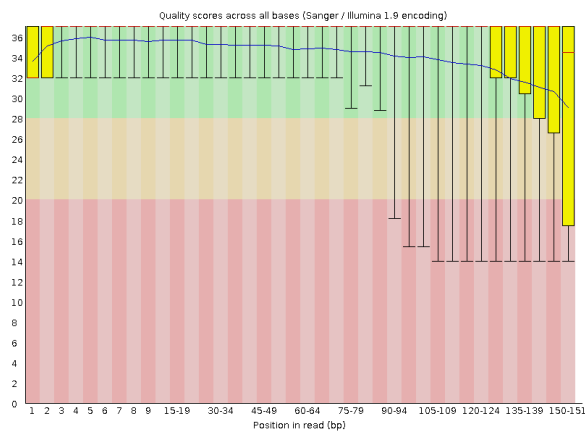
```
#install megahit
micromamba install -c conda-forge -c bioconda megahit
```

1. Quality control pre-filtering

```
fastqc -o qc_raw -f fastq ./training_material/fastq/sample04_1.fastq.gz ./training_material/fastq/sample04_2.fastq.gz
```

-o : output directory.

-f : fastq files, can be compressed fastq.



2. Quality filtering

```
#run trimmomatic
trimmomatic PE \
-threads 16 \
-phred33 \
-trimlog log.txt \
./training_material/fastq/sample04_1.fastq.gz \
./training_material/fastq/sample04_2.fastq.gz \
./clean/sample04paired_1.fastq.gz \
./clean/sample04unpaired_1.fastq.gz \
./clean/sample04paired_2.fastq.gz \
./clean/sample04unpaired_2.fastq.gz \
LEADING:20 \
TRAILING:20 \
SLIDINGWINDOW:4:20 \
AVGQUAL:20 \
MINLEN:100
```

PE → input is paired end reads

phred33 → use phred33 scoring system

trimlog → file to keep the output log

LEADING:20 → trim bases at the front if quality below threshold (20).

TRAILING:20 → trim bases at the end if quality below threshold (20).

SLIDINGWINDOW:4:20 → perform sliding window trimming: check the quality every 4 nucleotides, trim when quality falls below the threshold (20).

AVGQUAL:20 → remove read if the average base quality is below threshold (20).

MINLEN:100 → remove read if the length is shorter than threshold (100).

If adapter sequences need to be trimmed - add the **ILLUMINACLIP** option. I.e. ILLUMINACLIP:TruSeq3-

PE:2:30:10 → trim the TruSeq3-PE adapter, maximum 2 mismatches, palindrome clip threshold=30, simple clip threshold=10

3. Filter short reads and low complexity reads

```
fastp -i ./clean/sample04paired_1.fastq.gz -o ./clean/sample04clean_1.fastq.gz \
-I ./clean/sample04paired_2.fastq.gz -O ./clean/sample04clean_2.fastq.gz \
-e 20 -y 30 -l 100
```

-i and **-o**: input fastq files (paired end)

-I and **-O**: output fastq files (paired end)

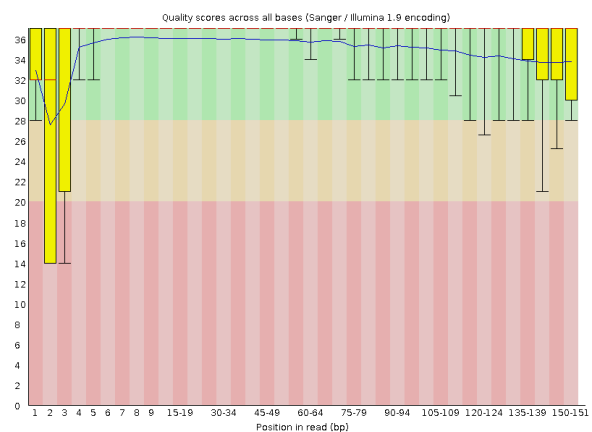
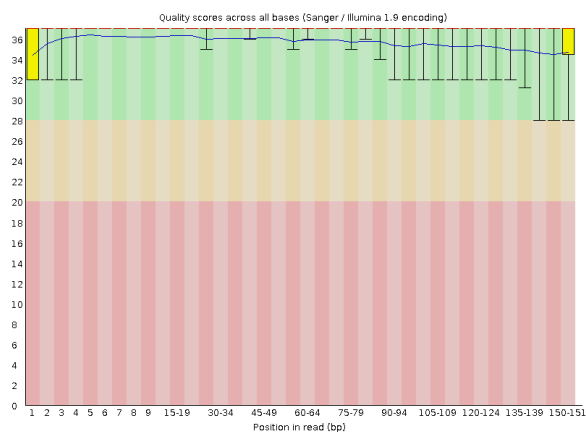
-e: mean read quality threshold

- y : complexity threshold
- l : minimum length threshold

4. Quality control post-filtering

```
fastqc -o qc_clean -f fastq \
./clean/sample04clean_1.fastq.gz \
./clean/sample04clean_2.fastq.gz
```

- o : output directory.
- f : fastq files, can be compressed fastq.



4. Host sequence removal

Get the index for human genome release 38.

```
#download premade index from bowtie https://benlangmead.github.io/aws-indexes/bowtie
wget https://genome-idx.s3.amazonaws.com/bt/GRCh38_noalt_as.zip

#unzip the index
unzip GRCh38_noalt_as.zip
```

Map the reads to the human genome

```
mkdir sample04
bowtie2 -x ./GRCh38_noalt_as/GRCh38_noalt_as -p 16 \
-1 ./clean/sample04paired_1.fastq.gz \
-2 ./clean/sample04paired_2.fastq.gz -S ./sample04/sample04.sam
```

Sam to bam - **which flag to filter?**

```
#SAM to BAM, output only unmapped reads.
samtools view -b -f XX ./sample04/sample04.sam | samtools sort \
> ./sample04/sample04unmapped.bam
```

```
#or if the file is Bam
samtools view -b -f XX ./sample04/sample04.bam > ./sample04/sample04unmapped.bam
```

BAM to fastq

```
samtools bam2fq ./sample04/sample04unmapped.bam \
-1 ./sample04/sample04hostremoved_1.fastq \
-2 ./sample04/sample04hostremoved_2.fastq
```

5. Assemble the non-host reads

```
#use megahit
megahit -1 ./sample04/sample04hostremoved_1.fastq \
-2 ./sample04/sample04hostremoved_2.fastq --min-contig-len 1500 \
-o ./sample04/megahit_res
```

-1 and **-2** : input fastq files (paired end)
--min-contig-length : minimum contig length.
-o : output directory

6. Remapping the reads to the contigs

Index the fasta file with contigs.

```
bowtie2 -x ./sample04/megahit_res/consensus_final.fasta \
./sample04/megahit_res/consensus_final.fasta
```

Map the host-decontaminated reads.

```
bowtie2 -x ./sample04/megahit_res/consensus_final.fasta \
-p 16 \
-1 ./sample04/sample04hostremoved_1.fastq \
-2 ./sample04/sample04hostremoved_2.fastq \
-S ./sample04/sample04_remapping_to_contig.sam
```


SAM to BAM.

```
#SAM to BAM, output only unmapped reads.
samtools view -b ./sample04/sample04_remapping_to_contig.aam | samtools sort \
> ./sample04/sample04_remapping_to_contig.bam
```

Run deeptools to judge the coverage

```
plotCoverage -o ./sample04/deeptools04.png -b ./sample04/sample04_remapping_to_contig.bam
```

7. Assembly parameters/evaluation

QUAST
QUAST assembly assessment tool
 <http://cab.cc.spbu.ru/quast/>

8. Annotations

<https://www.ncbi.nlm.nih.gov/orffinder/>