Project 1. What causes antibiotic resistance? - lab notebook

Downloading all necessary files:

- Reference genome sequence and annotation to genome

```
wget https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/005/845/GCF_000005845.2_ASM584v2/GCF_000005845.2_ASM584v2_genomic.fna.gz
wget https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/005/845/GCF_000005845.2_ASM584v2/GCF_000005845.2_ASM584v2_genomic.gff.gz
```

- Raw illumina sequencing reads of an E.coli strain [link] (unzipped for trimmomatic)

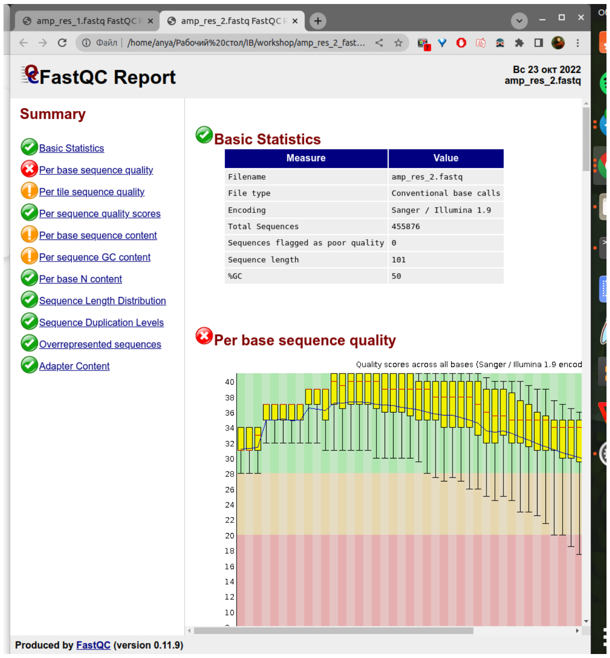
Inspecting data structure and counted reads

```
zless amp_res_1.fastq.gz
zless amp_res_1.fastq.gz


zcat amp_res_1.fastq.gz | wc -l
1823504  # 455 876 reads

zcat amp_res_2.fastq.gz | wc -l
1823504 # 455 876 reads
```

Downloading FastQC and running it

```
conda install -c bioconda fastqc
fastqc -o . '/pathtofile1/amp_res_1.fastq' '/pathtofile2/amp_res_2.fastq'
```

Inspecting FastQC report files


amp_res_1.fastq FastQC R
amp_res_2.fastq FastQC R
Файл | /home/anya/Рабочий%20стол/IB/workshop/amp_res_2_fast...

FastQC Report

Вс 23 окт 2022
amp_res_2.fastq

Summary

- ✓ 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

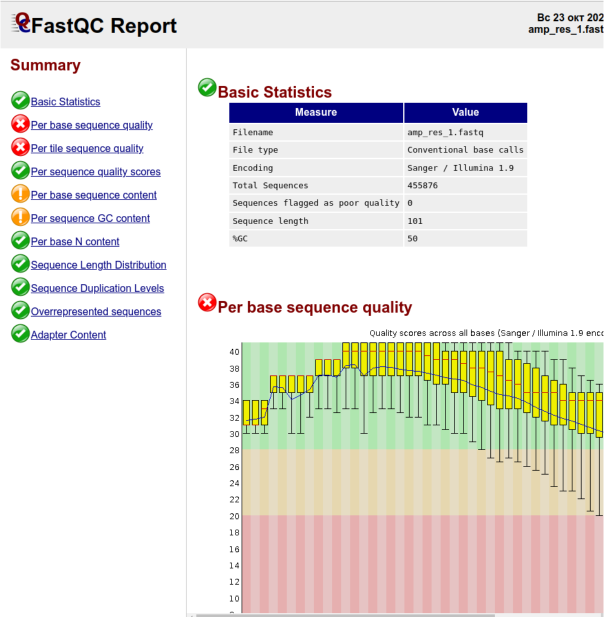
✓ Basic Statistics

Measure	Value
Filename	amp_res_2.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	455876
Sequences flagged as poor quality	0
Sequence length	101
%GC	50

✗ Per base sequence quality

Quality scores across all bases (Sanger / Illumina 1.9 encod
40
38
36
34
32
30
28
26
24
22
20
18
16
14
12
10

Produced by FastQC (version 0.11.9)



FastQC Report

Вс 23 окт 202
amp_res_1.fast

Summary

- ✓ 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

✓ Basic Statistics

Measure	Value
Filename	amp_res_1.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	455876
Sequences flagged as poor quality	0
Sequence length	101
%GC	50

✗ Per base sequence quality

Instructions about interpreting FastQC results [link]

Downloading trimmomatic and filtering reads

```
conda install -c bioconda trimmomatic
```

Trimmomatic README page [link]:

> Code was taken from README with some modifications:
>
> - Paired reads (PE)
> - Phred33 quality scale (-phred33)
> - Remove leading low quality or N bases (below quality 20) (LEADING:20)
> - Remove trailing low quality or N bases (below quality 20) (TRAILING:20)
> - Scan the read with a 10-base wide sliding window, cutting when the average quality per base drops below 20 (SLIDINGWINDOW:10:20)
> - Drop reads below the 20 bases long (MINLEN:20)

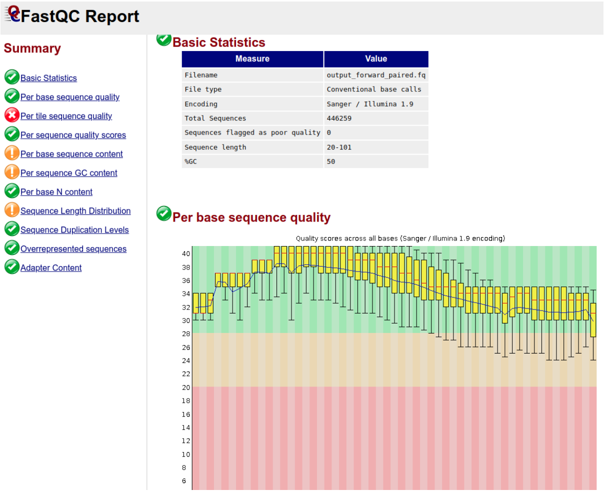
```
java -jar /pathtofile/trimmomatic-0.39.jar PE -phred33 amp_res_1.fastq
amp_res_2.fastq output_forward_paired.fq.gz
output_forward_unpaired.fq.gz output_reverse_paired.fq.gz
output_reverse_unpaired.fq.gz
LEADING:20 TRAILING:20 SLIDINGWINDOW:10:20 MINLEN:20
# RESULT
Input Read Pairs: 455876 Both Surviving: 446259 (97,89%) Forward Only Surviving: 9216 (2,02%) Reverse Only Surviving: 273 (0,06%) Dropped:
TrimmomaticPE: Completed successfully
```

Counting reads in file

```
zcat output_forward_paired.fq.gz | wc -l
1785036 # 446 259 reads
```

Repeating FastQC analysis on filtered reads

```
fastqc -o . '/pathtofile/output_forward_paired.fq' '/pathtofile/output_reverse_paired.fq'
```



FastQC Report

Summary

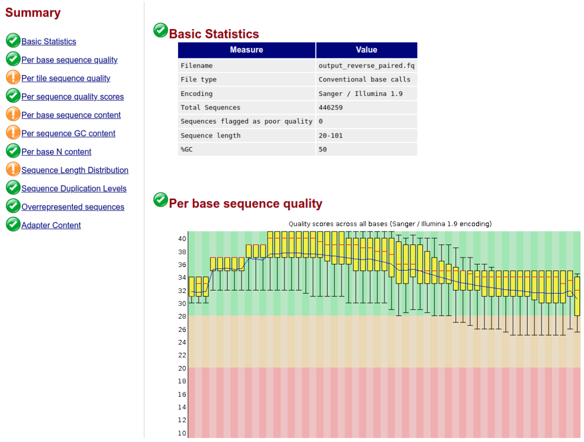
- ✓ Basic Statistics
- ✓ Per base sequence quality
- ✗ Per tile sequence quality
- ✓ Per sequence quality scores
- ! Per base sequence content
- ! Per sequence GC content
- ✓ Per base N content
- ! Sequence Length Distribution
- ✓ Sequence Duplication Levels
- ✓ Overrepresented sequences
- ✓ Adapter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

✓ Basic Statistics

Measure	Value
Filename	output_forward_paired.fq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	446259
Sequences flagged as poor quality	0
Sequence length	20-101
%GC	50

✓ Per base sequence quality

Quality scores across all bases (Sanger / Illumina 1.9 encoding)
40
38
36
34
32
30
28
26
24
22
20
18
16
14
12
10
8
6



Summary

- ✓ 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

✓ Basic Statistics

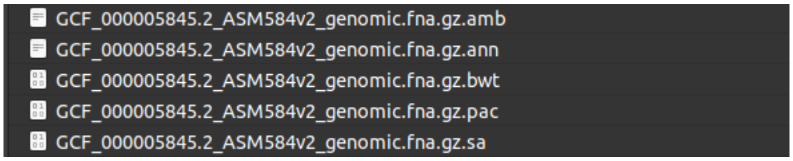
Measure	Value
Filename	output_reverse_paired.fq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	446259
Sequences flagged as poor quality	0
Sequence length	20-101
%GC	50

✓ Per base sequence quality

Quality scores across all bases (Sanger / Illumina 1.9 encoding)
40
38
36
34
32
30
28
26
24
22
20
18
16
14
12
10

Indexing reference genome with bwa

```
bwa index '/pathtofile/GCF_000005845.2_ASM584v2_genomic.fna.gz'
```



GCF_000005845.2_ASM584v2_genomic.fna.gz.amb
GCF_000005845.2_ASM584v2_genomic.fna.gz.ann
GCF_000005845.2_ASM584v2_genomic.fna.gz.bwt
GCF_000005845.2_ASM584v2_genomic.fna.gz.pac
GCF_000005845.2_ASM584v2_genomic.fna.gz.sa

Running bwa mem

```
bwa mem -t 4 '/pathtofile/GCF_000005845.2_ASM584v2_genomic.fna.gz' '/pathtofile/output_forward_paired.fq.gz' '/pathtofile/output_reverse_pa
```

Compressing .sam file into .bam and inspecting it’s statistics

SAM format specifications [link]

```
samtools view -S -b alignment.sam > alignment.bam
```

```
samtools flagstat alignment.bam
# RESULT
892776 + 0 in total (QC-passed reads + QC-failed reads)
892518 + 0 primary
0 + 0 secondary
258 + 0 supplementary
0 + 0 duplicates
0 + 0 primary duplicates
891649 + 0 mapped (99.87% : N/A)
891391 + 0 primary mapped (99.87% : N/A)
892518 + 0 paired in sequencing
446259 + 0 read1
446259 + 0 read2
888554 + 0 properly paired (99.56% : N/A)
890412 + 0 with itself and mate mapped
979 + 0 singletons (0.11% : N/A)
0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ>=5)
```

Sorting and indexing .bam file

```
samtools sort alignment.bam > alignment_sorted.bam
samtools index alignment_sorted.bam
```

Variant calling by creating a mpileup file and running VarScan

VarScan github [link]

VCF file specifications [link]

```
samtools mpileup -f '/pathtofile/GCF_000005845.2_ASM584v2_genomic.fna' '/pathtofile/alignment_sorted.bam' > my.mpileup
java -jar '/pathtofile/VarScan.v2.4.3.jar' mpileup2snp my.mpileup --min-var-freq 0.75 --variants --output-vcf 1 > VarScan_results.vcf
# RESULTS
Only SNPs will be reported
Warning: No p-value threshold provided, so p-values will not be calculated
Min coverage: 8
Min reads2: 2
Min var freq: 0.75
Min avg qual: 15
```

```
P-value thresh: 0.01
Reading input from my.mpileup
4641343 bases in pileup file
9 variant positions (6 SNP, 3 indel)
0 were failed by the strand-filter
6 variant positions reported (6 SNP, 0 indel)
(base) anya@anya-laptop:~/Рабочий стол/IB/wor
```

Automatic SNP annotation with SnpEff

SnpEff documentation [link]

- Installation of the SnpEff package archive (which is then unzipped)

```
wget https://snpeff.blob.core.windows.net/versions/snpEff_latest_core.zip
```

- Downloading sequence and annotation of reference (in the GenBank format)

```
wget https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/005/845/GCF_000005845.2_ASM584v2/GCF_000005845.2_ASM584v2_genomic.gbff.gz
```

- Creating empty `snpeff.config.txt` file and writing one line inside: `k12.genome : ecoli_K12`
- Creating folder for the database and putting .gbk file inside

```
mkdir -p data/k12
cp GCF_000005845.2_ASM584v2_genomic.gbff data/k12/genes.gbk
```

- Creating database

```
java -jar /pathtofile/snpEff.jar build -genbank -v k12
```

- Annotating

```
java -jar /pathtofile/snpEff.jar ann k12 VarScan_results.vcf > VarScan_results_annotated.vcf
```