
# Introduction to Sequence Data - practical

# AMR Bioinformatics Practical environment:

1. Load the NGS AMR 2022 virtual environment using Oracle VM VirtualBox.
2. Make sure the shortcut to the folder ”manager” is on the desktop
3. Open the terminal
4. Rename the “**Genome Assembly**” folder to “**Genome_Assembly**”

# Practical 1 – FastQC

1. Do quality check on all Ecoli and Styphi raw fastq data

   (hint: the Ecoli and Styhi data files are inside the **Genome_Assembly/Raw_fastq** folder)

Questions:

A) How many reads are in Ecoil-A forward read file?

B) Is there any adaptor sequence in Styphi-C reverse read file?

# Practical 1 – FastQC (Answers)

1. In your VM, open the terminal:

```
cd Genome_Assembly
cd Raw_FASTQs
cd Ecoli
mkdir fastqc_result
fastqc –o /home/manager/Genome_Assembly/Raw_FASTQs/Ecoli/fastqc_results  \
    /home/manager/Genome_Assembly/Raw_FASTQs/Ecoli/Ecoli-A_38843_1.fastq.gz
```

2. Check the output folder fastqc_results for the .html report
3. Repeat fastqc command with Ecoli-A_38843_2.fastq.gz

# Practical 1 – FastQC (Answers)

```
fastqc –o /home/manager/Genome_Assembly/Raw_FASTQs/Ecoli/fastqc_results  \
    /home/manager/Genome_Assembly/Raw_FASTQs/Ecoli/Ecoli-A_38843_2.fastq.gz
```

4. Repeat fastqc with the all the other Ecoli samples

```
fastqc –o /home/manager/Genome_Assembly/Raw_FASTQs/Ecoli/fastqc_results  \
    /home/manager/Genome_Assembly/Raw_FASTQs/Ecoli/Ecoli-*_37*.fastq.gz
```

5. Repeat fastqc with the all the Styphi samples

```
cd ../Styphi
mkdir fastqc_results
fastqc –o /home/manager/Genome_Assembly/Raw_FASTQs/Styphi/fastqc_results  \
    /home/manager/Genome_Assembly/Raw_FASTQs/Styphi/*.fastq.gz
```

# Practical 2 - MultiQC

1. Generate MultiQC report from Styphi raw .fastq fastQC results
2. Generate MultiQC report from Ecoli raw .fastq fastQC results

Questions

a) Which Ecoli sample has the most reads?

b) What is the GC content of the Styphi samples?

# Practical 2 – MultiQC (Answers)

1. In the terminal, change directory into the Styphi fastqc_result folder

```
cd /home/manager/Genome_Assembly/Raw_FASTQs/Styphi/fastqc_results
mkdir multiQC_result

multiQC  \
   -o /home/manager/Genome_Assembly/Raw_FASTQs/Styphi/fastqc_results/multiQC_result \
   /home/manager/Genome_Assembly/Raw_FASTQs/Styphi/fastqc_results/*
```

# Practical 2 – MultiQC (Answers)

2. Repeat the multiQC command on the Ecoli dataset

```
cd /home/manager/Genome_Assembly/Raw_FASTQs/Ecoli/fastqc_results
mkdir multiQC_result

multiQC  \
   -o/home/manager/Genome_Assembly/Raw_FASTQs/Ecoli/fastqc_results/multiQC_result \
   /home/manager/Genome_Assembly/Raw_FASTQs/Ecoli/fastqc_results/*
```

# Practical 3 – Trim galore

1. Generate MultiQC report from Ecoli raw fastq fastQC results
2. Generate MultiQC report from Styhi raw fastq fastQC results

Questions:

A) How many reads are in Ecoil-A forward read file?

B) Is there any adaptor sequence  in Styhi-C reverse read file?

# Practical 3 – Trim galore (Answer)

1. In your terminal, change directory to the Raw fastqs Ecoli folder

```
cd /home/manager/Genome_Assembly/Raw_FASTQs/Ecoli/
mkdir trim_result
```

2. Run Trim galore on Ecoli-A raw data files

```
trim_galore\
    --paired –fastqc –illumina \
    -o /home/manager/Genome_Assembly/Raw_FASTQs/Ecoli/trim_result/ \
    /home/manager/Genome_Assembly/Raw_FASTQs/Ecoli/Ecoli-A_38843_1.fq.gz \
    /home/manager/Genome_Assembly/Raw_FASTQs/Ecoli/Ecoli-A_38843_2.fq.gz
```

# Practical 3 – Trim galore (Answer)

3. Run Trim galore on Ecoli-B, Ecoli-C, Ecoli-D and Ecoli-E raw data files

```
trim_galore\
    --paired –fastqc –illumina \
    -o /home/manager/Genome_Assembly/Raw_FASTQs/Ecoli/trim_result/ \
    /home/manager/Genome_Assembly/Raw_FASTQs/Ecoli/Ecoli-B*.fq.gz \
    /home/manager/Genome_Assembly/Raw_FASTQs/Ecoli/Ecoli-C*.fq.gz \
    /home/manager/Genome_Assembly/Raw_FASTQs/Ecoli/Ecoli-D*.fq.gz \
    /home/manager/Genome_Assembly/Raw_FASTQs/Ecoli/Ecoli-E*.fq.gz
```

# Practical 3 – Trim galore (Answer)

4. Run Trim galore on Ecoli-B, Ecoli-C, Ecoli-D and Ecoli-E raw data files

```
cd /home/manager/Genome_Assembly/Raw_FASTQs/Styphi/
mkdir trim_result
trim_galore\
    --paired –fastqc –illumina \
    -o /home/manager/Genome_Assembly/Raw_FASTQs/Styphi/trim_result/ \
    /home/manager/Genome_Assembly/Raw_FASTQs/Styphi/*.fq.gz
```

# Practical 4 - BactInspector

1. Run bactinspector check_species on Ecoli-A trimmed data
2. Run bactinspector closest_match on Ecoli-A

Questions:

A) From the check_species result, what is the speice ID and the top hit distance?

B) From the closest_match result, what is the closet ReSeq match? (hint refseq_organism_name?)

# Practical 4 – BactInspector (Answers)

1. In your terminal, change directory to the Raw fastqs Ecoli folder

```
cd /home/manager/Genome_Assembly/Raw_FASTQs/Ecoli/
mkdir bactInspector_result/Ecoli_A
```

2. Run Bactinspector check_species on Ecoli-A trimmed data

```
bactinspector check_species \
        -i /home/manager/Genome_Assembly/Raw_FASTQs/Ecoli/trim_result/ \
       -o /home/manager/Genome_Assembly/Raw_FASTQs/Ecoli/bactInspector_result/Ecoli_A \
       –p 4 –fq Ecoli-A_38843_1_val_1.fq.gz
```

# Practical 4 – BactInspector (Answers)

3. Run bactinspector closest_match on Ecoli-A trimmed data

```
bactinspector closest_match  \
    -i /home/manager/Genome_Assembly/Raw_FASTQs/Ecoli/ bactInspector_result/Ecoli_A/ \
    -o /home/manager/Genome_Assembly/Raw_FASTQs/Ecoli/bactInspector_result/Ecoli_A \
    –p 4 –r –m Ecoli-A_38843_1_val_1.msh
```

# Practical 4 – BactInspector (Answers)

Species ID = Escherichia coli

Top_hit_distance = 0.00755945

ReSeq closest match = Escherichia coli 083:H1 str.