Computational practical 3

Short read assembly

Introduction

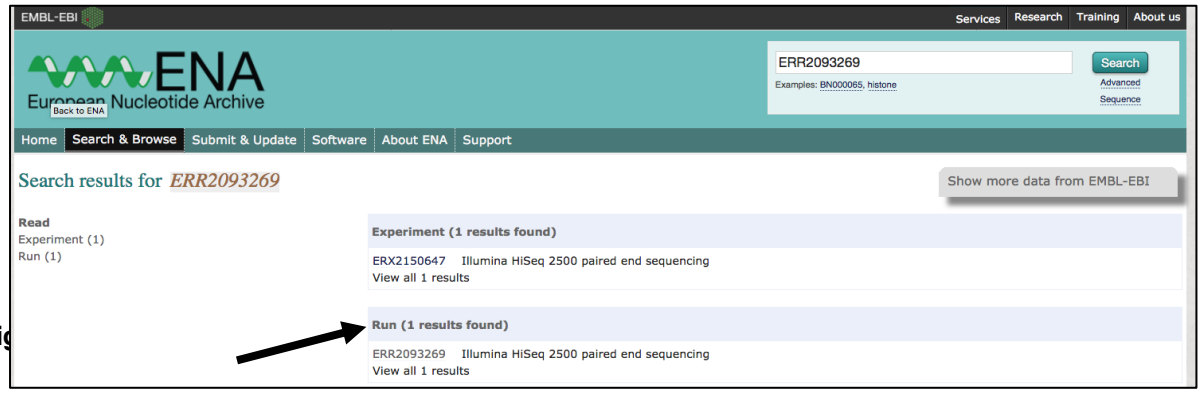
Whole genome sequencing is rapidly being used for understanding evolution and spread of antimicrobial resistance. This has fostered the development of various bioinformatics tools that are more user-friendly and requires minimum expertise. Through global efforts a number of antimicrobial resistance databases and tools have been developed that can help identify determinants of resistance from whole genome sequences.

Objective

In this chapter, we will gain an understanding of downloading publicly available genome sequences (raw reads and genome assemblies) from databases (European Nucleotide Archive), *de novo* assembly and detection of genetic determinants of resistance using web-based tools. We will start by learning how to access and download the assembled genome sequences/sequence reads from publicly available repositories. We will then assemble the reads to generate contigs (long contiguous stretch of nucleotides).

Downloading sequence reads

Open the ENA website (https://www.ebi.ac.uk/ena) in the browser and repeat step 1 from the above exercise with the accession number (**ERR2093269**). This will open a window as shown below. Click on the "Run" as pointed by the arrow in figure 1.



Fig

Now you can see all the information including sequencing read files associated with this accession number (figure2). Right click on the first read file "_1.fastq.gz" and copy the link.

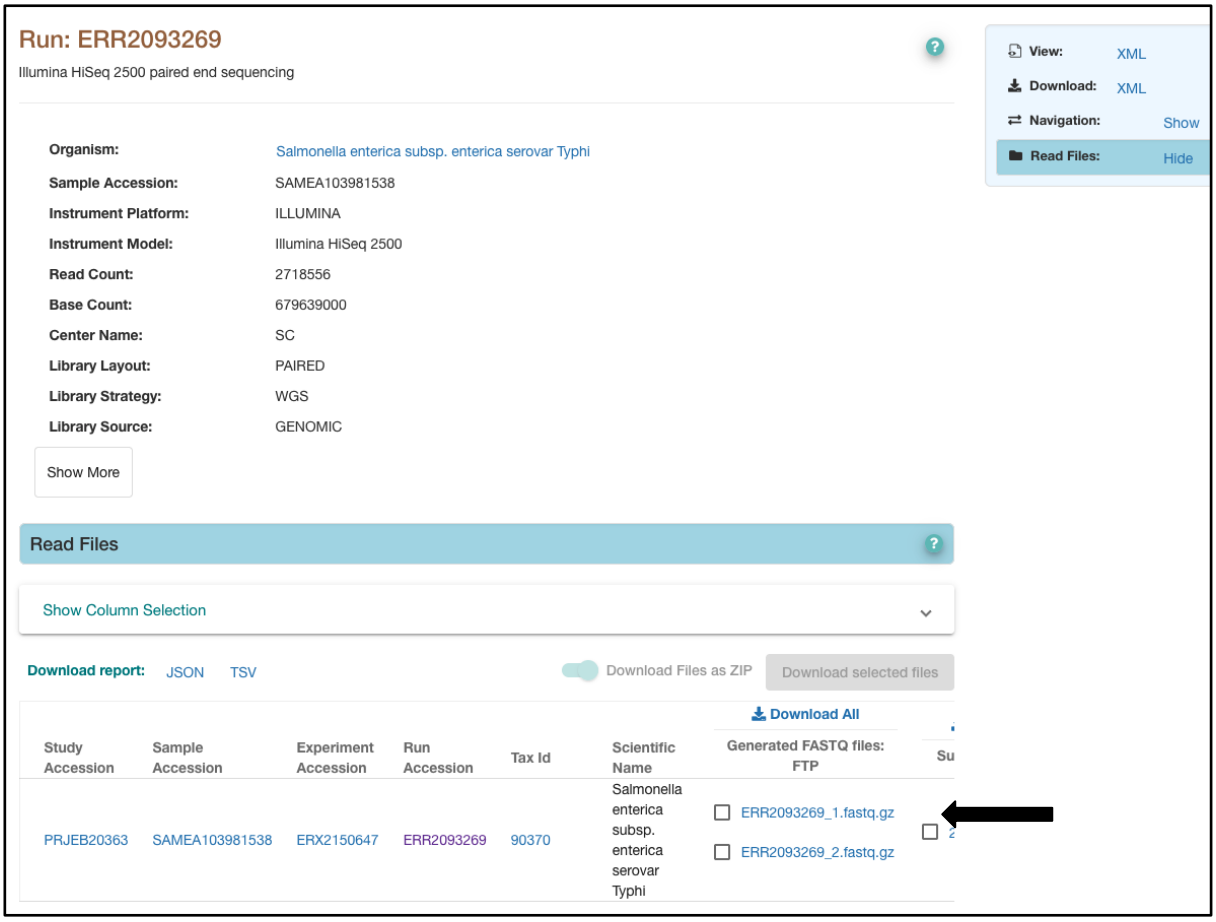


Figure 2: Sequencing information page

Downloading sequences using command line

Change to the working folder for this section **cp3**

cd /home/manager/course/cp3

After copying the fastq file link as shown in the figure above and using the command shown below:

wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR209/009/ERR2093269/ERR2093269_1.fastq.gz

Similarly you can download the other read file "_2.fastq.gz" by copying the link and then using wget command:

wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR209/009/ERR2093269/ERR2093269_2.fastq.gz

Genome assembly

Now we will learn to assemble the sequence reads using the command line. This method is convenient when handling a high number of isolates. There are many

tools available such as SPAdes, velvet etc. Here, we will use the tool shovill (https://github.com/tseemann/shovill) to assemble the sequence reads of the isolate ERR2093269. The fastq files are located in the folder cp3. Once we have entered into the folder type the following command

shovill --outdir ERR2093269assembly --R1 ERR2093269_1.fastq.gz --R2 ERR2093269_2.fastq.gz

In the command option --outdir refers to the name of the output folder, --R1 and --R2 refer to the read1 and read2 files. The process will take a while to run, once finished all the output files will be in the assembly folder. The output folder will have the following files:

Filename	**Description**
contigs.fa	The final assembly that is to be used
shovill.log	Full log file for bug reporting
shovill.corrections	List of post-assembly corrections
contigs.gfa	Assembly graph (spades)
contigs.fastg	Assembly graph (megahit)
contigs.LastGraph	Assembly graph (velvet)
skesa.fasta	Raw assembly (skesa)
spades.fasta	Raw assembled contigs (spades)
megahit.fasta	Raw assembly (megahit)
velvet.fasta	Raw assembly (velvet)

We can look into the output folder, the designated

ls -lh

We can also generate statistics for the assembled contigs, namely, number of contigs N50 and total assembled size using another tool “QUAST”. It can be run

using the following command:

```
quast.py contigs.fa
```

The tool will create a folder “quast_results” and the results will be within the folder prefixed “results”. In order to view the results by opening the “report.pdf” file.

Now we have learned about accessing the ENA database using accession Ids, downloading the sequence reads, performing assembly and generating basic assembly stats. These assemblies now are ready for further downstream analysis such as antimicrobial resistance detection that we will learn in the next practicals.