EnSE314 Public Health Microbiology II

2022/2023

Practical Session: Metagenomic investigation of bacteria in sewage

Contents

Step 0 Background

Step 1 Prerequisites of analysis

Step 2 Input datasets quality control (QC)

Step 3 Assembly of shotgun metagenomic reads

Step 4 Metagenome-assembled genome (MAG) binning

Step 5 Identification and annotation of MAGs

Step 6 Results and discussions

Pei-Ying Hong (Peiying.hong@kaust.edu.sa)

Changzhi Wang ([Changzhi.wang@kaust.edu.sa](mailto:Changzhi.wang@kaust.edu.sa))

WDRC, BESE, KAUST

**Step 0 Background**

For this session, we will work with one real metagenomic datasets from sewage. This is a shotgun metagenomic dataset with a SRA ID ERR1713378, which belong to a public ENA project PRJEB13831 (<https://www.ebi.ac.uk/ena/browser/view/PRJEB13831>). It is the basis for the paper “Global monitoring of antimicrobial resistance based on metagenomics analyses of urban sewage” on Nature communications (<https://doi.org/10.1038/s41467-019-08853-3>). This paper investigated the antimicrobial resistance (AMR) in untreated sewage from 60 countries. Their findings suggested the global AMR as a serious threat to global public health.

In this paper, authors didn’t investigate any pathogens from untreated sewage, therefore, our session aim to investigate pathogens through a metagenomic analysis pipeline. We chose ERR1713378 (called as NGA.50 in this paper), originally from Nigeria. This dataset is sequenced by paired-end illumina HiSeq and named as:

ERR1713378\_1.fastq.gz

ERR1713378\_2.fastq.gz

For the rest of session, we will follow a widely used metagenomic protocol to investigate pathogens from this dataset.

**Step 1 Prerequisites of analysis**

Before we start, we need to assure our work environment are ready:

- Required software (App) on local computer:

1. Terminal App (for Mac: Terminal, for Windows: Xshell)

2. scp/sftp App (for Mac: Filezilla, for Windows: Xftp)

- Access to KAUST high performance computer (hpc) Ibex:

Check <https://www.hpc.kaust.edu.sa/ibex/ibex_guide>

Quick login: ssh -X your\_portal\_loginname@ilogin.ibex.kaust.edu.sa

\*if you cannot login though above procedure, contact ibex team:

ibex@hpc.kaust.edu.sa

- Obtain datasets and sources from github:

Log in your ibex space, create a folder for this practical session:

mkdir ense314

cd ense314

git clone https://github.com/Chiongchih/EnSE3142023

download metagenomic dataset from ENA (<https://www.ebi.ac.uk/ena/browser/view/PRJEB13831>), find and download the target dataset ERR1713378 (sample alias: NGA.50). Upload dataset to your ibex space under the same folder.

- Tools used:

Throughout this session we will be using a variety of tools that listed below, along with the particular versions. You don’t need to install these tools because Ibex already installed for everybody.

FastQC v0.11.8; trimmomatic v0.38; megahit v1.2.9; metawrap v1.3; gtdb-tk v1.3...

**Step 2 Input datasets quality control (QC)**

The first step is to access the quality of input reads by running fastqc.sh. Fastqc scans fastq files to generate a broad overview of summarized statistics and has several screening modules that test for some commonly occurring problems. It produces an html output for samples. Let us make sure we put the fastqc.sh file with the same folder of metagenomic datasets, and then run fastqc.sh file in terminal (before doing it, open the sh file and change the email to your email address!):

vim fastqc.sh

#Change the email accordingly, and save fastqc.sh by press ‘esc’, ‘:’, ‘wq’.

sbatch fastqc.sh

The resulting html output files can be downloaded and opened by any browser on your local computer. So open it and check the results, and check the results.

After fastqc, we will remove reads with a low quality and remove overrepresented sequences in our dataset. Trimmomatic enables you to trim up your sequences based on quality thresholds and other requirement. You can now set up and run trimmomatic with default setting. This step will cost 1 hour.

sbatch trim\_run\_real.sh

The output will be delivered as 4 files: paired\_1, paired\_2, unpaired\_1, unpaired\_2. We will use the paired\_1 and paired\_2 for further analysis.

After trimmomatic, we should check how many reads were removed from our datasets. You can now set up and run multiqc.sh accordingly.

sbatch multiqc.sh

Same as fastqc, multiqc output the results into html file, you can download the html result and open it on your local browser to check the result of trimmomatic.

**Step 3 Assembly of shotgun metagenomic reads**

We are now ready to assemble our data into genomic sequences, using megahits\_t16.sh and set up accordingly. This step will cost 3-4 hours.

sbatch megahits\_t16.sh

Meaghit will cluster all its results in an independent folder (ERR\*\_megahit), open this folder, and you can find out our assembled contigs results named as “final.contigs.fa”.

**Step 4 Metagenome-assembled genome (MAG) binning**

Now we are going forward with the assembled contigs to recover bacterial genomes out of it. We call this step “Genome binning”, and we call the recovered genome as “metagenome-assembled genomes” or “MAGs”. To do genome binning, we will ask a tool “Metawrap” to do everything for us. It produced MAGs with proper quality control according to our settings. Before we get into details of Metawrap, let’s run it first. This step will take 2 hours approximately.

sbatch metawrap.sh

When this step completed, 2 folders are created: ERR\*\_metawrap and ERR\*\_metawrap\_refine. Open the second folder, and there are 2 key files we should focus on:

metawrap\_70\_10\_bins.stats, which listed the basic information of genomes passed QC. metawrap\_70\_10\_bins, which contained the sequences of these genomes in fasta format.

When you open the .stats file, you will find out genomes were classified, but they are not enough to determine what exactly bacteria it is and if the genome is a pathogen. Therefore, we need to use a better tool to classify these MAGs and further determine other characters of these MAGs (such as antibiotic resistant, etc.).

**Step 5 Identification and annotation of MAGs**

Now we will use GTDB-tk to classify the MAGs, which will tell us what exactly the bacteria is (likely in species level). Let us set up the .sh file and start the run, this step will take 50 min approximately:

sbatch GTDB.sh

When it completed, a folder named ERR\*\_GTDB will come out, enter this folder, download the file gtdbtk.bac120.summary.tsv to your local computer and open it with excel or similar app. Check the result and tell if there is any pathogens.

Now we successfully classified MAGs recovered from the metagenome. We are free to do some extra analysis according to our interest. In this paper (<https://doi.org/10.1038/s41467-019-08853-3>), authors focus on the AMR in sewage. In this protocol, we will continue their interest by investigating AMR of our MAGs: Are there any antibiotic resistant genes can be found in these MAGs? In particular, these pathogens? By doing that, we need to download these MAGs to our local computer first, then open an authorized ARG database CARD-RGI (<https://card.mcmaster.ca/analyze/rgi>). Upload your FASTA sequences files accordingly and submit your request. Then check the results.

**Step 6 Results and discussions**

Finally, we successfully recovered bacterial genomes from metagenomes and please answer the following questions:

1. Which pathogens did you find from this dataset?

2. Is there any MAGs carried ARG? If so, which are they? Are there any pathogens carried ARGs?

3. What is the advantages and disadvantages for metagenomics? (Compare to cultivation based and 16S amplicon-based analysis)