GNA5031 Applied Session 12 - TA's copy

Case study on the antibiotic resistome

Learning Objectives:

At the conclusion of this session, students should be able to:

- Examine different genomic sequence files and understand their content
- Query an antimicrobial resistance gene (ARG) database to identify likely ARGs within genomes and plasmids
- Use BLAST to identify an organism by its 16S rRNA genome sequence
- Explain the benefits and implications of searching chromosomes compared to plasmids for ARGs

Software required: `DIAMOND`, `seqkit`, text editing software (`BBedit` for Mac, `Notepad++` for PC), `Microsoft excel`

Introduction

The antibiotic resistome refers to the collection of all antimicrobial resistance genes (ARGs) in both pathogenic and non-pathogenic bacteria. To understand the resistome of a particular environment or setting, metagenomic sequencing can be used to examine the genomes of all microbial community members and profile the ARGs they contain. However, depending on the depth of sequencing and the success of metagenome assembly efforts, with this approach it may be challenging to link ARGs with the organisms that have them, to verify whether the presence of such genes corresponds to an antibiotic resistance phenotype, and to understand transmission dynamics.

In this case scenario, a collaborator has been studying antimicrobial resistant bacteria in agricultural runoff from a dairy farm. Recent infections in the cows have not resolved despite the farmer's use of antibiotics, and there is concern that the farm's runoff may be a hotspot for antimicrobial resistance that may be entering nearby waterways.

Using a range of antibiotic selection agar plates to culture bacteria from the runoff, your collaborator has isolated a number of bacterial strains that grow in the presence of antibiotics: specifically, they have isolated these strains on agar plates containing imipenem, an antibiotic belonging to the class **carbapenems**. Three of these isolates exhibited high levels of resistance and your collaborator wishes to investigate them further. They have done some initial analyses and provided you with three files for each isolate:

1. They initially conducted 16S rRNA sequencing of each isolate they found, to taxonomically identify them. They have provided you with the 16S rRNA sequence for each of the three concerning isolates in nucleotide format. `_16S.fna`
2. They have sequenced the genomes of these isolates and have provided you with the set of annotated genes from each genome. `_genome.fna`
3. To examine plasmid-borne resistance, they have isolated and purified plasmids from these isolates, and provided their sequences in nucleotide format. `_plasmid.fna`

In this workshop you will analyse this data to identify the ARGs in these isolates that are likely to confer their resistance phenotypes, and consider the implications of this resistance, so that you may provide direction to your collaborator.

Part 1: Identify the isolates and search their genomes for ARGs

Login to virtual machine

- list of VMs
- Instructions for VM login in detail (example from week 2)

```
ssh -x gnii0001@gna5031s1-gnii0001-01.rep.monash.edu # currently na
# next type in your passwords
```

Obtain data and software

Once logged in, use the following commands to obtain software and data.

```
conda activate gna5031

conda install -c bioconda diamond

diamond help # test if diamond has been installed

conda install -c bioconda seqkit
seqkit -v # test if seqkit has been installed

conda install pandas
pip list | grep pandas # Check pandas is installed

git clone https://github.com/ganiatgithub/GNA5031_applied12.git #
obtain all information needed for this session.
```

Inpect data

Let's start by checking the files you have been given. Use the following scripts and understand their contents.

```
cd data
head A_genome.fna
less A_genome.fna
seqkit stats A_genome.fna
```

It would be helpful to know which organisms we're dealing with, so let's check those 16S rRNA sequences to identify what these organisms are. Head to NCBI BLASTn. Copy the sequence into the query box, and change the search to BLAST against rRNA/ITS databases – specifically 16S ribosomal RNA sequences. You can leave the rest of the options at the default settings, and run the search. Examine the results.

Exercise 1

1. How many gene sequences are there in each genome, and how many gene sequences are their plasmids?

model answer

- A genome: 6522
- B genome: 5779
- C genome: 2911

A_plasmid has 539 B_plasmid has 141 C_plasmid has 5

2. What is the taxonomic identity of each of your isolates based on their 16S rRNA gene sequence? How similar is it to known reference strains?

model answer Genome A is Pseudomonas aerugonosa, 98 - 100% identical to reference strains (Pseudomonas aeruginosa HS18-89) Genome B is Klebsiella pneumoniae, 98-99% identical to reference strains (Klebsiella pneumoniae HS11286) Genome C is a Staphylococcus species, it's similar to several strains of aureus, but there's other hits with very similar sequences too. (Staphylococcus aureus Gv51)

Now we know what species these isolates are, we want to search their genes to determine if any of them are ARGs. For this we use DIAMOND, a sequence alignment tool which works similarly to BLAST but runs faster when searching a lot of sequences.

A copy of the Comprehensive Antibiotic Resistance Database (CARD) has been provided to you. CARD is a rigorously curated collection of known resistance determinants and associated antibiotics. To use it with DIAMOND, we need to make a DIAMOND database – a version of the database that can be recognised and searched by DIAMOND.

```
diamond makedb --in CARD.faa -d CARD.dmnd
```

Then, we can run DIAMOND using the new CARD file as a database. We are using the blastx function from DIAMOND, because we are comparing a nucleotide query (the genes from the isolate) with a protein database (CARD). The blastx command does this by translating the DNA sequence to a protein sequence in all 6 open reading frames, and aligns them to the protein database. To avoid running it three times, we'll run it in a loop on all three isolate files:

```
for isolate in *_genome.fna
do
  if [ -s "$isolate" ]; then
    echo "Processing $isolate"
    name="$(basename -- "$isolate" | sed 's/_genome.fna//')"
    echo "Name is $name"
    diamond blastx -d CARD.dmnd -q "$isolate" -o
results/"${name}_genome_CARD_results.txt" --outfmt 6 --max-target-
seqs 1 --max-hsps 1 --id 70
    fi
done
```

What's happening?

This command begins by running a loop on all three isolate files: for each one (which ends in .faa), do the following:

What it does for each file is define the “name” variable, which is simply the sample name. The basename command strips any directory names that might come before the filename, then sed removes the .faa extension. This leaves us with the isolate name in the $name variable, which we use to name the output files. Then, the diamond blastp command takes our new CARD.dmnd database and our isolate proteins, and searches those proteins in the database. The --out flag allows us to name the output file (which uses the isolate name, coming from the $name variable we defined above). We add a few additional options to the end to control our output:

- --outfmt 6: DIAMOND has several options for its output, this is a tabular format. See Glossary for details.
- --max-target-seqs 1: DIAMOND will output only one best hit when a protein matches something in the database
- --max-hsps-1: DIAMOND will output only one best ‘high scoring pair’ per alignment. This prevents matches appearing twice if the query protein happens to align equally well in more than one place on the same database protein.
- --id 70: This controls the minimum percentage identity between the query and database proteins. For proteins, 70% is considered a strict match to allow ARGs less similar to the reference sequences to be picked up, but not many proteins that are too distant from these ARGs.

The loop is closed with done, and you should have one DIAMOND results file for each isolate, named appropriately. Have a look at each of these files with `cat A_genome_CARD_results.txt`. What information do you find useful?

Many ARGs seem to be identified, but how to further interpret the data?

We have a helper tool: `annotate.py`, which uses the CARD Short Name from blast output (such as `A_genome_results.tsv`) to query the `CARD_metadata.tsv`, to obtain information such as Drug Class and Resistance Mechanism, summarised in such as `A_genome_summary.tsv`

 alt

To run this script:

```
./annotate.py ./results/A_genome_CARD_results.txt CARD_metadata.tsv
./results/A_genome_CARD_summary.tsv
./annotate.py ./results/B_genome_CARD_results.txt CARD_metadata.tsv
./results/B_genome_CARD_summary.tsv
./annotate.py ./results/C_genome_CARD_results.txt CARD_metadata.tsv
./results/C_genome_CARD_summary.tsv
```

Exercise 2

1. How many ARGs have been identified in each genome, and how similar are they to known reference ARG sequences?

model answer

- A: 59 ARGs, all above 70% threshold cut off, some are 100% identicial to what has been curated in CARD.
- B: 62 ARGs, all above 70% threshold cut off, some are 100% identicial to what has been curated in CARD.
- C: 28 ARGs, all above 70% threshold cut off, some are 100% identicial to what has been curated in CARD.

2. Summarise the ARGs for each isolate and the class of antibiotics they confer resistance to.

model answer

- A: disinfecting agents, aminocoumarin, **carbapenem**, diaminopyrimidine, monobactam, tetracycline
- B: disinfecting agents, aminocoumarin, **carbapenem**, peptide antibiotic, fluoroquinolone
- C: disinfecting agents, aminoglycoside antibiotic, glycylcycline;tetracycline antibiotic
  Note that disinfecting agents are not a class of antibiotic, but the presence of efflux pumps can make bacteria more resilient to disinfection as well as antibiotics and other drugs.

3. Based on your results, does this information match the information you received from your collaborator? Is anything unclear?

model answer

All three isolates contain resistance to multiple antimicrobial mechanisms. Isolates A and B both appear to be resistant to carbapenems, which coincides with what our collaborator has reported (that they isolated them on imipenem plates). However, while isolate C also grew on imipenem and has therefore demonstrated resistance to it, it has no genes related to carbapenem resistance in its genome.

Part 2: Identifying plasmid-borne ARGs and making recommendations

During your analysis you notice something strange – even though the isolate C came from antibiotic selection on imipenem, it doesn't seem to contain any ARGs corresponding to carbapenem resistance. What's going on? Let's do some forensic bioinformatics.

First, let's rule out a simple mistake – perhaps your collaborator got the files mixed up and they've given you the gene sequences from the wrong organism. Let's check some of the genes from the isolate to confirm that it is what we think it is.

Exercise 3

Take the first few proteins in the file: `head -n 10 C_genome.fna`

Copy these and check them in NCBI BLASTn. This time, leave all of the settings at their default.

model answer

The closest match correspond with the 16S rRNA sequence file that identifies this isolate as Staphylococcus aureus. Therefore, there shouldn't be a mix up.

It's likely that the resistant phenotype in this isolate comes from an ARG on the plasmid. Let's check all the plasmids for ARGs. Using the CARD database we made, let's query it with the plasmid sequences the collaborator provided. These are nucleotide sequences (you can see this when you check the file with head), so again we use the blastx command.

```
for isolate in *_plasmid.fna
do
  if [ -s "$isolate" ]; then
    echo "Processing $isolate"
    name="$(basename -- "$isolate" | sed 's/_plasmid.fna//')"
    echo "Name is $name"
    diamond blastx -d CARD.dmnd -q "$isolate" -o
results/"${name}_plasmid_CARD_results.txt" --outfmt 6 --max-target-
seqs 1 --max-hsps 1 --id 70
    fi
done
```

This command works exactly the same as the blastx command before. Use less to view the results for each plasmid. Again, we are using the annotate.py to contextualize the results:

```
./annotate.py ./results/A_plasmid_CARD_results.txt CARD_metadata.tsv
./results/A_plasmid_CARD_summary.tsv
./annotate.py ./results/B_plasmid_CARD_results.txt CARD_metadata.tsv
./results/B_plasmid_CARD_summary.tsv
./annotate.py ./results/C_plasmid_CARD_results.txt CARD_metadata.tsv
./results/C_plasmid_CARD_summary.tsv
```

Exercise 4

1. What information did you find in the plasmids of these organisms?

model answer

A: The plasmid of A contains broadly similar categories of antimicrobial resistance to its genome. Therefore, A is likely a potent multi-drug resistant bacterium that has potential to transfer its resistance to other organisms. B: The plasmid of B contains minimal antimicrobial resistance, so has less capacity to transfer this to other organisms, but its genome indicates it is still a multi-drug resistant organism and would be problematic if this is pathogenic. C: Carbapenem resistance genes have been identified in the plasmid of C, which likely suggests that its resistance to carbapenem as reported by our collaborator is due to horizontally acquired carbapenem resistance gene.

2. Let's look more closely at the few ARGs from the plasmid of isolate C, with `cat C_plasmid_card_results.txt`. What is the taxonomy associated with the CARD database hit for the ARGs in these plasmids? (Hint: search for the ID or name in the CARD.faa file.) Is this the organism you expected? Is that the case for all plasmids?

model answer Three of the four belong to *Staphylococcus aureus*. One of them most closely matches something else entirely (Limosilactobacillus reuteri) so this could indicate some transfer of genes, or a homologous gene that is more divergent from known Staph aureus resistance genes. If the students check some genes from other plasmids, sometimes the gene

belongs to the same species as the isolate they have - sometimes not. This is generally indicative of horizontal gene transfer, similar genes carried by different organisms, or the 'best match' being a different organism because the gene in their isolate is more novel or distant from other known genes in that species.

3. If you had metagenomic data in addition to these isolates, how would you make use of both datasets? What further analyses could be done to assist your collaborator in understanding where the antimicrobial resistance is coming from, and how they may recommend the farmer to approach treatment for the animals?

model answer Many answers would be acceptable here. Metagenomics can help us understand the full resistome in that environment (i.e. not just organisms that were culturable by our collaborator) and this information could help to streamline or target further surveillance for the most critical pathogens and AMR. Genomic information from the cows with antibiotic-resistant infection will help to determine what is causing their infections, and assessment of those ARGs followed by laboratory confirmation of resistance and susceptibility phenotypes will guide the usage of antibiotics at the farm. With metagenomes and isolate genomes, including plasmid sequences, you could determine which ARGs are at risk of being horizontally transmitted. It is concerning that multiple organisms with multi-drug resistance are present in the agricultural runoff from this farm, as it could enter waterways, infect people who work at or visit the farm, or be transmitted via dairy products from the farm - interventions are likely needed to curb transmission along several different pathways.

Glossary

Diamond Tabular Output format

The following fields are the column headers of Diamond Blast tabular output format 6:

`qseqid sseqid pident length mismatch gapopen qstart qend sstart send evalue bitscore`

Explanation:

- `qseqid`: query or source sequence id
- `sseqid`: subject or reference sequence id
- `pident`: percentage of identical positions
- `length`: alignment length (sequence overlap)
- `mismatch`: number of mismatches
- `gapopen`: number of gap openings
- `qstart`: start of alignment in query
- `qend`: end of alignment in query
- `sstart`: start of alignment in subject
- `send`: end of alignment in subject
- `evalue`: expect value
- `bitscore`: bit score

Carbapenem

Carbapenems, among the beta-lactams, are the most effective against Gram-positive and Gram-negative bacteria presenting a broad spectrum of antibacterial activity. Carbapenems are considered to be the most reliable last-resort treatment for bacterial infections, therefore, the emergence and rapid spread through all continents of carbapenem resistance, mainly among Gram-negative bacteria, constitutes a global public-healthcare problem of major importance. Meletis 2016

Tetracycline

Tetracycline antibiotics are well known for their broad spectrum of activity, spanning a wide range of Gram-positive and -negative bacteria, spirochetes, obligate intracellular bacteria, as well as protozoan parasites. Several of tetracyclines remain in clinical use for the treatment of uncomplicated respiratory, urogenital, gastrointestinal, and other rare and serious infections; however, the dissemination of tetracycline-resistant mechanisms has narrowed their utility, limiting use to only infections with confirmed susceptibility. Grossman 2016