Identification of Antimicrobial Resistance Genes - A technical report submitted to HackBio in partial fulfillment of the Genomics workshop completion requirement.

By Aanuoluwa Adekoya - Team Crick/Project1
(Link to GitHub Repo)

Introduction

Antimicrobial resistance occurs when microbes - bacteria, fungi, viruses, and parasites are able to evade the effectiveness of antimicrobials used to combat the infections caused by these microbes. It takes full effect when a mutation occurs in the genome of the microbe and it is able to synthesize proteins that can perform this function. According to WHO, antimicrobial resistance remains a threat to the management of diseases as it threatens the effective prevention and treatment of an ever-increasing range of infections caused by microbes. Despite active research in this area of biology, microbes have developed resistance to almost all of the commonly used antimicrobials and the mechanism used is specific to each microbe.

Aim: This project aims to identify the beta-lactam, tetracycline, and aminoglycoside resistance pattern present in *Staphylococcus aureus* in five countries in the world. Our results will provide clues on the resistance pattern in each of these countries and the types of antibiotics that should be prescribed to patients infected by *S. aureus* in these regions.

Importance: Despite the relief that the discovery of antimicrobials and the advent of antimicrobial therapy brought, antimicrobial resistance has remained a threat to the core of modern medicine and global public health. Knowledge of the prevalence of antimicrobial resistance genes will help in optimizing antimicrobials usage in medicine.

Methodology

Data collection: Whole Genome Sequencing (WGS) of *S. aureus* from five countries - Argentina, Canada, China, Nigeria, and Russia were retrieved. The SRR/ERR IDs, that is, the run accessions for the actual sequencing data of each experiment to be used per country were retrieved from NCBI (https://www.ncbi.nlm.nih.gov/guide/sitemap) a public biological database. The SRR/ERR IDs were then used to retrieve the datasets and their metadata from the Sequence Read Archive (SRA), an open and free access repository of high throughput sequencing data (https://sra-explorer.info/). The selected countries in this project represent the most populated or second most populated countries in five continents - South America, North America, Asia, Africa, and Europe respectively and ninety-eight reads were retrieved in total.

Computation: (FastP, SPAdes, ResFinder, Matplotlib-Python)
Retrieved data were screened for quality control and fastp (Chen S., *et al*., 2018; https://github.com/OpenGene/fastp/blob/master/README.md) was used to trim adapter

sequences. The trimmed sequences were assembled using the genome assembler - SPAdes (Bankevich *et al*., 2012; https://github.com/ablab/spades). Post assembly, the contigs of each sequence were renamed to have their SRR/ERR IDs for easy identification and they were then uploaded to ResFinder (Florensa *et al*., 2022; https://cge.cbs.dtu.dk/services/ResFinder/), an open online resource for identification of antimicrobial resistance genes in high-throughput sequencing data and prediction of phenotypes from genotypes. A 90% ID threshold was used alongside a 60% minimum length. The resistance genes for beta-lactam, tetracycline, and aminoglycosides were compiled into a CSV file.

- A bash script containing the curl command alongside the URL to the SRA for each sequence was used to **get** the dataset into the Linux pipeline for further computation.
- To make the sequences accessible, a text file containing the sequence names was created and called in every future command.
- Another bash script was written to contain a for loop to execute **FastP** on all of the sequences.

```
#!/bin/bash

mkdir argentinafastp_reads

for SAMPLE in $(cat argentina.txt) ; do

fastp \
   -i "$PWD/${SAMPLE}_1.fastq.gz" \
   -I "$PWD/${SAMPLE}_2.fastq.gz" \
   -o "argentinafastp_reads/${SAMPLE}_1.fastq.gz" \
   -O "argentinafastp_reads/${SAMPLE}_2.fastq.gz" \
   --html "argentinafastp_reads/${SAMPLE}_fastp.html"
done
```

- The next step was to run SPAdes on the trimmed sequences and a bash script for SPAdes was written inside the directory of the trimmed reads. The text file containing the sequence names was copied into this directory as well. We specified that the name of the directory for each sequence's SPAdes result should be the sequence's name itself.
  - ○
```
#!/bin/bash


for sample in $(cat argentina.txt) ; do
             spades.py   -1   "$PWD/$sample"_1.fastq.gz   -2   "$PWD/$sample"_2.fastq.gz   -o
/home/einstein/project1/Aanuoluwa/Argentina/argentinafastp_reads/argentina_assembly/$sample
 done
```

- We needed the contigs.fasta file for each sequence to input in the RESfinder database to search for antimicrobial resistance genes, however, all of the sequences had this file named the same way so it was necessary that we rename each sequence's contig with its original sequence identity. We wrote a bash script within the genome assembly directory to rename each of these contigs and copy them into one contigs directory for easy download.
  - ○
```
#!/bin/bash

#rename all the contigs.fasta files with their SRR IDs.
```

```
for x in */contigs.fasta; do
   d=$(dirname "$x")
   p=$(echo $d | cut -d_ -f1)
   mv "$d/contigs.fasta" "$d/$p.fasta"
   cp "$d/$p.fasta"  ./argentina_contigs
done
```

- The renamed contigs.fasta files were downloaded and then uploaded onto the ResFinder webpage to screen for the presence of antimicrobial resistance genes to beta-lactams, tetracycline, and aminoglycosides.
- The compiled result from ResFinder was visualized using python’s matplotlib module. The table containing the final distribution was uploaded to google collab (https://colab.research.google.com/) and the data frame was called as ‘data’. The country column was used as the index and a bar graph comparing the distribution of the resistance genes in the five countries was plotted.

```
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline

data=pd.read_csv(‘/country_stat’)
data.head(7)

data2=data.set_index('Country')
data2.head(7)
data2.plot.bar(figsize=(8,8))
plt.title('ARG Ditribution in the 5 Countires')
plt.ylabel('Frequency')
plt.xlabel('Country')
plt.savefig('Country_stat.png')
```

- Since China had the highest distribution of the selected antimicrobial resistance genes, we plotted a pie chart to show the distribution of these genes in China and it was shown that penicillin and cefoxitin resistance genes were the most prevalent in the 20 samples.

```
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline

china = pd.read_csv('/content/China Stat.csv')
china.head()
china2 = china.set_index('ARG')
china2.head(9)
fig, ax = plt.subplots(figsize=(8, 8))
ax.pie(china2, labels=labels, autopct='%.1f%%', wedgeprops={'linewidth': 3.0, 'edgecolor': 'white'},
       textprops={'size': 'x-large'})
ax.set_title('China ARG Distribution', fontsize =20)
plt.savefig('China_pie.png')
```

Note: In the Linux pipeline, the datasets were worked on country by country, hence the directories were created based on country identity and the major codes remained the same.

Results and Discussion:

This analysis revealed a prevalence of *mecA*, a beta-lactam resistant gene against cefoxitin, and *blaZ*, a beta-lactam resistant gene against penicillin, and variants of *tet* genes against tetracycline in the five countries (Link to SRR IDs of the dataset and the distribution of the ARGs where 1 means positive and 0 means negative). The prevalence of penicillin-resistant genes confirms the report of previous studies that have addressed the reduced potency of penicillin as a drug of choice. Reads from China (Link to the table showing country by the country distribution of genes) contained a high amount (up to 50%) of aminoglycoside resistance genes which were specific to amikacin, tobramycin, and gentamicin when compared to the other countries (Fig. 1, Fig 2). The project showed that while aminoglycosides can be used as a drug of choice in combating staphylococcal infection in Russia, Canada, Argentina, and Nigeria, the use of beta-lactams should be reduced to avoid the spread of antimicrobial resistance genes. Table 2 (Link to SRR IDs of the dataset and the distribution of the ARGs where 1 means positive and 0 means negative) shows the occurrence of each resistance gene in all of the sample datasets. It is worthy to note that the resistance genes in each country's dataset were of the same or similar origin and these had between 99.89% and 100% identity. This project also showed that there is a need for more optimization of antimicrobial use in China and countries alike to limit the spread of antimicrobial genes.

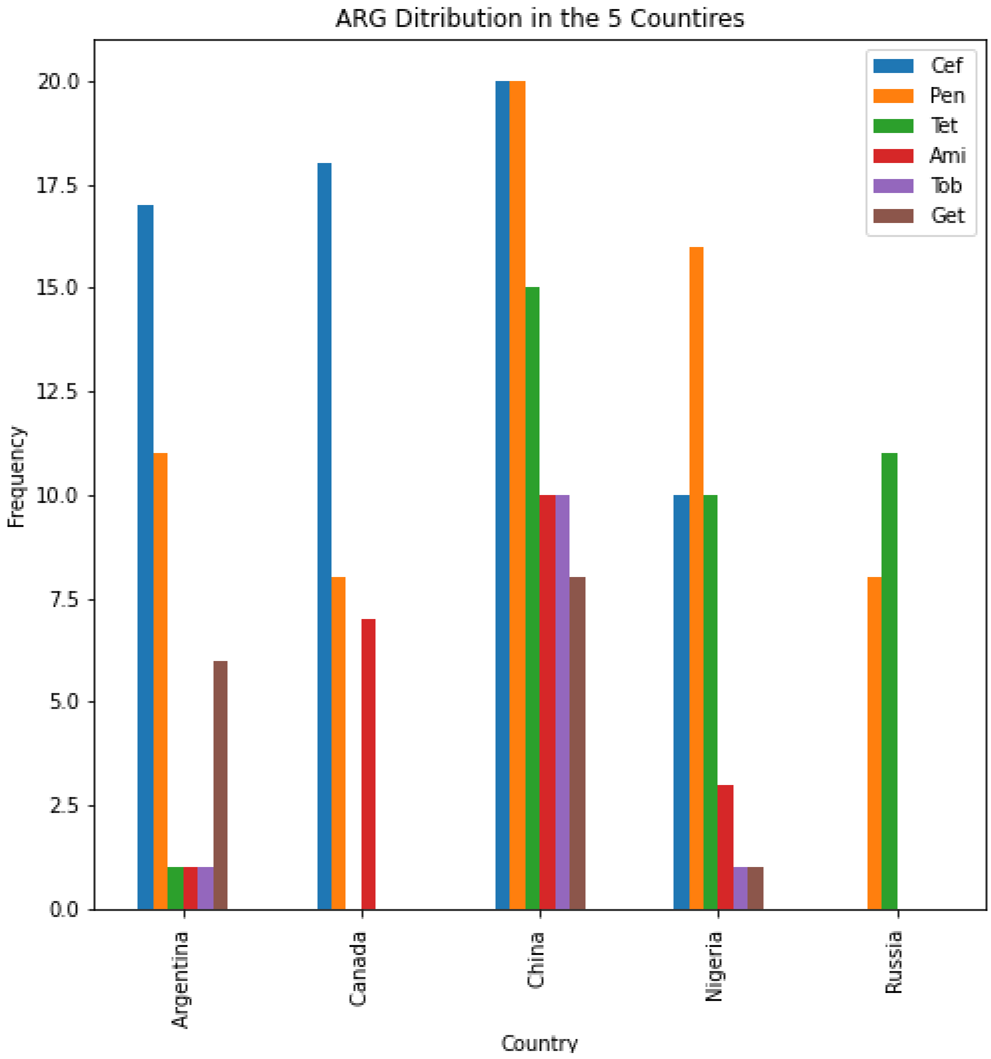

Fig. 1: Prevalence of antimicrobial resistance genes in sample reads. Cef=Cefoxitin, Pen=Penicillin, Tet=Tetracycline, Ami=Amikacin, Tob=Tobramycin, Get=Gentamicin. Sample size for Argentina and Nigeria = 19 each and Canada, China, and Nigeria = 20 each.

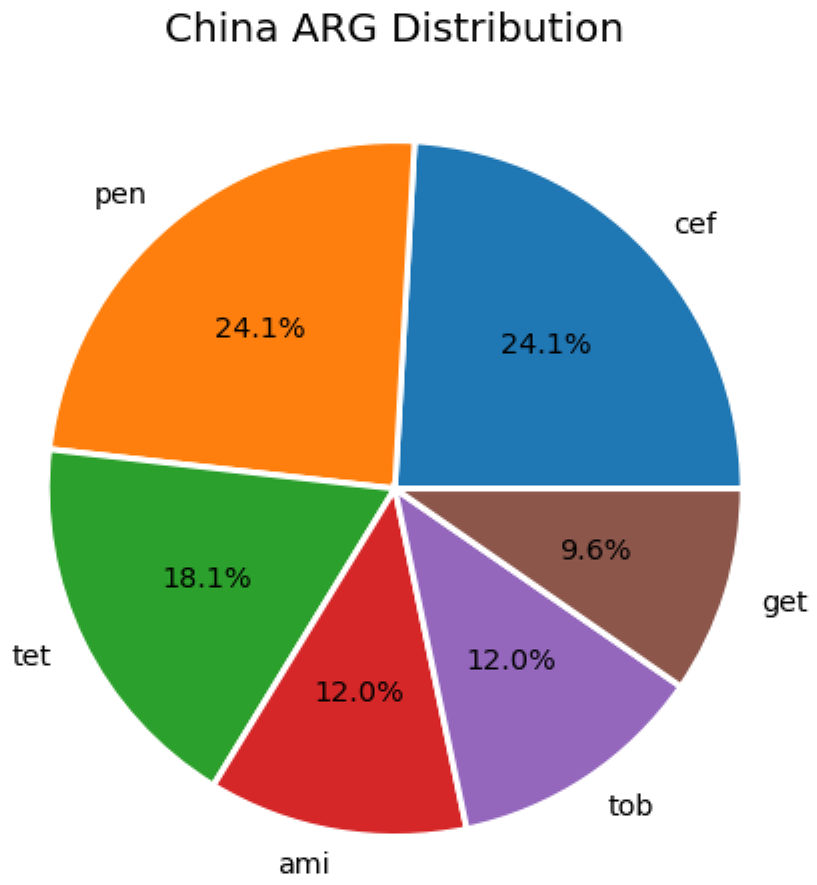

Fig. 2: Distribution of antimicrobial resistance genes in China sample reads. Cef=Cefoxitin, Pen=Penicillin, Tet=Tetracycline, Ami=Amikacin, Tob=Tobramycin, Get=Gentamicin. Sample size = 20.

References

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A., &

Pevzner, P. A. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. Journal of computational biology : a journal of computational molecular cell biology, 19(5), 455–477. https://doi.org/10.1089/cmb.2012.0021

Florensa, A. F., Kaas, R. S., Clausen, P., Aytan-Aktug, D., & Aarestrup, F. M. (2022). ResFinder - an open online resource for identification of antimicrobial resistance genes in next-generation sequencing data and prediction of phenotypes from genotypes. Microbial genomics, 8(1), 000748. https://doi.org/10.1099/mgen.0.000748

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). Fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics (Oxford, England), 34(17), i884–i890. https://doi.org/10.1093/bioinformatics/bty560