Computational practical 8: Metagenomics & AMR analysis using CZ ID

Module Developers: Dr. Elizabeth Fahsbender

Table Of Contents

Introduction

Metagenomics is a powerful tool that can be used to define microbiomes, identify novel and unexpected pathogens, and characterize AMR genes. CZ ID is a free, open-source, no-code analysis platform for analyzing sequencing data. Researchers from anywhere in the world can analyze their data straight from their laptops, regardless of computational power or bioinformatic skills. Our publication can be found here.

The tool contains three modules:

Module	**Input data**	**Sequencing technology support**
Metagenomics (mNGS)	**Unbiased metagenomics data *(no amplicon data/16s/18s)***	**Illumina & Nanopore**
Antimicrobial resistance	**Whole genome sequencing & mNGS**	**Illumina**
SARS-CoV-2	Amplicon sequencing	Illumina & Nanopore
Viral consensus genome	Amplicon sequencing & WGS	Illumina

The web application provides visualizations and downstream analyses to help the user interpret their data. **This practical will focus on the AMR and mNGS modules.**

Learning Outcomes

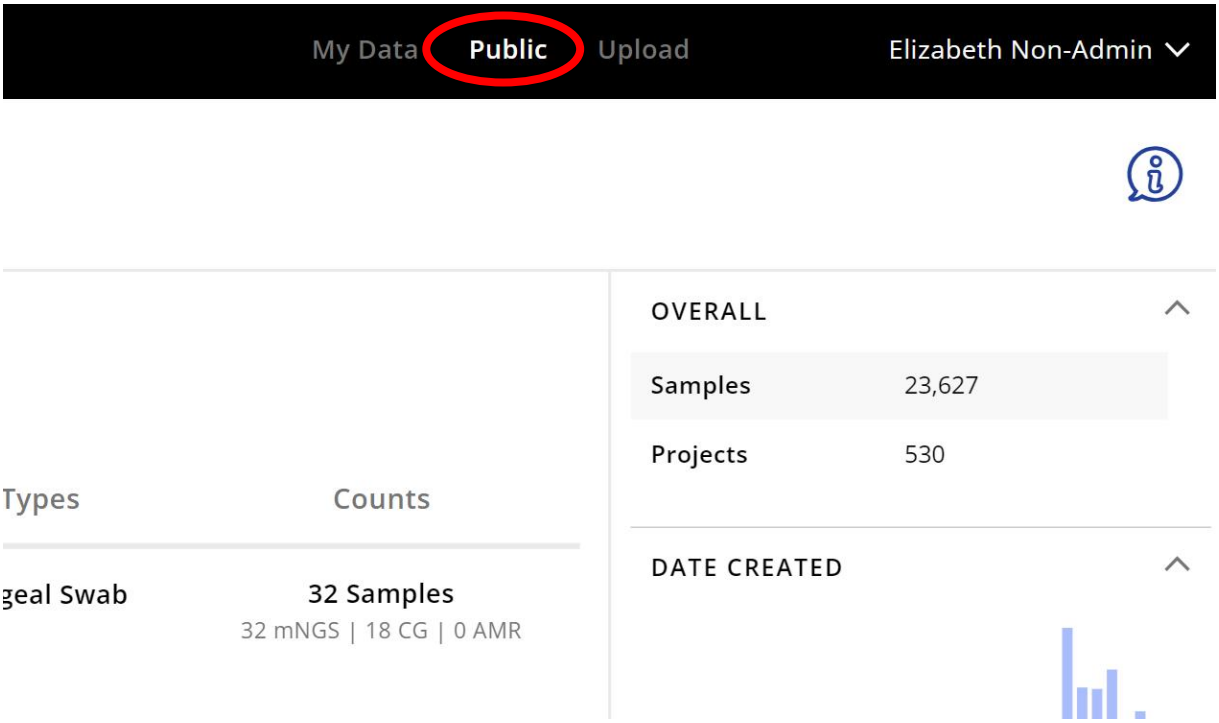
After completing this practical you will be able to:

- Navigate CZ ID successfully
- Perform an AMR analysis on individual samples
- Perform a metagenomic analysis on individual samples
- Take outputs from CZ ID into downstream analyses

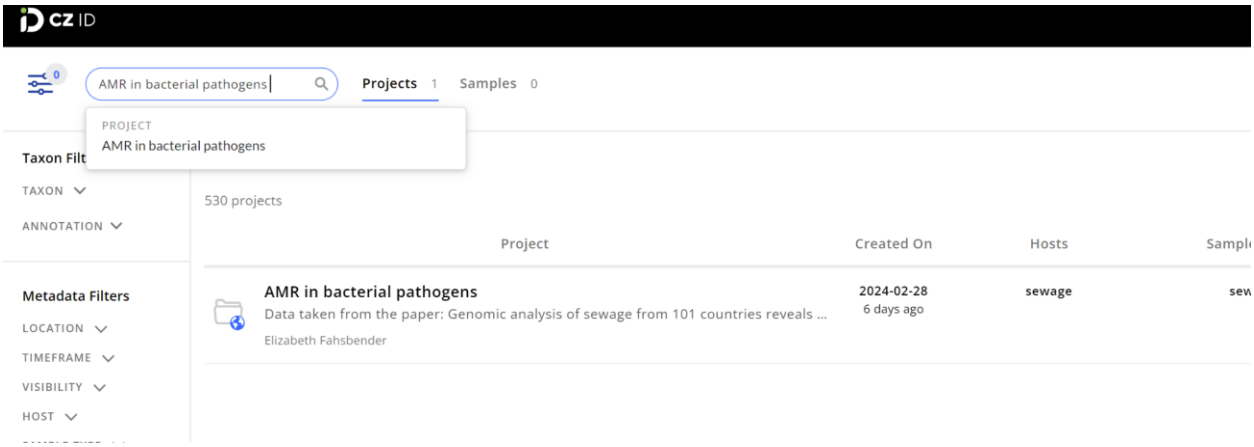
The raw fastq files have been uploaded to CZ ID and run through the metagenomic and AMR pipelines. You will be viewing the results.

Accessing CZ ID

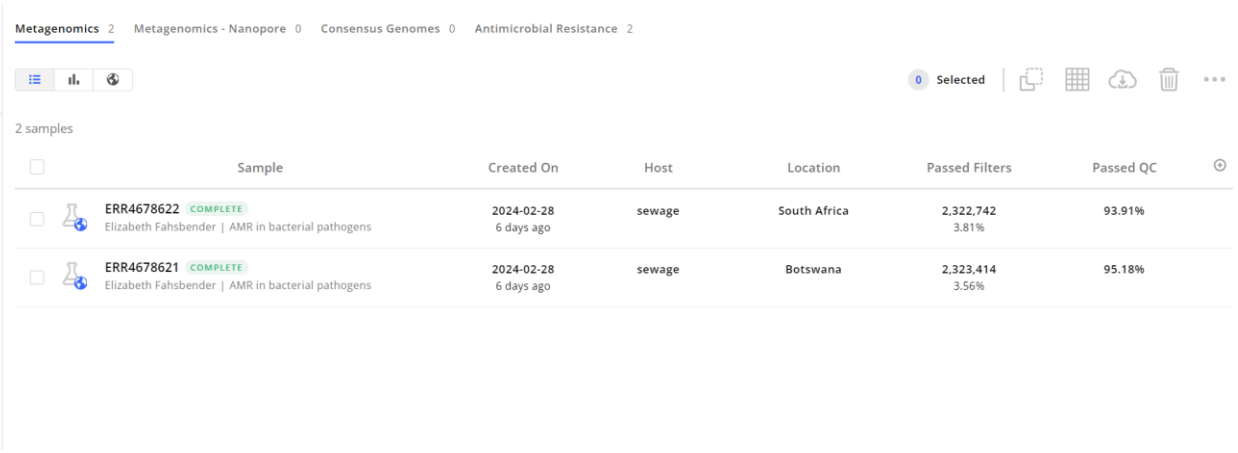
1. Go to czid.org
2. Click **sign in**, in the upper right corner.
3. Enter your email and password.
4. Click on the **public** tab in the upper right corner



5. Once you are in the public tab, type the project name **“AMR in bacterial pathogens”** in the search bar.



6. Click on the project.
7. You should see two samples



Viewing the metagenomic results

8. Click on sample ERR4678622. You will be in the metagenomic analysis tab. Here, you will see which taxa are present in your sample.

The sample report table provides information about taxa hits so you can determine the validity of the hit. You can read more about the sample report and how to interpret it here.

Report Metric	Definition
Score	Score is only calculated if a background model is applied. It is CZ ID's heuristic for ranking microbial hits. The score is intended to combine the following aspects of the evidence for a hit: (a) species-level information, (b) genus-level information, (c) information about relative abundance within the sample, (d) information about abundance relative to the chosen background controls. The score is calculated as follows: ((abs(genus NT Z) * species NT Z * species NT rPM) + (abs(genus NR Z) * species NR Z * species NR rPM))
Z	Z-score statistic is only computed if a background model is applied. It is used for evaluating the prevalence of microbes in your sample as compared to background controls. The Z-score is computed based on the specified background model.
rPM	Number of reads aligning to the taxon in the NCBI NT/NR database, per million reads sequenced
r	Number of reads aligning to the taxon in the NCBI NT/NR database
contig	Number of assembled contigs aligning to the taxon in the NCBI NT/NR database
contig r	Total number of reads aligning to all assembled contigs for this taxon
%id	Average percent-identity of alignments to NCBI NT/NR

L	Average length of the local alignment for all contigs and reads assigned to this taxon
E value	Average Expect value (e-value) of alignments to NCBI NT/NR. The Expect value (e-value) is a parameter that describes the number of hits one can "expect" to see by chance when searching a database of a particular size. The closer to 0 the better.

Metagenomics can be a semiqualitative tool, meaning that the more reads aligning to a specific taxa, the more abundant it most likely is in the sample. Of course, there are exceptions to the rule, but this is generally how you can think about taxa abundance.

In CZ ID the normalized value we use is reads per million (rPM). This normalized value takes sequencing depth into account so you can compare taxon abundance between samples.

9. Name the top 4 most abundant taxa at the genus level and how many reads per million (rPM) were sequences. Identify where they are found by clicking on the genus name. A modal will slide out on the right side of the screen with information about the genus and links to google, PubMed, and more.

Genus	**Where are they found? (Click on the genus name to learn more)**	**rPM**
1.		
2.		

3.		
4.		

10. Do the same for the next sample. Name the top 4 most abundant taxa at the genus level and how many reads per million (rPM) were sequences. Identify where they are found by clicking on the genus name. A modal will slide out on the right side of the screen with information about the genus and links to google, PubMed, and more.

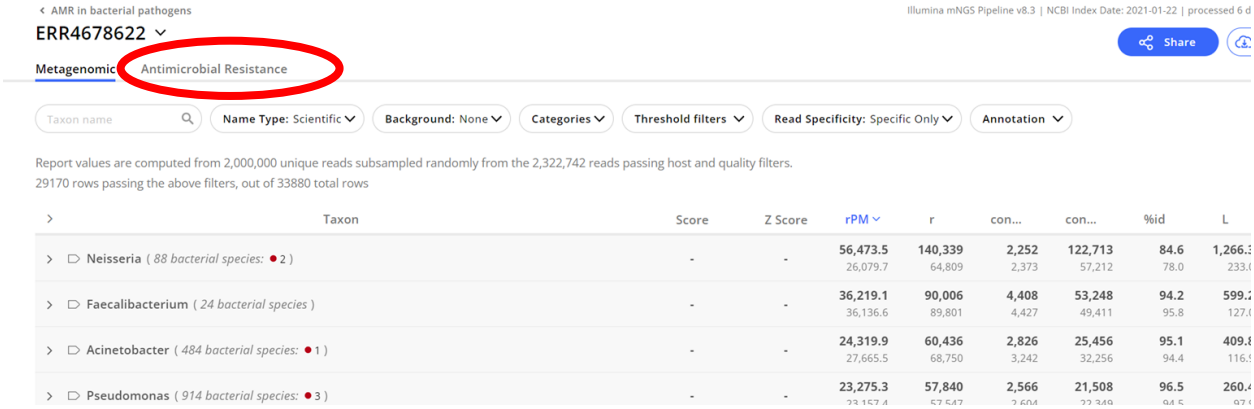
Genus	**Where are they found? (Click on the genus name to learn more)**	**rPM**
1.		
2.		
3.		

4.		

11. Of the top four taxa, which taxa appear in both samples? Which are different?

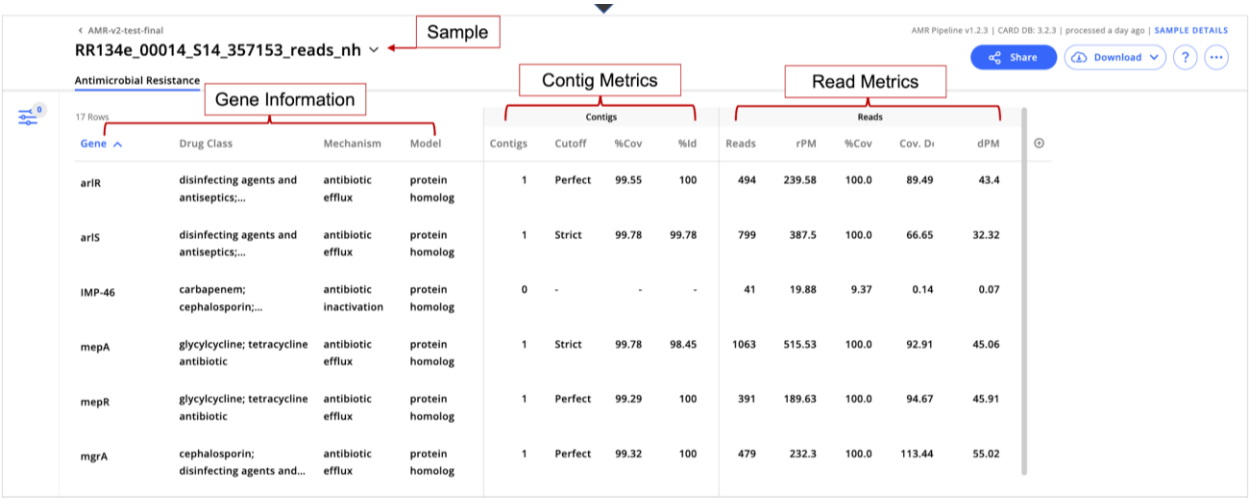
Viewing AMR results

12. Now, let's view which AMR genes are present. Switch to the AMR tab.



13. The table is organized by AMR gene detected. Similar to the metagenomics tab, if you click the AMR gene you can read more about it.
14. AMR Sample Report Layout
    a. By default, the AMR Sample Report Table will be sorted by the Gene column and show a set of columns containing AMR gene information and metrics for contigs

and reads. However, you can customize the number of columns, sort, and filter the table to suit your needs.



15. You may notice that some AMR genes are not supported by both contigs and reads. This is expected due to differences in the pipeline workflow for contigs and reads. See AMR FAQs for more details.

16. Take a look at the column headers. If you are unfamiliar with what a header is, hover over it to read the definition. Write the definition for the following:

Column Header	**Definition**
Mechanism	
Model	

Cutoff	

17. In a public health setting, using the column **"cutoff"** can be valuable to to identify perfect hits. As the name suggests, these hits are perfectly identical the the curated database. Sort the cutoff column by clicking it until all of the **perfect** hits are at the top.

18. How many perfect hits do you have?

19. Which **high-level drug classes** have perfect hits?

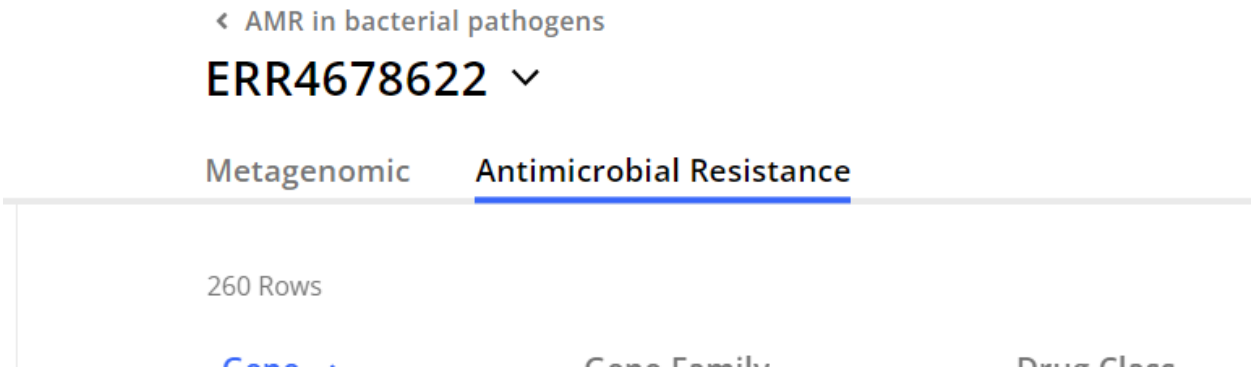
20. For the next sample, which **high-level drug classes** have perfect hits?

21. Which high-level drug classes with perfect cutoffs appear in both samples? Which are different?

Take your analysis off of CZ ID

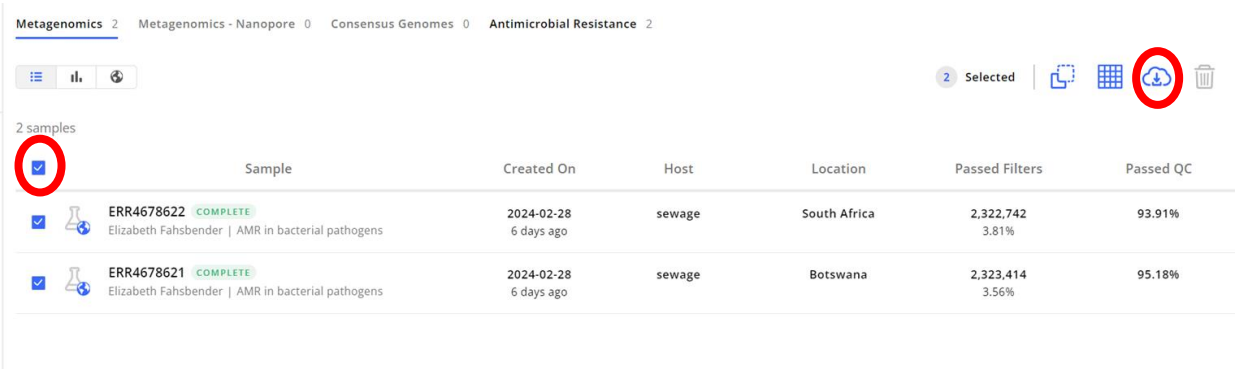
22. You may want to take your analysis off CZ ID to calculate diversity metrics, figures, or other downstream analyses.

23. To download metagenomic data. Go back to the project page by clicking on the project name **AMR in bacterial pathogens.**

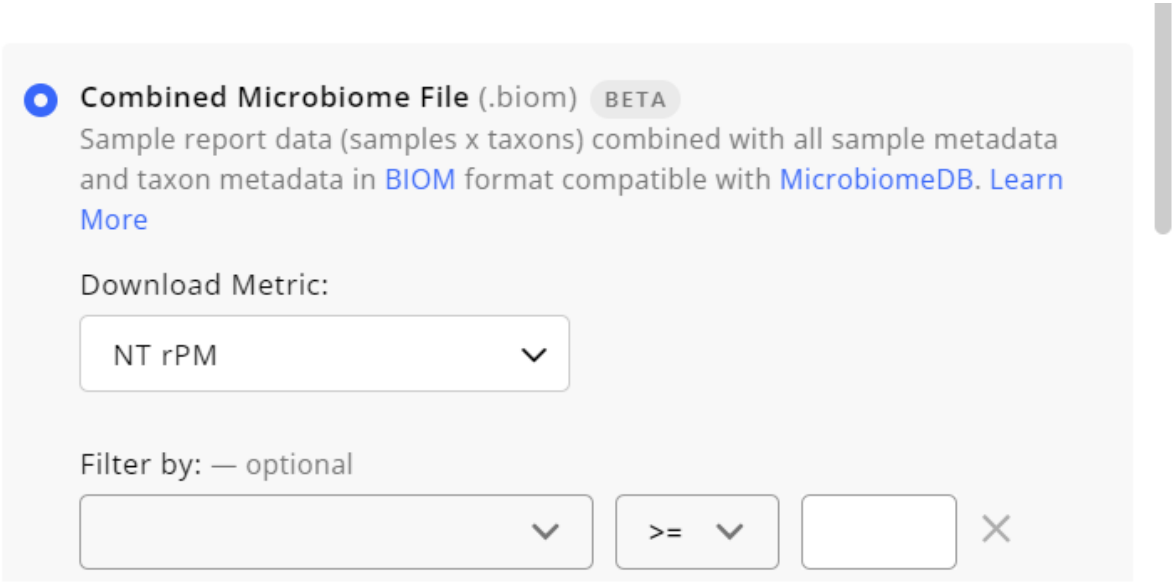


24. Click on the metagenic tab.
25. Select both samples by checking the box.

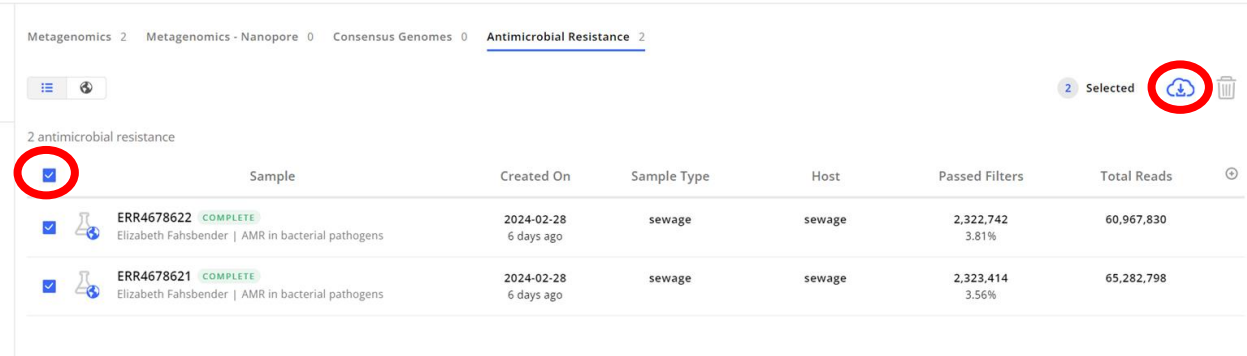
26. Select the cloud icon.



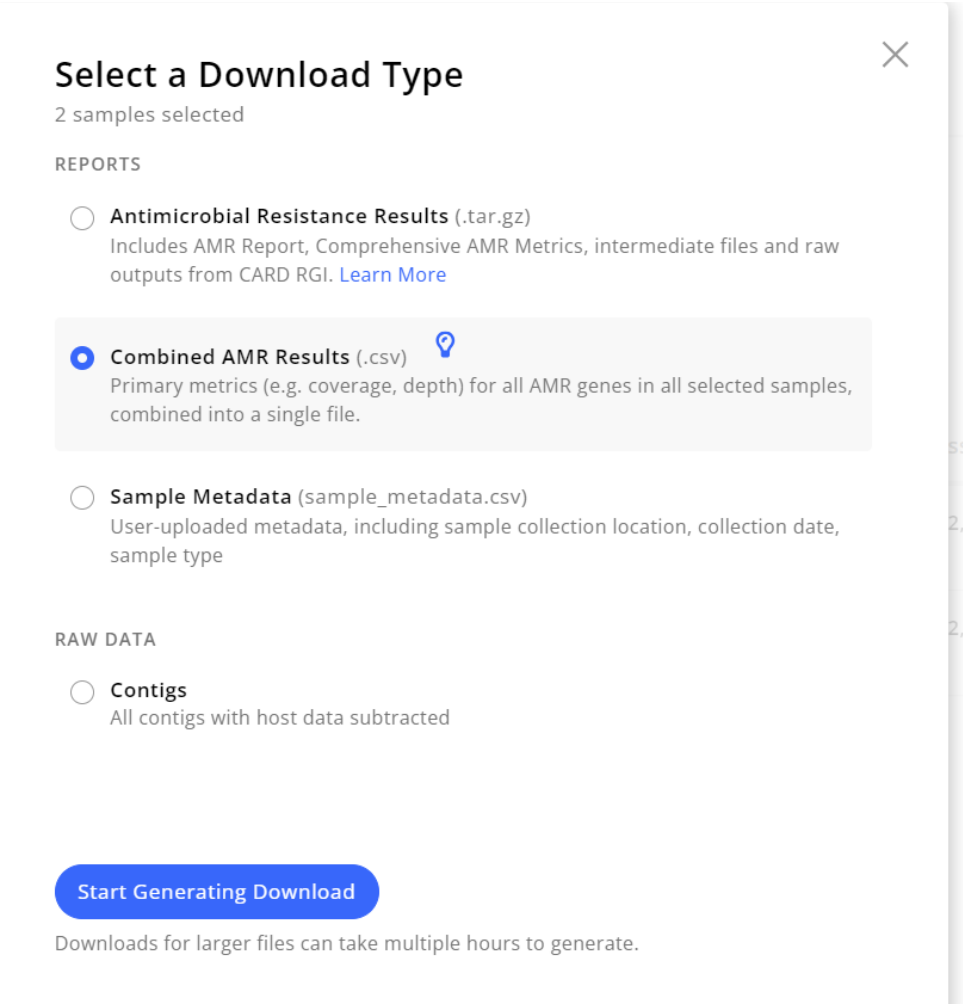
27. There are multiple downloads you can generate, but for calculating diversity metrics you can download the Biom file.



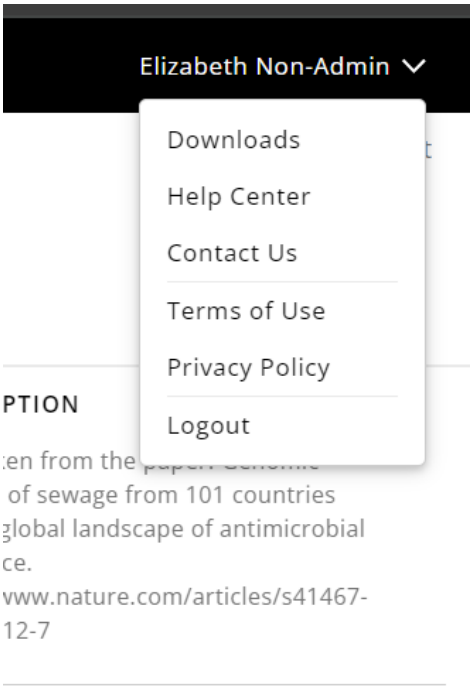
28. You can read about the different download types here.
29. To generate an AMR download, navigate back the AMR tab.
30. Check the box and click the cloud icon.



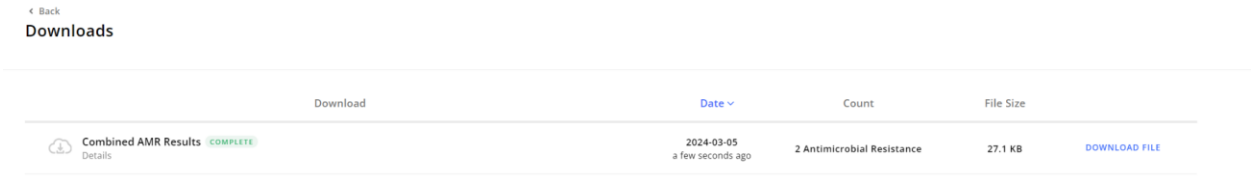
31. Choose **combined AMR results.**



32. Now click on your name in the upper right corner and click **downloads.**



33. Download those files locally by clicking **download file**. You can now use these inputs in other tools!



34. Access additional resources and tutorials here.