

Computational practical 8: Metagenomics & AMR analysis using CZ ID

Module Developers: Dr. Elizabeth Fahsbender

Table Of Contents

Introduction	1
Learning Outcomes.....	2
Accessing CZ ID	2
Viewing the metagenomic results.....	3
Viewing AMR results.....	5
Take your analysis off of CZ ID	6

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

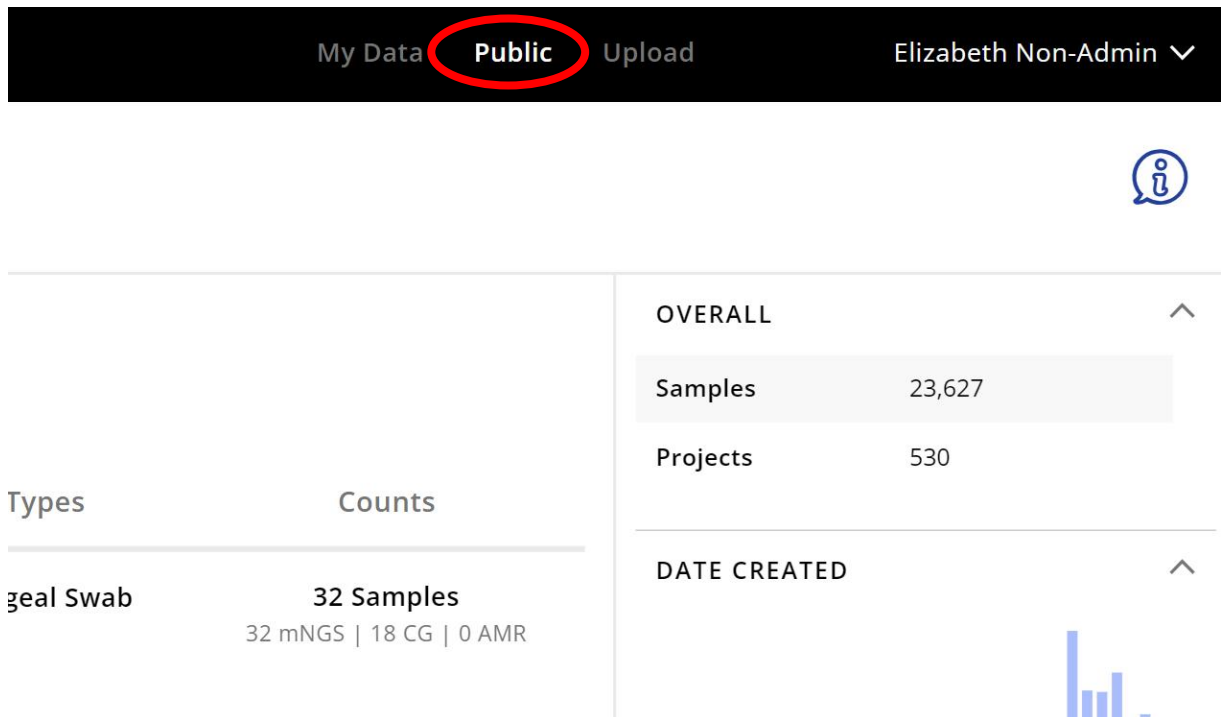
After completing this practical you will be able to:

- 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.

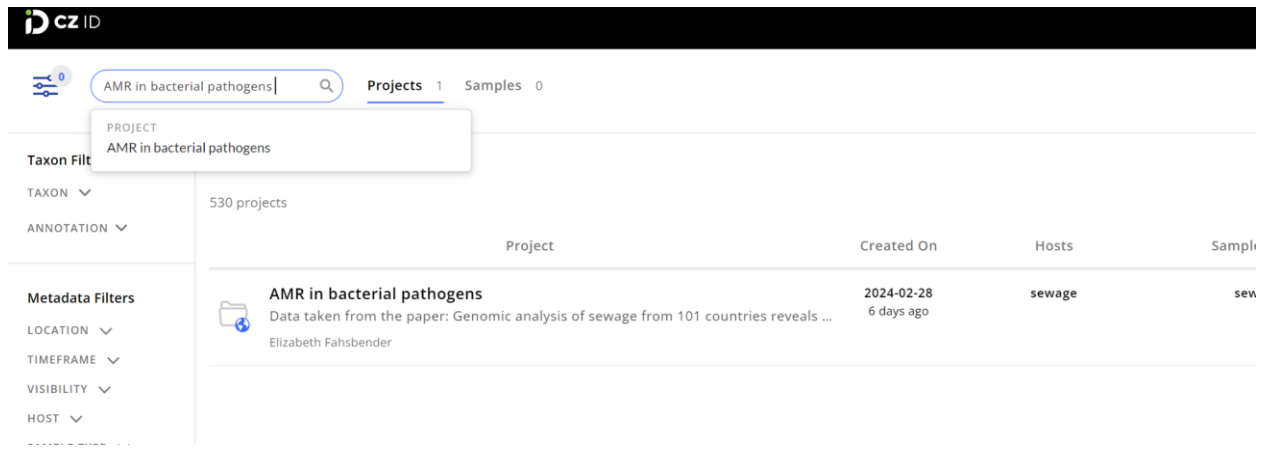
Accessing CZ ID

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



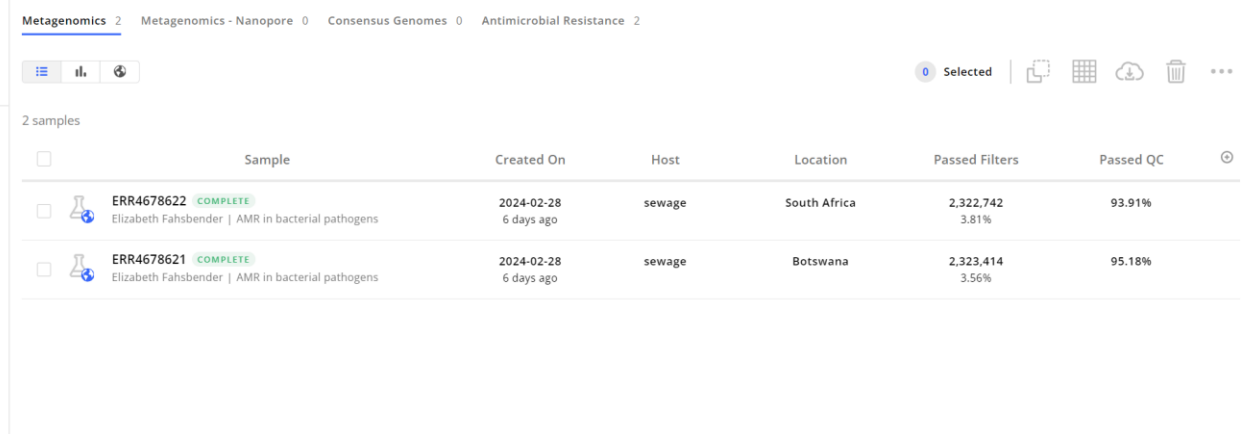
The screenshot shows the CZ ID Public tab interface. At the top, there is a navigation bar with 'My Data', 'Public' (highlighted with a red circle), and 'Upload' tabs. The user is logged in as 'Elizabeth Non-Admin'. Below the navigation bar, there is a table with two columns: 'Types' and 'Counts'. The table shows 'geal Swab' with '32 Samples' and '32 mNGS | 18 CG | 0 AMR'. To the right of the table, there is a summary section with 'OVERALL' and 'DATE CREATED' headers. The 'OVERALL' section shows 'Samples: 23,627' and 'Projects: 530'. A small bar chart is visible at the bottom right of the summary section.

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



The screenshot shows the CZ ID interface. At the top, there's a search bar with 'AMR in bacterial pathogens' entered. Below the search bar, a dropdown menu shows 'PROJECT AMR in bacterial pathogens'. On the left, there are filters for 'Taxon Filter', 'TAXON', 'ANNOTATION', and 'Metadata Filters' (LOCATION, TIMEFRAME, VISIBILITY, HOST). The main area displays a list of projects. The first project is 'AMR in bacterial pathogens', created on '2024-02-28' (6 days ago), with the host 'sewage' and a sample 'sew'.

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



The screenshot shows the 'Metagenomics' tab in the CZ ID interface. It displays two samples: 'ERR4678622' and 'ERR4678621'. Both samples are 'COMPLETE' and were created on '2024-02-28' (6 days ago). The host for both is 'sewage'. The location for 'ERR4678622' is 'South Africa' and for 'ERR4678621' is 'Botswana'. The 'Passed Filters' and 'Passed QC' columns show the percentage of samples that passed the respective filters and quality control.

Sample	Created On	Host	Location	Passed Filters	Passed QC
ERR4678622 COMPLETE Elizabeth Fahsbender AMR in bacterial pathogens	2024-02-28 6 days ago	sewage	South Africa	2,322,742 3.81%	93.91%
ERR4678621 COMPLETE Elizabeth Fahsbender AMR in bacterial pathogens	2024-02-28 6 days ago	sewage	Botswana	2,323,414 3.56%	95.18%

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: $((\text{abs}(\text{genus NT Z}) * \text{species NT Z} * \text{species NT rPM}) + (\text{abs}(\text{genus NR Z}) * \text{species NR Z} * \text{species NR rPM}))$
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 semiquantitative 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.

AMR in bacterial pathogens Illumina mNGS Pipeline v8.3 | NCBI Index Date: 2021-01-22 | processed 6 d

ERR4678622 ▼

Metagenomic **Antimicrobial Resistance** Share Download

Taxon name Name Type: Scientific ▼ Background: None ▼ Categories ▼ Threshold filters ▼ Read Specificity: Specific Only ▼ Annotation ▼

Report values are computed from 2,000,000 unique reads subsampled randomly from the 2,322,742 reads passing host and quality filters.
29170 rows passing the above filters, out of 33880 total rows

>	Taxon	Score	Z Score	rPM ▼	r	con...	con...	%id	L
>	Neisseria (88 bacterial species: ● 2)	-	-	56,473.5 26,079.7	140,339 64,809	2,252 2,373	122,713 57,212	84.6 78.0	1,266.3 233.0
>	Faecalibacterium (24 bacterial species)	-	-	36,219.1 36,136.6	90,006 89,801	4,408 4,427	53,248 49,411	94.2 95.8	599.7 127.0
>	Acinetobacter (484 bacterial species: ● 1)	-	-	24,319.9 27,665.5	60,436 68,750	2,826 3,242	25,456 32,256	95.1 94.4	409.8 116.5
>	Pseudomonas (914 bacterial species: ● 3)	-	-	23,275.3 73,167.4	57,840 67,647	2,566 7,604	21,508 77,349	96.5 94.5	260.4 97.1

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

- 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.

AMR-v2-test-final Sample RR134e_00014_S14_357153_reads_nh AMR Pipeline v1.2.3 | CARD DB: 3.2.3 | processed a day ago | [SAMPLE DETAILS](#)

[Share](#) [Download](#) [?](#) [...](#)

Antimicrobial Resistance				Contig Metrics				Read Metrics				
Gene Information				Contigs				Reads				
Gene	Drug Class	Mechanism	Model	Contigs	Cutoff	%Cov	%Id	Reads	rPM	%Cov	Cov. Di	dPM
arlR	disinfecting agents and antiseptics;...	antibiotic efflux	protein homolog	1	Perfect	99.55	100	494	239.58	100.0	89.49	43.4
arlS	disinfecting agents and antiseptics;...	antibiotic efflux	protein homolog	1	Strict	99.78	99.78	799	387.5	100.0	66.65	32.32
IMP-46	carbapenem; cephalosporin;...	antibiotic inactivation	protein homolog	0	-	-	-	41	19.88	9.37	0.14	0.07
mepA	glycylcycline; tetracycline antibiotic	antibiotic efflux	protein homolog	1	Strict	99.78	98.45	1063	515.53	100.0	92.91	45.06
mepR	glycylcycline; tetracycline antibiotic	antibiotic efflux	protein homolog	1	Perfect	99.29	100	391	189.63	100.0	94.67	45.91
mgrA	cephalosporin; disinfecting agents and...	antibiotic efflux	protein homolog	1	Perfect	99.32	100	479	232.3	100.0	113.44	55.02

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**.

< AMR in bacterial pathogens

ERR4678622 ▾

Metagenomic

Antimicrobial Resistance

260 Rows

Gene

Gene Family

Drug Class



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

26. Select the cloud icon.

Metagenomics 2 Metagenomics - Nanopore 0 Consensus Genomes 0 Antimicrobial Resistance 2

2 Selected

2 samples

Sample	Created On	Host	Location	Passed Filters	Passed QC
<input checked="" type="checkbox"/>  ERR4678622 COMPLETE Elizabeth Fahsbender AMR in bacterial pathogens	2024-02-28 6 days ago	sewage	South Africa	2,322,742 3.81%	93.91%
<input checked="" type="checkbox"/>  ERR4678621 COMPLETE Elizabeth Fahsbender AMR in bacterial pathogens	2024-02-28 6 days ago	sewage	Botswana	2,323,414 3.56%	95.18%

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

☒ **Combined Microbiome File (.biom)** BETA

Sample report data (samples x taxons) combined with all sample metadata and taxon metadata in **BIOM** format compatible with **MicrobiomeDB**. [Learn More](#)

Download Metric:

NT rPM


Filter by: — optional

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.

Metagenomics 2 Metagenomics - Nanopore 0 Consensus Genomes 0 Antimicrobial Resistance 2

2 Selected 

2 antimicrobial resistance

<input checked="" type="checkbox"/>	Sample	Created On	Sample Type	Host	Passed Filters	Total Reads	
<input checked="" type="checkbox"/>	ERR4678622 COMPLETE Elizabeth Fahsbender AMR in bacterial pathogens	2024-02-28 6 days ago	sewage	sewage	2,322,742 3.81%	60,967,830	
<input checked="" type="checkbox"/>	ERR4678621 COMPLETE Elizabeth Fahsbender AMR in bacterial pathogens	2024-02-28 6 days ago	sewage	sewage	2,323,414 3.56%	65,282,798	


31. Choose **combined AMR results**.

Select a Download Type

2 samples selected

REPORTS

☐ **Antimicrobial Resistance Results** (.tar.gz)
Includes AMR Report, Comprehensive AMR Metrics, intermediate files and raw outputs from CARD RGI. [Learn More](#)

☒ **Combined AMR Results** (.csv) 
Primary metrics (e.g. coverage, depth) for all AMR genes in all selected samples, combined into a single file.

☐ **Sample Metadata** (sample_metadata.csv)
User-uploaded metadata, including sample collection location, collection date, sample type

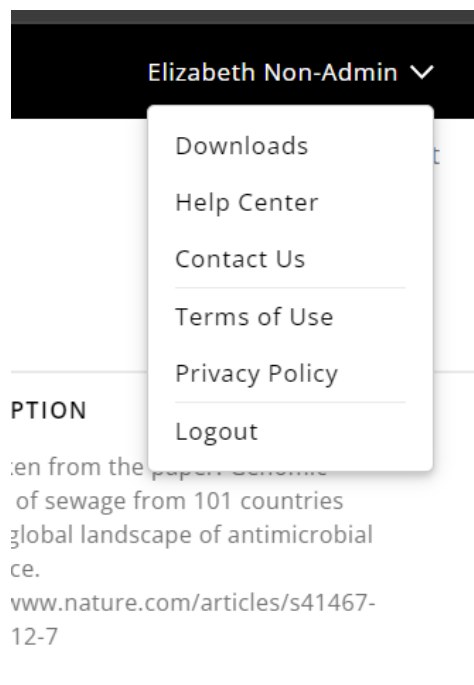
RAW DATA

☐ **Contigs**
All contigs with host data subtracted

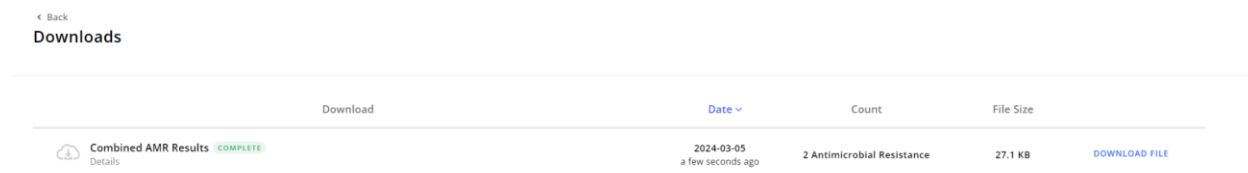
[Start Generating Download](#)

Downloads for larger files can take multiple hours to generate.

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](#).