CZ ID

AMR Pipeline and Sample Report

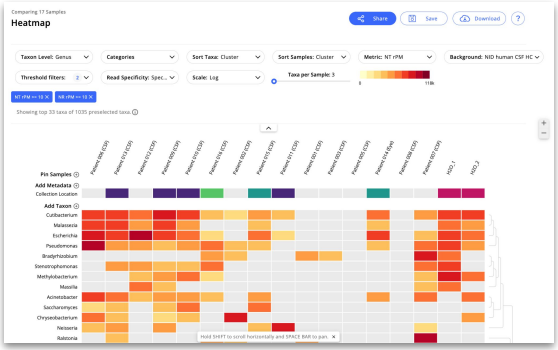
CHAN ZUCKERBERG BIOHUB

CHAN ZUCKERBERG INITIATIVE

cz ID core analysis workflows

Metagenomics

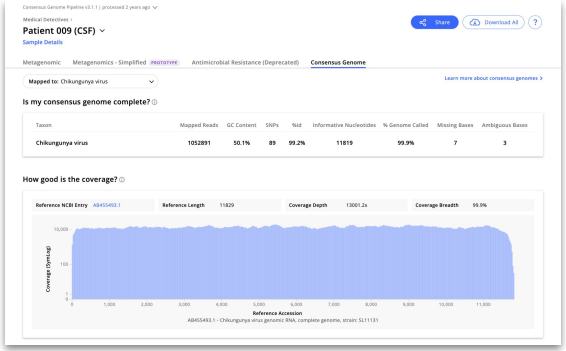
Understand what microbes are present in a sample and their relative abundances.



Also supports Nanopore mNGS

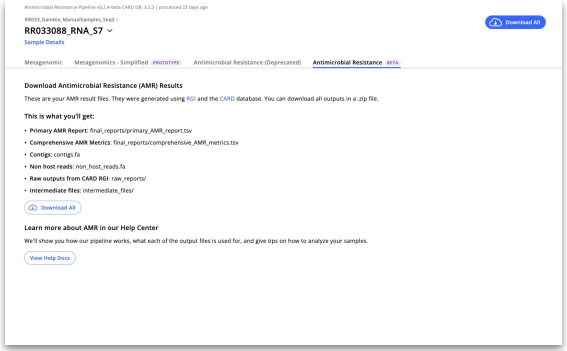
Consensus Genomes

Generate consensus sequences for viruses found in mNGS samples, to support downstream phylogenetics.



Antimicrobial Resistance

Understand the presence and abundance of AMR genes present in mNGS or WGS samples.



Workflow code is available on github at https://github.com/chanzuckerberg/czid-workflows/

mNGS vs AMR

Pipeline	Unknown sequence	Database of known sequences	Main goal
mNGS	Metagenomic reads	Known reference genomes (complete or partial) and translated sequences	Taxonomic identification

mNGS vs AMR

Pipeline	Unknown sequence	Database of known sequences	Main goal
mNGS	Metagenomic reads	Known reference genomes (complete or partial) and translated sequences	Taxonomic identification
AMR	Metagenomic or whole genome sequence reads	Known reference AMR genes	AMR gene detection

AMR Pipeline Highlights

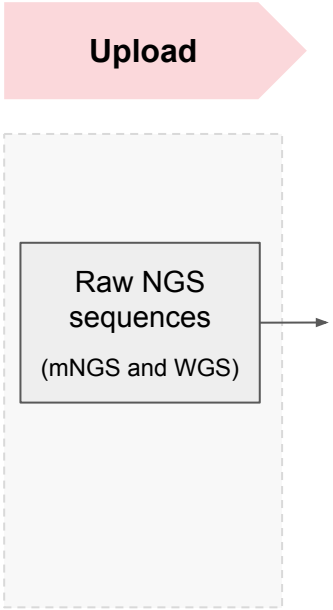
- Supports whole genome sequence (WGS) and metagenomic data.

- Uses the Comprehensive Antibiotic Resistance Database (CARD).
  - Combines the Antibiotic Resistance Ontology (ARO) with curated AMR gene sequences and resistance-conferring mutations
  - Routinely updated

- Uses both sequence reads and assembled sequences (contigs) for AMR gene detection.

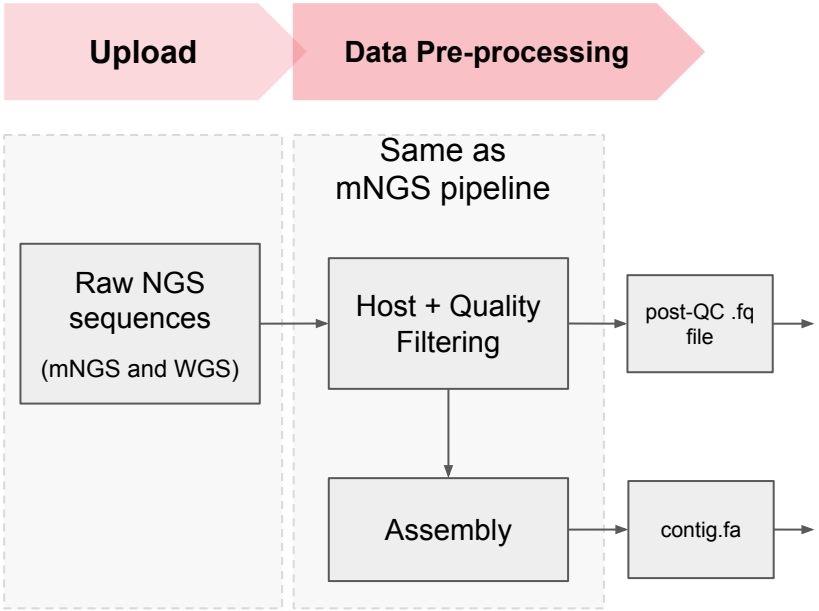
- Performs pathogen-of-origin prediction by matching query sequences to known pathogen sequences found in CARD (beta)

AMR Pipeline Workflow



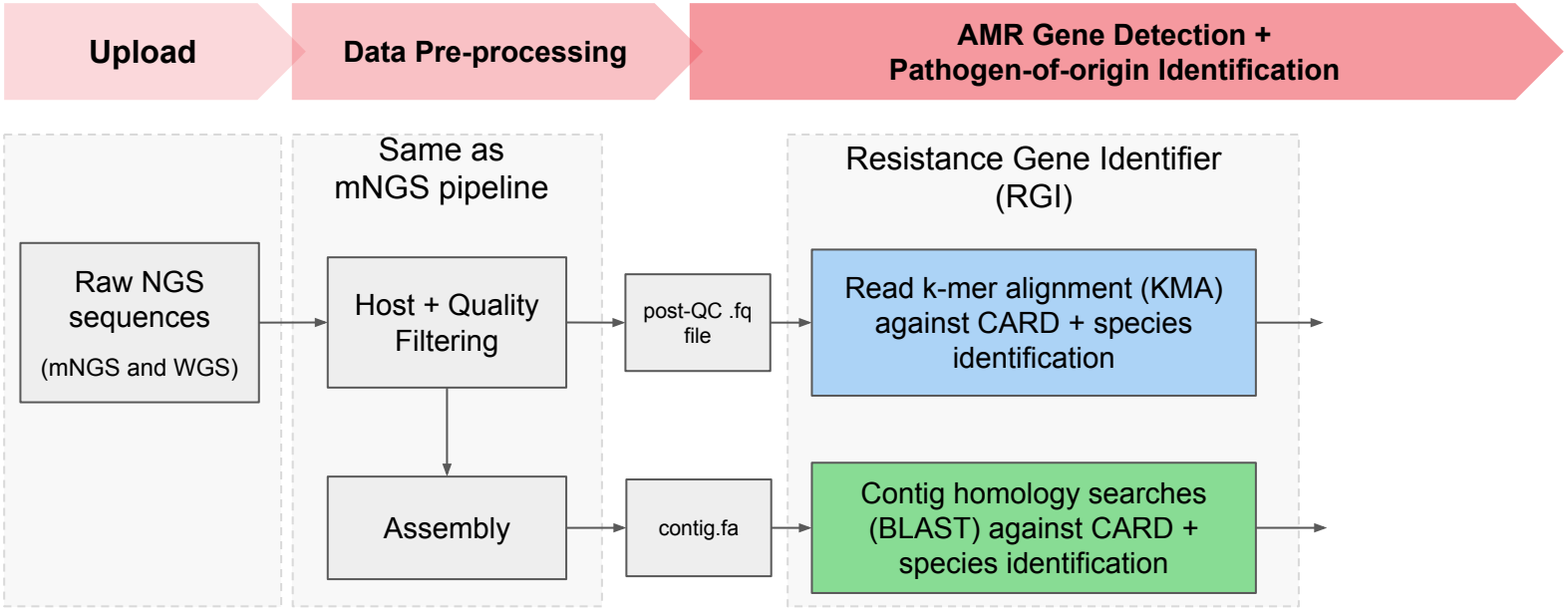
CARD DB and RGI
https://card.mcmaster.ca/

AMR Pipeline Workflow



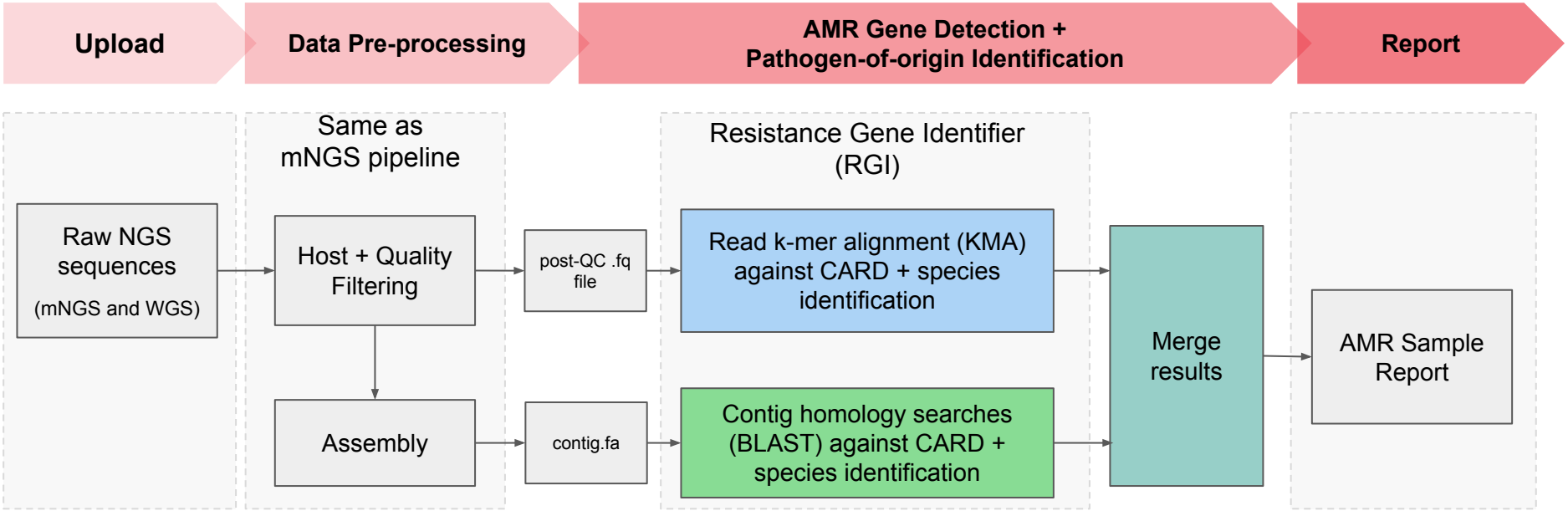
CARD DB and RGI
https://card.mcmaster.ca/

AMR Pipeline Workflow



CARD DB and RGI
https://card.mcmaster.ca/

AMR Pipeline Workflow



CARD DB and RGI
https://card.mcmaster.ca/

AMR Sample Report

‹ mNGS_Demo

AMR Pipeline v1.2.15 | CARD DB: 3.2.6 | processed 34 minutes ago | SAMPLE DETAILS

AMR_test1_reads_nh

Share | Download | ? | ···

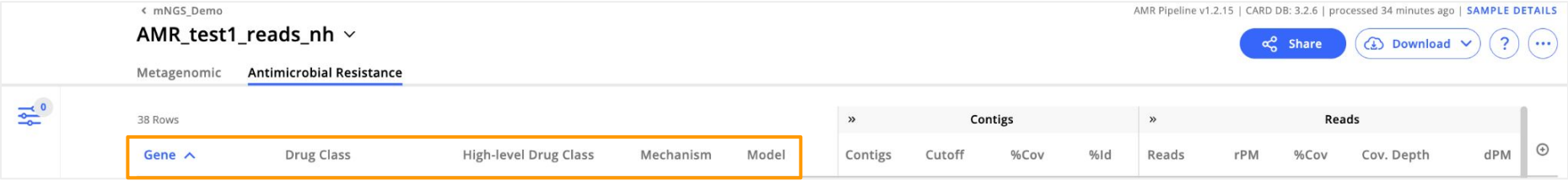
Metagenomic | Antimicrobial Resistance

0

38 Rows

					» Contigs				» Reads				
Gene	Drug Class	High-level Drug Class	Mechanism	Model	Contigs	Cutoff	%Cov	%Id	Reads	rPM	%Cov	Cov. Depth	dPM
AAC(6')-Ib7	aminoglycoside antibiotic	aminoglycoside antibiotic	antibiotic inactivation	protein homolog	0	-	-	-	4	1.81	24.05	0.56	0.25
aadA10	aminoglycoside antibiotic	aminoglycoside antibiotic	antibiotic inactivation	protein homolog	0	-	-	-	2	0.9	24.22	0.27	0.12
aadA27	aminoglycoside antibiotic	aminoglycoside antibiotic	antibiotic inactivation	protein homolog	0	-	-	-	1	0.45	13.31	0.14	0.06

AMR Sample Report: Gene Information



- **Gene:** Gene name for best match in CARD
- **Gene Family:** Refers to the ARO category for gene family
- **Drug Class:** Refers to the ARO category specifying resistance to a given antibiotic molecule, including antibiotic/adjuvant combination medications.
- **High-level Drug Class**: Refers to the antibiotic family of identified drug class.
- **Resistance Mechanism:** Refers to the ARO category specifying the resistance mechanism for a given AMR gene.
- **Model:** Specifies the model used for AMR detection. Models distinguish underlying molecular categories for detected AMR sequences (e.g., presence of genes vs mutations). More on models on the next slide…

Models for AMR Gene Detection

- **Protein Homolog Models (PHM)**: Detect dedicated AMR genes (“presence/absence”)
    - Represent ~⅔ of CARD
    - Contigs - BLASTP against CARD database (protein:protein alignment).
    - Reads - KMA against PHM sequences (**only model used for calling reads**)

Models for AMR Gene Detection

- **Protein Homolog Models (PHM)**: Detect dedicated AMR genes

- **Protein Variant Models** (PVM): Detect AMR acquired via mutation of house-keeping genes or antibiotic targets

  - 2nd largest in CARD

  - PVMs screen query sequences for curated sets of mutations that could differentiate them from antibiotic susceptible genotypes.

```
Query  7   SELNHCDECFALMNPLMILVKIIKLRWIHILSYDQMVSRKINNQTTRCANSIFYTTLFTN  66
           S+ N C+  FALMNPLMILVKIIKLRWIHILSYDQMVSRKINNQTTRCANSIFYTTLFTN
Sbjct  7   SQSNICNRDFALMNPLMILVKIIKLRWIHILSYDQMVSRKINNQTTRCANSIFYTTLFTN  66

Query  67  SAPNYT  72
           SAPNYT
Sbjct  67  SAPNYT  72
```

Models for AMR Gene Detection

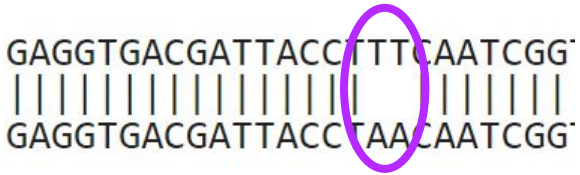
- **Protein Homolog Model (PHM)**- Detect dedicated AMR genes

- **Protein Variant Model (PVM)** - Designed to detect AMR acquired via mutation of house-keeping genes or antibiotic targets

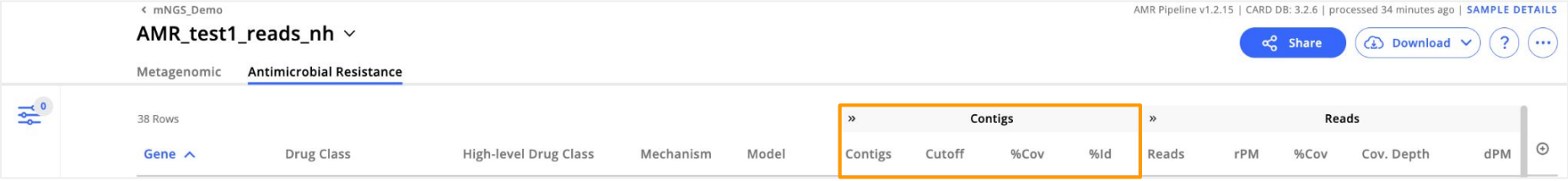
- **Protein Overexpression Models (POM)** - Detect mutations in regulatory proteins.
    - Similar to PVM but restricted to regulatory proteins.
    - POM screen for mutations that may lead to overexpression of efflux complexes.

Models for AMR Gene Detection

- **Protein Homolog Model (PHM)**- Detect dedicated AMR genes
- **Protein Variant Model (PVM)** - Detect AMR acquired via mutation of house-keeping genes or antibiotic targets.
- **Protein Overexpression Model (POM)** - Detect mutations in regulatory proteins.
- **rRNA Gene Variant Models (RVM)** - Detect AMR acquired via mutation of genes encoding rRNA (nucleotide:nucleotide alignment).
  - Screens query sequences for curated sets of mutations that could differentiate them from antibiotic susceptible genotypes

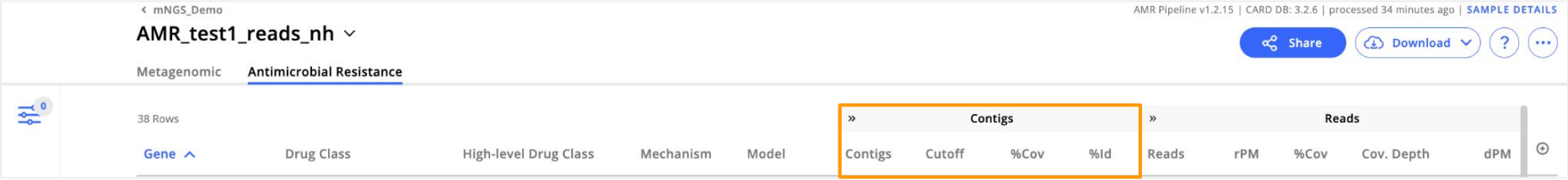


AMR Sample Report: Contig Metrics



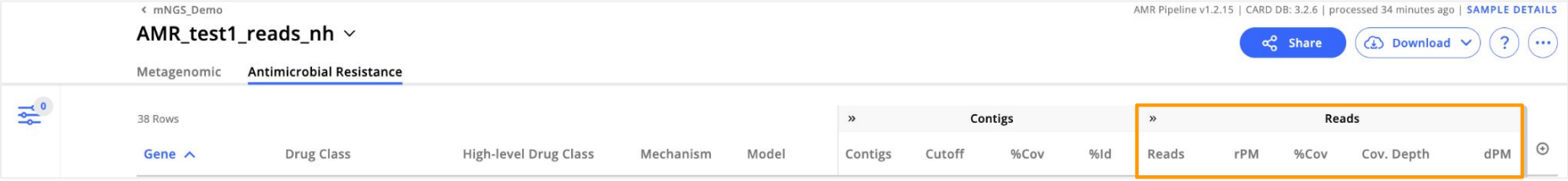
- **Contigs:** Total number of contigs matching a given AMR gene
- **Cutoff:** Cutoff used to detect AMR-associated contigs based on curated BLAST bit-score thresholds.
  - **Perfect:** Detects perfect or identical matches to the curated reference sequences in CARD
  - **Strict:** Detects previously unknown variants of known AMR genes and includes a secondary screen for key mutations (predicts functional variants)
  - **Loose**: Detects new and more distant homologs of AMR genes by working outside of established BLAST bit-score cut-offs allowing for AMR gene discovery. CZ ID only reports Loose contig matches that have at least 95% identity to known AMR genes (identified as "**Nudged**")
    - Nudged cut off does NOT account for alignment length

AMR Sample Report: Contig Metrics (cont.)

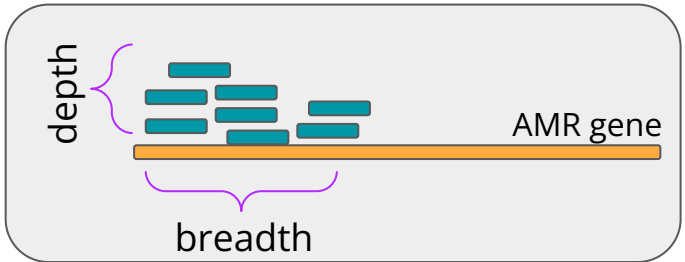


- **Contig Coverage Breath (% Cov):** Percentage length of the reference sequence that was covered by contigs
- **Contig Percent Identity (% ID):** Average percent identity between contigs and their top match in CARD
- **Contig Species:** Pathogen-of-origin prediction based on AMR-associated contig sequences.

AMR Sample Report: Read Metrics



- **Reads:** Total number of reads mapping to a given AMR reference sequence.
- **Read Coverage Breadth (% Cov):** Percent length of the reference sequence covered by read sequences.
- **Read Coverage Depth (Cov. Depth):** Mean read depth across the reference sequence.
- **Read Species:** Pathogen-of-origin prediction based on AMR-associated reads.



Normalized Read Metrics

When comparing across samples is important to **normalize** the data to make sure what we see is not just an artifact of one sample having more reads than another.

- **Reads per million (rPM)**

$$\text{rPM} = \frac{\text{Reads matching gene or allele}}{(\text{Total reads} - \text{ERCC reads}) \times \text{Subsampled fraction}} \times 10^6$$

- **Depth per million (dPM)**

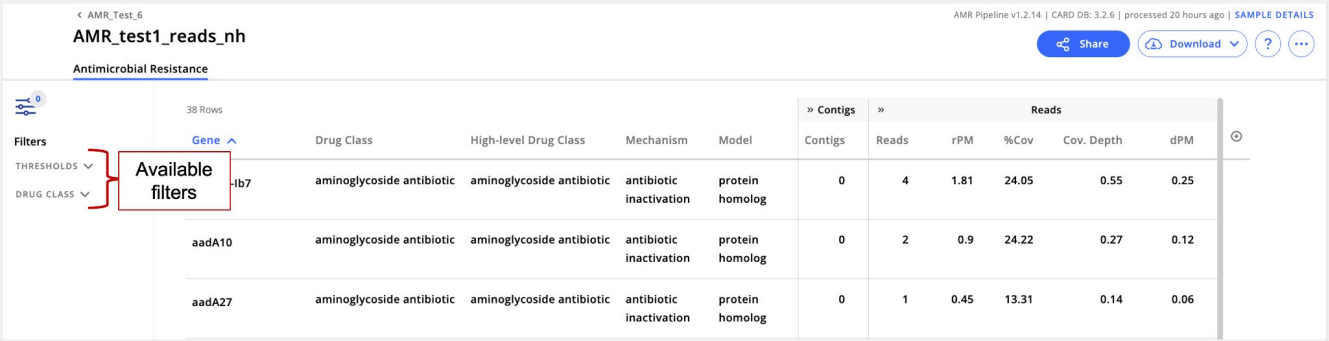
$$\text{dPM} = \frac{\left(\dfrac{\text{Bases mapped to gene or allele}}{\text{Length of gene or allele}}\right)}{(\text{Total reads} - \text{ERCC reads}) \times \text{Subsampled fraction}} \times 10^6$$

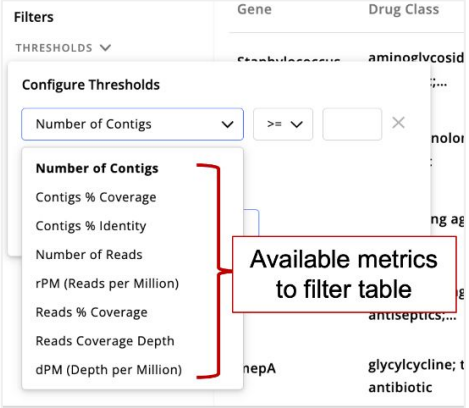
https://www.nature.com/articles/s41598-018-24280-8.pdf?origin=ppub

Applying Filters to AMR Sample Report

AMR genes tend to be at low abundances unless they are enriched (e.g., FLASH protocol). Use relaxed parameters compared to filters used for mNGS analysis

- Suggested threshold:
  - AMR genes with at least 10% coverage.
  - At least 5 reads matching AMR gene.





Notes

- May be challenging to distinguish sequencing error/artifacts from true SNPs that confer resistance
- There may be differences in AMR genes and species detected with contig data versus those identified with read data
  - KMA (reads) vs Assembly/BLAST (contigs)
- CZ ID mNGS & AMR modules may report different species. This is due to differences in databases & alignment methods

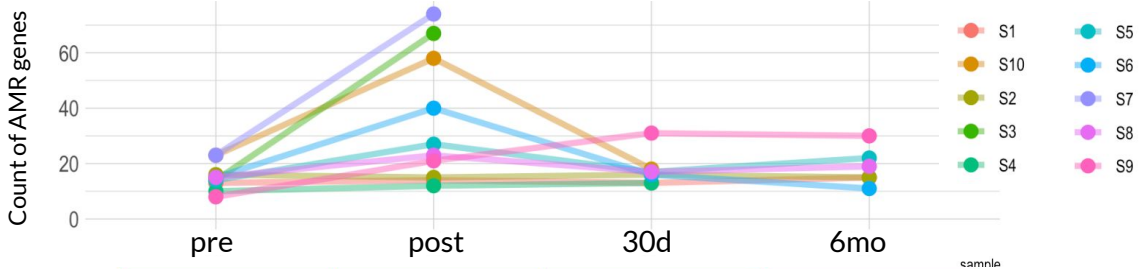
Reference Guides

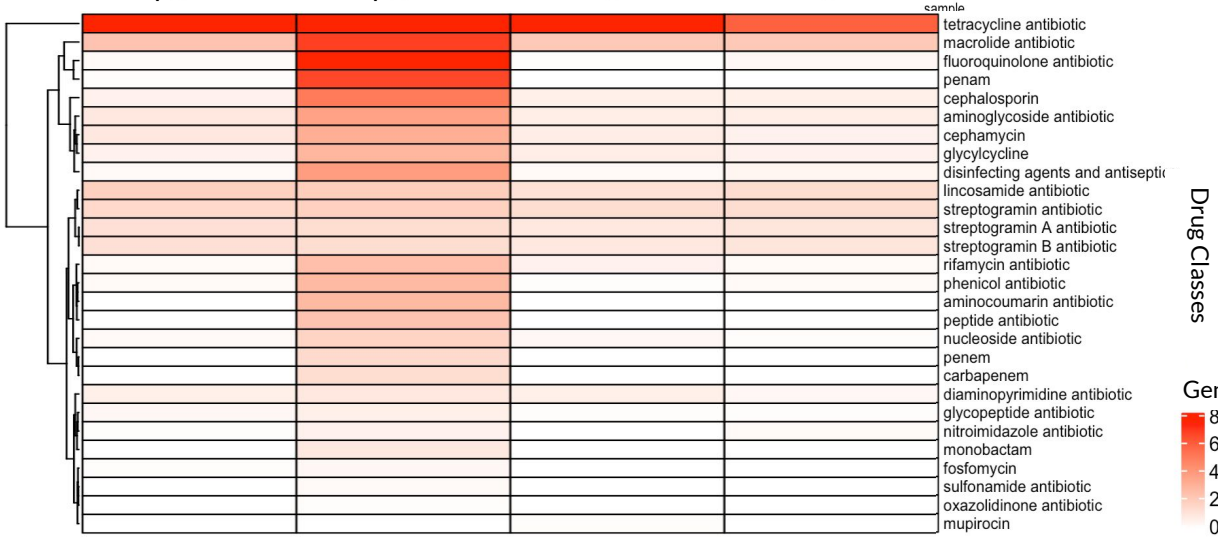
- CZ ID AMR Resources
  - Guides for how to upload and download AMR gene data, how to interpret and customize the AMR Sample Report, and general information about the AMR pipeline

- Video: AMR Detection and Analysis Using CARD & RGI
  - Presentation by Andrew McArthur (CARD/RGI PI) providing background information and details about CARD and RGI workflow

Examples of what you can do with AMR & mNGS data

AMR gene trends detected across **multiple travellers**

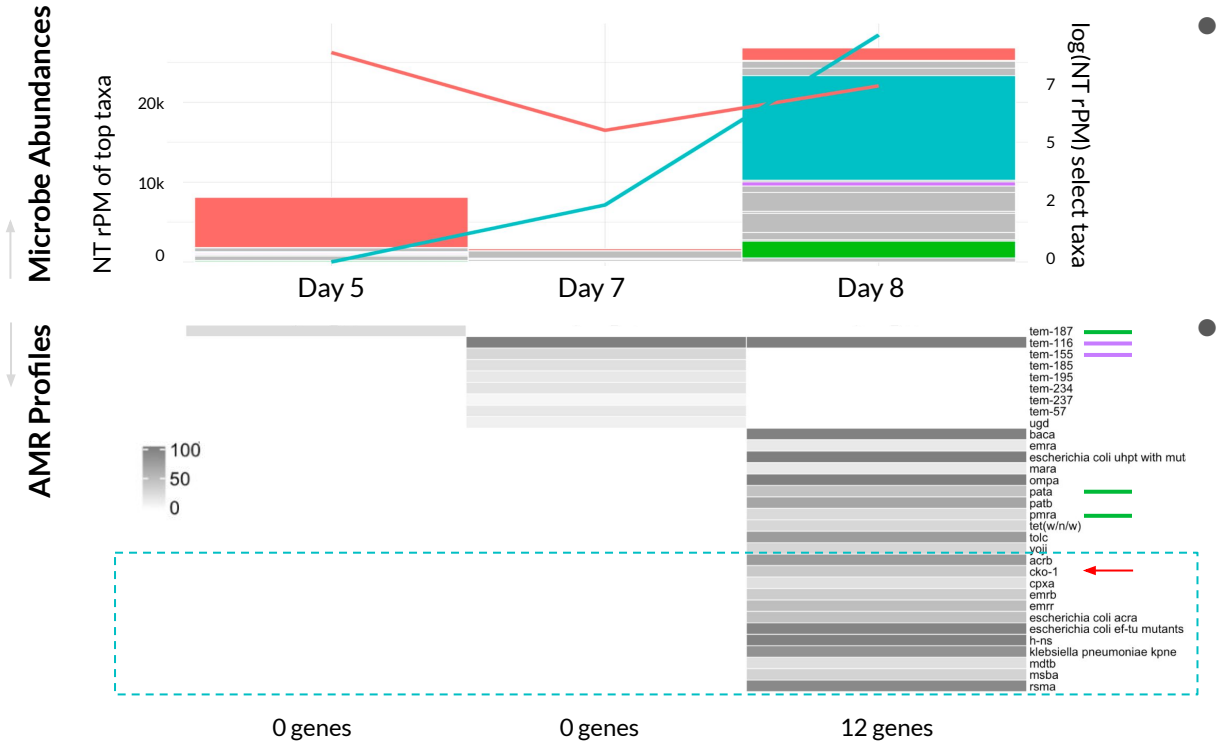




- Travel correlates with an increase in the total number of detected AMR genes
- Macrolide, Fluoroquinolone, Penam, and Cephalosporin drug classes drive the increase in AMR gene burden across the cohort

Katrina Kalantar

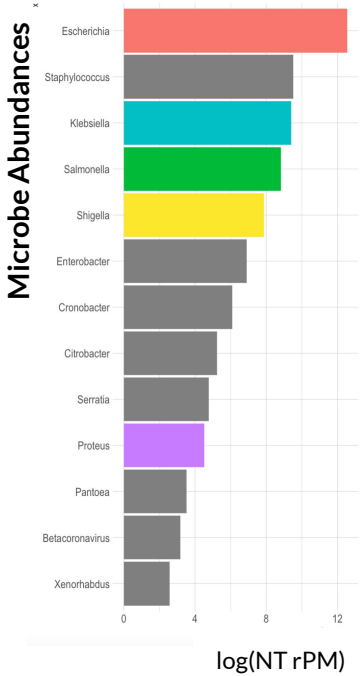
Time Series Case: SARS-CoV-2 patient with *Citrobacter* VAP, CKO-1 gene

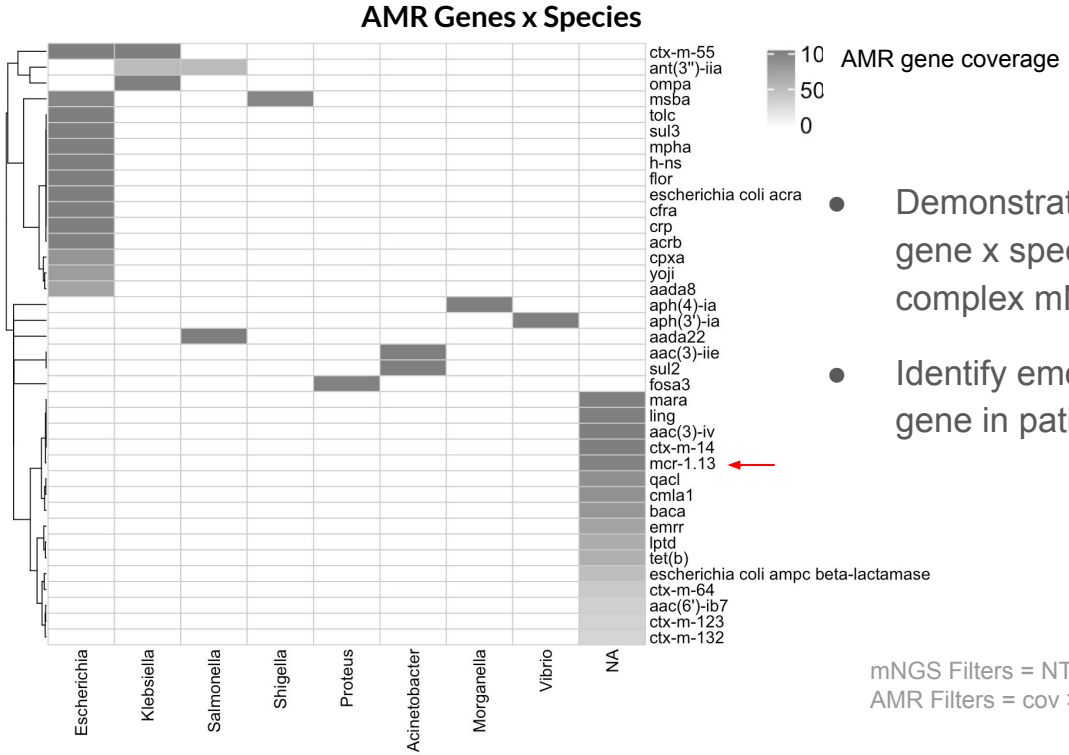


- mNGS time series analysis shows dynamics of primary viral and secondary bacterial (*Citrobacter*) infection in a single patient
- Combining AMR and mNGS data enables detection of concerning CKO-1 gene associated with *Citrobacter* VAP

Katrina Kalantar

Surveillance for emerging AMR pathogens and genes





- Demonstrates utility of AMR gene x species information in complex mNGS context
- Identify emerging MCR-1 gene in patient

mNGS Filters = NT rPM > 10, NR rPM > 10
AMR Filters = cov > 20, depth > 1x

Katrina Kalantar

CZ ID

Metagenomic pipeline

CHAN ZUCKERBERG BIOHUB

CHAN ZUCKERBERG INITIATIVE

Sequence Databases

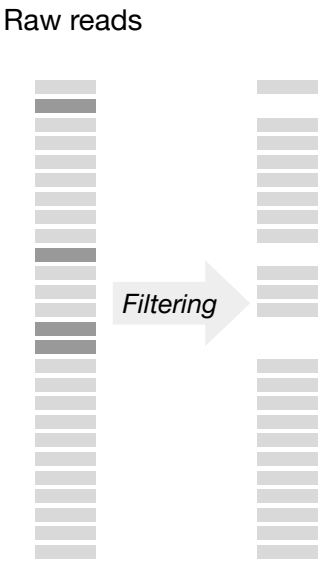
The NCBI houses a series of databases relevant to biotechnology and biomedicine and is an important resource for bioinformatics tools and services.

	NCBI NT Database	NCBI NR database
What does NT/NR mean?	NT = nucleotide	NR = non-redundant protein
What is in the database?	Contains **all nucleotide sequences** (RefSeq RNA records plus all GenBank sequences except for those from the EST, GSS, STS and HTG divisions).	Contains **non-redundant set of all CDS translations from GenBank** along with all RefSeq, UniProtKB/Swiss-Prot, PDB and PRF proteins. Note: **does not contain rRNA sequences** given that these do not create protein
How does it influence analysis?	NT hits **tend to be more accurate for most organisms** (bacteria, eukaryotes, viruses with close relatives in the DB)	NR alignments are **especially good for detecting divergent viruses**, since mutations in the NT sequence accumulate faster than the amino acid sequence

Raw reads

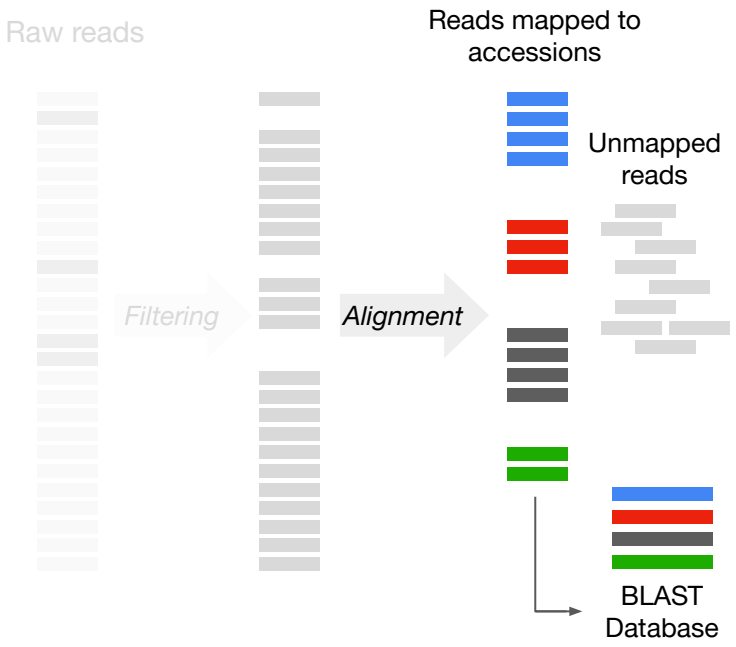
Raw Reads

- Start with raw reads from FASTQ file
- The pipeline goal is to assign reads to taxa using complete NCBI NT and NR reference databases
- This mapping help users identify pathogens in their samples



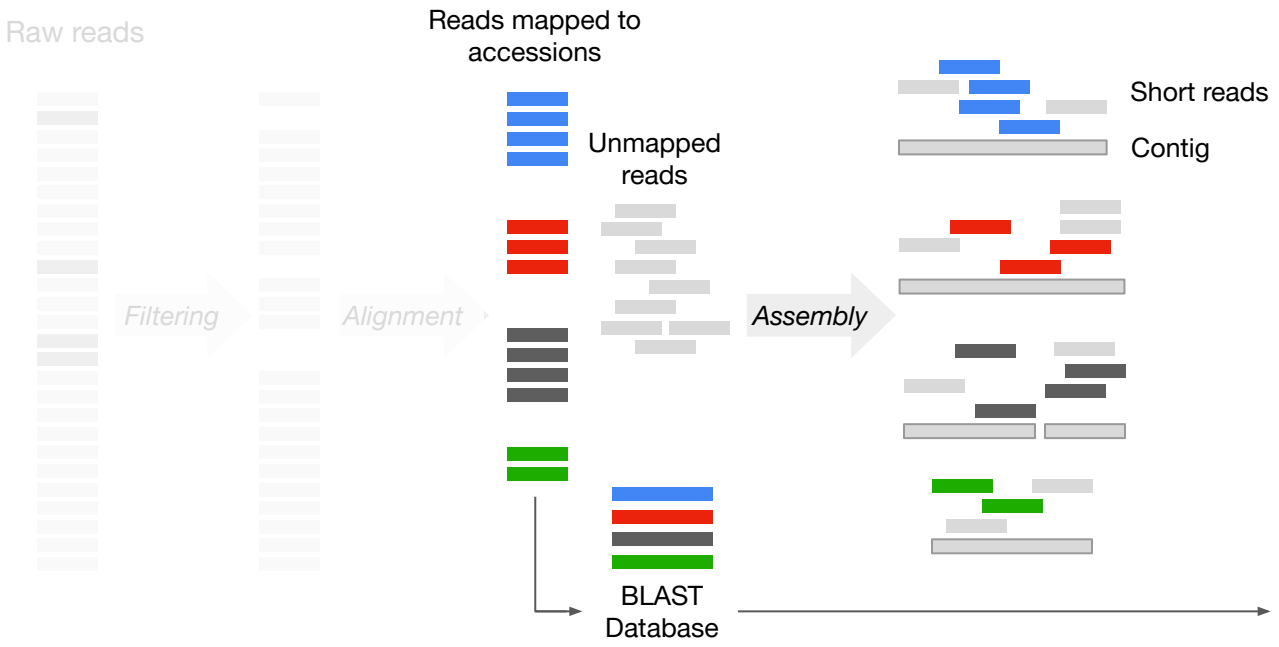
Host Filtering and QC Substeps

- Truncate to 150 million reads (75 million read-pairs)
- fastp to filter out low quality reads, low complexity reads, and adapters
- Bowtie followed by HIASAT2 for host read removal
- Bowtie followed by HIASAT2 for human read removal
- CZID-dedup to collapse duplicate reads
- Subsampling to 2 million reads (or 1 million read-pairs)



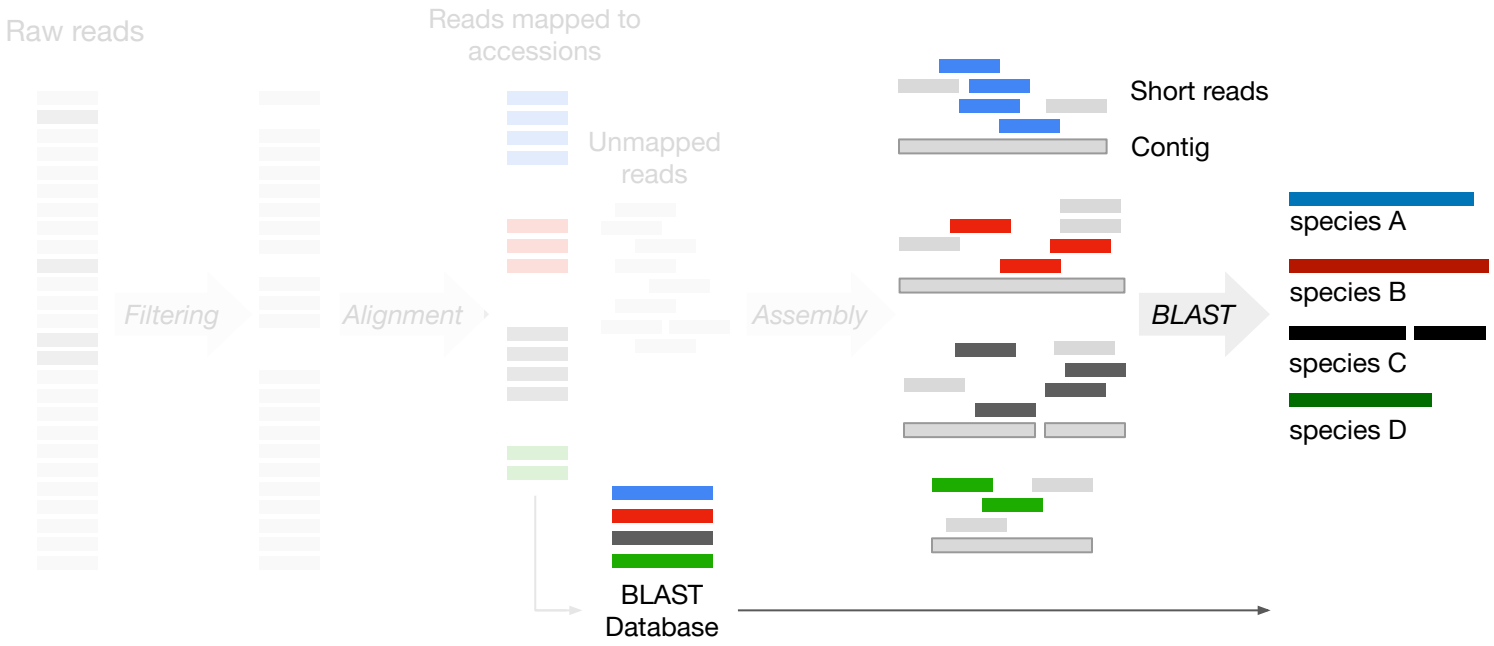
Alignment Step

- Reads are mapped against NCBI NT (nucleotide) and NR (protein) databases to obtain a preliminary accession for each read.
- Generate BLAST database containing all taxa identified as preliminary accessions.



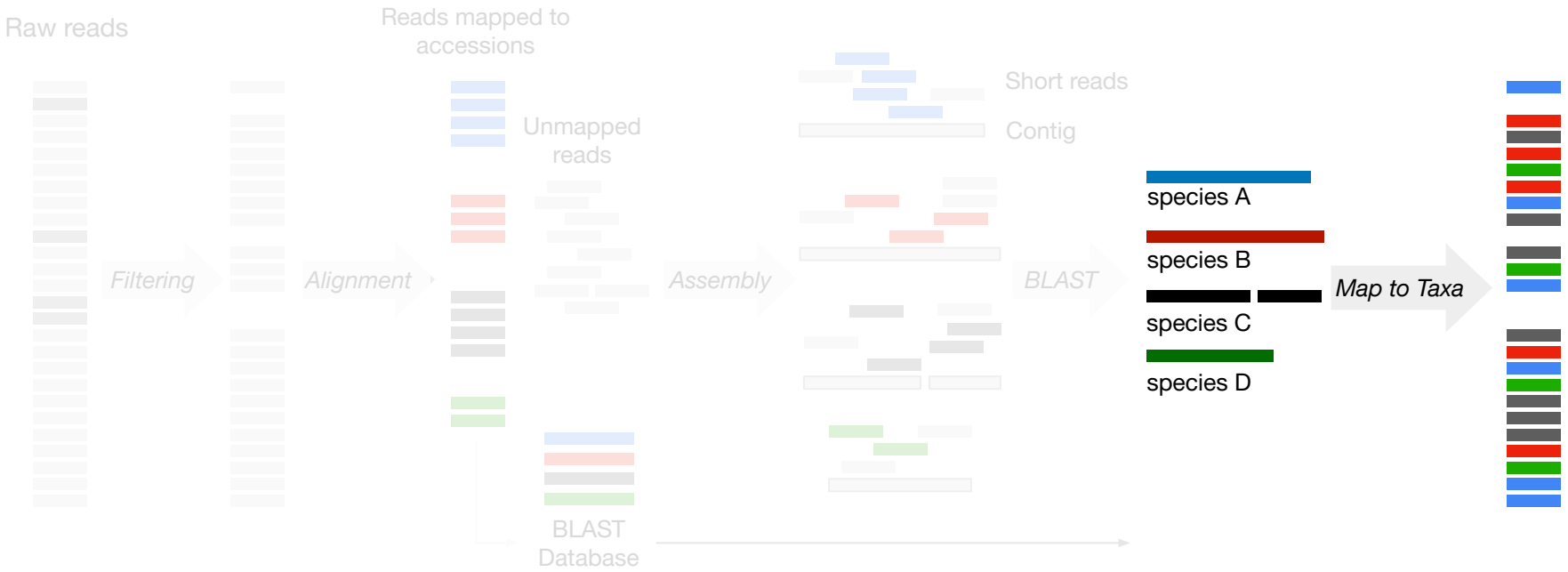
Assembly-based Alignment Step

- Assemble reads into contigs using de novo assembly via SPADES.
- Align reads to assembled contigs.



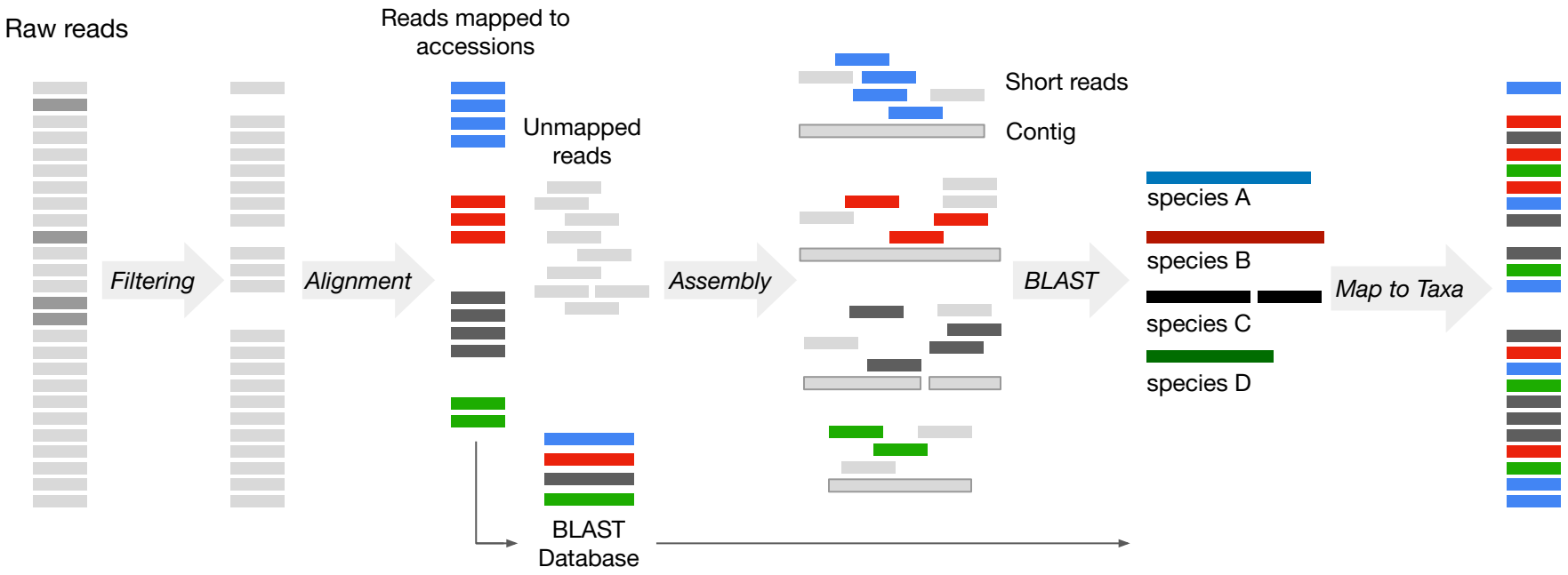
Assembly-based Alignment Step

- Assemble reads into contigs using de novo assembly via SPADES.
- Align reads to assembled contigs.
- BLAST contigs against database of putative accessions generated in the previous step.



Mapping to Taxa

- Assign each read a final accession based on 1. the contig accession, or if the read did not assemble into a contig, 2. the initial read accession.
- Use NCBI's accession to taxon database to assign a taxon per contig
- Compute statistics, reads per million per microbe.


Raw reads
Filtering
Alignment
Reads mapped to accessions
Unmapped reads
BLAST Database
Assembly
Short reads
Contig
BLAST
species A
species B
species C
species D
Map to Taxa