Rapid Identification of AMR Determinants from Metagenomic Samples

AMRtime Progress Report

Finlay Maguire

June 22, 2018

Faculty of Computer Science, Dalhousie University

Table of contents

1. Overview
2. Training Data
3. Read filtering
4. Sensitive Homology Search
5. Variant Models
6. Summary
7. Acknowledgements

Overview

Comprehensive Antibiotic Resistance Database

- `https://card.mcmaster.ca/` (Jia et al., 2016) as of June 2018:

Comprehensive Antibiotic Resistance Database

- `https://card.mcmaster.ca/` (Jia et al., 2016) as of June 2018:
- Built around Antibiotic Resistance Ontology (ARO): 3996 terms

Comprehensive Antibiotic Resistance Database

- `https://card.mcmaster.ca/` (Jia et al., 2016) as of June 2018:
- Built around Antibiotic Resistance Ontology (ARO): 3996 terms
- 2536 AMR Detection Models with manually curated criteria:

Comprehensive Antibiotic Resistance Database

- `https://card.mcmaster.ca/` (Jia et al., 2016) as of June 2018:
- Built around Antibiotic Resistance Ontology (ARO): 3996 terms
- 2536 AMR Detection Models with manually curated criteria:
    - Homology e.g. NDM beta-lactamases, aminoglycoside acetyltransferase

Comprehensive Antibiotic Resistance Database

- `https://card.mcmaster.ca/` (Jia et al., 2016) as of June 2018:
- Built around Antibiotic Resistance Ontology (ARO): 3996 terms
- 2536 AMR Detection Models with manually curated criteria:
    - Homology e.g. NDM beta-lactamases, aminoglycoside acetyltransferase
    - Protein Variant e.g. GyrA fluoroquinolone mutation, FolP sulfonamide mutation

Comprehensive Antibiotic Resistance Database

- `https://card.mcmaster.ca/` (Jia et al., 2016) as of June 2018:
- Built around Antibiotic Resistance Ontology (ARO): 3996 terms
- 2536 AMR Detection Models with manually curated criteria:
    - Homology e.g. NDM beta-lactamases, aminoglycoside acetyltransferase
    - Protein Variant e.g. GyrA fluoroquinolone mutation, FolP sulfonamide mutation
    - rRNA gene variants e.g. Mycobacterium aminoglycoside resistance

Comprehensive Antibiotic Resistance Database

- `https://card.mcmaster.ca/` (Jia et al., 2016) as of June 2018:
- Built around Antibiotic Resistance Ontology (ARO): 3996 terms
- 2536 AMR Detection Models with manually curated criteria:
    - Homology e.g. NDM beta-lactamases, aminoglycoside acetyltransferase
    - Protein Variant e.g. GyrA fluoroquinolone mutation, FolP sulfonamide mutation
    - rRNA gene variants e.g. Mycobacterium aminoglycoside resistance
    - Efflux pump e.g. AcrAB-TolC, MexAB-OprM mutations

Comprehensive Antibiotic Resistance Database

- `https://card.mcmaster.ca/` (Jia et al., 2016) as of June 2018:
- Built around Antibiotic Resistance Ontology (ARO): 3996 terms
- 2536 AMR Detection Models with manually curated criteria:
  - Homology e.g. NDM beta-lactamases, aminoglycoside acetyltransferase
  - Protein Variant e.g. GyrA fluoroquinolone mutation, FolP sulfonamide mutation
  - rRNA gene variants e.g. Mycobacterium aminoglycoside resistance
  - Efflux pump e.g. AcrAB-TolC, MexAB-OprM mutations
  - Gene cluster e.g. Van glycopeptide resistance clusters

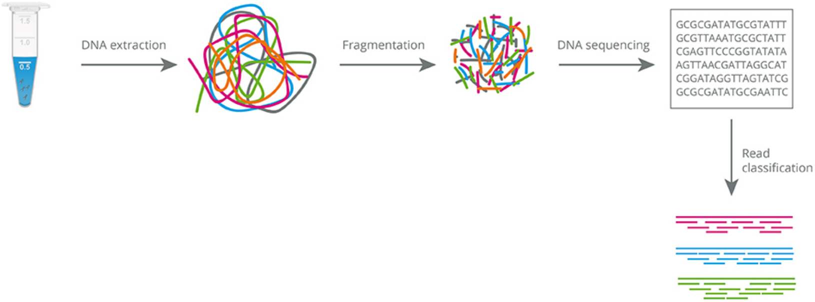
Comprehensive Antibiotic Resistance Database

- `https://card.mcmaster.ca/` (Jia et al., 2016) as of June 2018:
- Built around Antibiotic Resistance Ontology (ARO): 3996 terms
- 2536 AMR Detection Models with manually curated criteria:
    - Homology e.g. NDM beta-lactamases, aminoglycoside acetyltransferase
    - Protein Variant e.g. GyrA fluoroquinolone mutation, FolP sulfonamide mutation
    - rRNA gene variants e.g. Mycobacterium aminoglycoside resistance
    - Efflux pump e.g. AcrAB-TolC, MexAB-OprM mutations
    - Gene cluster e.g. Van glycopeptide resistance clusters
- Resistance Gene Identifier (RGI): contigs, predicted genes and merged metagenomic reads

Comprehensive Antibiotic Resistance Database

- `https://card.mcmaster.ca/` (Jia et al., 2016) as of June 2018:
- Built around Antibiotic Resistance Ontology (ARO): 3996 terms
- 2536 AMR Detection Models with manually curated criteria:
  - Homology e.g. NDM beta-lactamases, aminoglycoside acetyltransferase
  - Protein Variant e.g. GyrA fluoroquinolone mutation, FolP sulfonamide mutation
  - rRNA gene variants e.g. Mycobacterium aminoglycoside resistance
  - Efflux pump e.g. AcrAB-TolC, MexAB-OprM mutations
  - Gene cluster e.g. Van glycopeptide resistance clusters
- Resistance Gene Identifier (RGI): contigs, predicted genes and merged metagenomic reads
- CARDPredicted prevalence dataset

Metagenomic Analysis



DNA extraction

Fragmentation

DNA sequencing

```
GCGCGATATGCGTATTT
GCGTTAAATGCGCTATT
CGAGTTCCCGGTATATA
AGTTAACGATTAGGCAT
CGGATAGGTTAGTATCG
GCGCGATATGCGAATTC
```

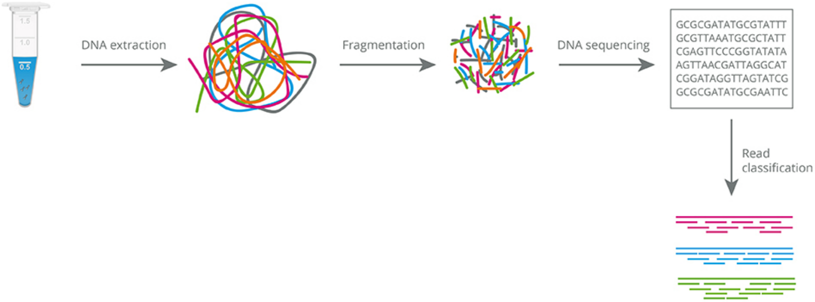
Read classification

modified from `https://www.gatc-biotech.com/en/expertise/genomics/metagenome-analysis.html`

Key difficulties:

- Variation in abundance and diversity

Metagenomic Analysis



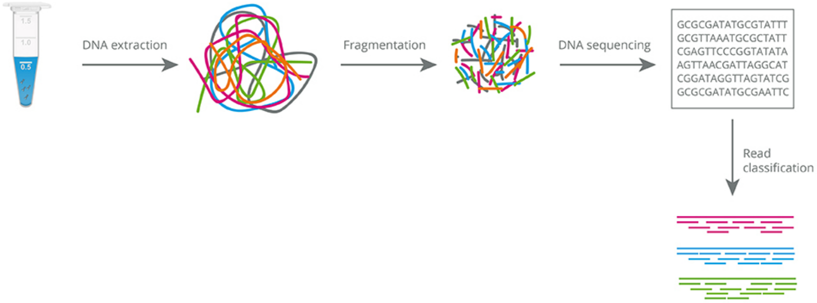
GCGCGATATGCGTATTT
GCGTTAAATGCGCTATT
CGAGTTCCCGGTATATA
AGTTAACGATTAGGCAT
CGGATAGGTTAGTATCG
GCGCGATATGCGAATTC

modified from `https://www.gatc-biotech.com/en/expertise/genomics/metagenome-analysis.html`

Key difficulties:

- Variation in abundance and diversity
- Short fragmentary data

Metagenomic Analysis

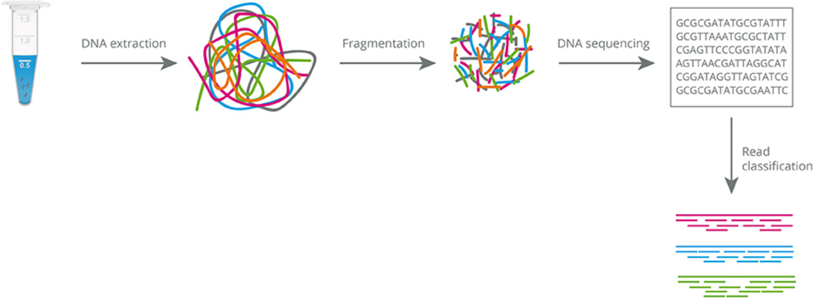


modified from `https://www.gatc-biotech.com/en/expertise/genomics/metagenome-analysis.html`

Key difficulties:

- Variation in abundance and diversity
- Short fragmentary data
- Large amounts of data

Metagenomic Analysis

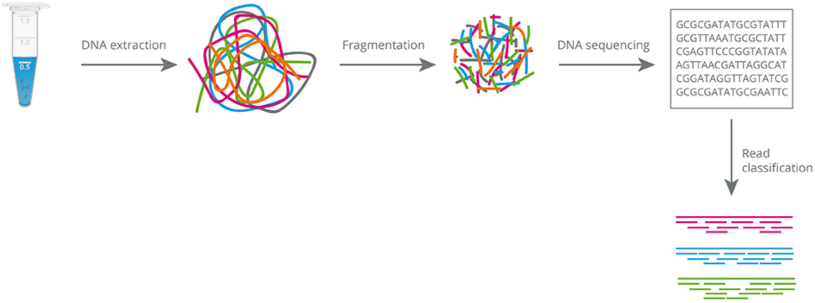


modified from `https://www.gatc-biotech.com/en/expertise/genomics/metagenome-analysis.html`

Key difficulties:

- Variation in abundance and diversity
- Short fragmentary data
- Large amounts of data
- Compositionality

Metagenomic Analysis

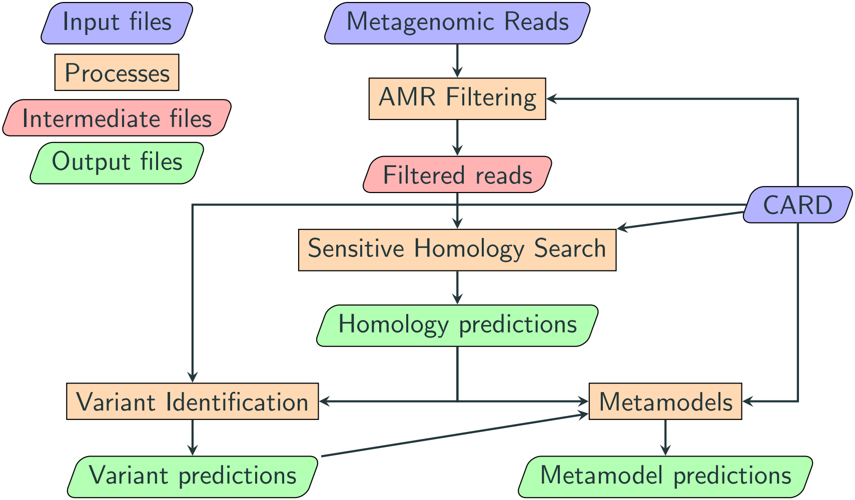


modified from `https://www.gatc-biotech.com/en/expertise/genomics/metagenome-analysis.html`

Key difficulties:

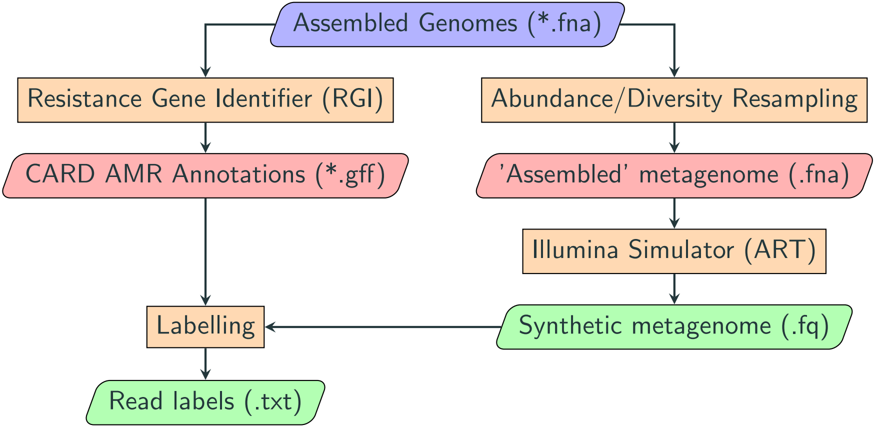
- Variation in abundance and diversity
- Short fragmentary data
- Large amounts of data
- Compositionality
- Spare and imbalanced labels

AMRtime Structure

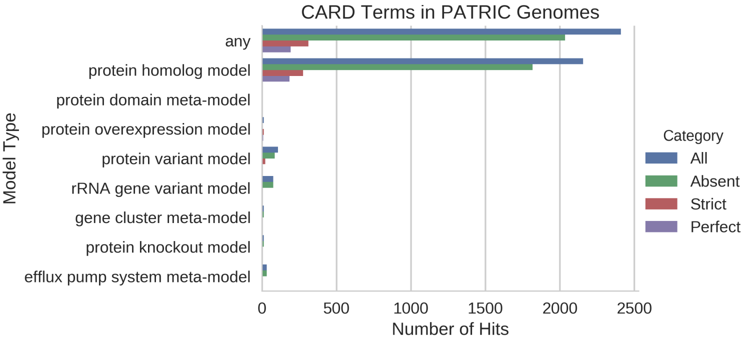


Training Data

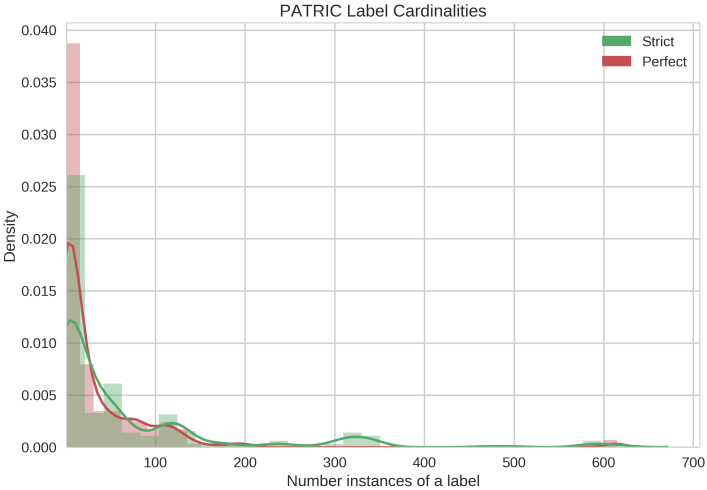
Dataset Generator


Assembled Genomes (*.fna)
Resistance Gene Identifier (RGI)
Abundance/Diversity Resampling
CARD AMR Annotations (*.gff)
'Assembled' metagenome (.fna)
Illumina Simulator (ART)
Labelling
Synthetic metagenome (.fq)
Read labels (.txt)

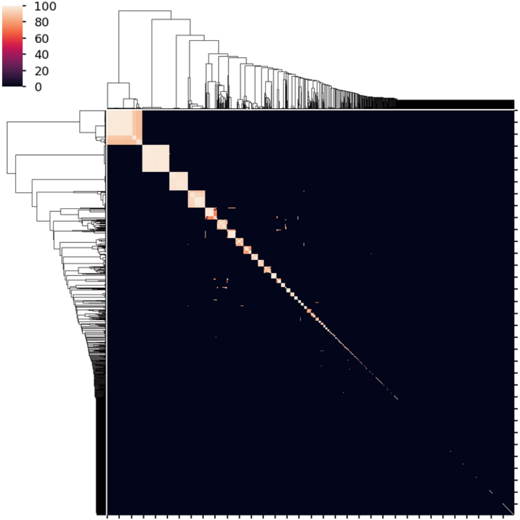
Determinants are scarce



Determinants are imbalanced



AMR sequence space is biased

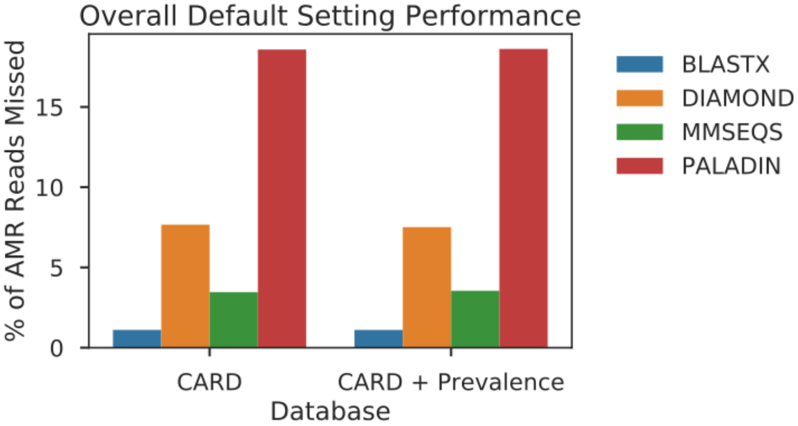


Read filtering

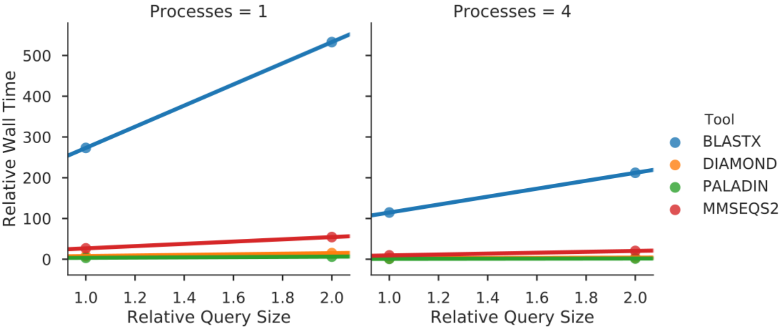
Homology Filter Approaches

- BLASTX (Gish et al., 1993)
- DIAMOND (Buchfink et al., 2015)
- PALADIN (Westbrook et al., 2017)
- MMSeqs2 (Steinegger and Söding, 2017)

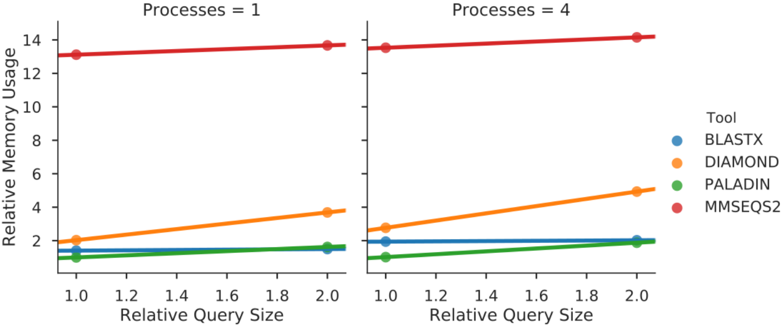
Performance at defaults?



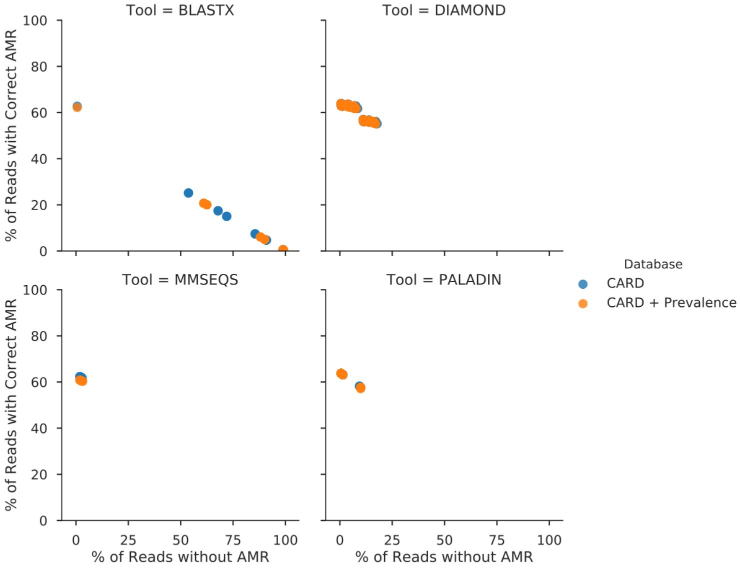
How computationally efficient are they?



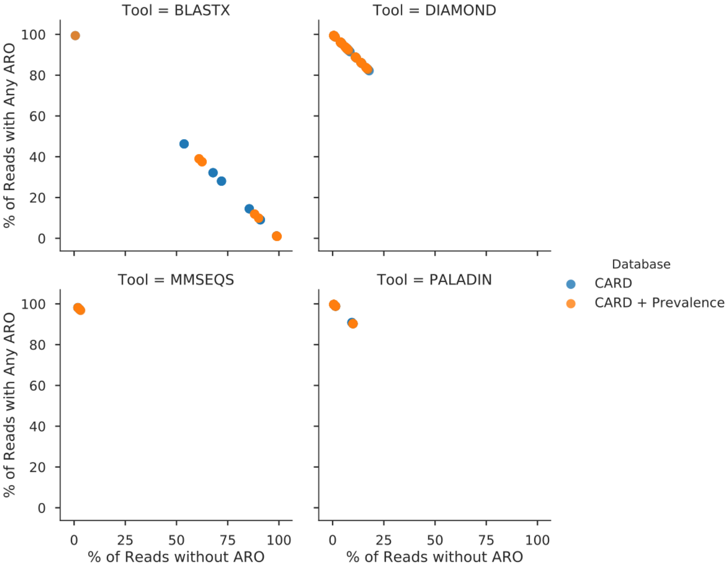
What about in terms of memory?



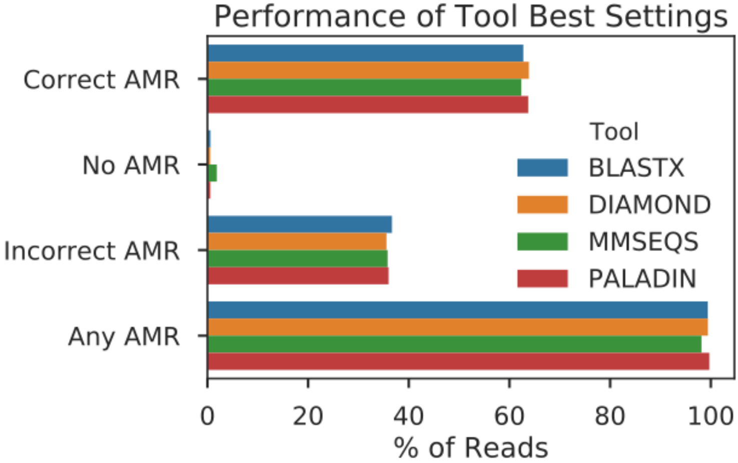
Is there a cap on overall performance?



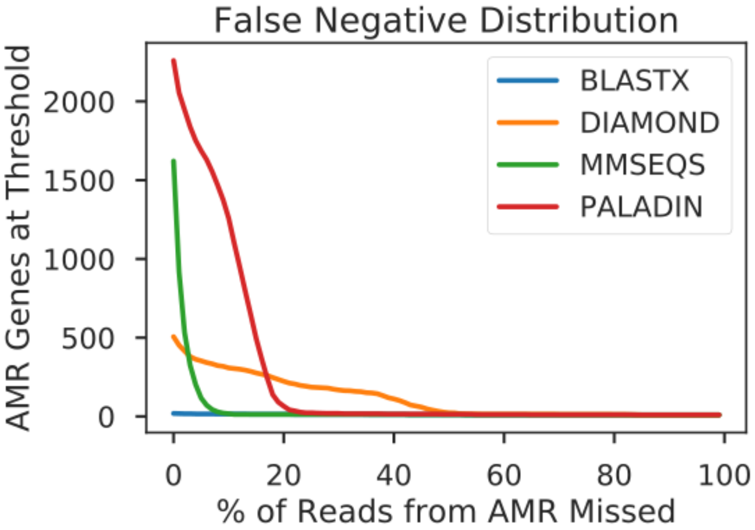
What about to hit any ARO?



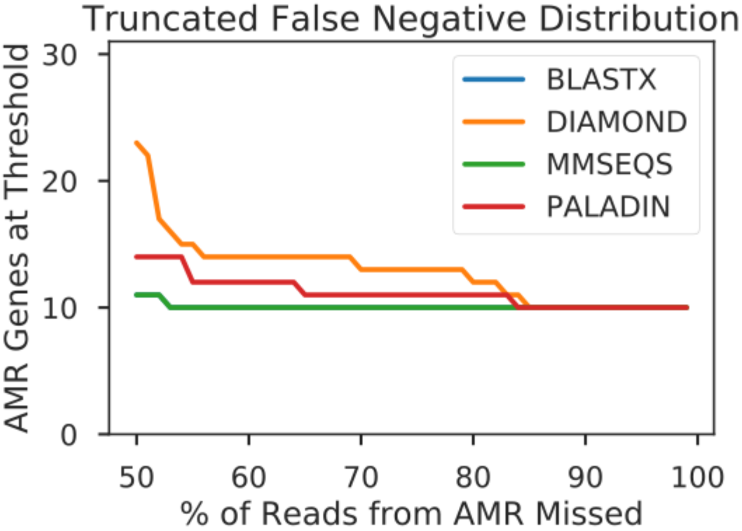
Performance for best setting per tool



But what about individual ARO performance?



Systematically missing AROs



Why are these 10 always missed?

- Enterococcus faecalis liaS mutant conferring daptomycin resistance (AE016830.1):

Why are these 10 always missed?

- Enterococcus faecalis liaS mutant conferring daptomycin resistance (AE016830.1):
    - Protein 2790824-2789724

Why are these 10 always missed?

- Enterococcus faecalis liaS mutant conferring daptomycin resistance (AE016830.1):
  - Protein 2790824-2789724
  - DNA 1-732

Why are these 10 always missed?

- Enterococcus faecalis liaS mutant conferring daptomycin resistance (AE016830.1):
  - Protein 2790824-2789724
  - DNA 1-732
- OXA-2 (M95287.4):

Why are these 10 always missed?

- Enterococcus faecalis liaS mutant conferring daptomycin resistance (AE016830.1):
  - Protein 2790824-2789724
  - DNA 1-732
- OXA-2 (M95287.4):
  - Protein 2456-3280

Why are these 10 always missed?

- Enterococcus faecalis liaS mutant conferring daptomycin resistance (AE016830.1):
  - Protein 2790824-2789724
  - DNA 1-732
- OXA-2 (M95287.4):
  - Protein 2456-3280
  - DNA 1-828

Why are these 10 always missed?

- Enterococcus faecalis liaS mutant conferring daptomycin resistance (AE016830.1):
  - Protein 2790824-2789724
  - DNA 1-732
- OXA-2 (M95287.4):
  - Protein 2456-3280
  - DNA 1-828
- Acinetobacter OprD conferring resistance to imipenem (CP006768.1):

Why are these 10 always missed?

- Enterococcus faecalis liaS mutant conferring daptomycin resistance (AE016830.1):
  - Protein 2790824-2789724
  - DNA 1-732
- OXA-2 (M95287.4):
  - Protein 2456-3280
  - DNA 1-828
- Acinetobacter OprD conferring resistance to imipenem (CP006768.1):
  - Protein 3513470-3514777

Why are these 10 always missed?

- Enterococcus faecalis liaS mutant conferring daptomycin resistance (AE016830.1):
  - Protein 2790824-2789724
  - DNA 1-732
- OXA-2 (M95287.4):
  - Protein 2456-3280
  - DNA 1-828
- Acinetobacter OprD conferring resistance to imipenem (CP006768.1):
  - Protein 3513470-3514777
  - DNA 3514887-3515414

CARD Full Length Alignment QC

- 11 AROs protein not detected from DNA

CARD Full Length Alignment QC

- 11 AROs protein not detected from DNA
- 2 AROs different top protein hit from DNA

CARD Full Length Alignment QC

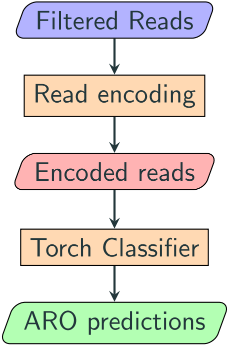
- 11 AROs protein not detected from DNA
- 2 AROs different top protein hit from DNA
- Warnings: 119 AROs with different top protein but $ID\% > 99$

CARD Full Length Alignment QC

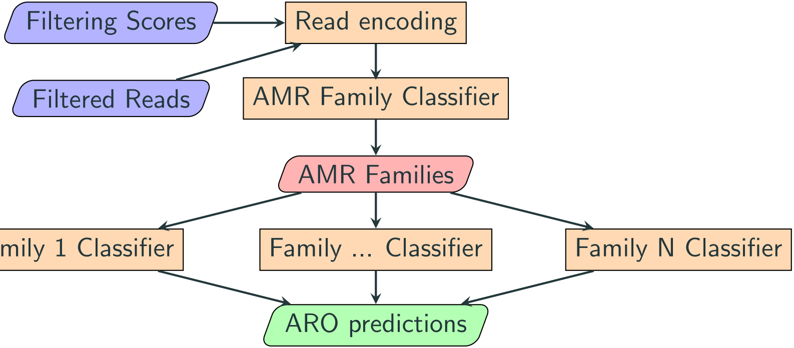
- 11 AROs protein not detected from DNA
- 2 AROs different top protein hit from DNA
- Warnings: 119 AROs with different top protein but $ID\% > 99$
- Warnings: 2 AROs with $ID\% < 99$ to correct protein

Sensitive Homology Search

First attempt at sensitive classification



Revised classifier structure



Encodings

- Raw sequence

Encodings

- Raw sequence
- Filtering homology search family similarity/dissimilarity

Encodings

- Raw sequence
- Filtering homology search family similarity/dissimilarity
- Manual feature extraction (GC/TNF/compositional)

Encodings

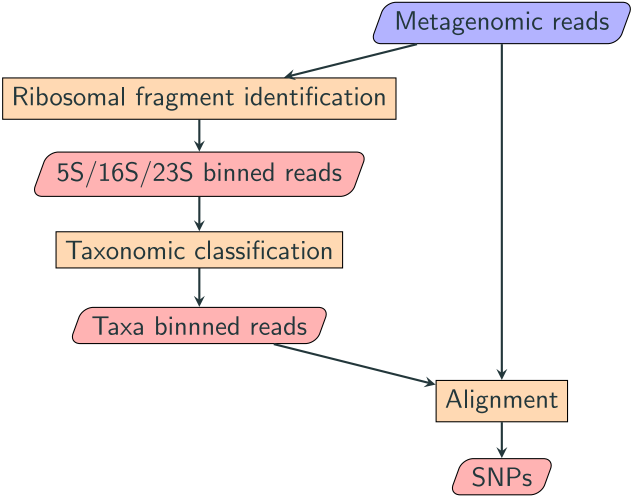
- Raw sequence
- Filtering homology search family similarity/dissimilarity
- Manual feature extraction (GC/TNF/compositional)
- One-hot K-mer representation

Encodings

- Raw sequence
- Filtering homology search family similarity/dissimilarity
- Manual feature extraction (GC/TNF/compositional)
- One-hot K-mer representation
- K-mer embeddings (DNA2vec/BioVec)

Variant Models

Ribosomal Variant Models



Metagenomic reads

Ribosomal fragment identification

5S/16S/23S binned reads

Taxonomic classification

Taxa binnned reads

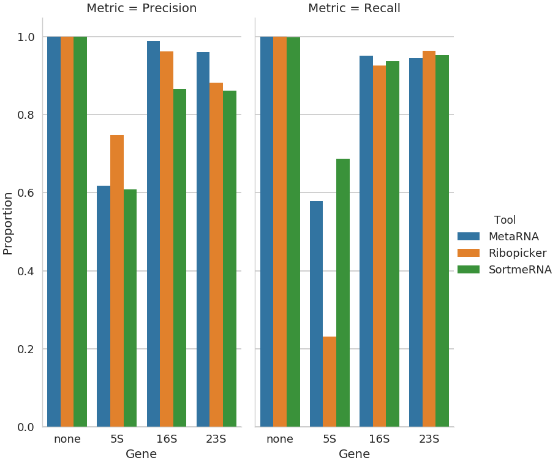
Alignment

SNPs

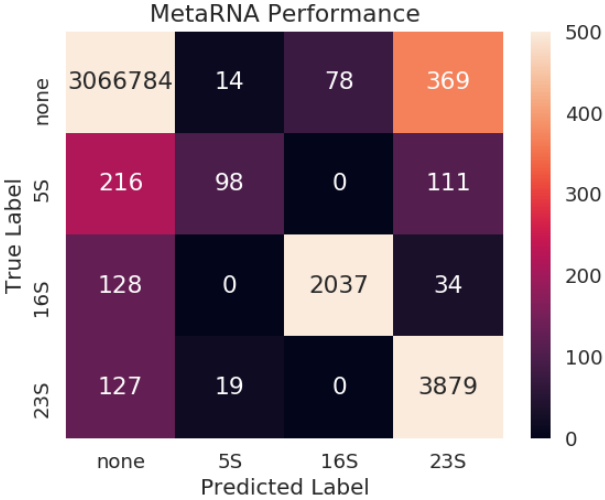
Identifying Ribosomal Reads

- MetaRNA (Huang et al., 2009)
- Ribopicker (Schmieder et al., 2011)
- SortmeRNA (Kopylova et al., 2012)
- 77 models
- Reads simulated from the underlying 30 species reference genomes

Identifying Ribosomal Reads



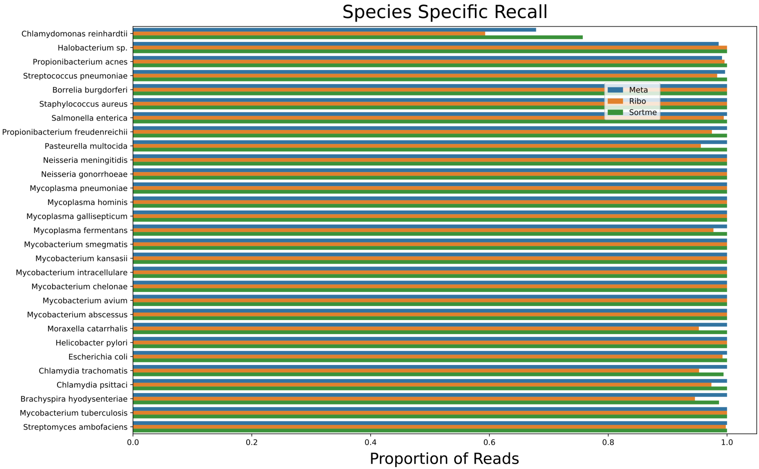
Identifying Ribosomal Reads



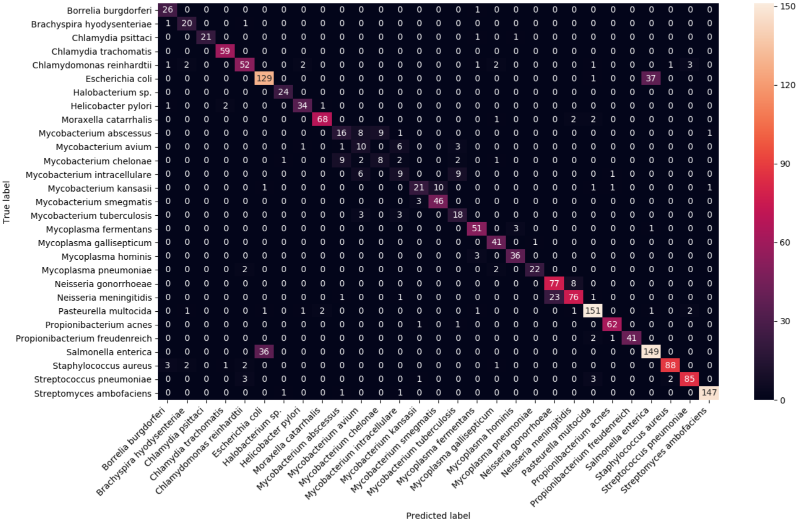
MetaRNA Performance

True Label \ Predicted Label	none	5S	16S	23S
none	3066784	14	78	369
5S	216	98	0	111
16S	128	0	2037	34
23S	127	19	0	3879

Identifying Ribosomal Reads

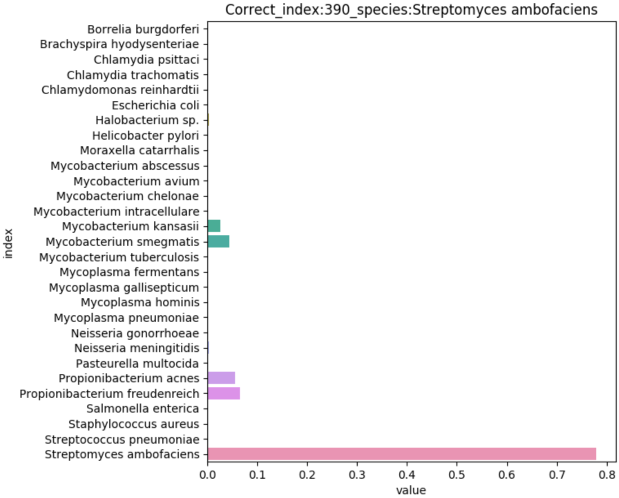


Identifying Taxonomy

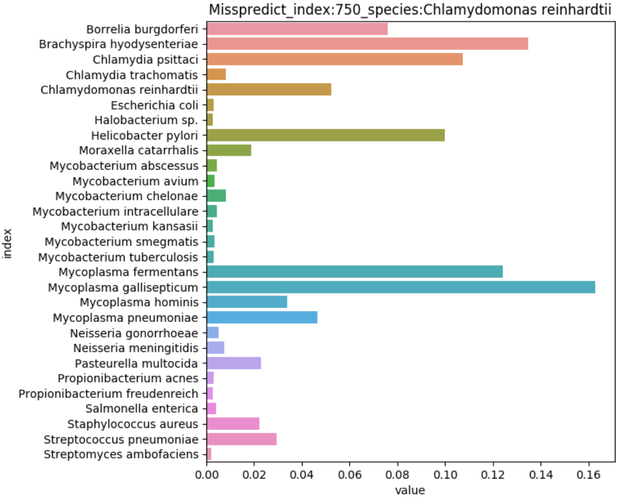


	Borrelia burgdorferi	Brachyspira hyodysenteriae	Chlamydia psittaci	Chlamydia trachomatis	Chlamydomonas reinhardtii	Escherichia coli	Halobacterium sp.	Helicobacter pylori	Moraxella catarrhalis	Mycobacterium abscessus	Mycobacterium avium	Mycobacterium chelonae	Mycobacterium intracellulare	Mycobacterium kansasii	Mycobacterium smegmatis	Mycobacterium tuberculosis	Mycoplasma fermentans	Mycoplasma gallisepticum	Mycoplasma hominis	Mycoplasma pneumoniae	Neisseria gonorrhoeae	Neisseria meningitidis	Pasteurella multocida	Propionibacterium acnes	Propionibacterium freudenreich	Salmonella enterica	Staphylococcus aureus	Streptococcus pneumoniae	Streptomyces ambofaciens
Borrelia burgdorferi	26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
Brachyspira hyodysenteriae	1	20	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Chlamydia psittaci	0	0	21	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0
Chlamydia trachomatis	0	0	0	59	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Chlamydomonas reinhardtii	1	2	0	0	52	0	0	2	0	0	0	0	0	0	0	0	1	2	0	0	0	0	1	0	0	0	1	3	0
Escherichia coli	0	0	0	0	0	129	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	37	0	0	0
Halobacterium sp.	0	0	0	0	0	0	24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Helicobacter pylori	1	0	0	2	0	0	0	34	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Moraxella catarrhalis	0	0	0	0	0	0	0	0	68	0	0	0	0	0	0	0	0	1	0	0	0	2	2	0	0	0	0	0	0
Mycobacterium abscessus	0	0	0	0	0	0	0	0	0	16	8	9	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Mycobacterium avium	0	0	0	0	0	0	0	1	0	1	10	0	6	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0
Mycobacterium chelonae	0	0	0	0	0	0	1	0	0	9	2	8	2	0	0	2	0	1	0	0	0	0	0	0	0	0	0	0	0
Mycobacterium intracellulare	0	0	0	0	0	0	0	0	0	0	6	0	9	0	0	9	0	0	0	0	0	0	0	1	0	0	0	0	0
Mycobacterium kansasii	0	0	0	0	0	1	0	0	0	0	0	0	0	21	10	0	0	0	0	0	0	0	1	1	0	0	0	0	1
Mycobacterium smegmatis	0	0	0	0	0	0	0	0	0	0	0	0	0	3	46	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Mycobacterium tuberculosis	0	0	0	0	0	0	0	0	0	0	3	0	3	0	0	18	0	0	0	0	0	0	0	0	0	0	0	0	0
Mycoplasma fermentans	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	51	0	3	0	0	0	0	0	0	1	0	0	0
Mycoplasma gallisepticum	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	41	0	1	0	0	0	0	0	0	0	0	0
Mycoplasma hominis	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	36	0	0	0	0	0	0	0	0	0	0
Mycoplasma pneumoniae	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	2	0	22	0	0	0	0	0	0	0	0	0
Neisseria gonorrhoeae	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	77	8	0	0	0	0	0	0	0
Neisseria meningitidis	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	23	76	1	0	0	0	0	0	0
Pasteurella multocida	0	1	0	0	0	1	0	1	2	0	0	0	0	0	0	0	1	0	0	0	0	1	151	0	0	1	0	2	0
Propionibacterium acnes	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	62	0	0	0	0	0
Propionibacterium freudenreich	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	1	41	0	0	0	0
Salmonella enterica	0	0	0	0	0	36	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	149	0	0	0
Staphylococcus aureus	3	2	0	1	2	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	88	0	0
Streptococcus pneumoniae	0	0	0	0	3	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	3	0	0	0	2	85	0
Streptomyces ambofaciens	0	0	0	0	0	0	1	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	147

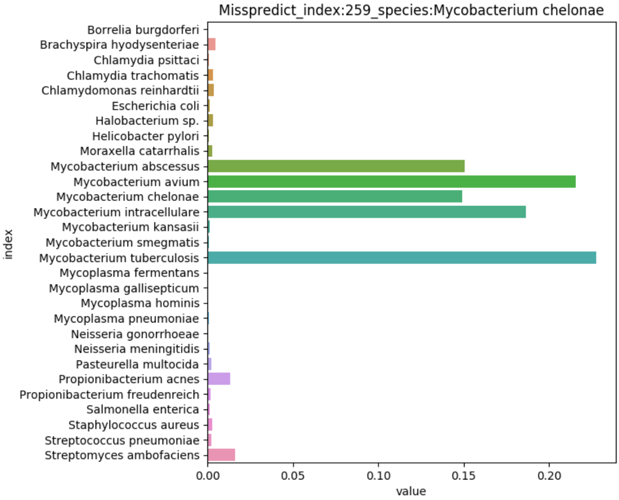
Some are relatively easy



Others are a mess

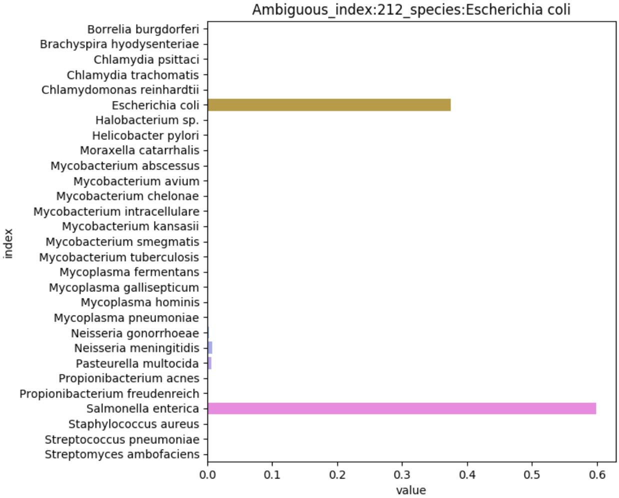


Some are group ambiguous

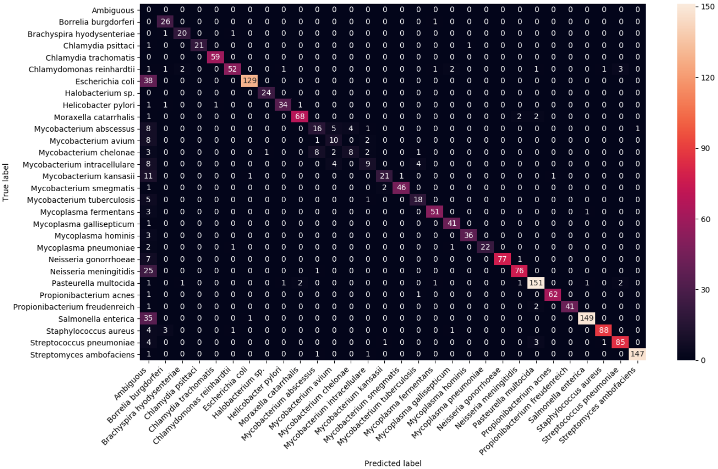


Probably a Mycobacterium?

Others are just a toss-up



Ambiguity in classification



	Ambiguous	Borrelia burgdorferi	Brachyspira hyodysenteriae	Chlamydia psittaci	Chlamydia trachomatis	Chlamydomonas reinhardtii	Escherichia coli	Halobacterium sp.	Helicobacter pylori	Moraxella catarrhalis	Mycobacterium abscessus	Mycobacterium avium	Mycobacterium chelonae	Mycobacterium intracellulare	Mycobacterium kansasii	Mycobacterium smegmatis	Mycobacterium tuberculosis	Mycoplasma fermentans	Mycoplasma gallisepticum	Mycoplasma hominis	Mycoplasma pneumoniae	Neisseria gonorrhoeae	Neisseria meningitidis	Pasteurella multocida	Propionibacterium acnes	Propionibacterium freudenreich	Salmonella enterica	Staphylococcus aureus	Streptococcus pneumoniae	Streptomyces ambofaciens
Ambiguous	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Borrelia burgdorferi	0	26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
Brachyspira hyodysenteriae	0	1	20	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Chlamydia psittaci	1	0	0	21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
Chlamydia trachomatis	0	0	0	0	59	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Chlamydomonas reinhardtii	1	1	2	0	0	52	0	0	1	0	0	0	0	0	0	0	0	1	2	0	0	0	0	1	0	0	0	1	3	0
Escherichia coli	38	0	0	0	0	0	129	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Halobacterium sp.	0	0	0	0	0	0	0	24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Helicobacter pylori	1	1	0	0	1	0	0	0	34	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Moraxella catarrhalis	1	0	0	0	0	0	0	0	0	68	0	0	0	0	0	0	0	0	0	0	0	0	2	2	0	0	0	0	0	0
Mycobacterium abscessus	8	0	0	0	0	0	0	0	0	0	16	5	4	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Mycobacterium avium	8	0	0	0	0	0	0	0	0	0	1	10	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Mycobacterium chelonae	3	0	0	0	0	0	0	1	0	0	8	2	8	2	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
Mycobacterium intracellulare	8	0	0	0	0	0	0	0	0	0	0	4	0	9	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0
Mycobacterium kansasii	11	0	0	0	0	0	1	0	0	0	0	0	0	0	21	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0
Mycobacterium smegmatis	1	0	0	0	0	0	0	0	0	0	0	0	0	0	2	46	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Mycobacterium tuberculosis	5	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	18	0	0	0	0	0	0	0	0	0	0	0	0	0
Mycoplasma fermentans	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	51	0	0	0	0	0	0	0	0	1	0	0	0
Mycoplasma gallisepticum	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	41	0	0	0	0	0	0	0	0	0	0	0
Mycoplasma hominis	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	36	0	0	0	0	0	0	0	0	0	0
Mycoplasma pneumoniae	2	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	22	0	0	0	0	0	0	0	0	0
Neisseria gonorrhoeae	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	77	1	0	0	0	0	0	0	0
Neisseria meningitidis	25	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	76	0	0	0	0	0	0	0
Pasteurella multocida	1	0	1	0	0	0	0	0	1	2	0	0	0	0	0	0	0	1	0	0	0	0	1	151	0	0	1	0	2	0
Propionibacterium acnes	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	62	0	0	0	0	0
Propionibacterium freudenreich	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	41	0	0	0	0
Salmonella enterica	35	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	149	0	0	0
Staphylococcus aureus	4	3	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	88	0	0
Streptococcus pneumoniae	4	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	3	0	0	0	1	85	0
Streptomyces ambofaciens	1	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	147

Next Steps

- Mapping reads to reference to assess presence or absence of mutation related SNP
- Comparison of whole pipeline with just direct mapping to database of ribosomal sequences and SNP calling approaches.
- Tuning of sensitivity for number of potential SNPs required to make a prediction of AMR.

Summary

Conclusions

- AMRtime still not a ‘fait accompli‘

Conclusions

- AMRtime still not a ‘fait accompli‘
- Filtering analysis possibly needs redone for fixed CARD

Conclusions

- AMRtime still not a ‘fait accompli‘
- Filtering analysis possibly needs redone for fixed CARD
- False positive analysis pending for best settings

Conclusions

- AMRtime still not a ‘fait accompli‘
- Filtering analysis possibly needs redone for fixed CARD
- False positive analysis pending for best settings
- Framework and code developed for sensitive homology classification but optimisation and evaluation work still required

Conclusions

- AMRtime still not a ‘fait accompli‘
- Filtering analysis possibly needs redone for fixed CARD
- False positive analysis pending for best settings
- Framework and code developed for sensitive homology classification but optimisation and evaluation work still required
- Not shown but preliminary family level classification shows 100x improvements over previous ARO attempts

Conclusions

- AMRtime still not a ‘fait accompli‘
- Filtering analysis possibly needs redone for fixed CARD
- False positive analysis pending for best settings
- Framework and code developed for sensitive homology classification but optimisation and evaluation work still required
- Not shown but preliminary family level classification shows 100x improvements over previous ARO attempts
- Ribosomal Variant Model work progressing well with full pipeline metrics available soon.

Acknowledgements

Acknowledgements

- Zhou Zhilei
- Brian Alcock, Amos Raphenya, Kara Tsang
- Rob Beiko, Fiona Brinkman and Andrew McArthur
- Funding: Genome Canada and a NERC Undergraduate Student Research Award

References

Buchfink, B., Xie, C., and Huson, D. H. (2015). Fast and sensitive protein alignment using diamond. *Nature methods*, 12(1):59.

Gish, W. et al. (1993). Identification of protein coding regions by database similarity search. *Nature genetics*, 3(3):266.

Huang, Y., Gilna, P., and Li, W. (2009). Identification of ribosomal rna genes in metagenomic fragments. *Bioinformatics*, 25(10):1338–1340.

Jia, B., Raphenya, A. R., Alcock, B., Waglechner, N., Guo, P., Tsang, K. K., Lago, B. A., Dave, B. M., Pereira, S., Sharma, A. N., et al. (2016). Card 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic acids research*, page gkw1004.

Kopylova, E., Noé, L., and Touzet, H. (2012). Sortmerna: fast and accurate filtering of ribosomal rnas in metatranscriptomic data. *Bioinformatics*, 28(24):3211–3217.

Schmieder, R., Lim, Y. W., and Edwards, R. (2011). Identification and removal of ribosomal rna sequences from metatranscriptomes. *Bioinformatics*, 28(3):433–435.

Steinegger, M. and Söding, J. (2017). Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature biotechnology*, 35(11):1026.

Westbrook, A., Ramsdell, J., Schuelke, T., Normington, L., Bergeron, R. D., Thomas, W. K., and MacManes, M. D. (2017). Paladin: protein alignment for functional profiling whole metagenome shotgun data. *Bioinformatics*, 33(10):1473–1478.