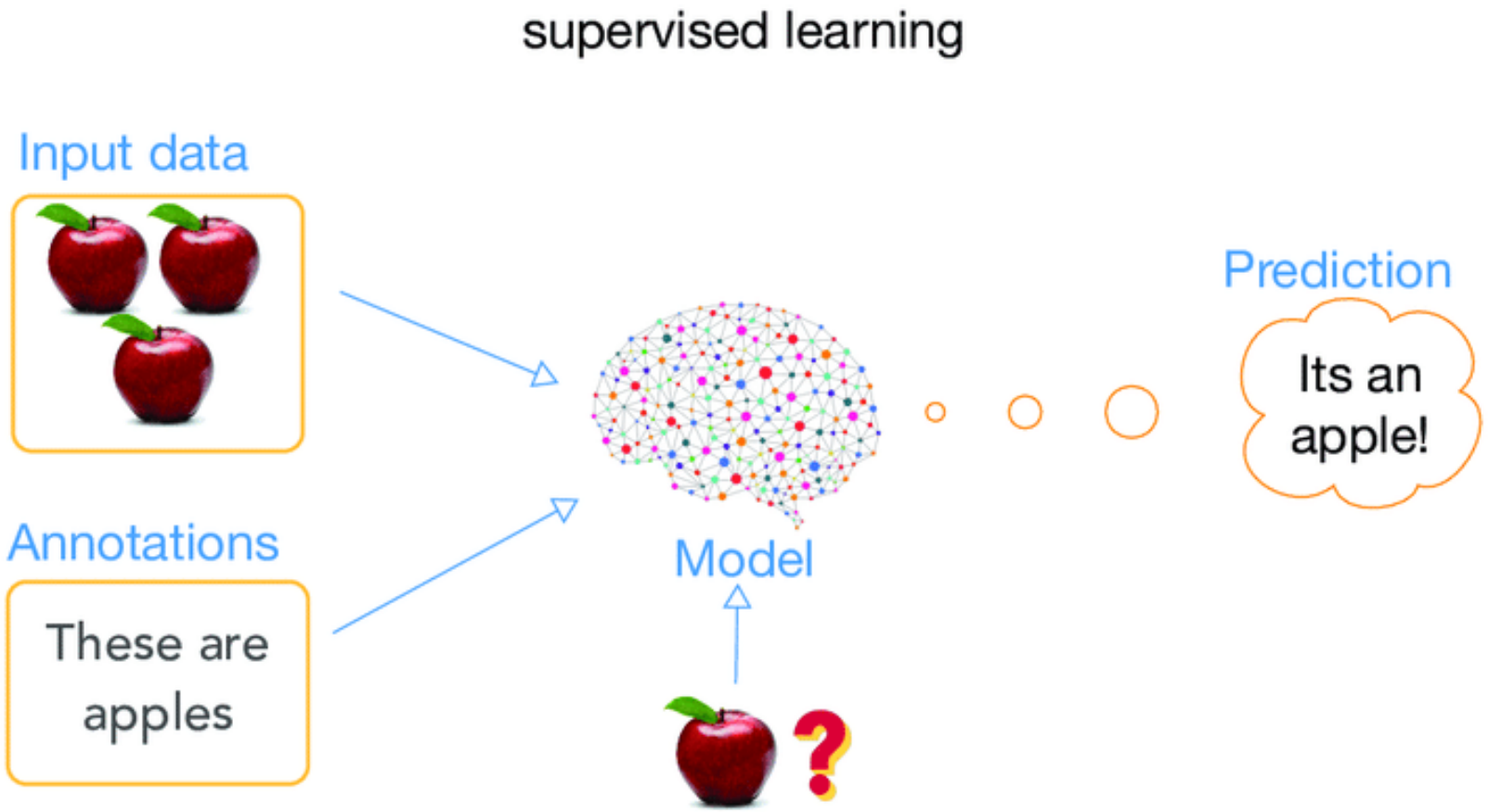
Validation of machine learning models

Derya Aytan-Aktug

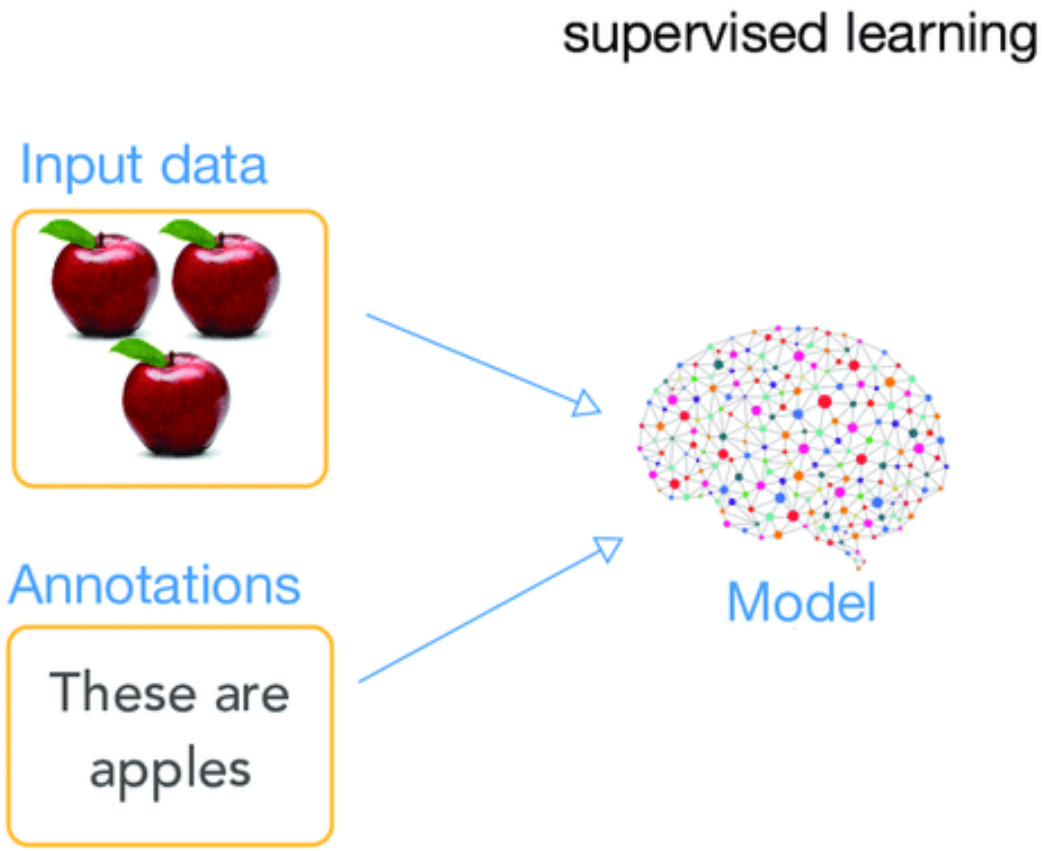
Machine learning is cool!

- Automated pattern detection method
- Unsupervised/supervised

Very briefly supervised machine learning



Feature importance



Discoveries with machine learning

- Novel antimicrobial resistance genes
- Novel plasmid hosts

Discoveries with machine learning

- Novel antimicrobial resistance genes
- Novel plasmid hosts

Are they?

Discoveries with machine learning

- Novel antimicrobial resistance genes
- Novel plasmid hosts

Are they?

- **Linkage disequilibrium**
- **Epistatic interactions**
- **Wrong**
- **…**

Project-1: Machine learning based ResFinder

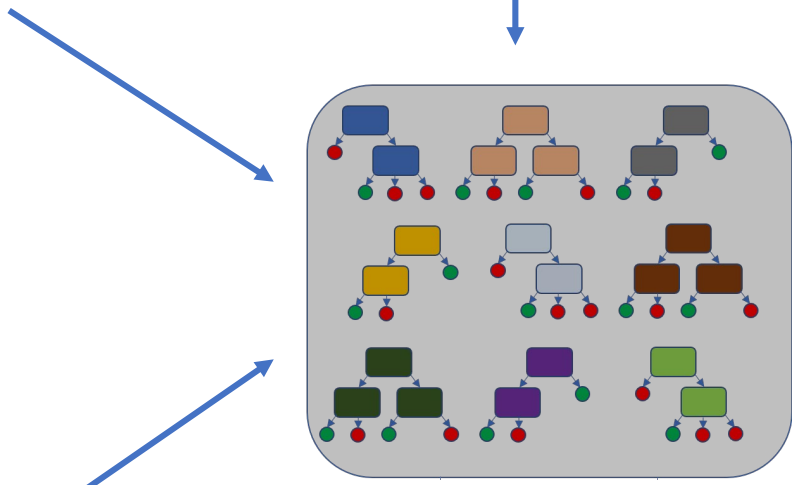
- **Goals:**
  - Unknown effects of known resistance genes?
  - In-vitro validation of the models with TWIW isolates

Method

	Gene-1	Gene-2	Gene-3
Isolate-1	coverage	coverage	coverage
Isolate-2	coverage	coverage	coverage
…	…	…	…

	Gene-1	Gene-2	Gene-3
Isolate-1000	coverage	coverage	coverage
Isolate-10001	coverage	coverage	coverage
…	…	…	…

Pheno-1	Pheno-2	Pheno-3
1/0	1/0	1/0
1/0	1/0	1/0
…	…	…



Pheno-1	Pheno-2	Pheno-3
0	1	0
1	1	0
…	…	…

Organism	Antibiotic	ResFinder - MCC	PointFinder - MCC	ResFinder + PointFinder - MCC	5-mers - MCC	8-mers-MCC
Neisseria	Ciprofloxacin	0,207	0,749	0,751	0,759	0,792
Salmonella	Ampicillin	0,955	0,36	0,954	0,761	0,877
Salmonella	Streptomycin	0,892	0,446	0,864	0,746	0,836
Salmonella	Tetracycline	0,955	0,504	0,951	0,843	0,904
Campylobacter	Ciprofloxacin	0,251	0,825	0,787	0,576	0,633
Campylobacter	Tetracycline	0,954	0,273	0,951	0,859	0,949
Escherichia	Amikacin	0,526	0,472	0,618	0,51	0,526
Escherichia	Ampicillin	0,857	0,388	0,75	0,418	0,574
Escherichia	Cefoxitin	0,757	0,192	0,676	0,411	0,725
Escherichia	Chloramphenicol	0,537	0,062	0,377	0,148	0,311
Escherichia	Colistin	0,458	0,452	0,518	0,383	0,625
Escherichia	Gentamicin	0,669	0,303	0,582	0,425	0,466
Escherichia	Imipenem	0,62	0,369	0,568	0,56	0,59
Escherichia	Tetracycline	0,7	0,324	0,542	0,271	0,508
Escherichia	Tigecycline	0,479	0,531	0,584	0,326	0,598
Escherichia	Tobramycin	0,745	0,069	0,66	0,351	0,504
Klebsiella	Amikacin	0,877			0,762	0,809
Klebsiella	Cefoxitin	0,52			0,525	0,591
Klebsiella	Colistin	0,182			0,189	0,208
Klebsiella	Gentamicin	0,781			0,445	0,594
Klebsiella	Imipenem	0,813			0,74	0,845
Klebsiella	Minocycline	0,328			0,387	0,359
Klebsiella	Tetracycline	0,698			0,414	0,54
Klebsiella	Tigecycline	0,065			0,122	0,202
Klebsiella	Tobramycin	0,853			0,54	0,688
Mycobacterium	Amikacin	0	0,745	0,652	0,17	0,561

Feature importance results

- aadA1_4_JQ480156 imipenem
- aadA1b_1_M95287 amikacin
- ant(3'')-Ia_1_X02340 amikacin, cefoxitin, gentamicin, minocycline, tigecycline, tobramycin
- aph(6)-Id_1_M28829 amikacin, tetracycline
- blaACT-5_1_FJ237369 amikacin
- blaMIR-1_1_M37839 amikacin, chloramphenicol, colistin, tigecycline
- blaOXA-9_1_KQ089875 imipenem
- blaTEM-1A_1_HM749966 imipenem
- catB3_2_U13880 tobramycin
- formA_1_X73835 amikacin, ampicillin, cefoxitin, colistin, gentamicin, imipenem, tigecycline
- mdf(A)_1_Y08743 cefoxitin, gentamicin, imipenem, minocycline, amikacin
- …

Feature importance results

- aadA1_4_JQ480156 imipenem
- aadA1b_1_M95287 amikacin
- ant(3'')-Ia_1_X02340 amikacin, cefoxitin, gentamicin, minocycline, tigecycline, tobramycin
- aph(6)-Id_1_M28829 amikacin, tetracycline
- blaACT-5_1_FJ237369 amikacin
- blaMIR-1_1_M37839 amikacin, chloramphenicol, colistin, tigecycline
- blaOXA-9_1_KQ089875 imipenem
- blaTEM-1A_1_HM749966 imipenem
- **catB3_2_U13880 tobramycin**
- formA_1_X73835 amikacin, ampicillin, cefoxitin, colistin, gentamicin, imipenem, tigecycline
- mdf(A)_1_Y08743 cefoxitin, gentamicin, imipenem, minocycline, amikacin
- …

How the validation could be done?

Target:

catB3-2 → tobramycin

Validation isolates:

Should not include any known tobramycin resistance genes or mutations

Should include catB3-2 gene

Should be tested for tobramycin

But there might no case…

- A subset from TWIW isolates tested in-house:
  - CatB3-2 including *E.coli* and *K.pneumoniae*
    - All of them resistance against tobramycin
      - In the same contig, blaOXA-1 and/or aac(6')-Ib-cr present

- CatB3-2, not responsible for tobramycin resistance
  - In linkage disequilibrium with aac(6')-Ib-cr

Take home messages

- Important features of prediction model should be interpreted in caution.
- In-vitro experiments are required.

Weakness of machine learning

- Showing the associations between the features

How project-1 will continue?

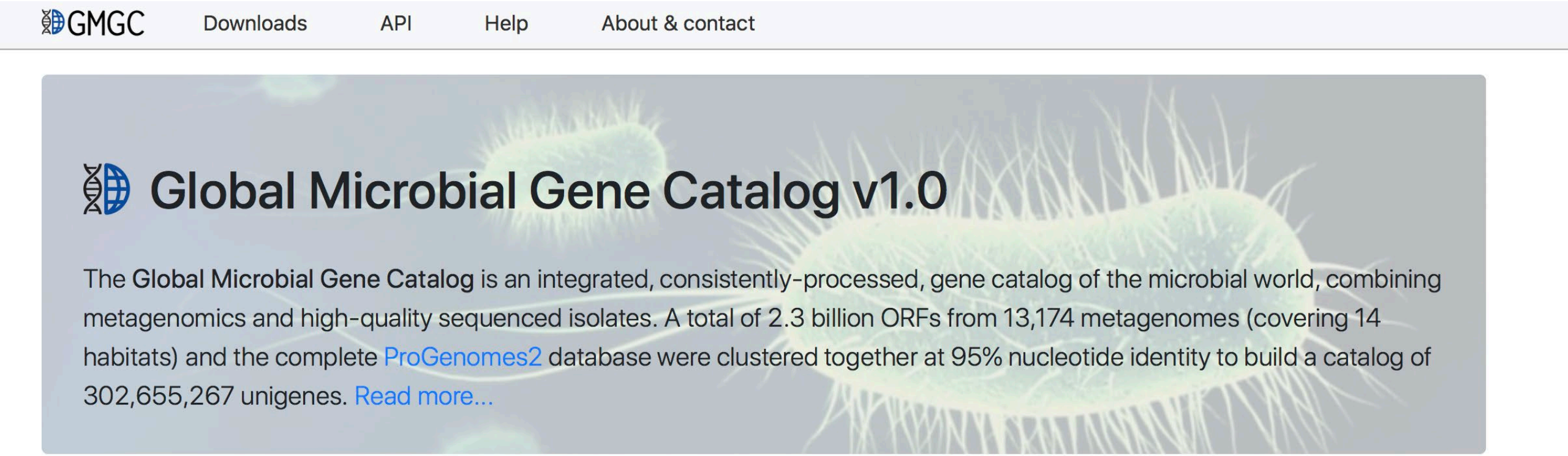
- Another model will be developed from pangenomes.
  - To identify novel resistance genes
    - Models and novel genes will be validated using the in-house data.

Previous pangenome studies:

- Pangenome annotation using prodigal
- Clustering annotated genes using CD-HIT using aa identity

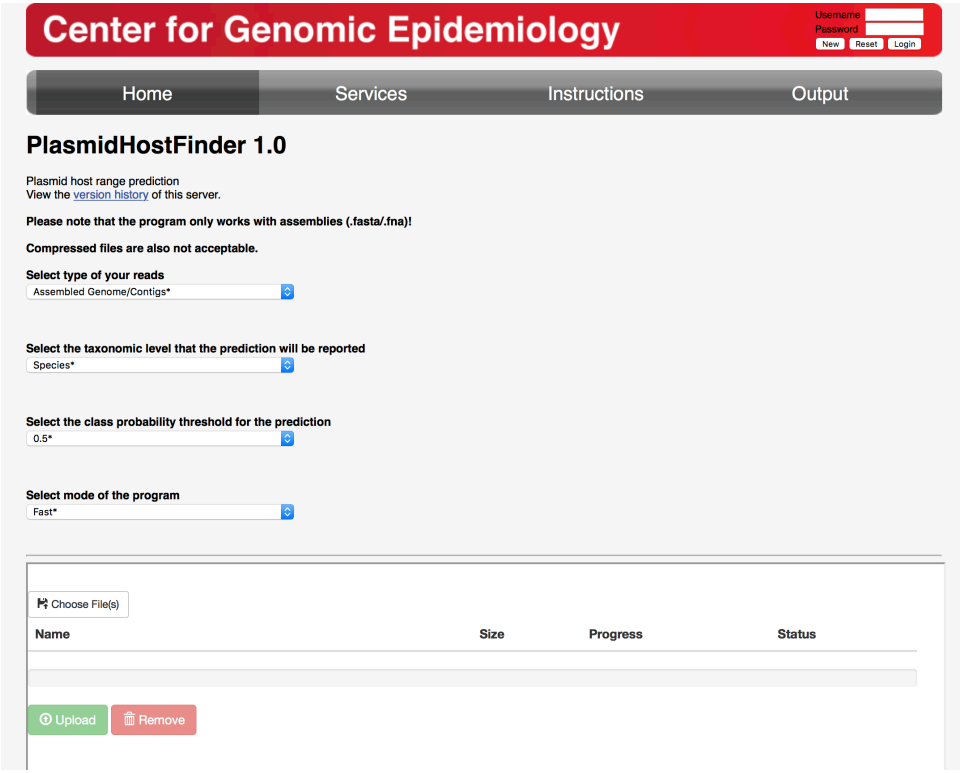
My approach:

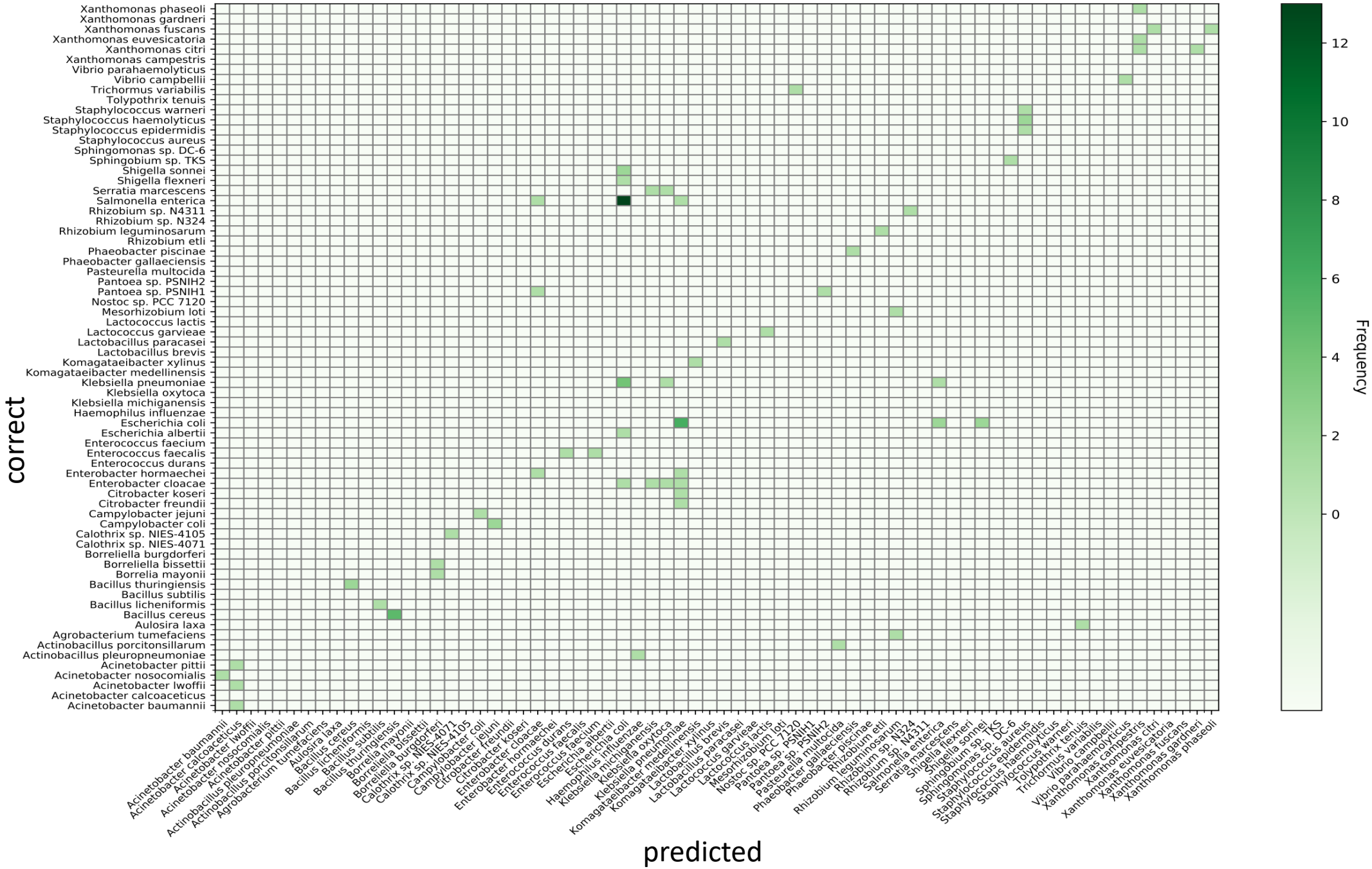
- Aligning against global gene catalog (homology reduced) using KMA
  - Annotate the important features later



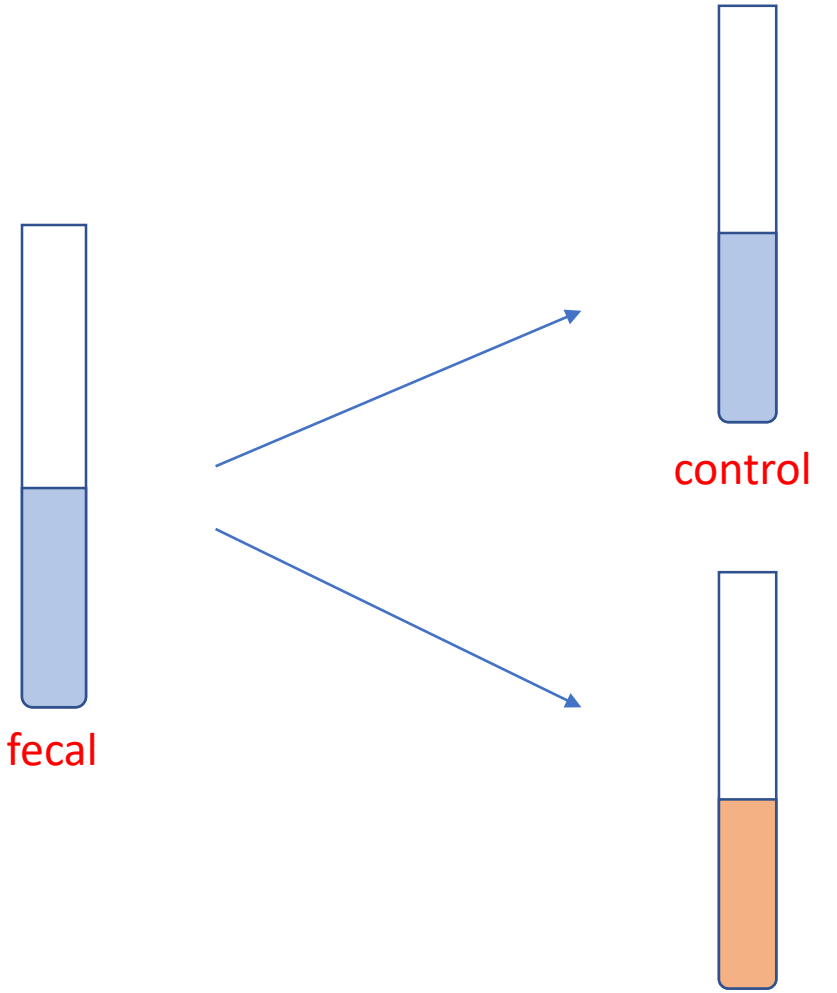
Project-2: Validation of PlasmidHostFinder

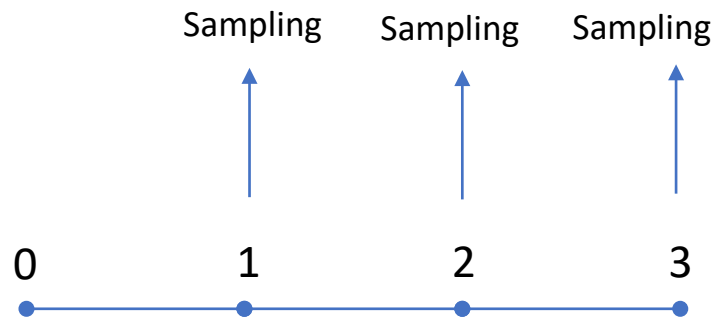
- Prediction of plasmid hosts
  - K-mer based model
    - At different taxonomy levels
- False positives




correct
predicted
Frequency
0
2
4
6
8
10
12
Xanthomonas phaseoli
Xanthomonas gardneri
Xanthomonas fuscans
Xanthomonas euvesicatoria
Xanthomonas citri
Xanthomonas campestris
Vibrio parahaemolyticus
Vibrio campbellii
Trichormus variabilis
Tolypothrix tenuis
Staphylococcus warneri
Staphylococcus haemolyticus
Staphylococcus epidermidis
Staphylococcus aureus
Sphingomonas sp. DC-6
Sphingobium sp. TKS
Shigella sonnei
Shigella flexneri
Serratia marcescens
Salmonella enterica
Rhizobium sp. N4311
Rhizobium sp. N324
Rhizobium leguminosarum
Rhizobium etli
Phaeobacter piscinae
Phaeobacter gallaeciensis
Pasteurella multocida
Pantoea sp. PSNIH2
Pantoea sp. PSNIH1
Nostoc sp. PCC 7120
Mesorhizobium loti
Lactococcus lactis
Lactococcus garvieae
Lactobacillus paracasei
Lactobacillus brevis
Komagataeibacter xylinus
Komagataeibacter medellinensis
Klebsiella pneumoniae
Klebsiella oxytoca
Klebsiella michiganensis
Haemophilus influenzae
Escherichia coli
Escherichia albertii
Enterococcus faecium
Enterococcus faecalis
Enterococcus durans
Enterobacter hormaechei
Enterobacter cloacae
Citrobacter koseri
Citrobacter freundii
Campylobacter jejuni
Campylobacter coli
Calothrix sp. NIES-4105
Calothrix sp. NIES-4071
Borreliella burgdorferi
Borreliella bissettii
Borrelia mayonii
Bacillus thuringiensis
Bacillus subtilis
Bacillus licheniformis
Bacillus cereus
Aulosira laxa
Agrobacterium tumefaciens
Actinobacillus porcitonsillarum
Actinobacillus pleuropneumoniae
Acinetobacter pittii
Acinetobacter nosocomialis
Acinetobacter lwoffii
Acinetobacter calcoaceticus
Acinetobacter baumannii

Project-3: Tracking metagenomic samples under antimicrobial stress





- **Track organisms**
- **Track resistance**
  - **HGT**
  - **Spontaneous mutations**

Side project: Machine learning-based taxonomy identifier

- Predicting taxonomies from complete and fragmented assemblies
  - Tried for *Klebsiella spp*., *Escherichia spp*., *Salmonella spp*., *Campylobacter spp*.
- Fragment model is less sensitive at the species level.