AMR and machine-learning

Prediction of AMR from metagenomes among other things

Finlay Maguire
finlaymaguire@gmail.com

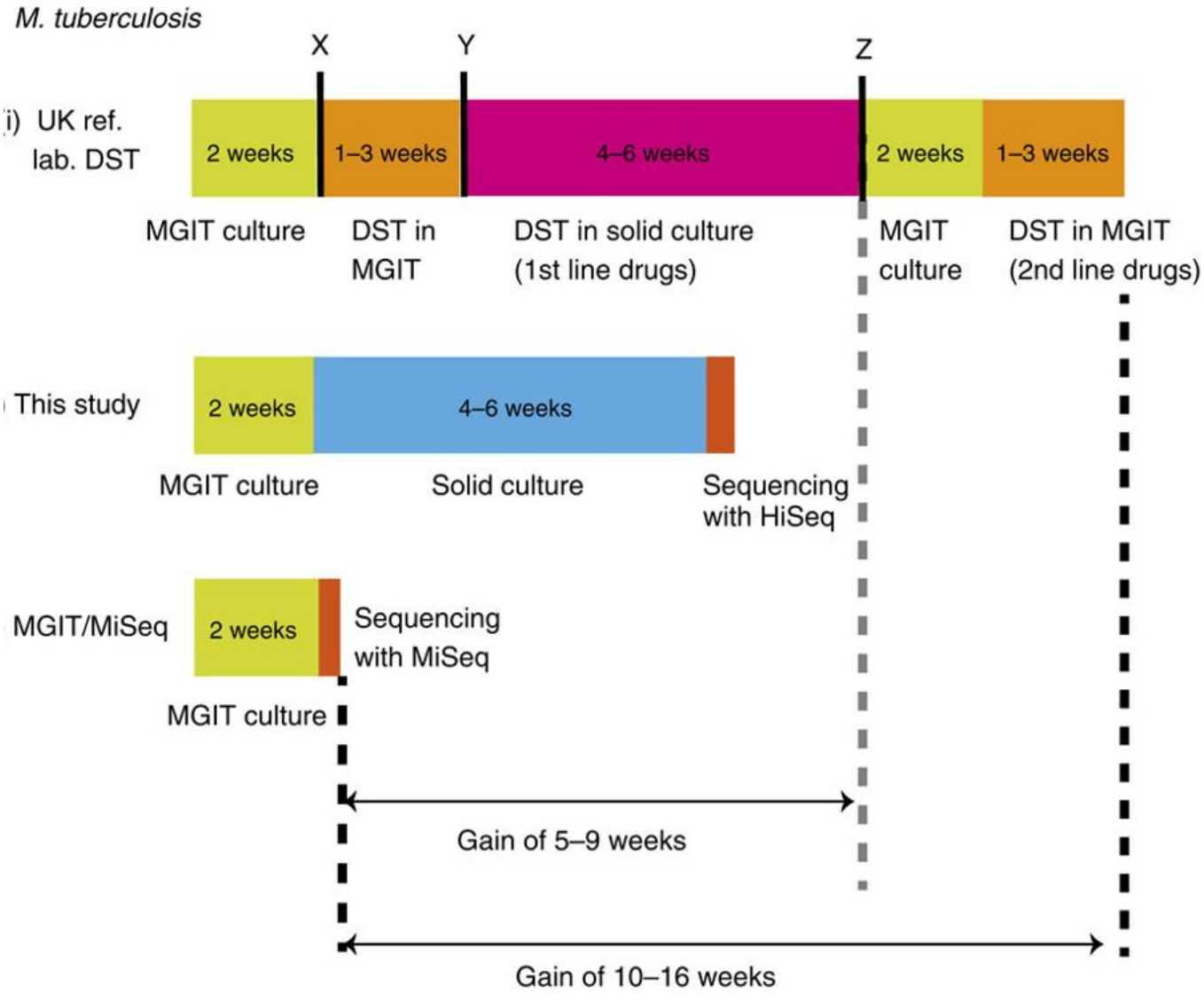
December 3, 2019

Faculty of Computer Science, Dalhousie University

Table of contents

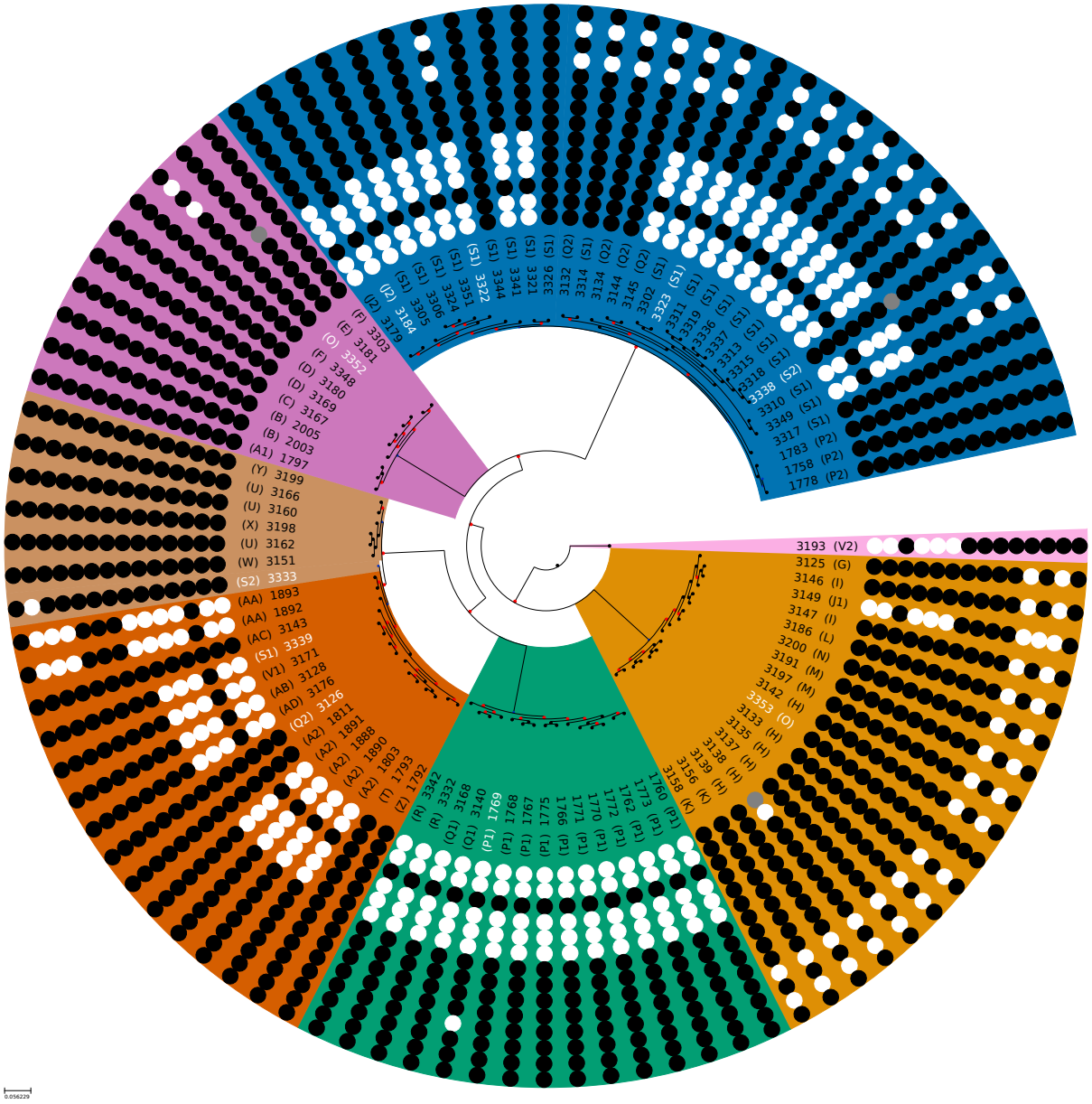
Genomic Phenotype Prediction

Antibiotic Susceptibility Testing

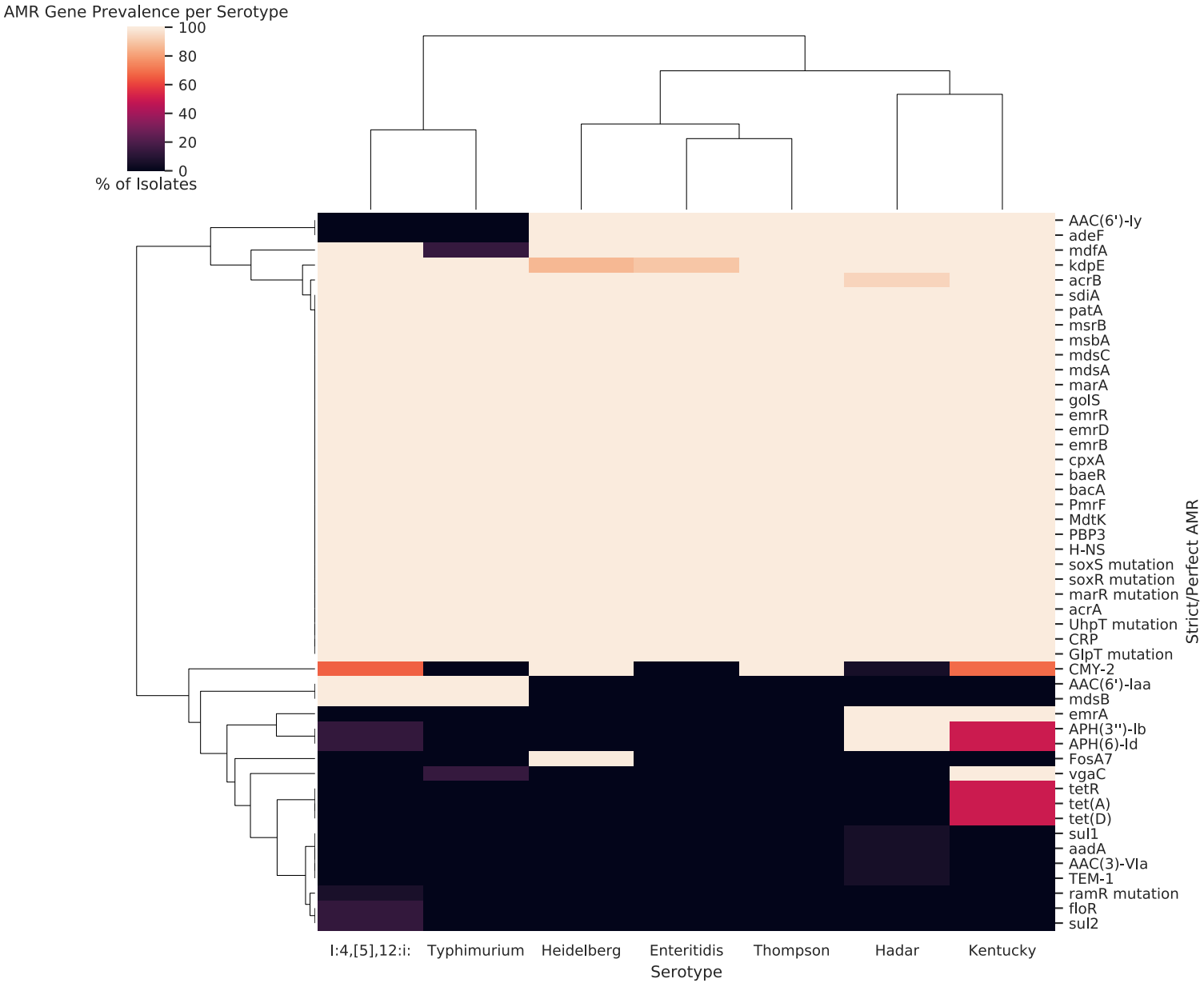


Bradley et al. (2015)

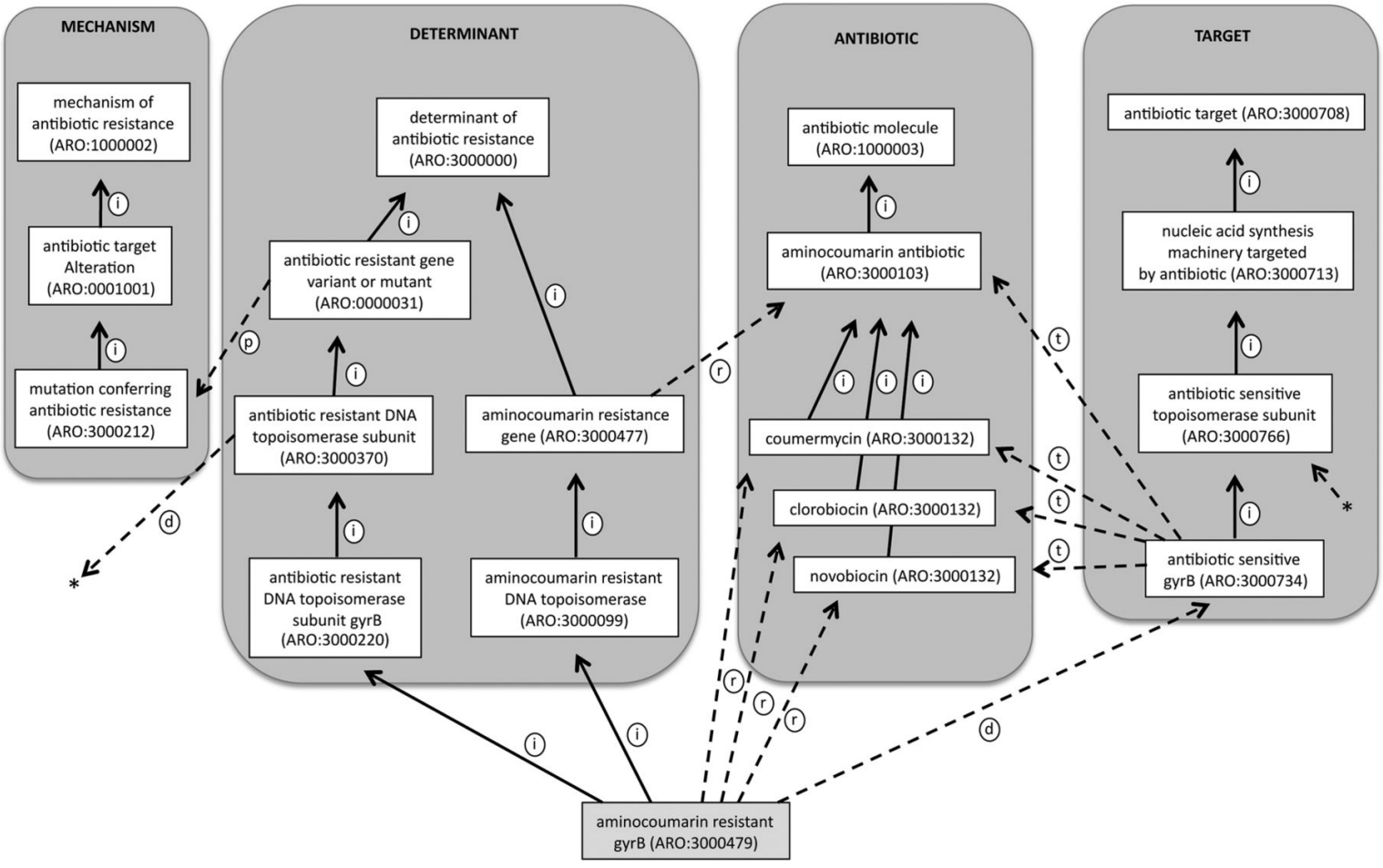
AAFC Salmonella Data-set



Genomic RGI Predictions



Linking AMR determinants to Phenotype



McArthur et al. (2013)

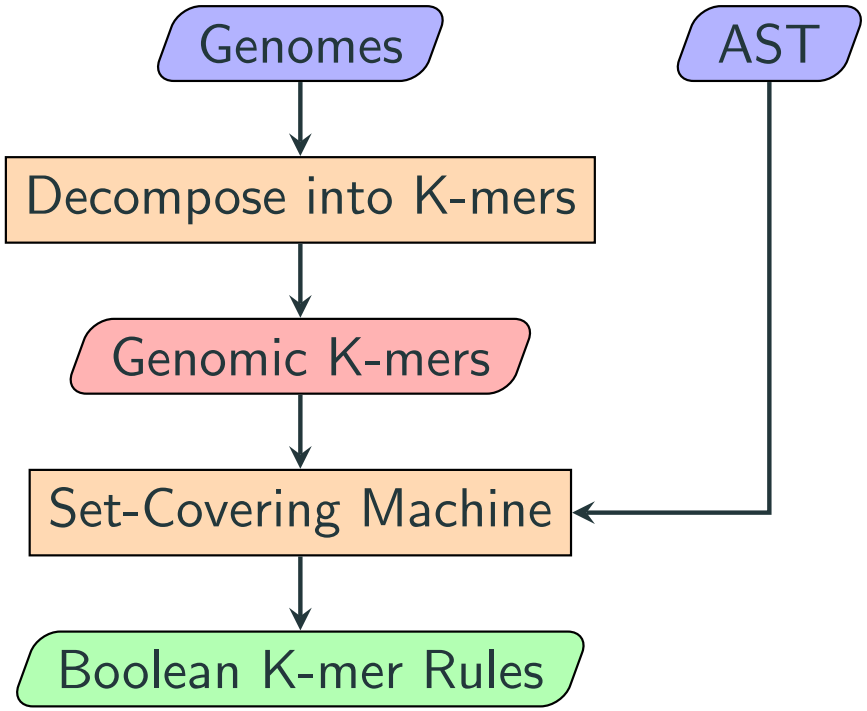
Logistic Regression

$$RGI = \begin{array}{c} \\ genome_1 \\ genome_2 \\ \dots \\ genome_I \end{array} \begin{array}{c} \begin{array}{cccc} amr_1 & amr_2 & \dots & amr_J \end{array} \\ \begin{pmatrix} 1 & 0 & \dots & 1 \\ 0 & 1 & \dots & 1 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 \end{pmatrix} \end{array}$$

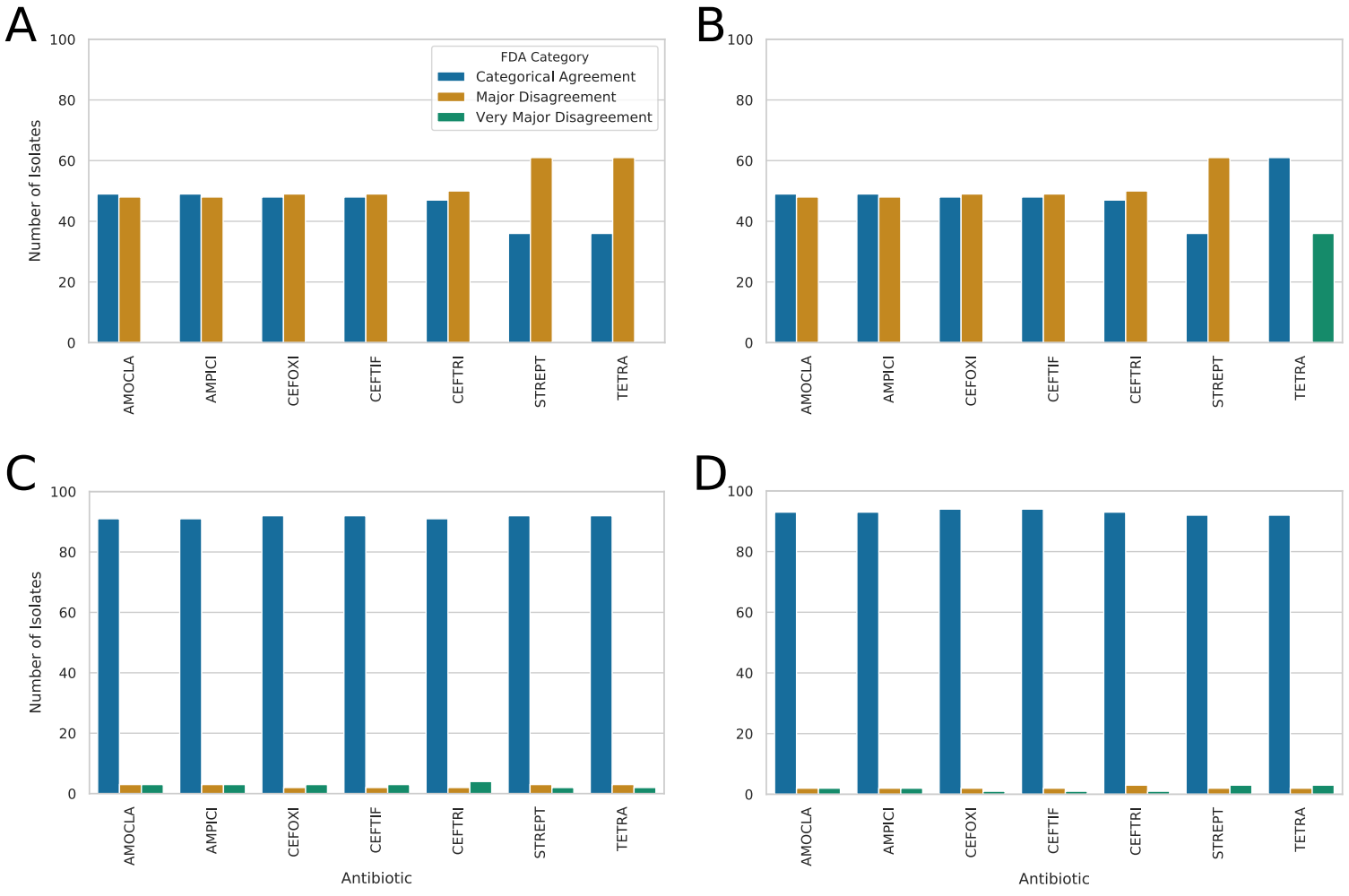
$$AST = \begin{array}{c} \\ genome_1 \\ genome_2 \\ \dots \\ genome_I \end{array} \begin{array}{c} \begin{array}{cccc} abx_1 & abx_2 & \dots & abx_K \end{array} \\ \begin{pmatrix} S & S & \dots & R \\ R & R & \dots & S \\ \dots & \dots & \dots & \dots \\ S & S & \dots & S \end{pmatrix} \end{array}$$

$$\beta RGI = AST$$

Set-Covering Machines



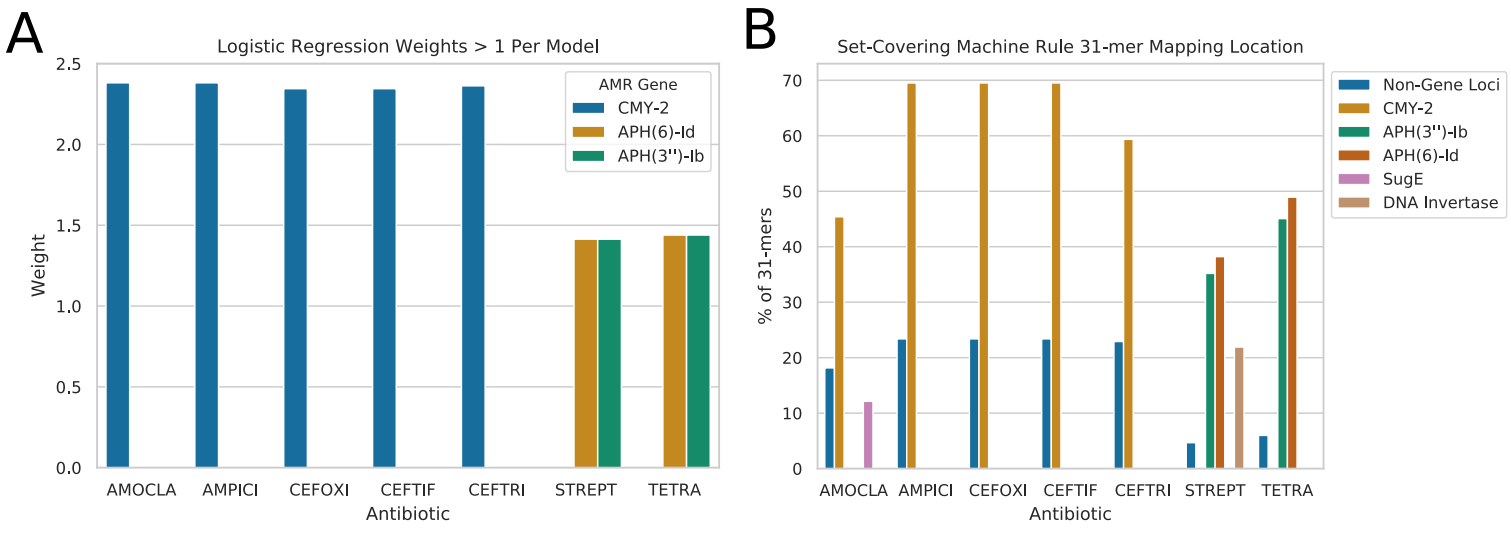
AST Prediction Performance



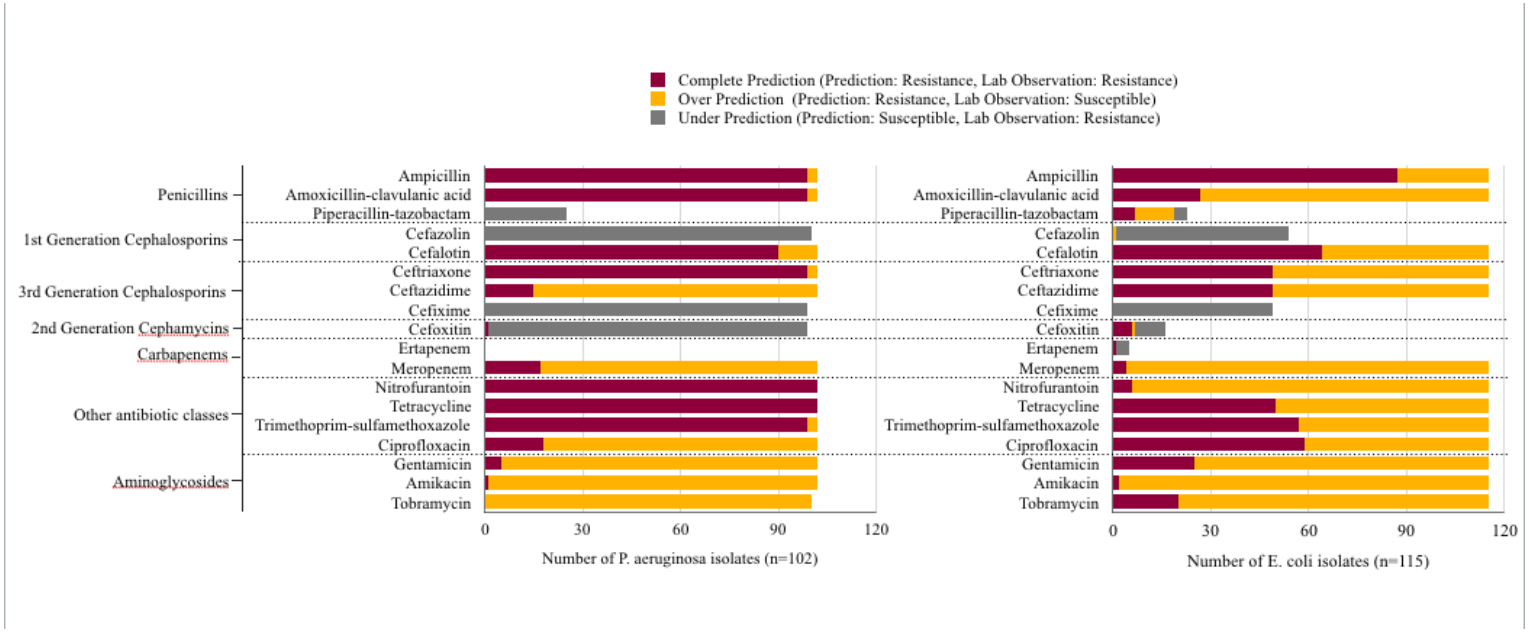
A: RGI, **B**: RGI-efflux, **C**: Logistic Regression, **D**: Set Covering Machines.

Major Disagreement is overprediction of resistance, Very Major Disagreement is underprediction

Learnt features/weights

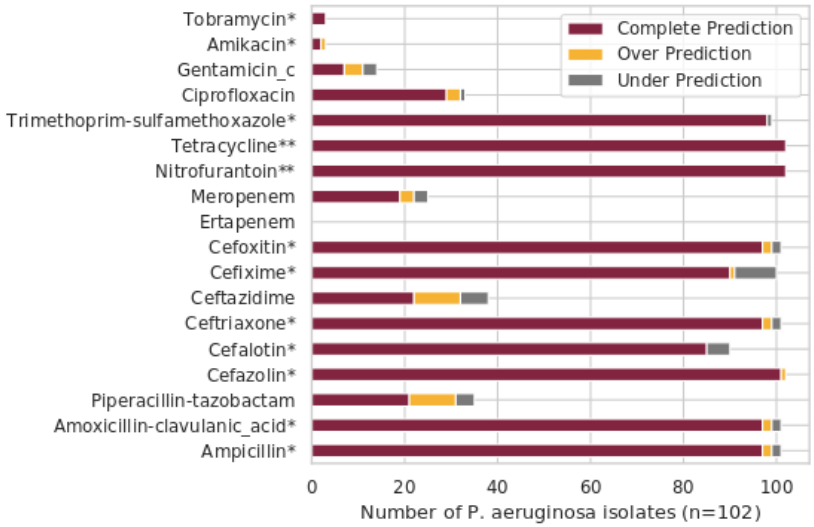


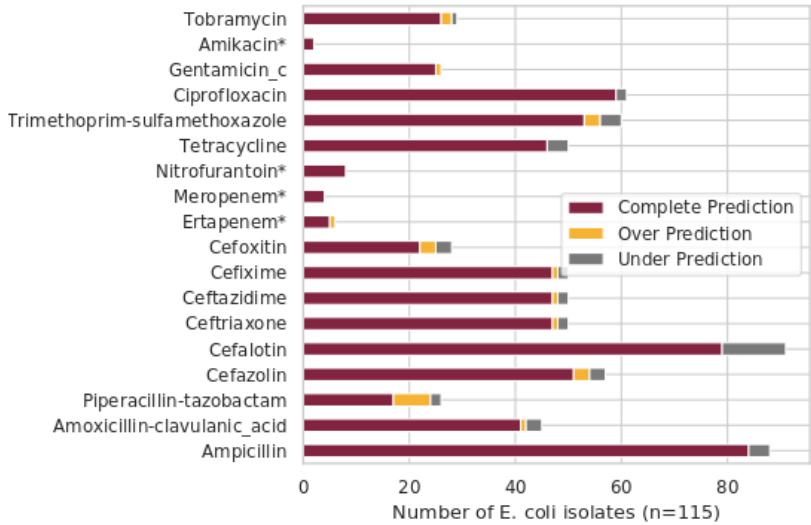
Extending beyond Salmonella



ARO Predictions (Kara Tsang)

Extending beyond Salmonella





Logistic Regression

Genomic AST Prediction

- Using direct annotations works very poorly across different organisms and resistance mechanisms.

Genomic AST Prediction

- Using direct annotations works very poorly across different organisms and resistance mechanisms.
- Even very simple logistic regression models greatly improve predictions.

Genomic AST Prediction

- Using direct annotations works very poorly across different organisms and resistance mechanisms.
- Even very simple logistic regression models greatly improve predictions.
- Investigation of learnt weights and features can be very scientifically informative.

Non-Bioinformatics Interlude

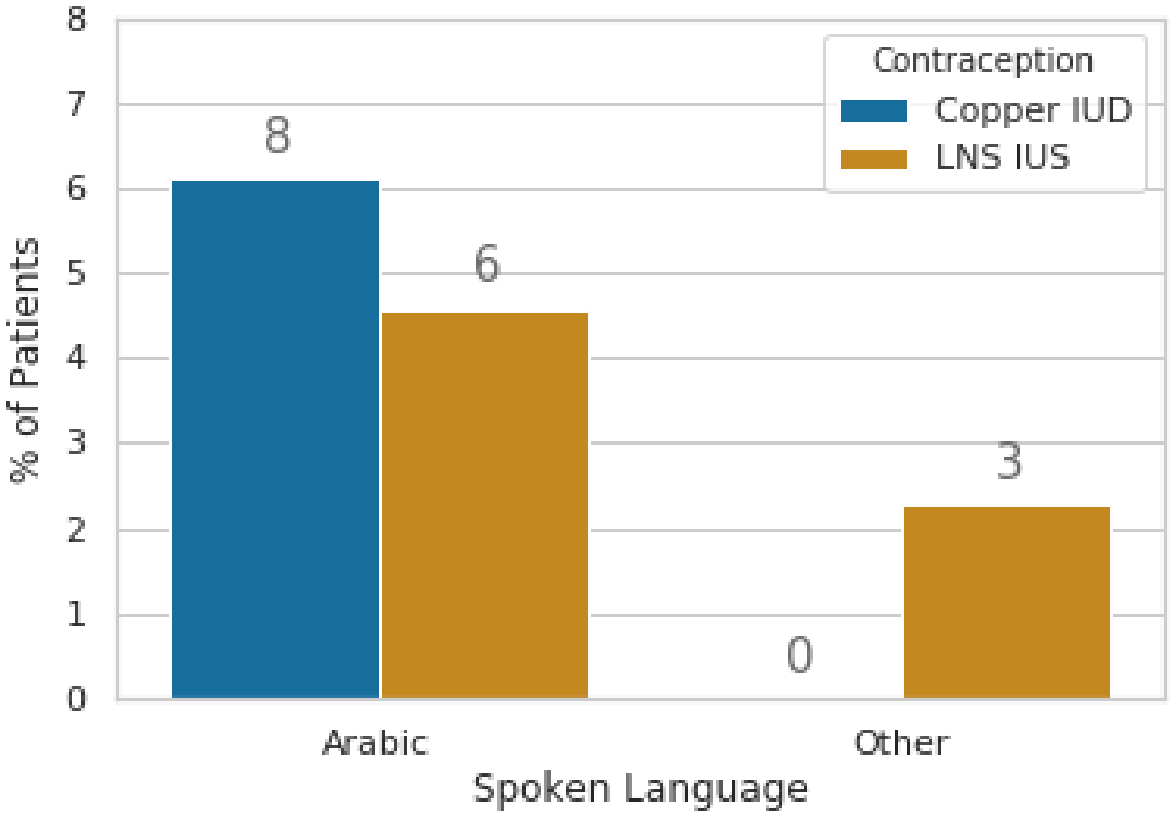
- Non-profits have data and lots of contextualising knowledge.

- Non-profits have data and lots of contextualising knowledge.
- No time or resources to analyse or use it

- Non-profits have data and lots of contextualising knowledge.
- No time or resources to analyse or use it
- Informaticians have the skills and resources but no specific understanding of the context.

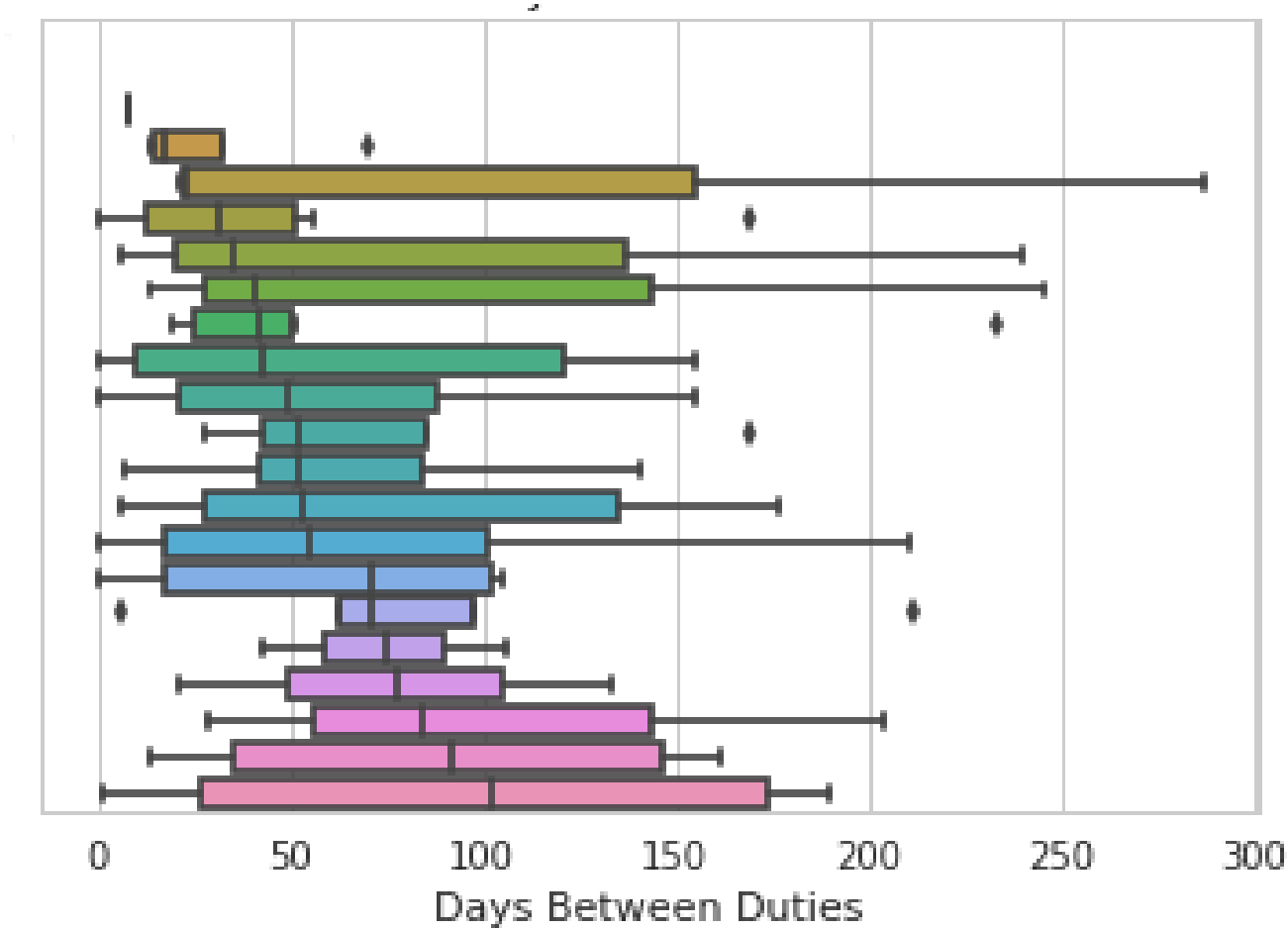
- Non-profits have data and lots of contextualising knowledge.
- No time or resources to analyse or use it
- Informaticians have the skills and resources but no specific understanding of the context.
- Many low-hanging fruit that can make big differences.

Refugee Women's Health Clinic



LEAF FAEJ
Women's Legal Education and Action Fund
Fonds d'action et d'éducation juridiques pour les femmes

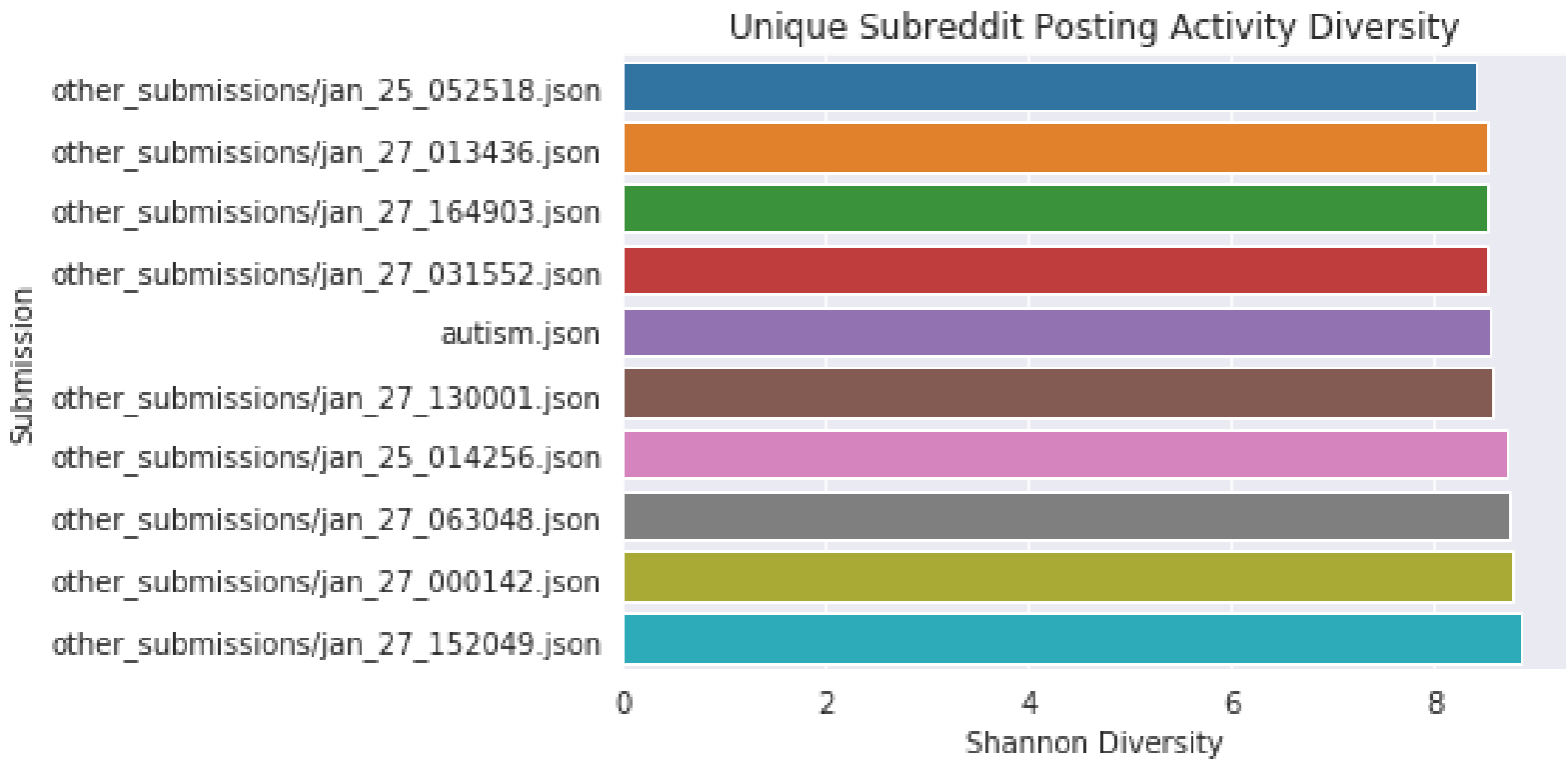
Staff Scheduling



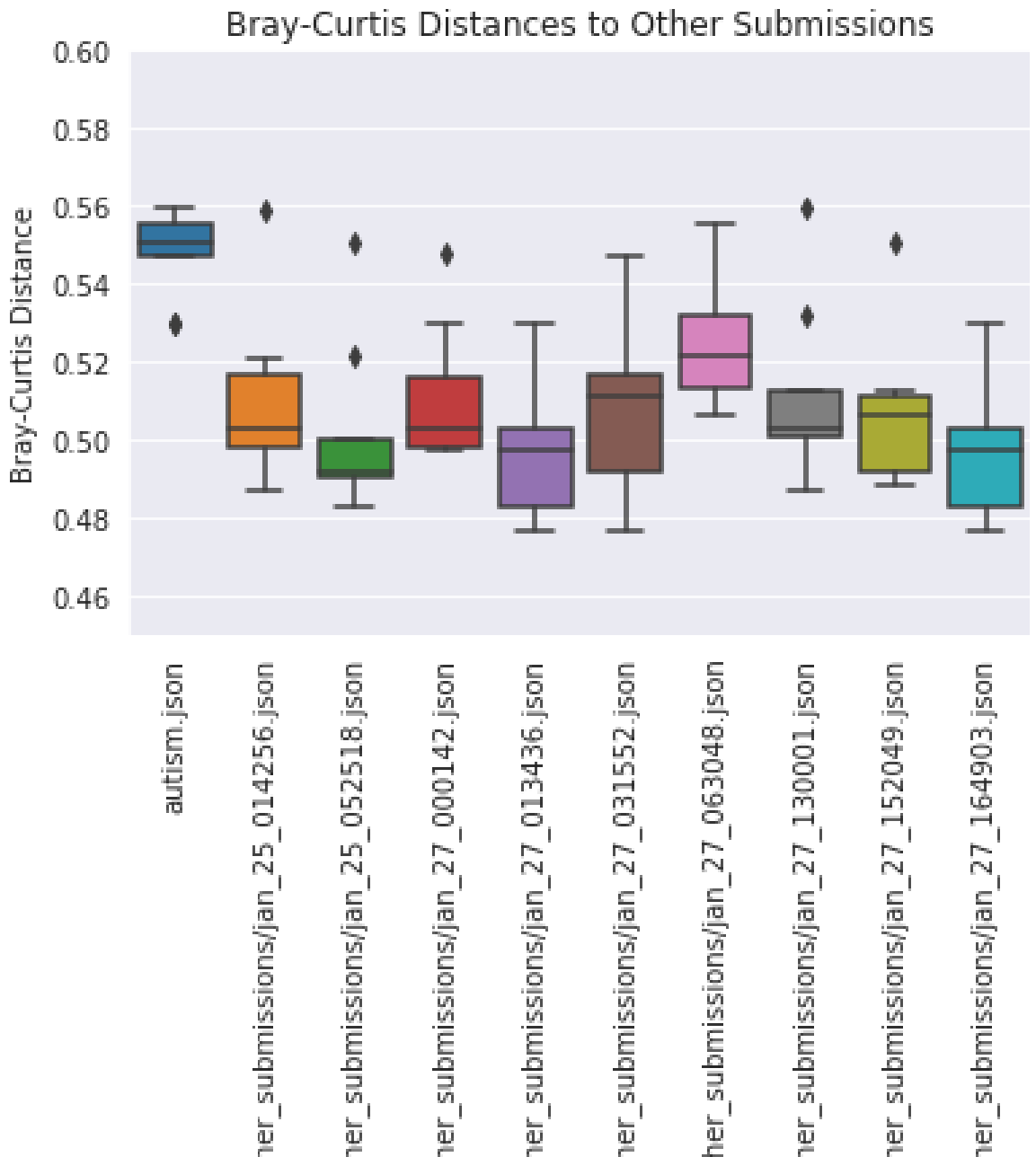
Language Development in Autism

Qualitative Social Media Analysis (Tamara Sorenson-Duncan)

Alpha Diversity of Posting Activity



Beta Diversity of Posting Activity

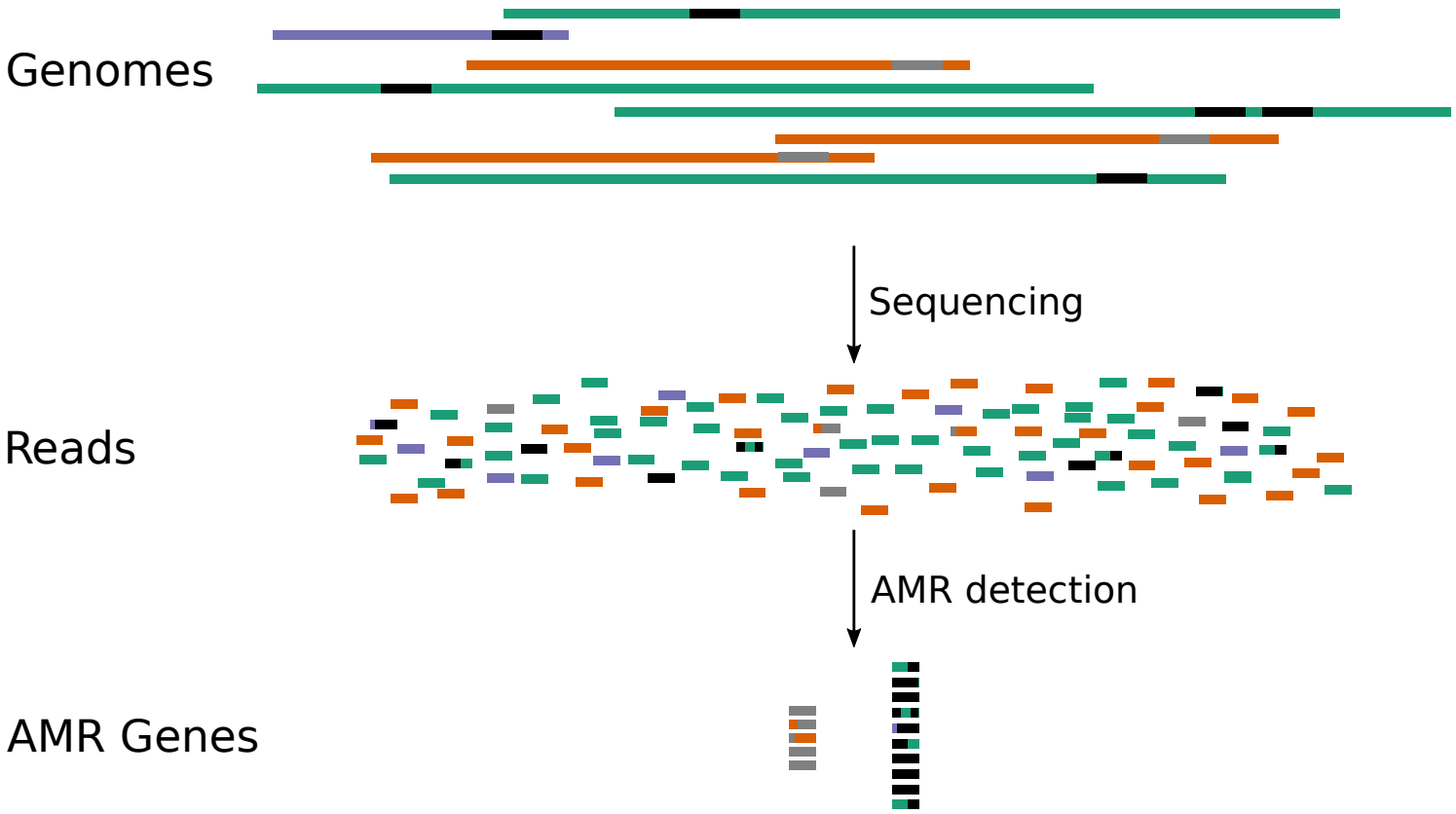


Other on-going Projects

- Halifax Community Learning Network
- Shelter Nova Scotia
- 211 Nova Scotia

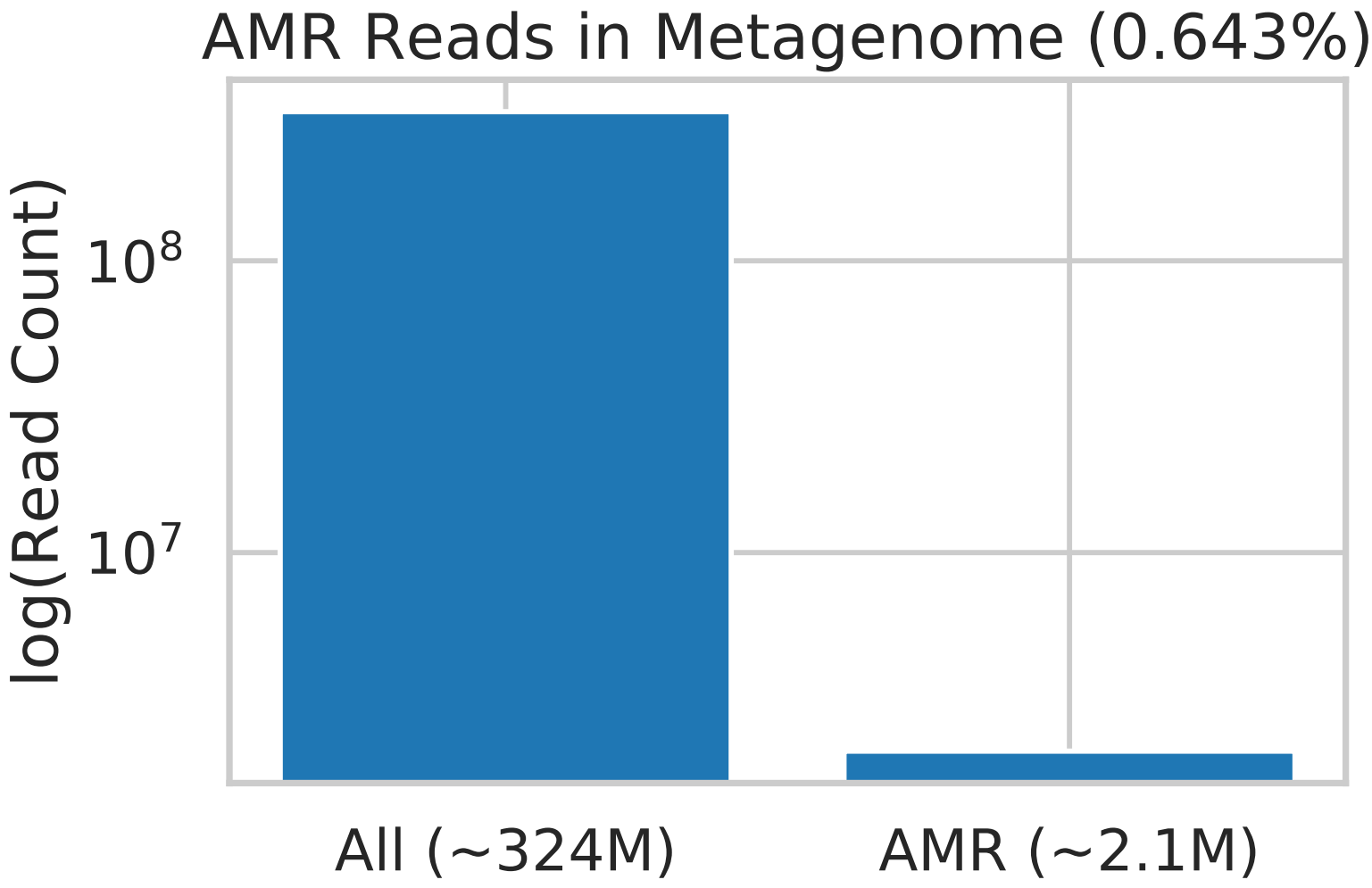
AMRtime

AMR-metagenomics



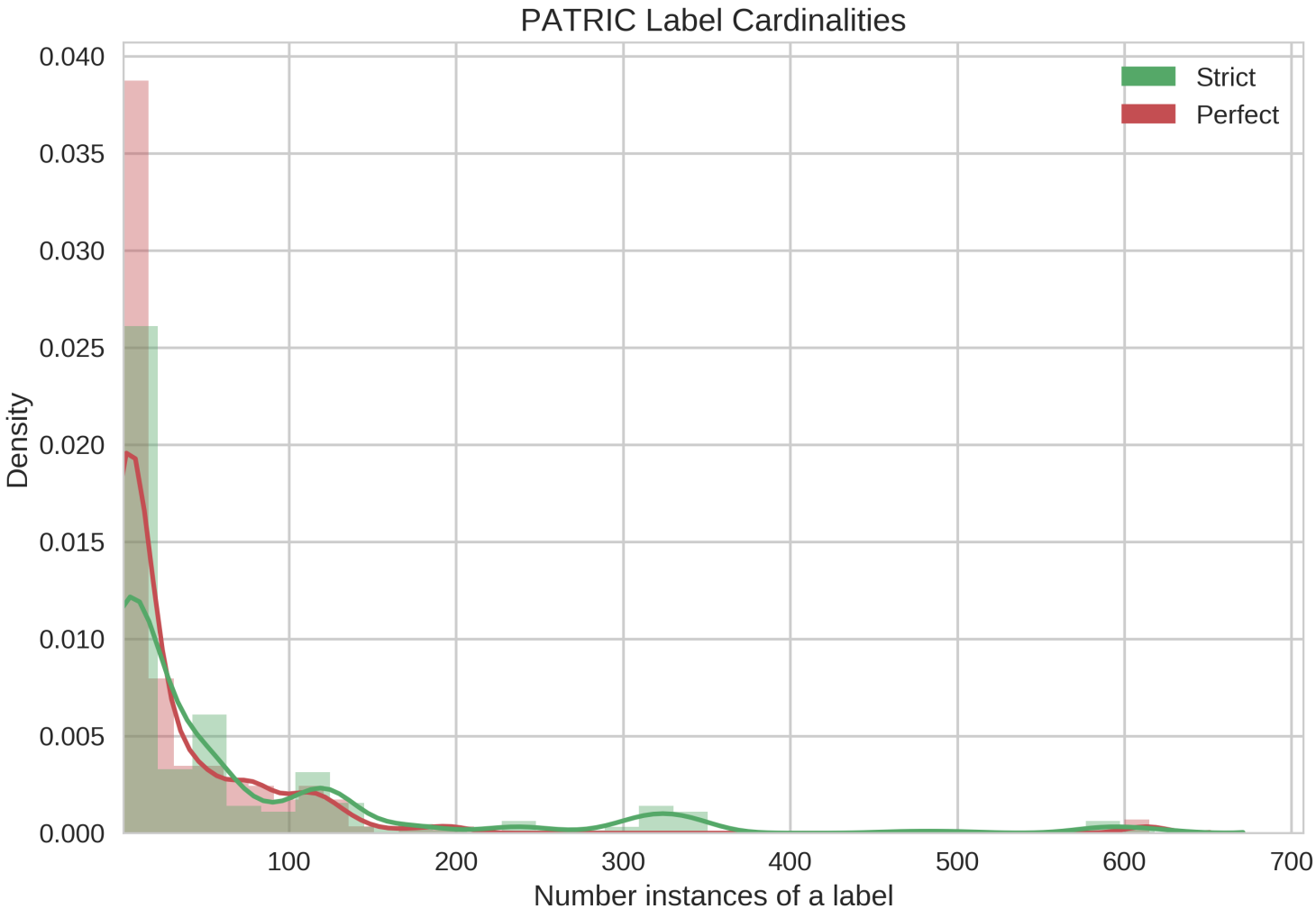
Why is this difficult?

AMR genes are rare genomically



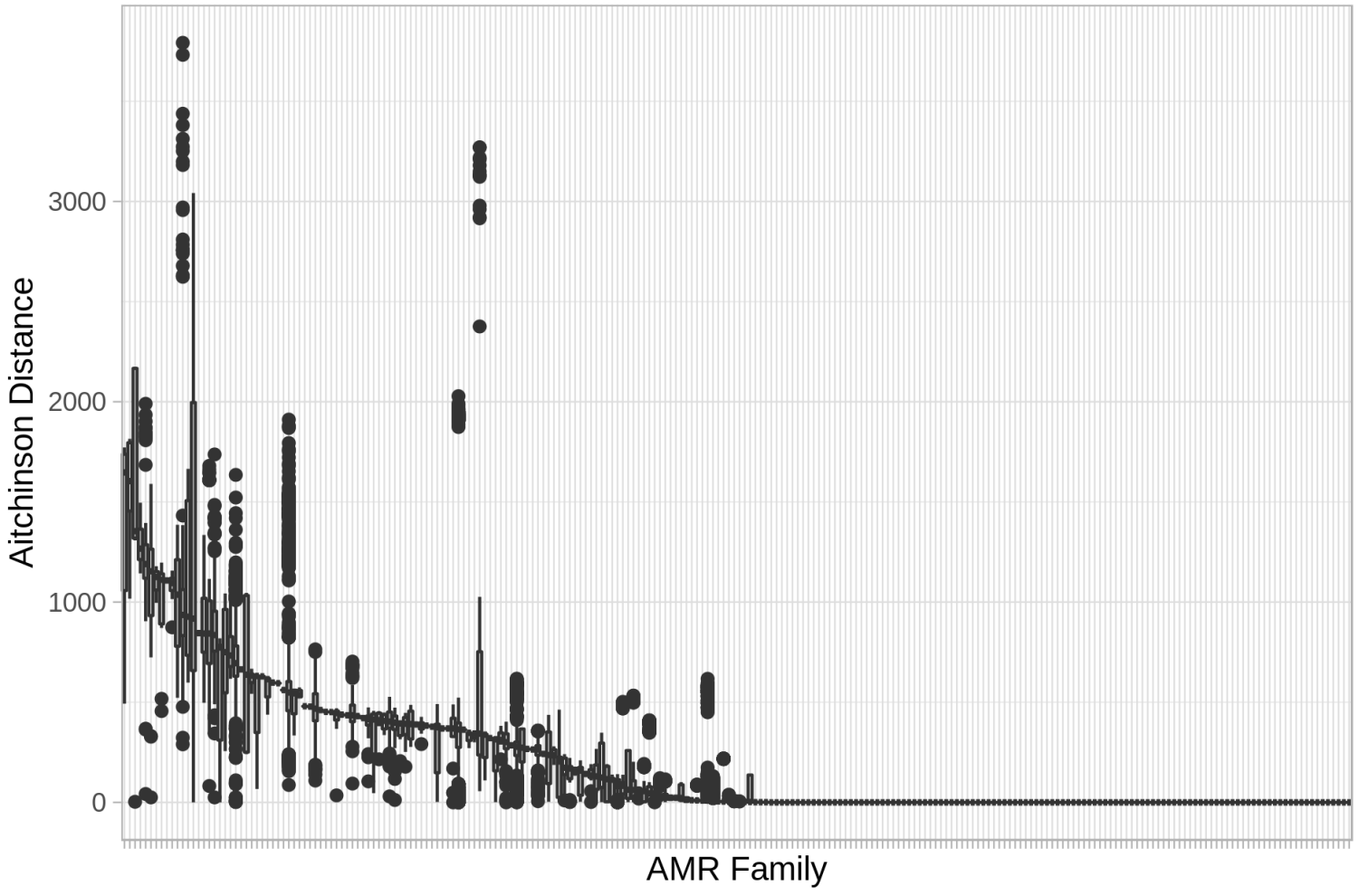
2184 CARD-Prevalence Genomes at 1-10X abundance

AMR genes have wildly different abundances

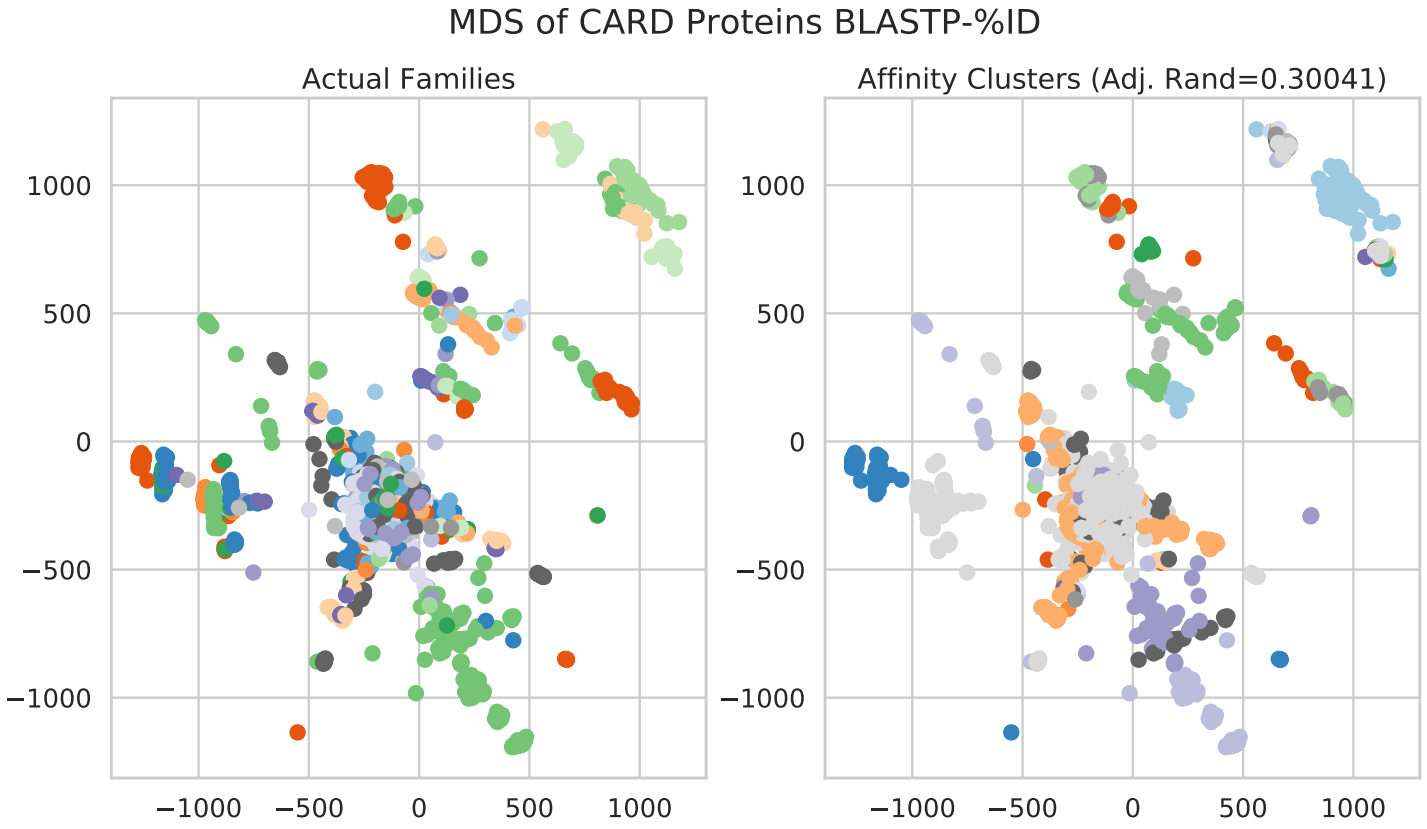


1236 AMR PATRIC genomes

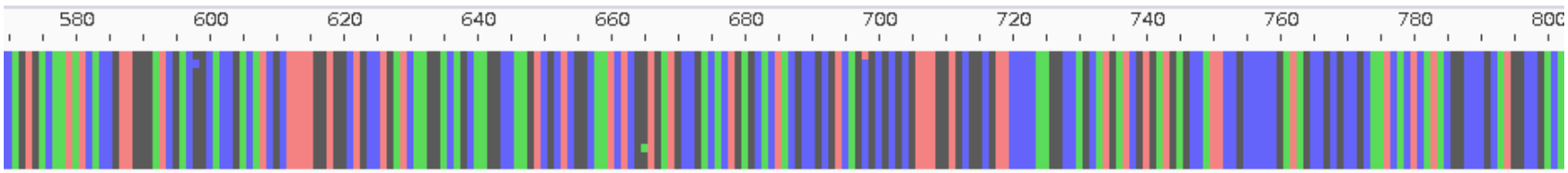
AMR genes have highly variable diversity



AMR sequence space overlaps

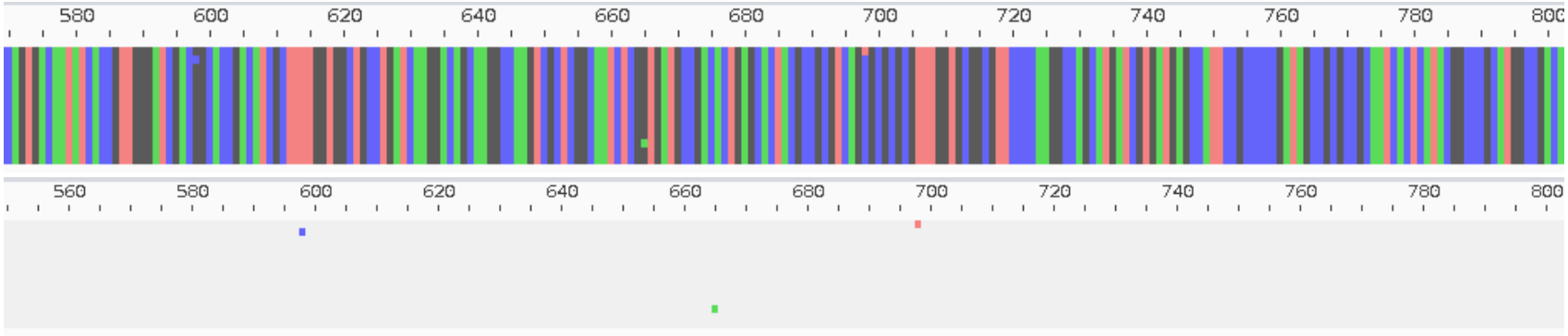


Insufficient Signal in 250bp Fragments



NDM Multiple Sequence Alignment

Insufficient Signal in 250bp Fragments



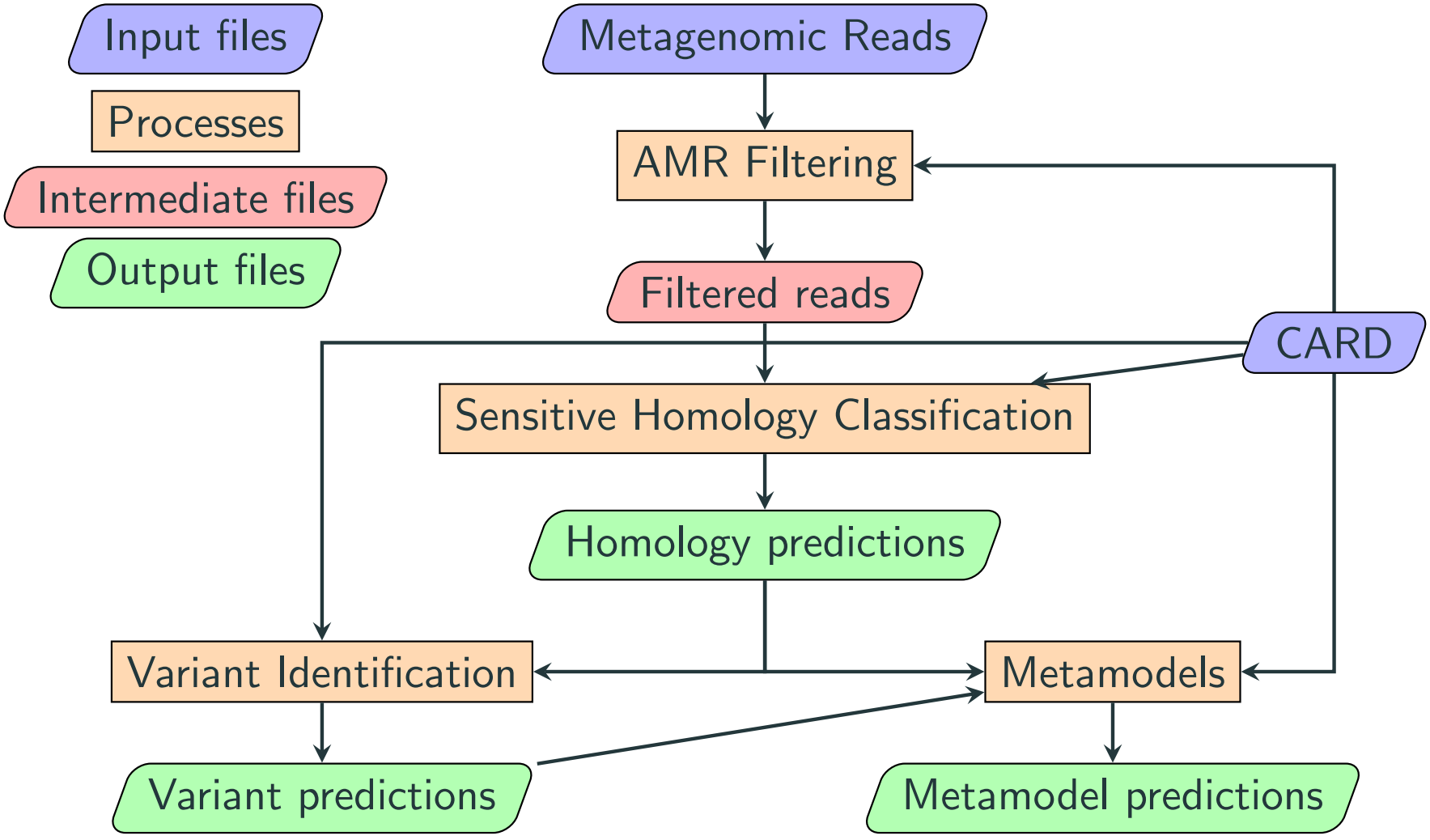
NDM Multiple Sequence Alignment

Other constraints

- No point doing what we do if people can't use it.
- Limited hardware requirements (a standard workstation or instance $< 8 - 12$Gb, $1 - 8$ cores).
- Fast enough (< 12 hours).
- Easy to install/configure.
- Easy to use.
- Easy to update.

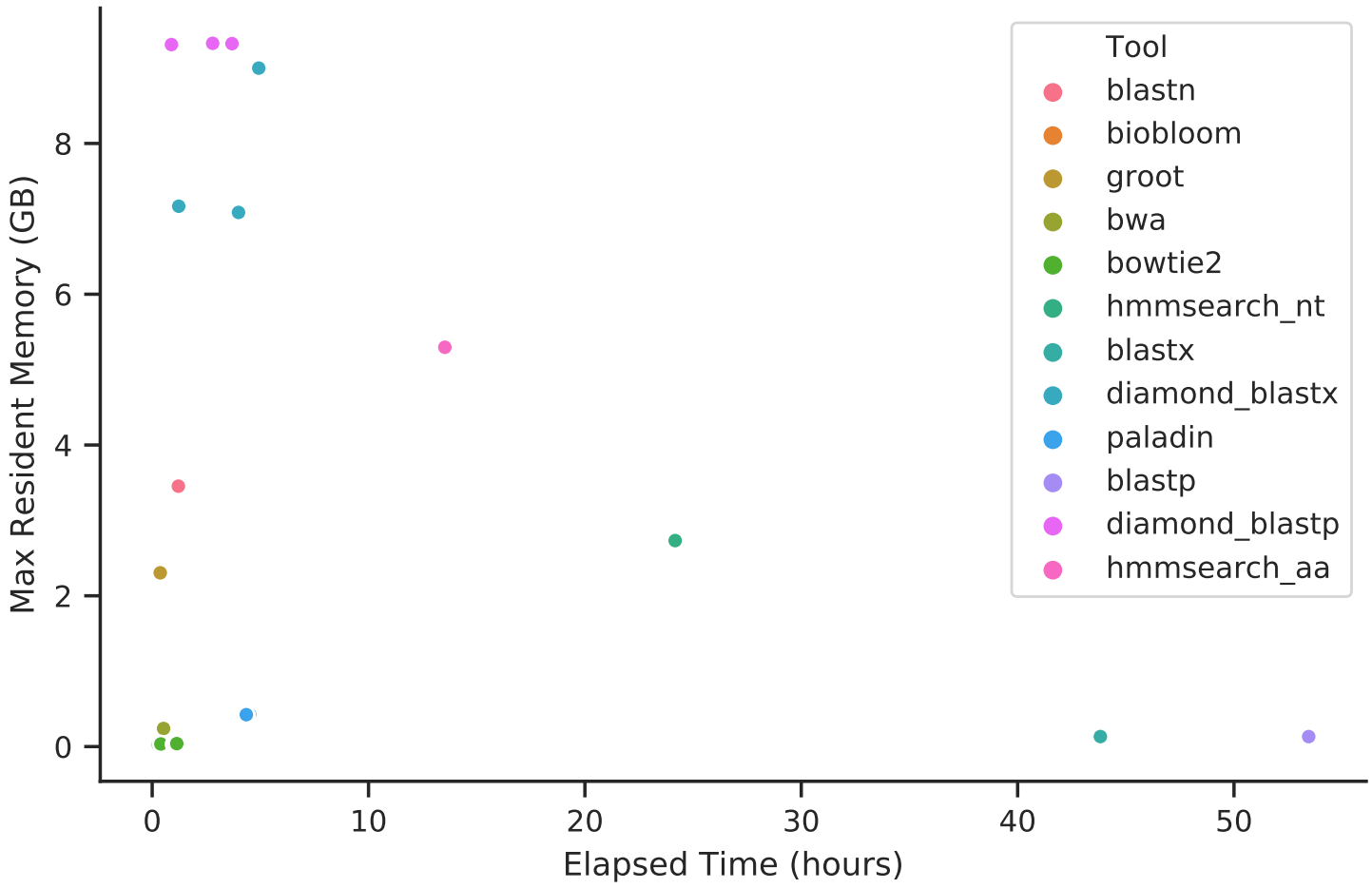
AMRtime

AMRtime structure



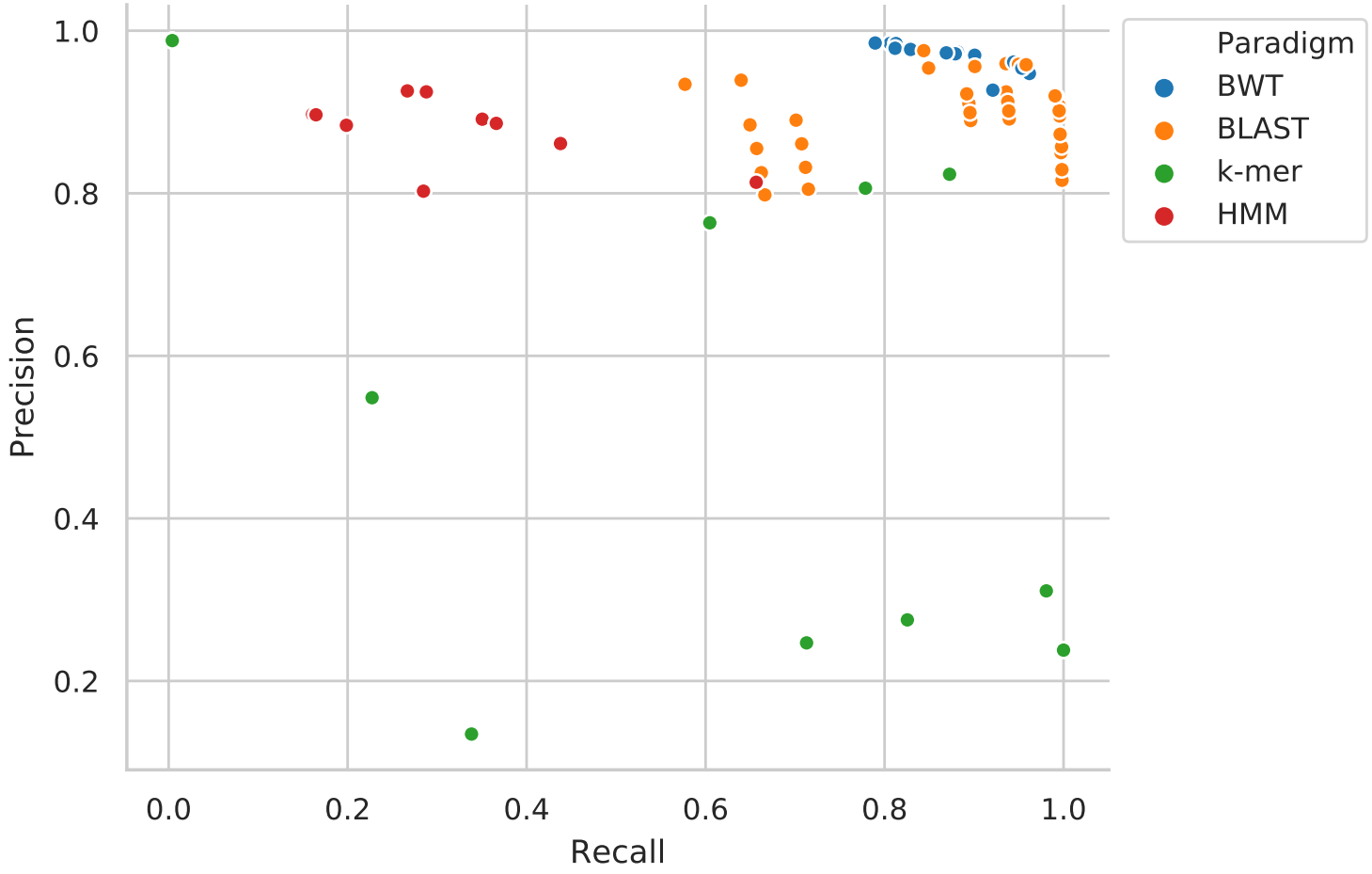
Read filtering

Homology Filter Approaches

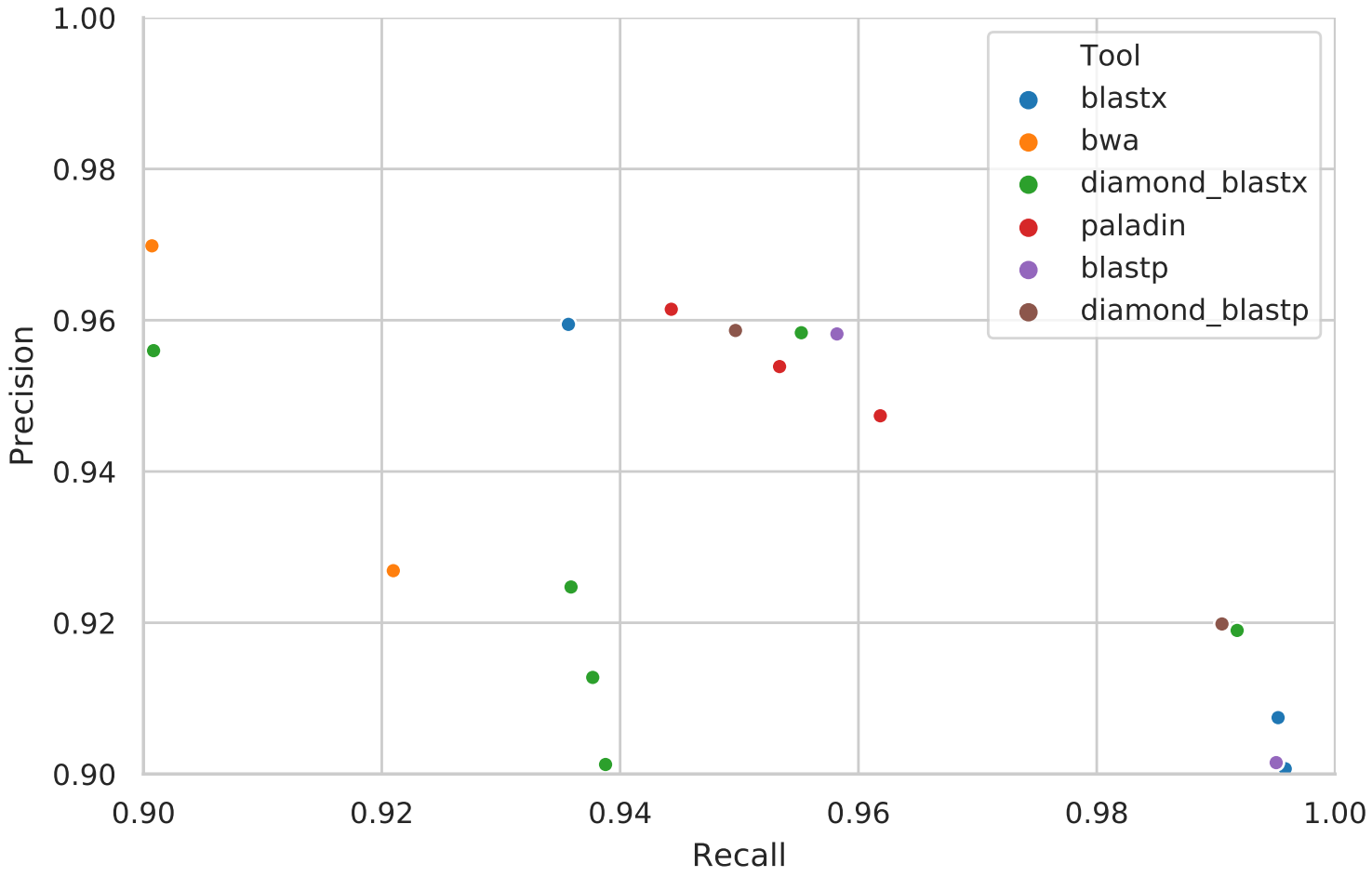


Relative Computational Demands

Precision-Recall of Homology Search

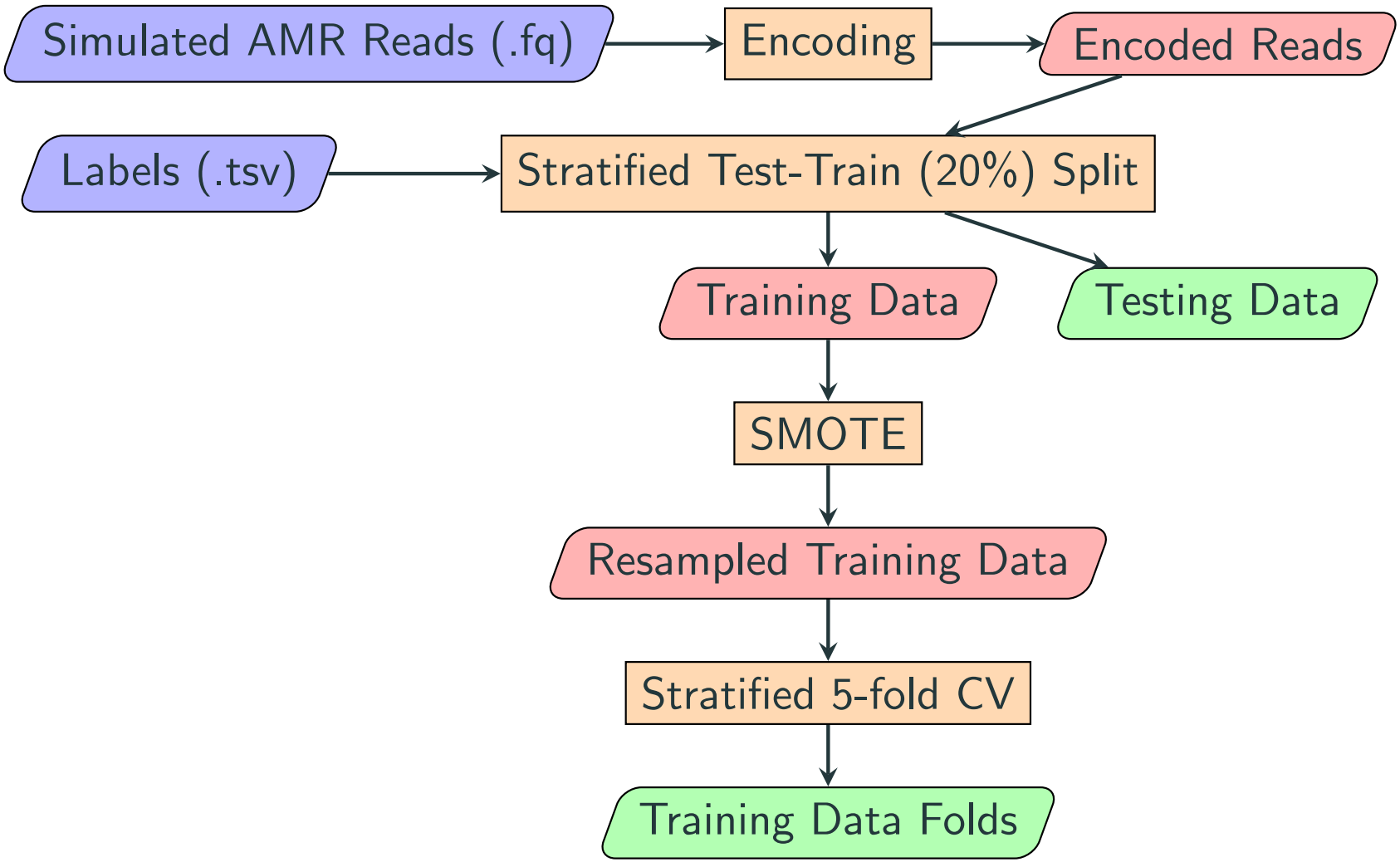


Optimising for recall



Sensitive Homology Classification

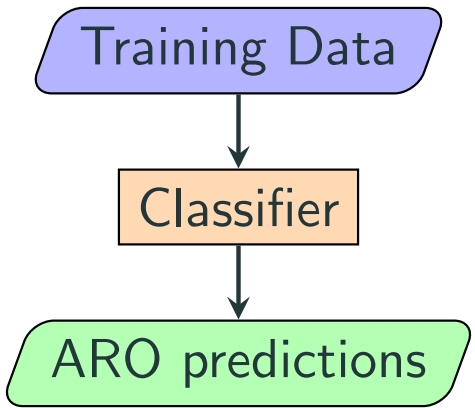
Dealing with imbalanced training data



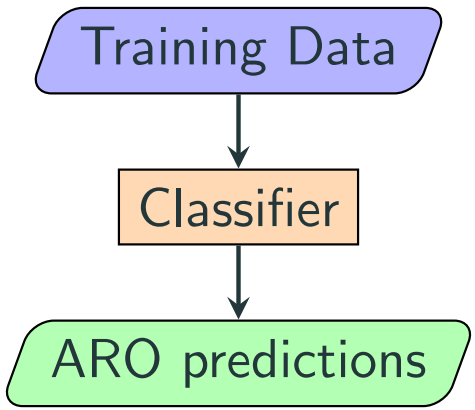
What is balance?

- Different gene lengths within families (coverage vs read number)?
- Different family sizes?
- Different family diversity?
- Using a generator to improve on SMOTE.

Initial classifier

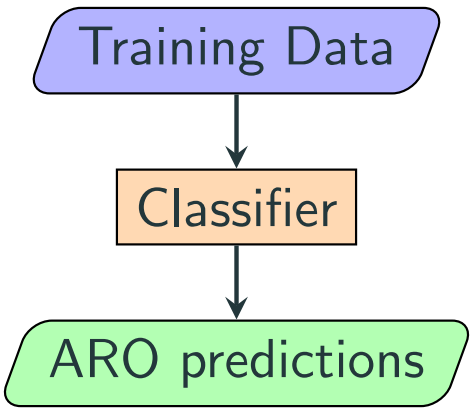


Initial classifier



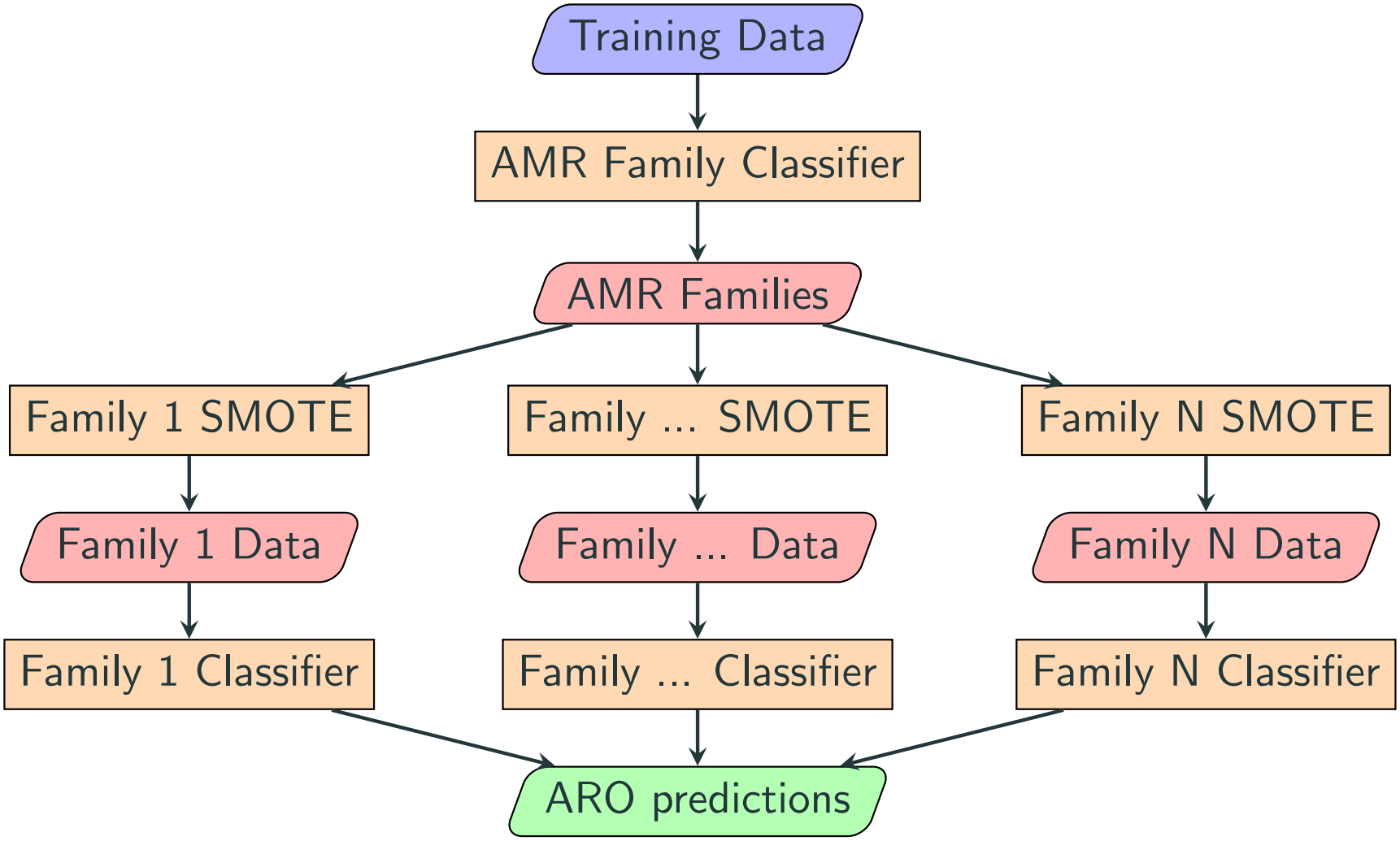
NB 7-mer Average Precision: 0.63

Initial classifier



NB 7-mer Average Precision: 0.63 %

Revised classifier structure: exploiting the ARO



Sequence similarity encoding

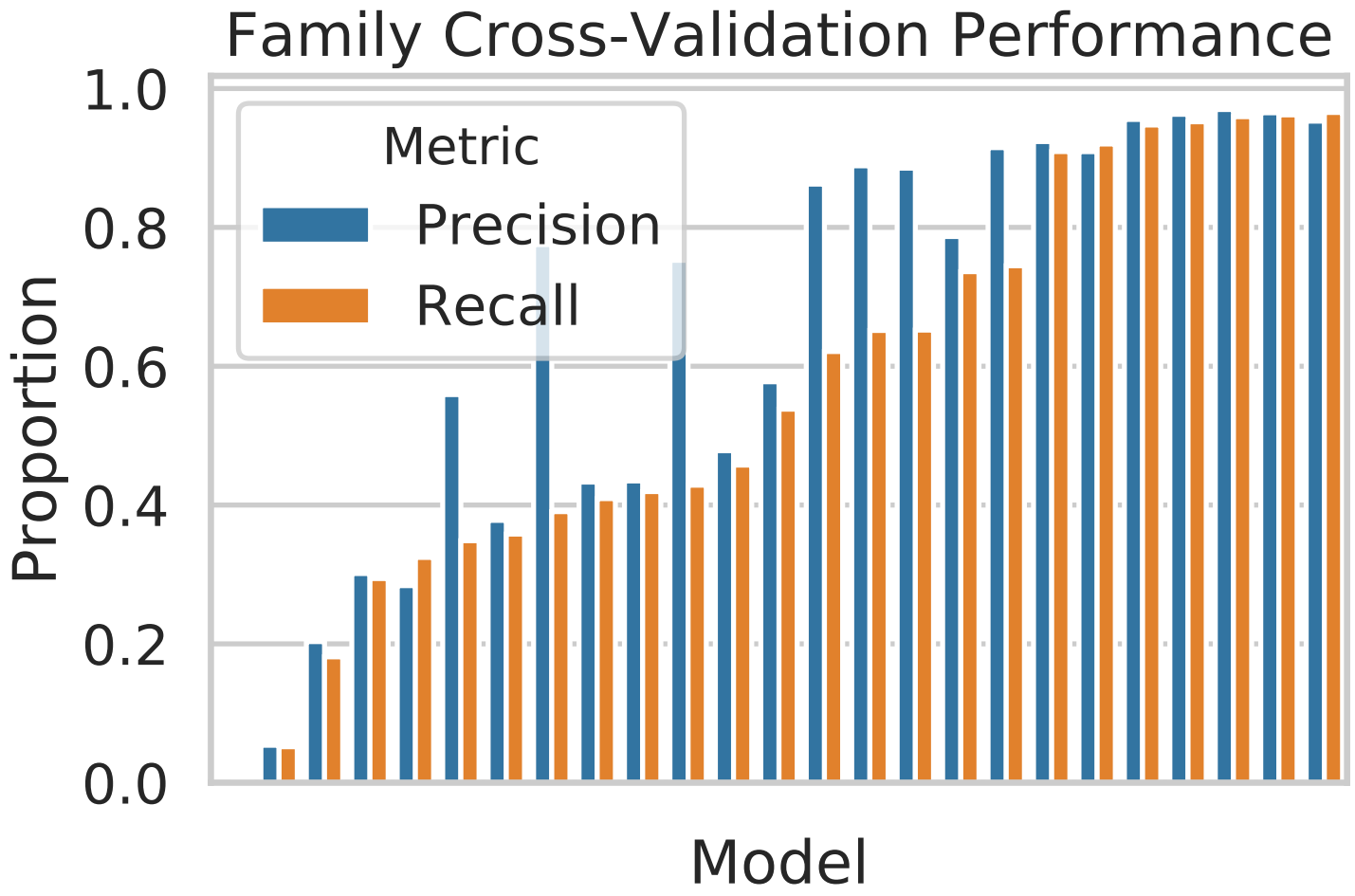
$$\text{Sequence bitscore matrix} = \begin{array}{c} \\ read_1 \\ read_2 \\ \dots \\ read_{i-1} \\ read_i \end{array} \begin{array}{c} \begin{array}{ccccc} gene_1 & gene_2 & \dots & gene_{j-1} & gene_j \end{array} \\ \begin{pmatrix} 1256 & 0 & \dots & 0 & 63 \\ 0 & 0 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 512 & \dots & 0 & 0 \\ 0 & 0 & \dots & 785 & 129 \end{pmatrix} \end{array}$$

Advantages: read length invariant, low dimensionality, uses filtering data computation

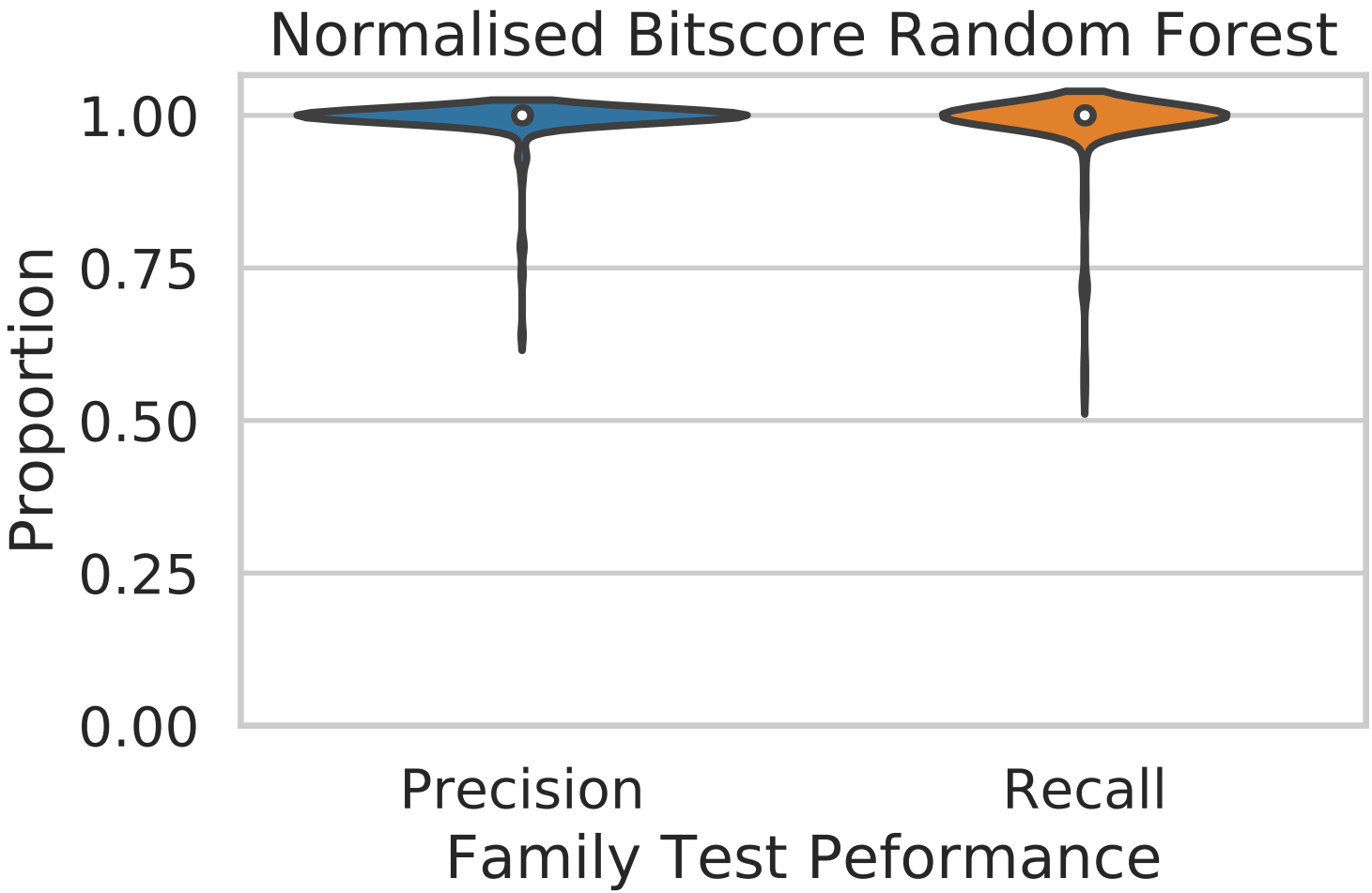
Cross-Validation

- 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)
- Classifiers:
  - Random Forests
  - Naive Bayes
  - Logistic Regression
  - Neural Networks of varying architecture (Torch)

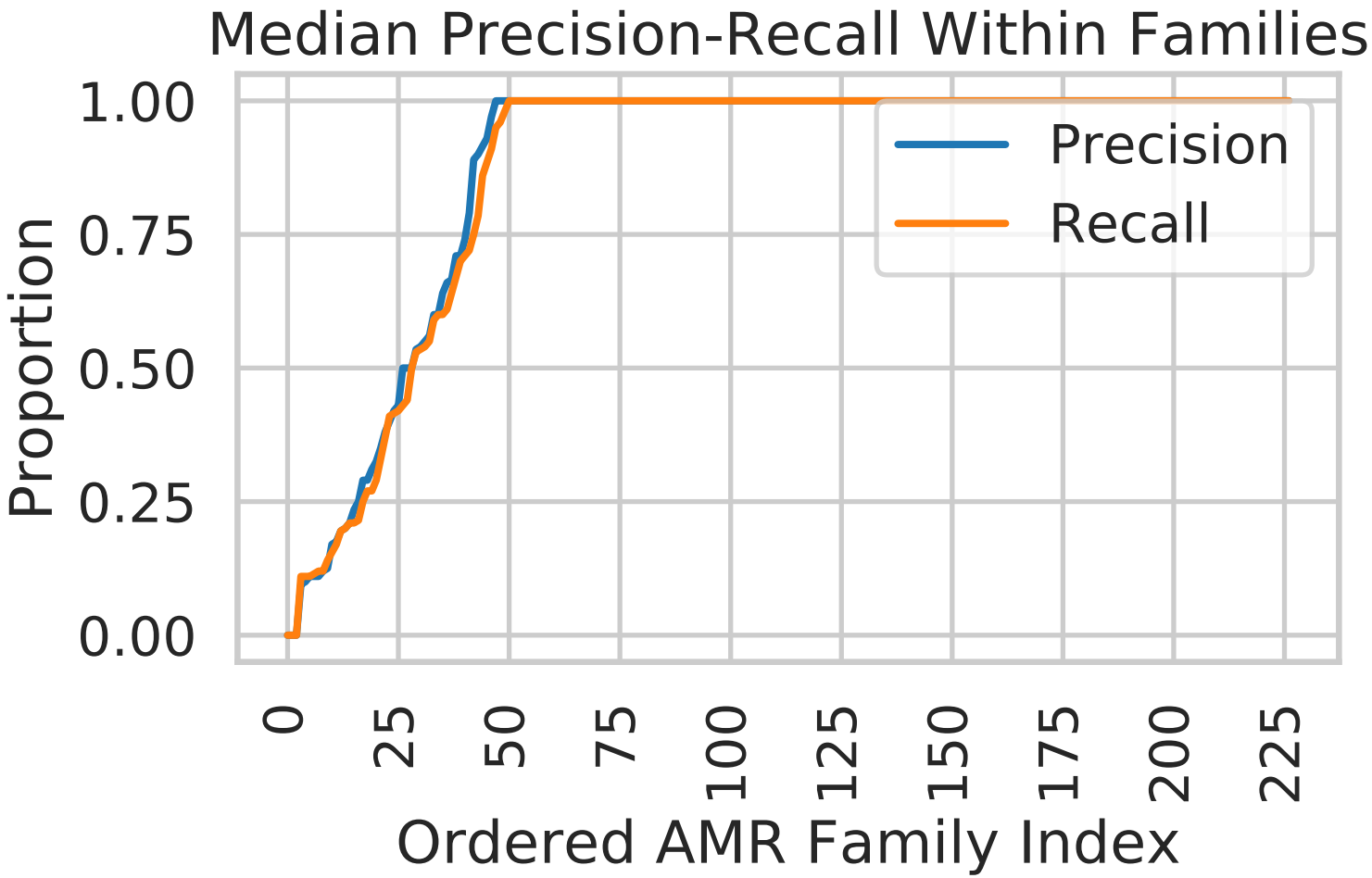
Cross-validation



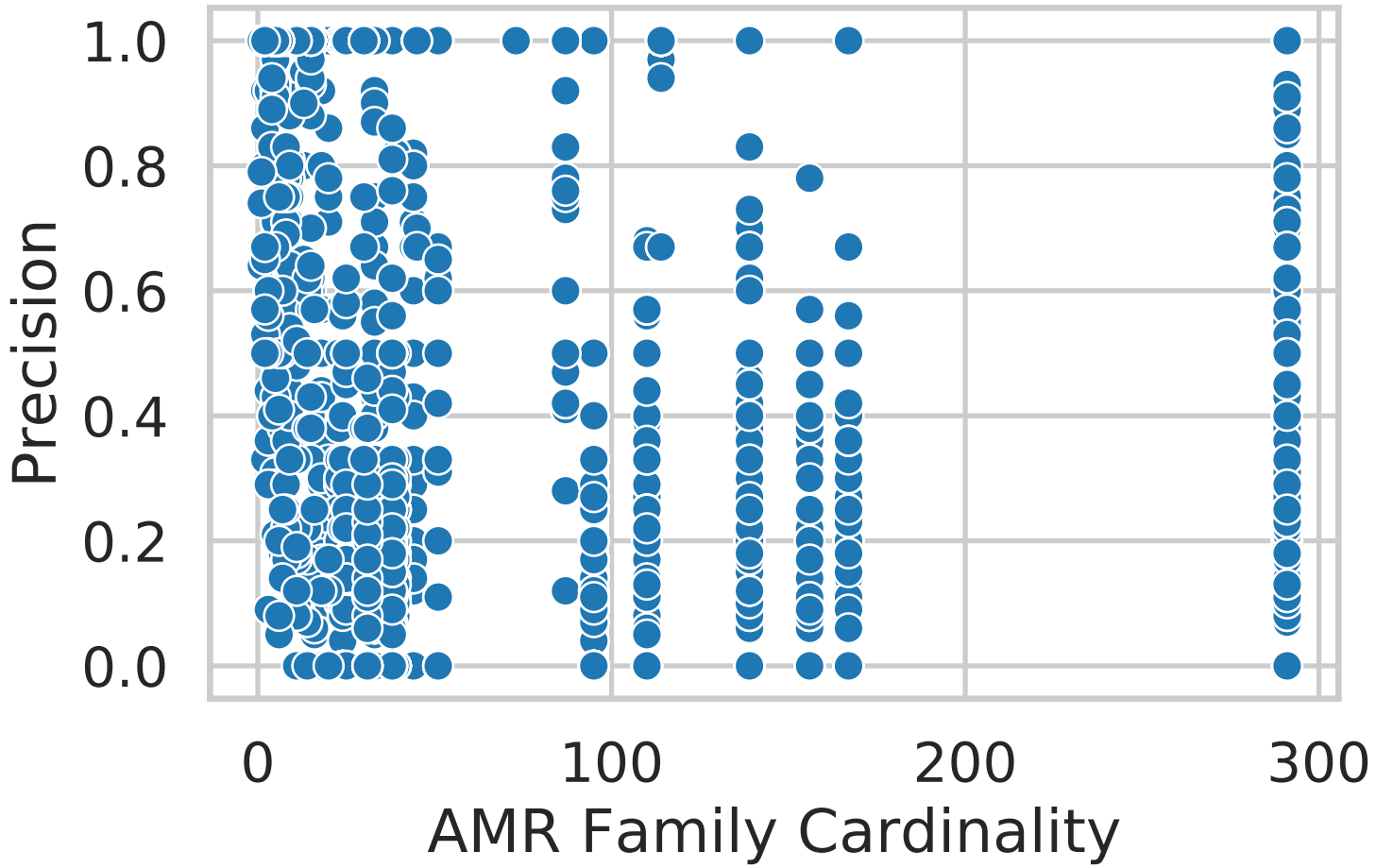
Held-out test results



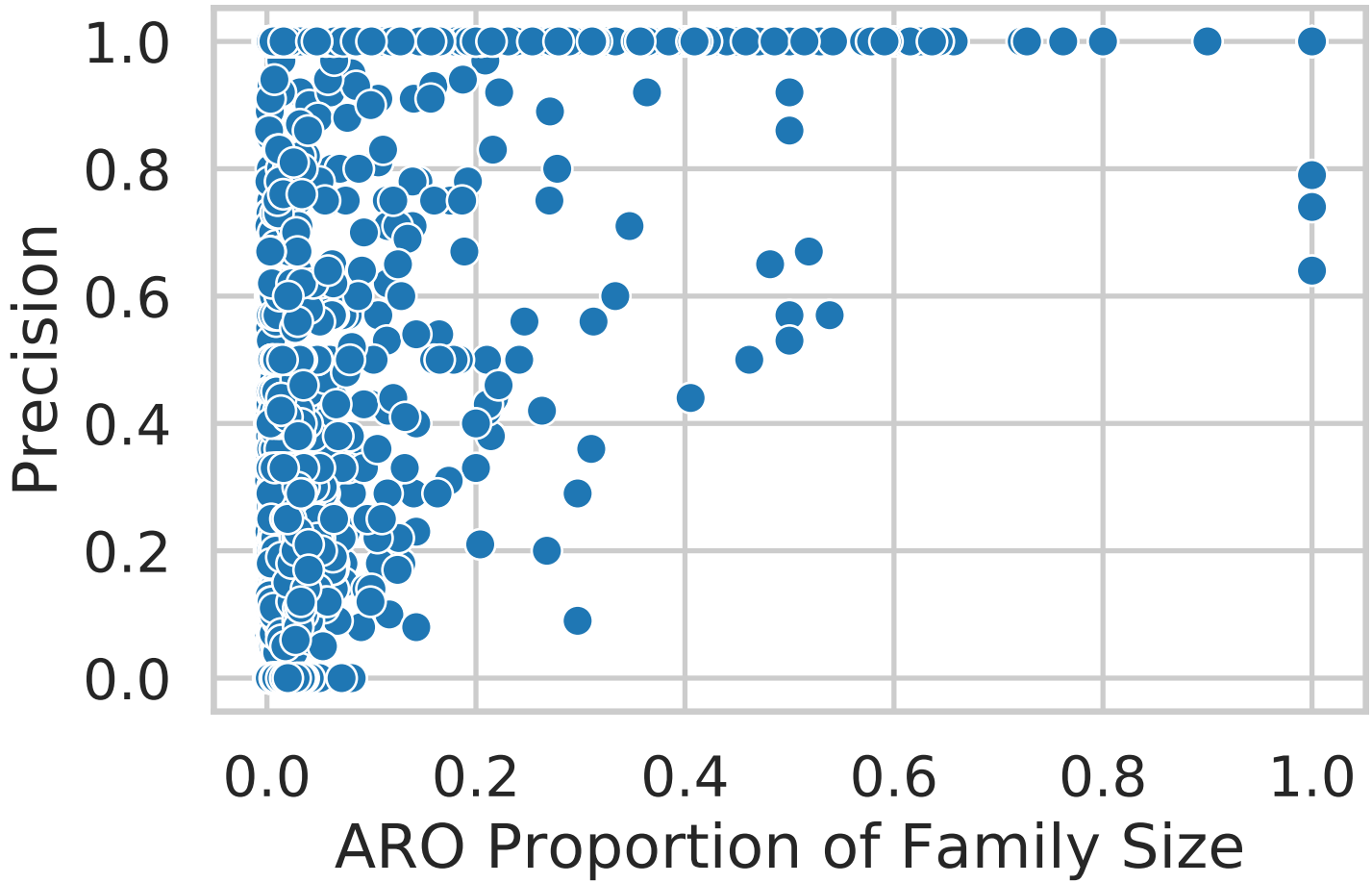
ARO level classification more variable



Family diversity as explanation?



Within family label imbalance



On-going Problems

- Multiset prediction when insufficient signal.
- Systematic benchmarking.
- Full end-to-end comparisons with other approaches (soliciting ideas!)
- rRNA and variant models (not discussed here).
- Integration into CARD platform and IRIDA.

Acknowledgements

Acknowledgements

- McMaster University: Kara Tsang, Brian Alcock, and Andrew McArthur
- Simon Fraser University: Fiona Brinkman
- Dalhousie University: Robert Beiko
- Funding: Genome Canada, NERC Undergraduate Student Research Award, Donald Hill Family Fellowship

References

Bradley, P., Gordon, N. C., Walker, T. M., Dunn, L., Heys, S., Huang, B., Earle, S., Pankhurst, L. J., Anson, L., De Cesare, M., et al. (2015). Rapid antibiotic-resistance predictions from genome sequence data for staphylococcus aureus and mycobacterium tuberculosis. *Nature communications*, 6:10063.

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

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

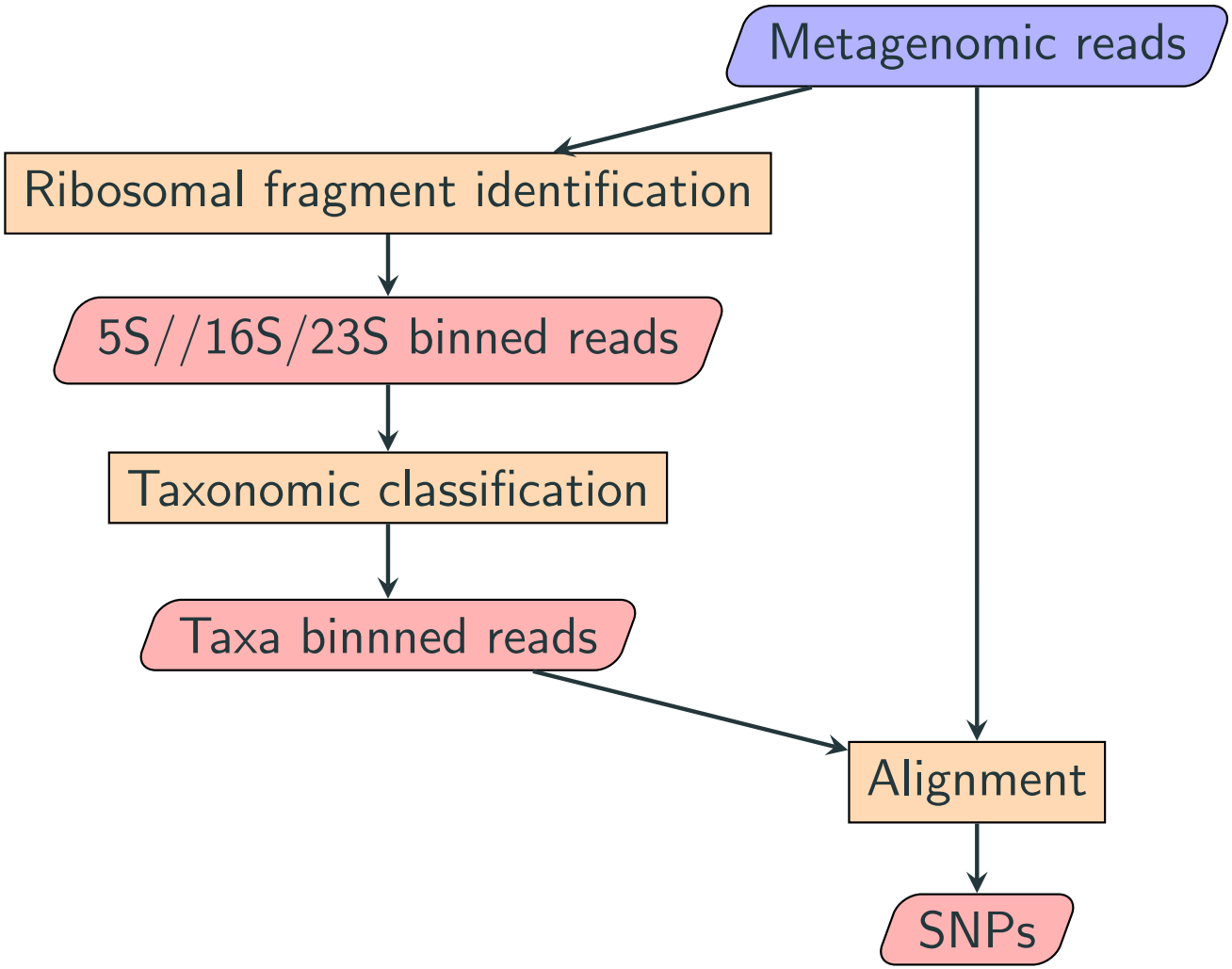
McArthur, A. G., Waglechner, N., Nizam, F., Yan, A., Azad, M. A., Baylay, A. J., Bhullar, K., Canova, M. J., De Pascale, G., Ejim, L., et al. (2013). The comprehensive antibiotic resistance database. *Antimicrobial agents and chemotherapy*, 57(7):3348–3357.

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

Backup

Variant Models

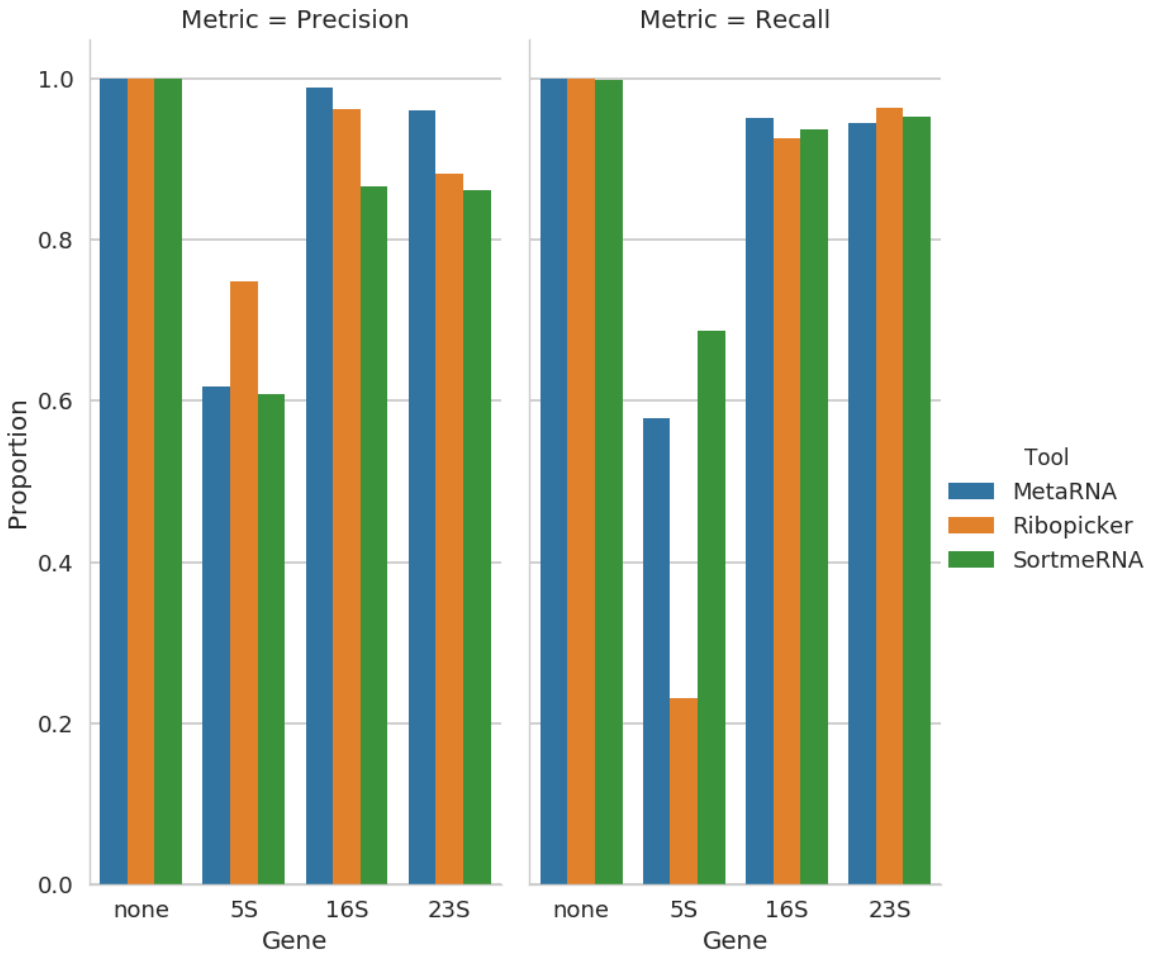
Ribosomal Variant Models



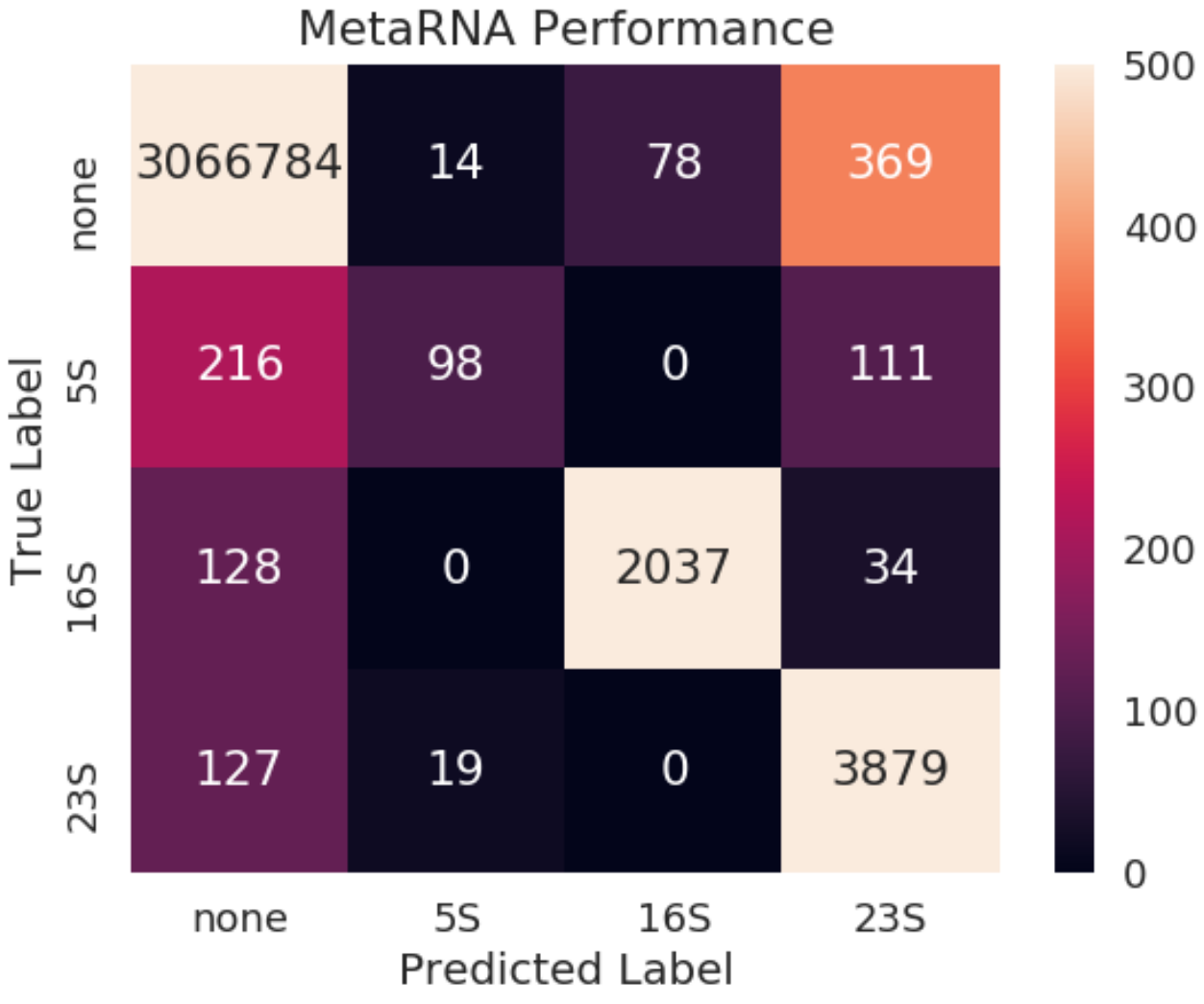
Identifying Ribosomal Reads

- 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

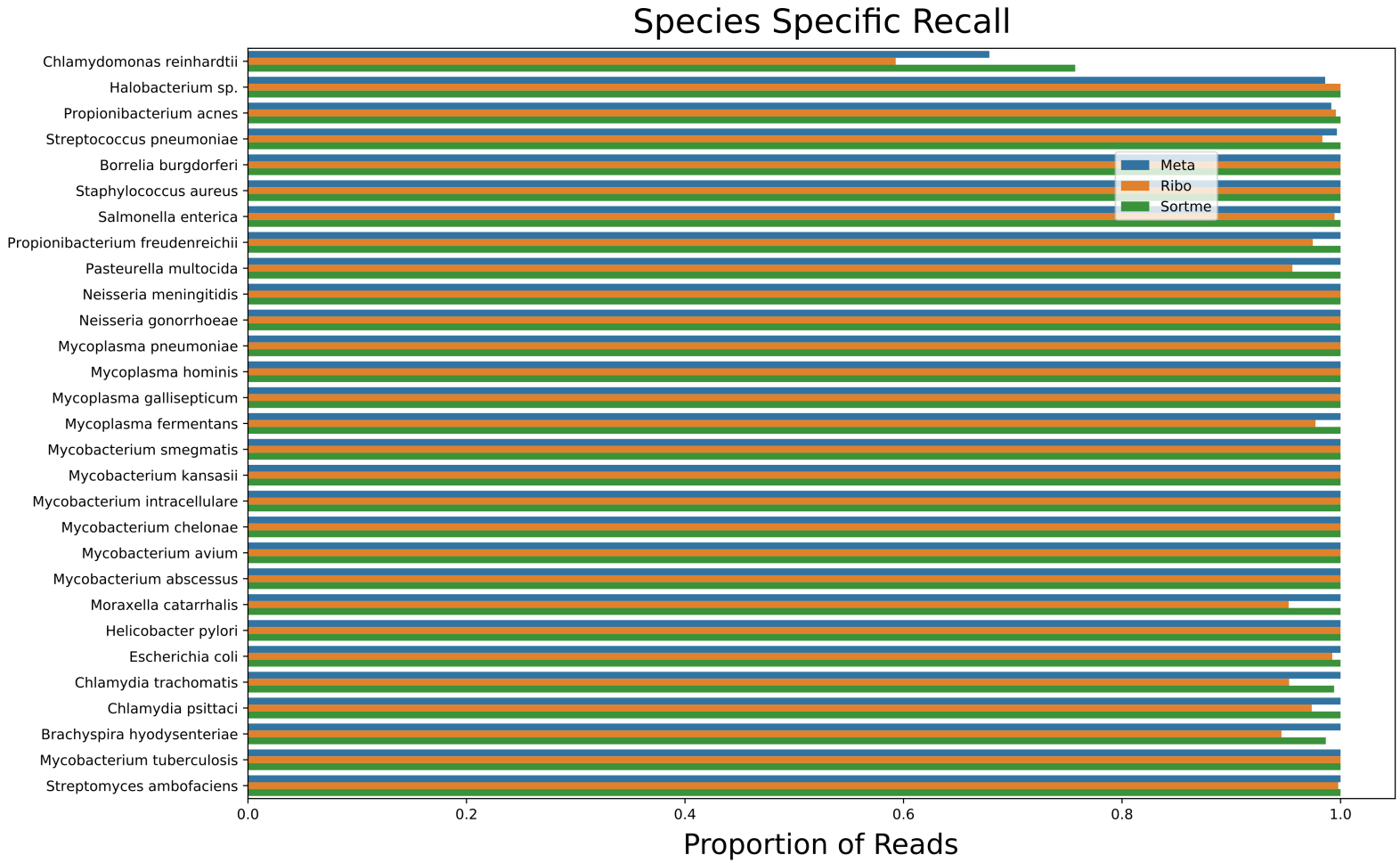
Identifying Ribosomal Reads



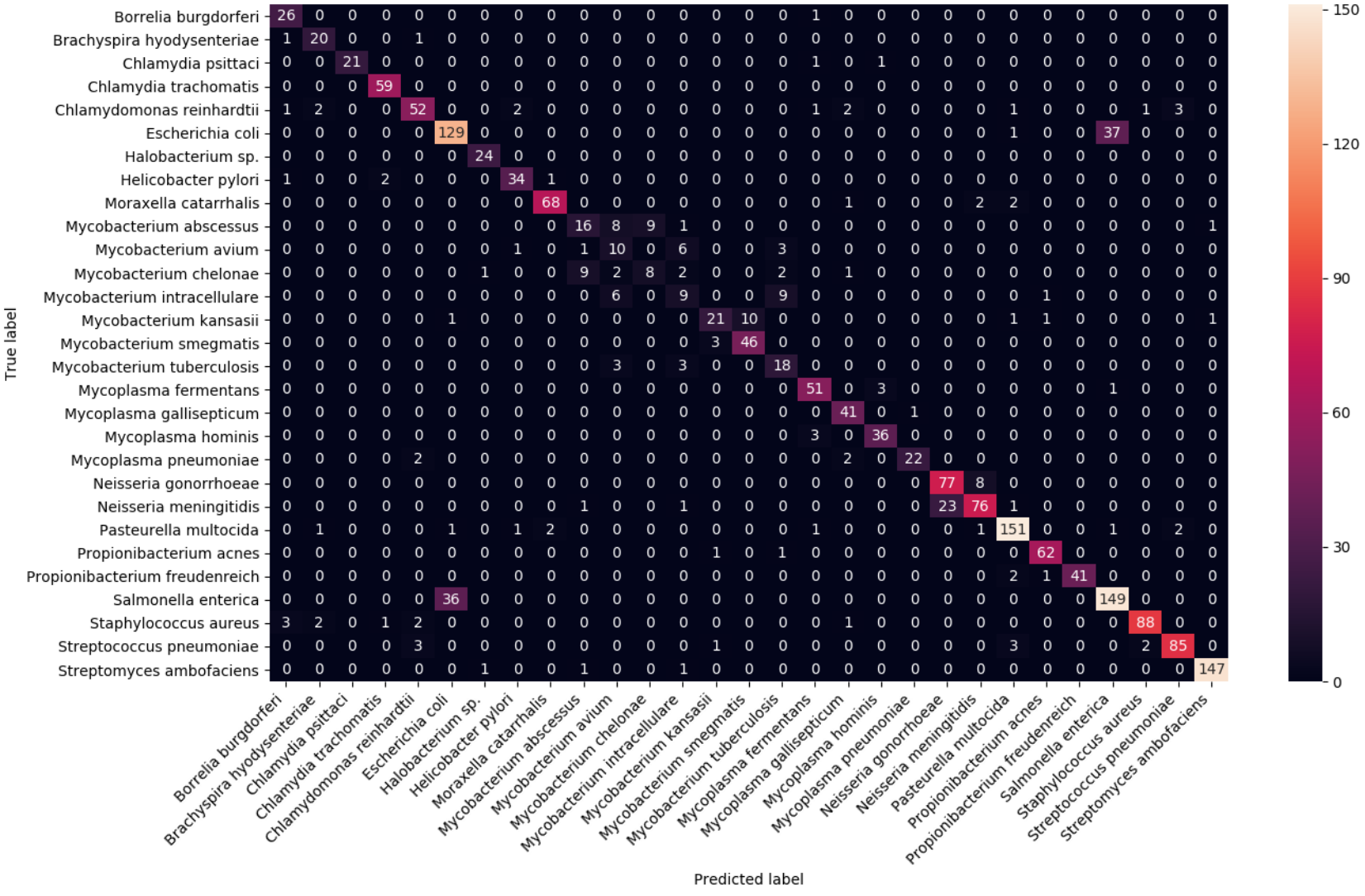
Identifying Ribosomal Reads



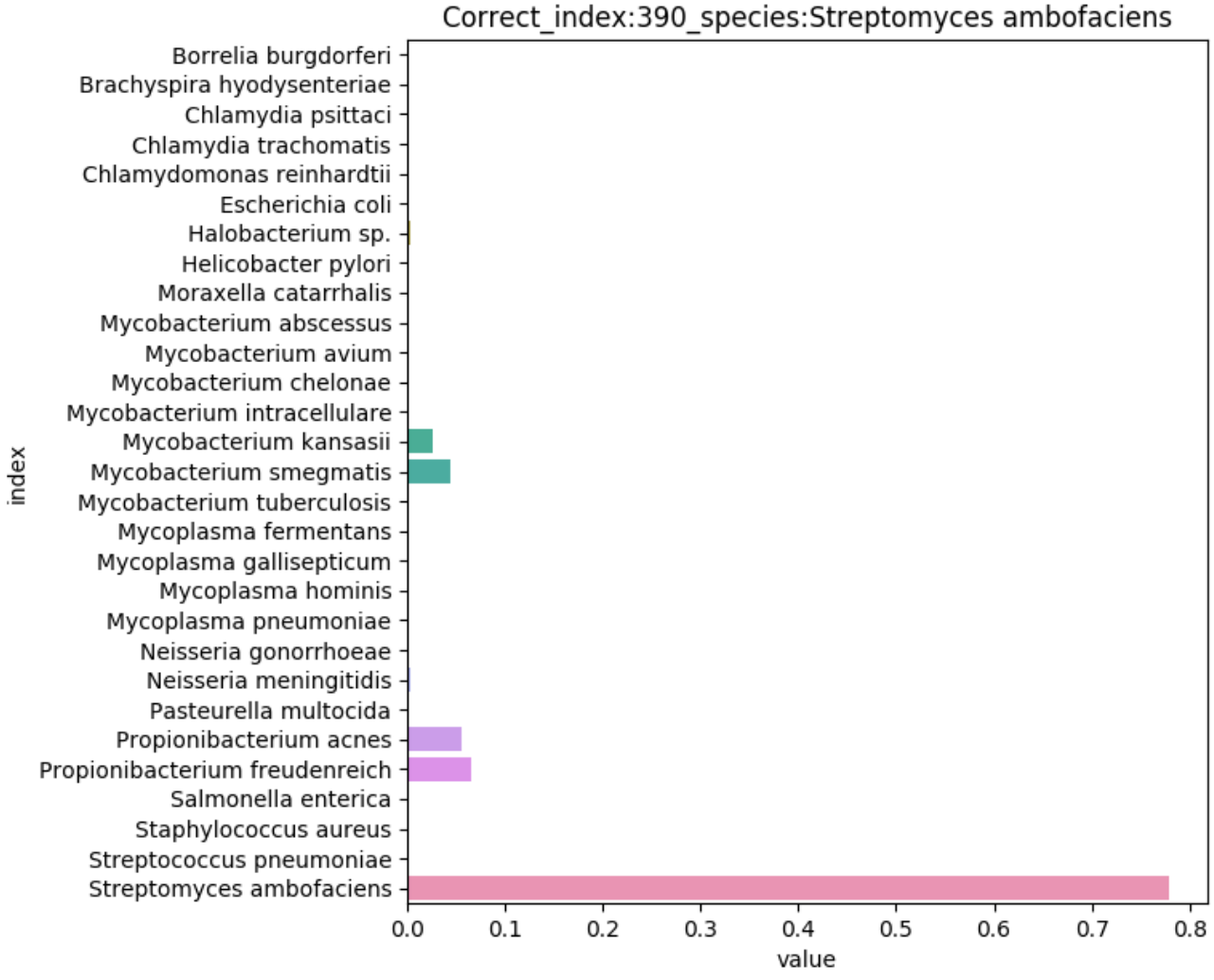
Identifying Ribosomal Reads



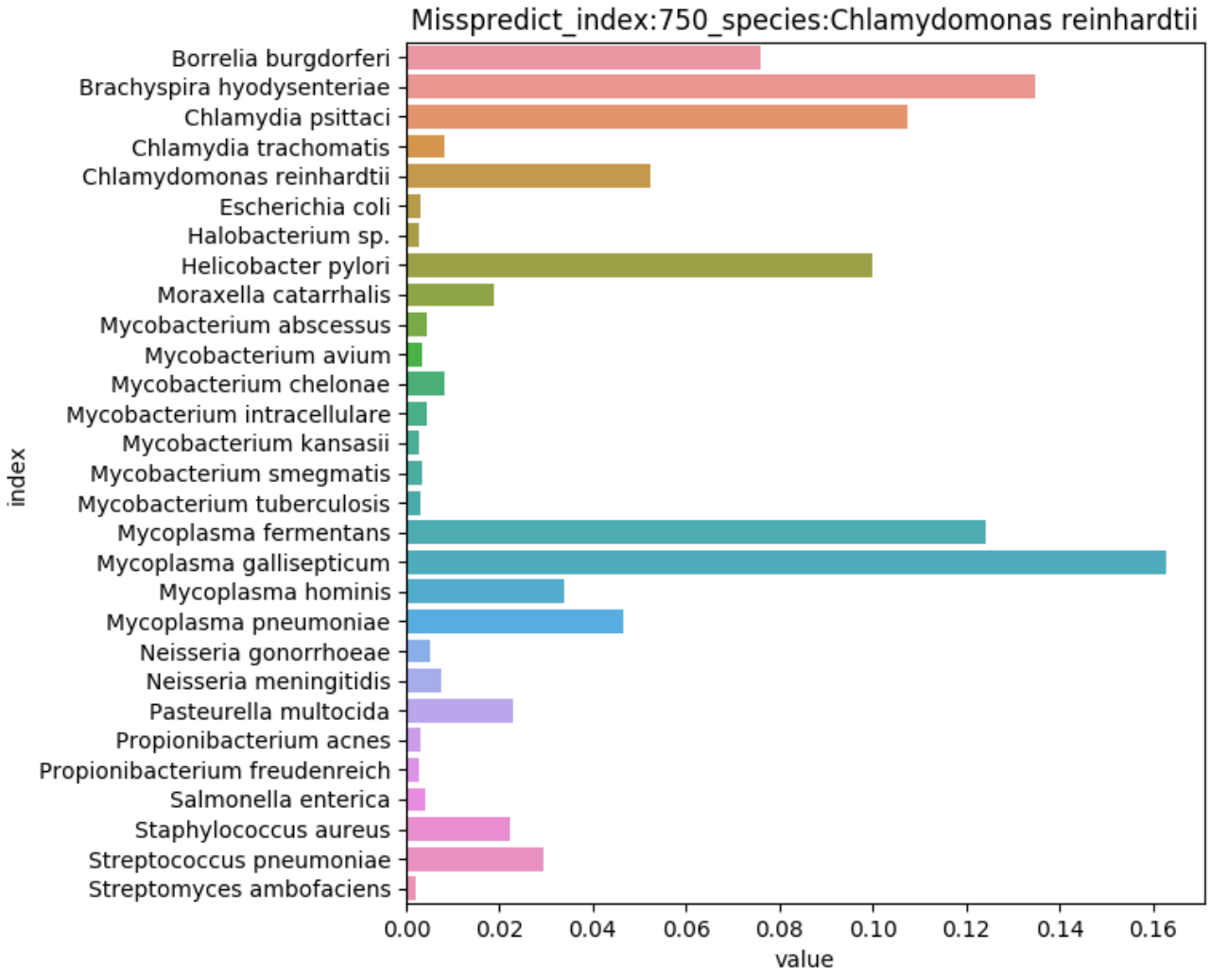
Identifying Taxonomy



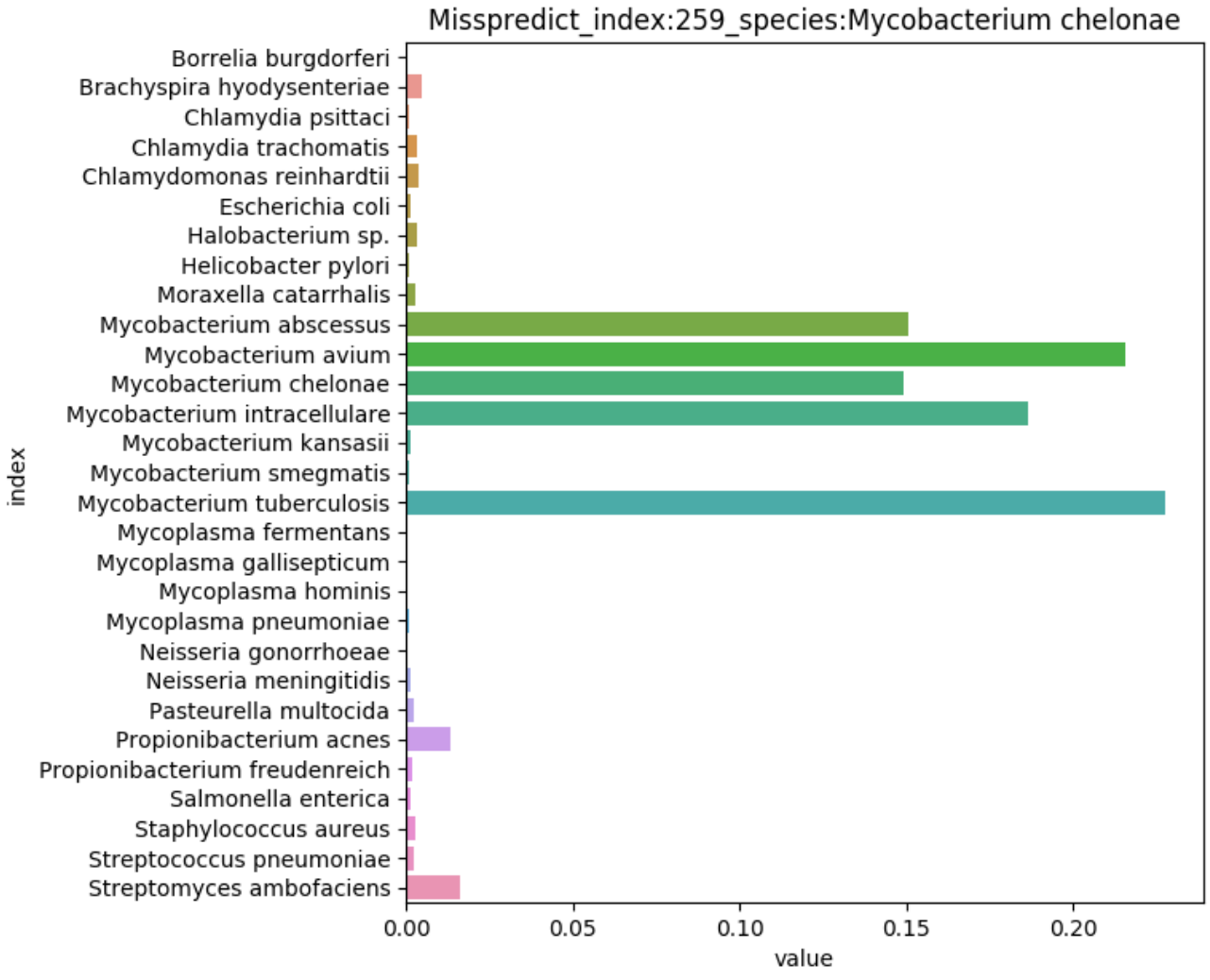
Some are relatively easy



Others are a mess

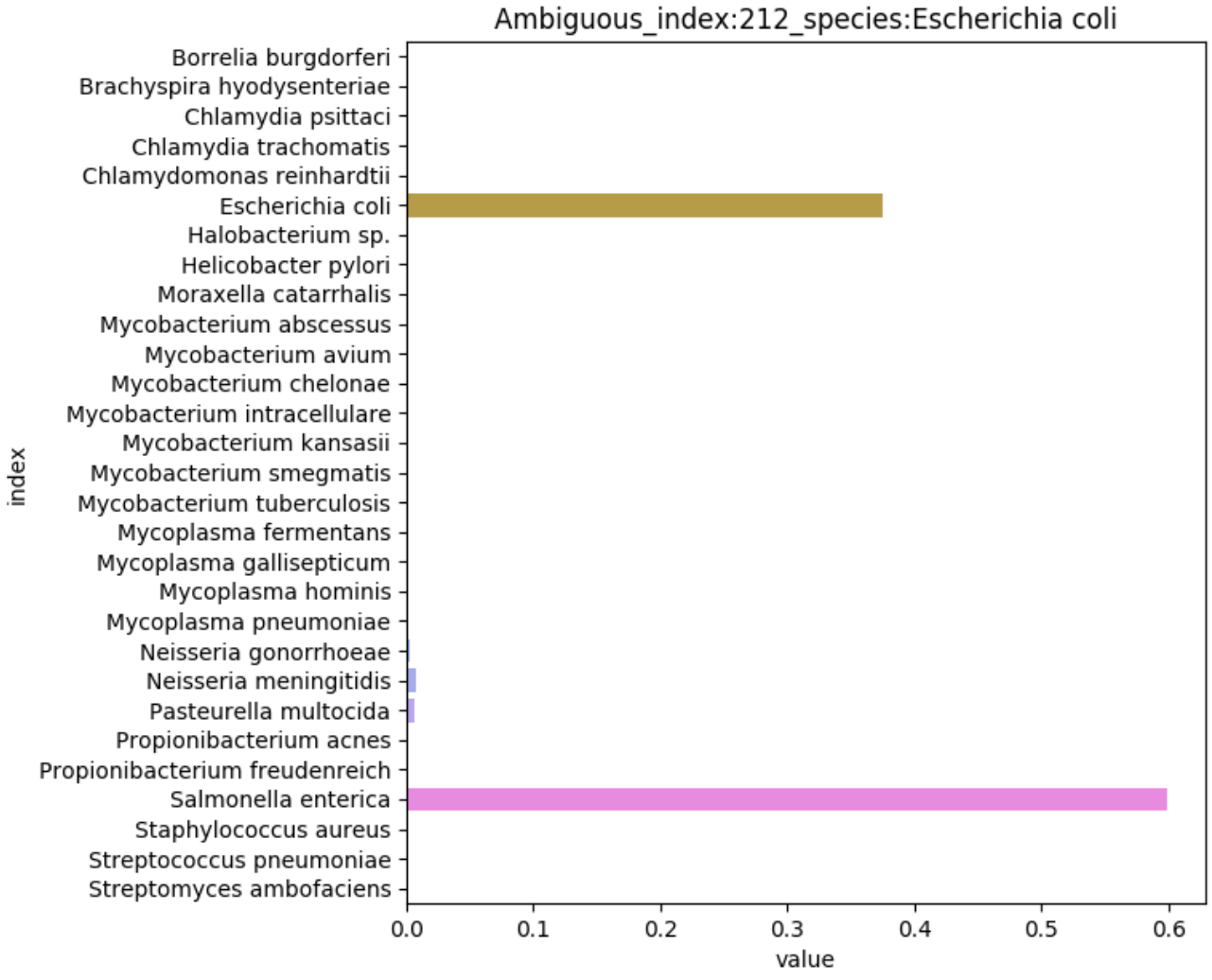


Some are group ambiguous

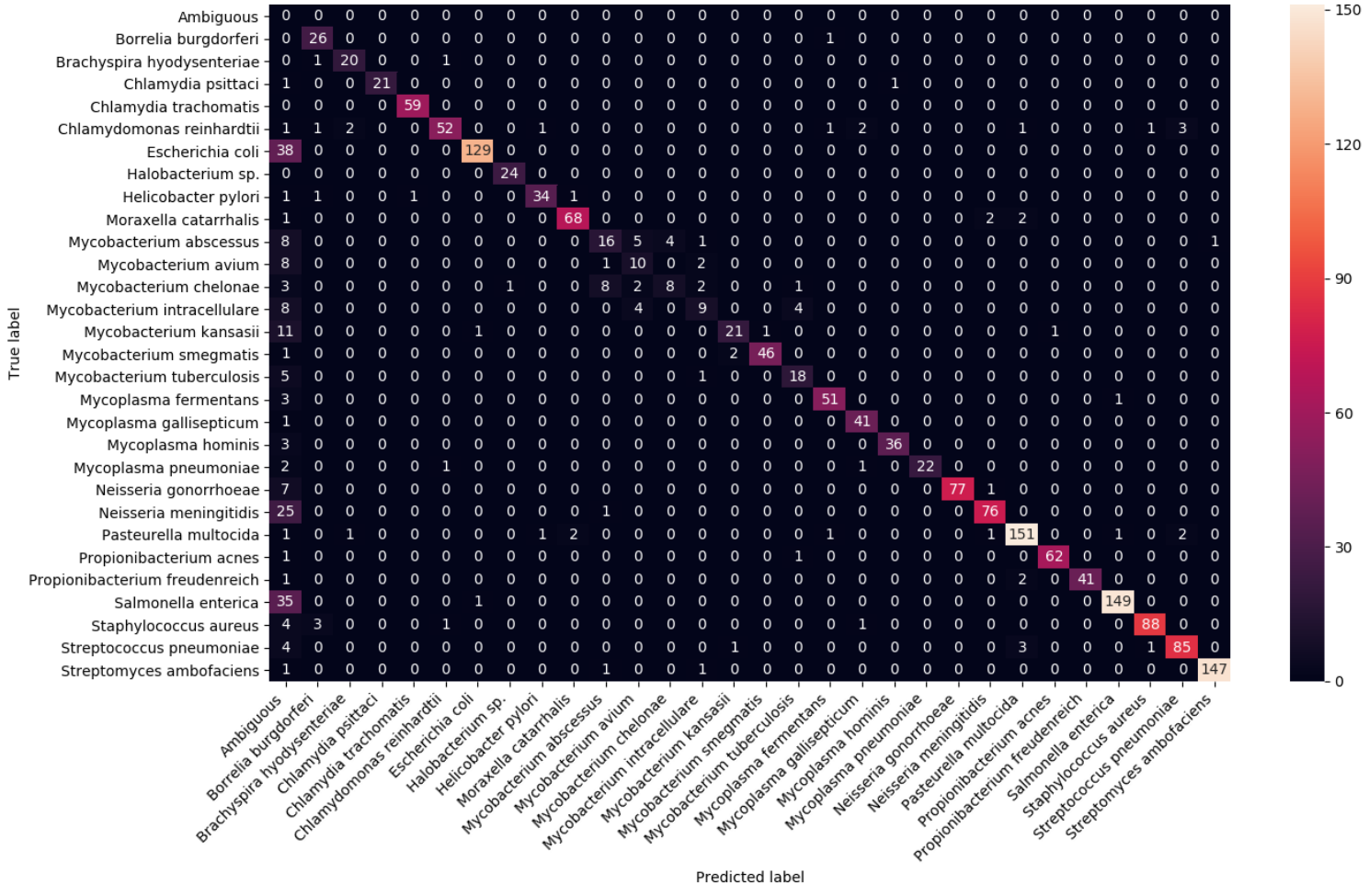


Probably a Mycobacterium?

Others are just a toss-up



Ambiguity in classification



Meta-models

Meta-models

- Efflux Pump
- Gene Cluster