AMRtime

Precise identification of antimicrobial resistance determinants from metagenomic data

---

Finlay Maguire
finlaymaguire@gmail.com

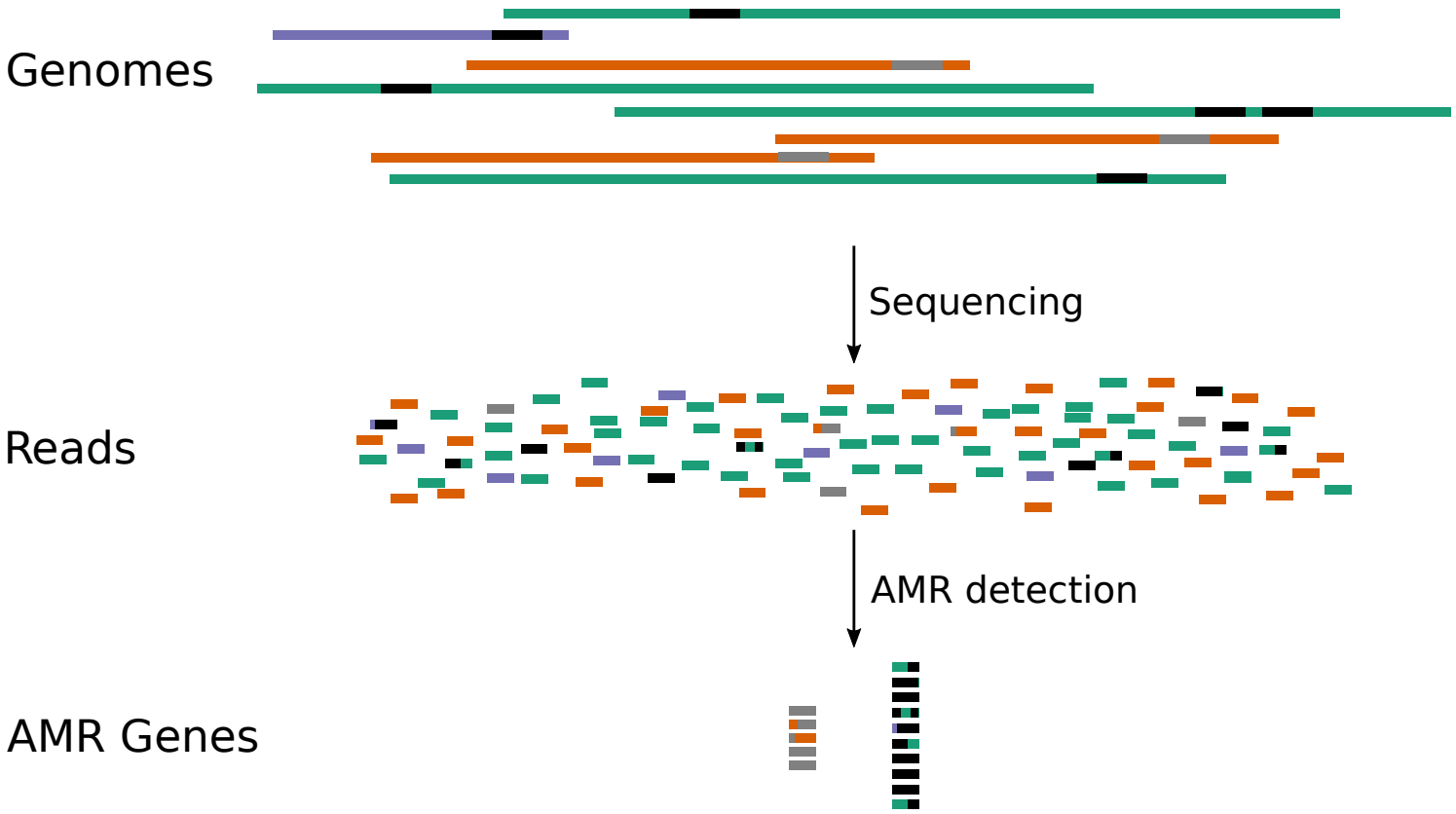
June 11, 2019

Faculty of Computer Science, Dalhousie University

Table of contents

Background

AMR-metagenomics



Comprehensive Antibiotic Resistance Database

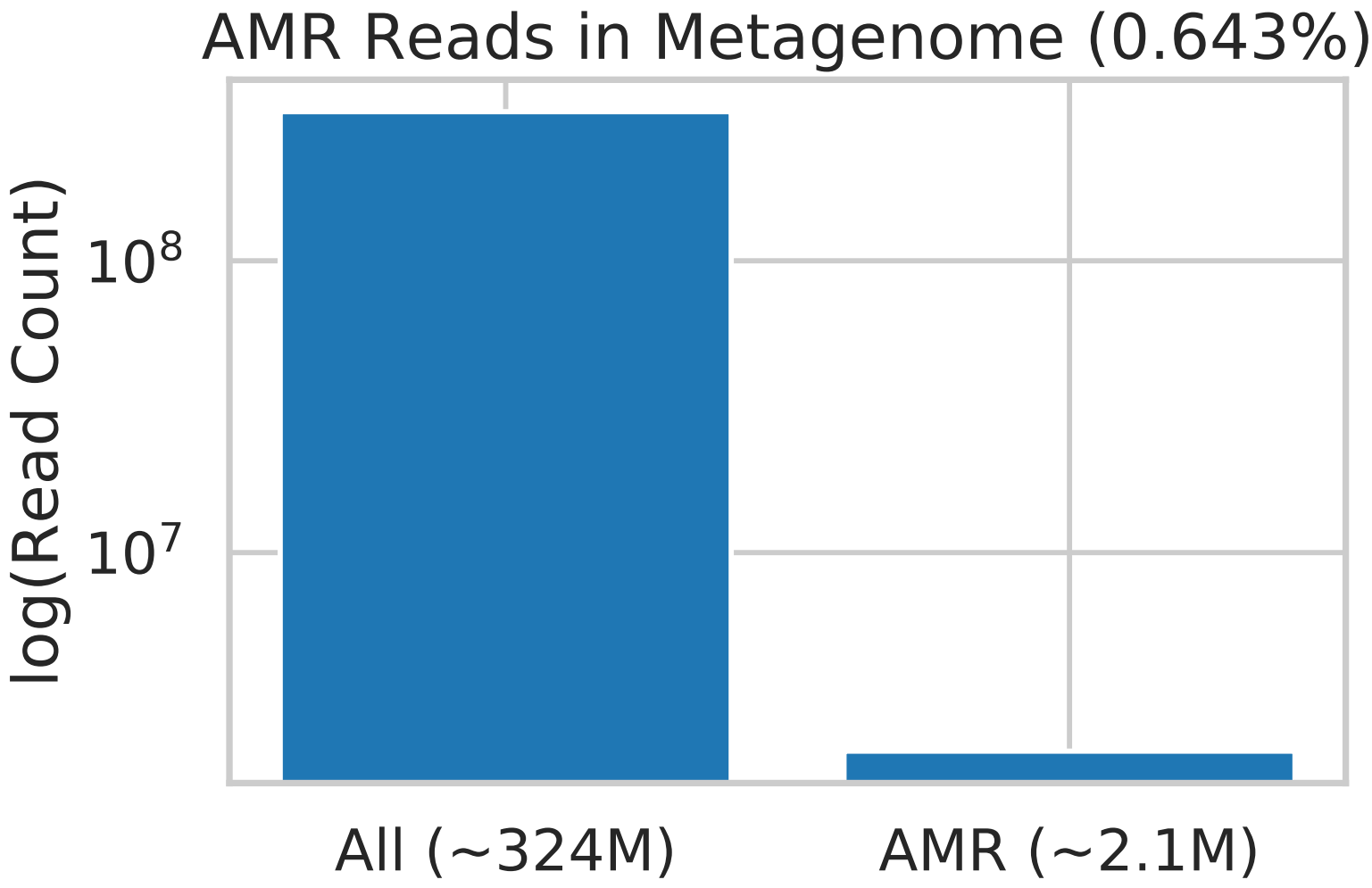
cmlA1

Download Sequences

Accession	ARO:3002693
Definition	cmlA1 is a plasmid or transposon-encoded chloramphenicol exporter that is found in Pseudomonas aeruginosa and Klebsiella pneumoniae
AMR Gene Family	major facilitator superfamily (MFS) antibiotic efflux pump
Drug Class	phenicol antibiotic
Resistance Mechanism	antibiotic efflux
Efflux Component	efflux pump complex or subunit conferring antibiotic resistance
Classification	**7 ontology terms** | Hide + process or component of antibiotic biology or chemistry + mechanism of antibiotic resistance + determinant of antibiotic resistance + antibiotic molecule + antibiotic efflux [Resistance Mechanism] + phenicol antibiotic [Drug Class] + efflux pump complex or subunit conferring antibiotic resistance [Efflux Component]
Parent Term(s)	**2 ontology terms** | Hide + major facilitator superfamily (MFS) antibiotic efflux pump [AMR Gene Family] + *confers_resistance_to_drug* chloramphenicol [Antibiotic]
Publications	Bissonnette L, et al. 1991. J Bacteriol 173(14): 4493-4502. Characterization of

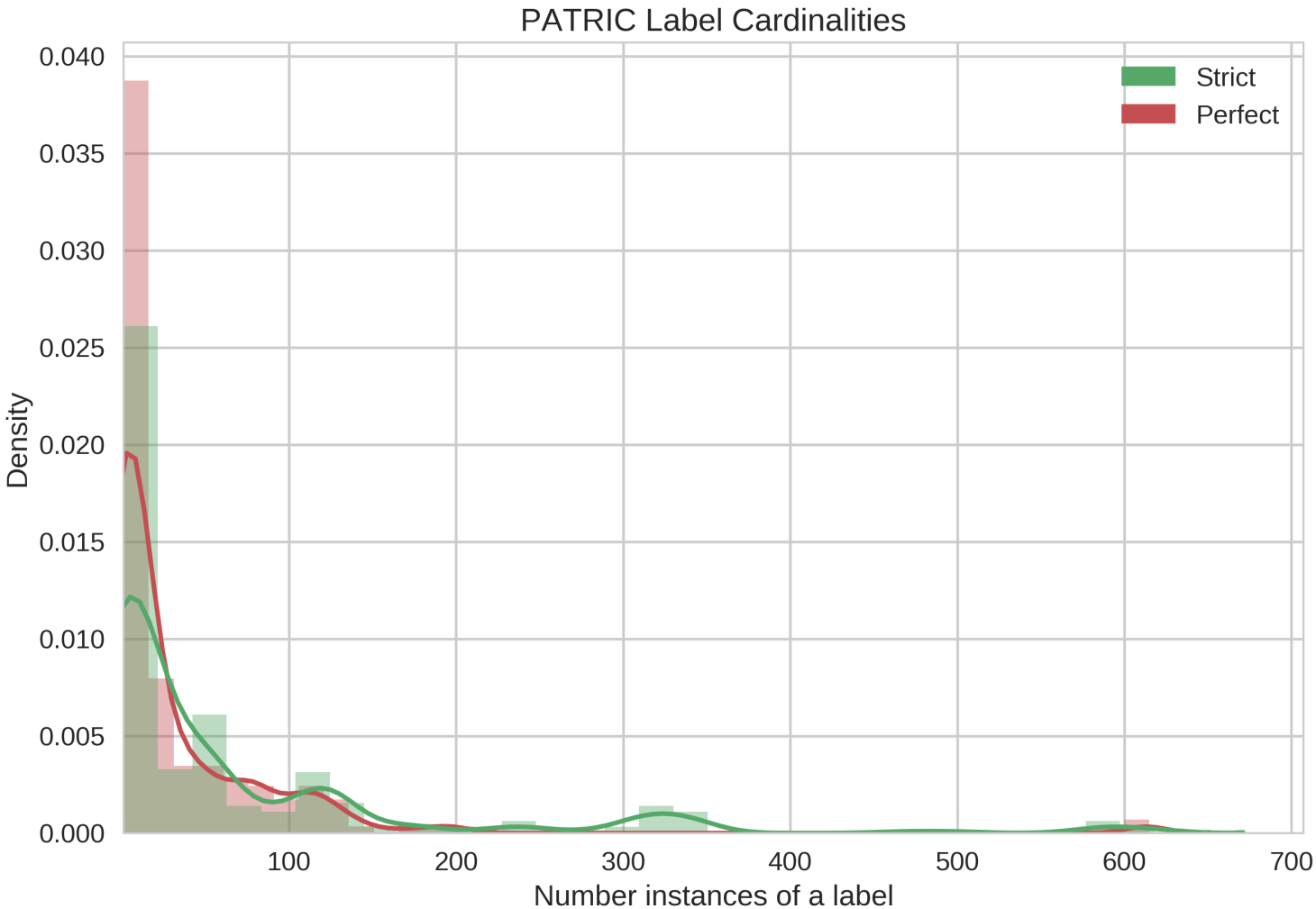
Why is AMR metagenomics difficult?

AMR genes are rare genomically



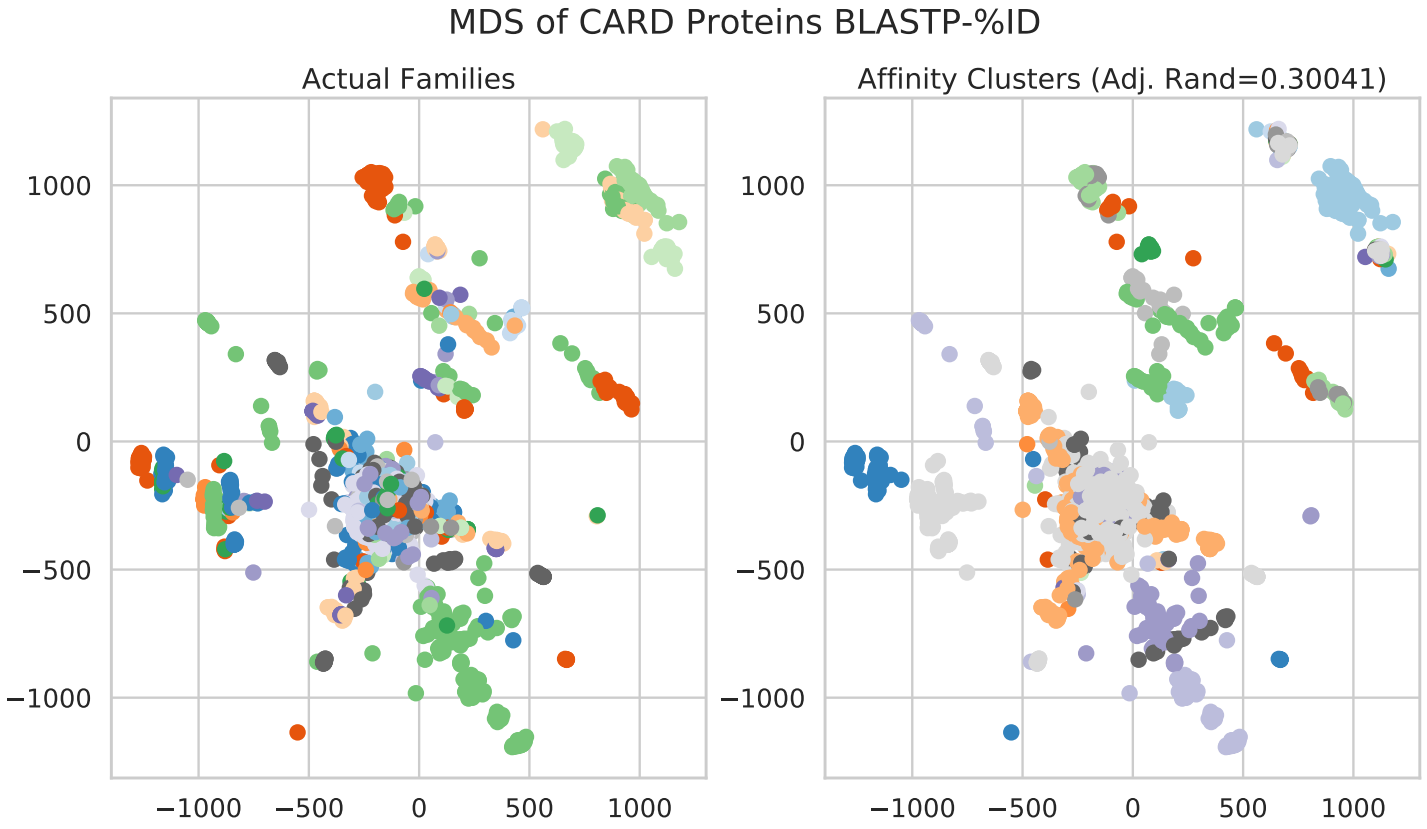
2184 CARD-Prevalence Genomes at 1-10X abundance

AMR genes have wildly different abundances



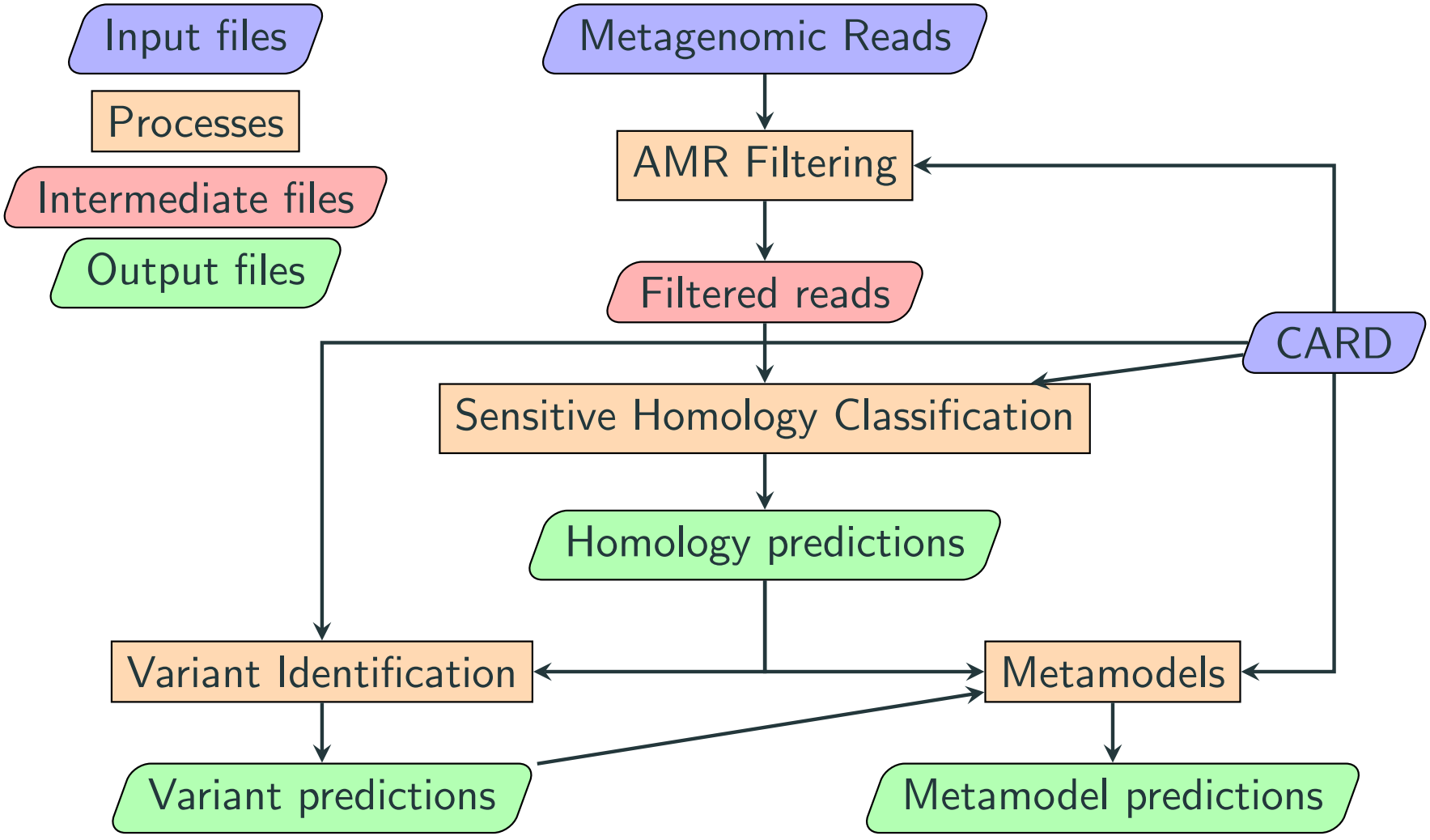
1236 AMR PATRIC genomes

AMR sequence space overlaps

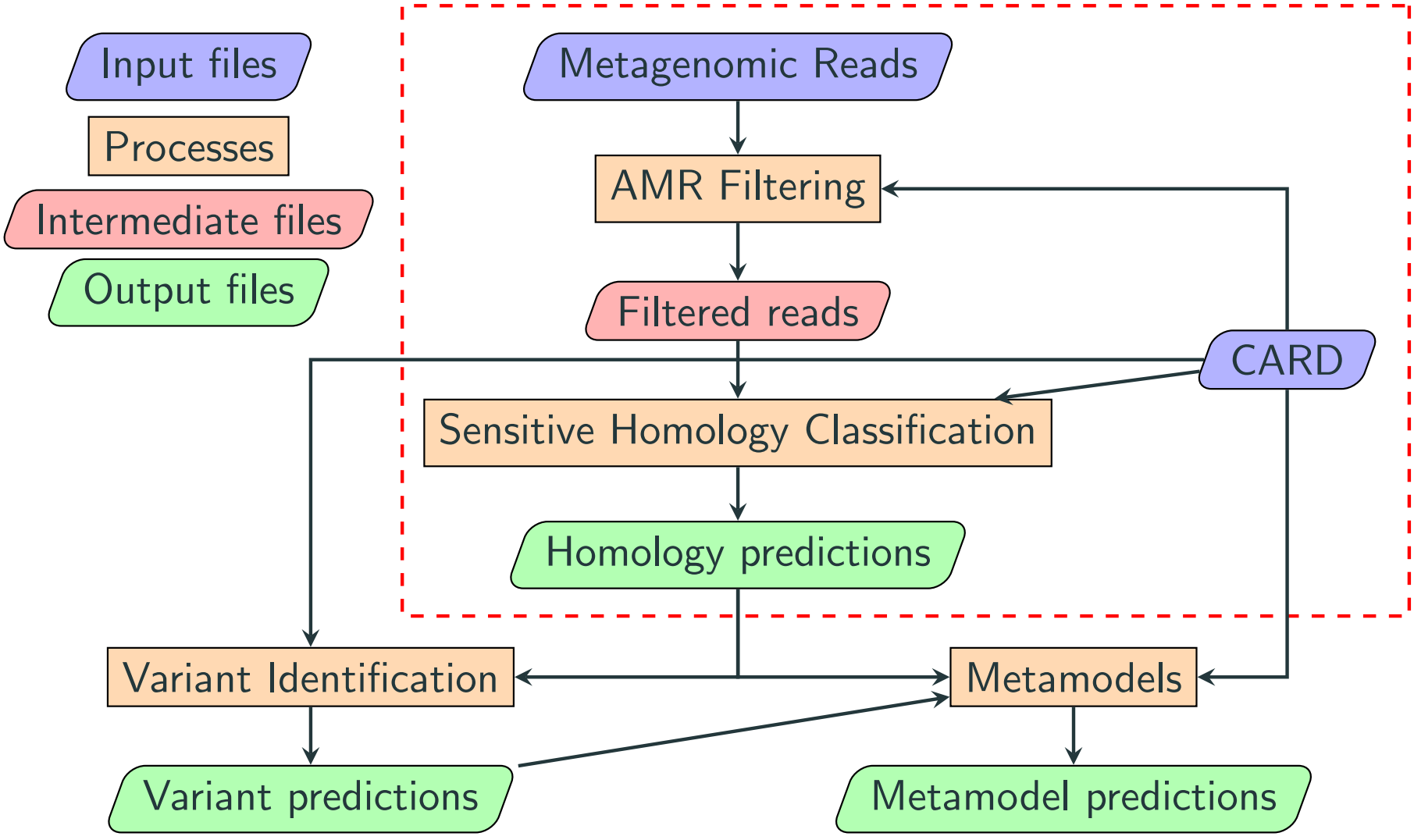


AMRtime Overview

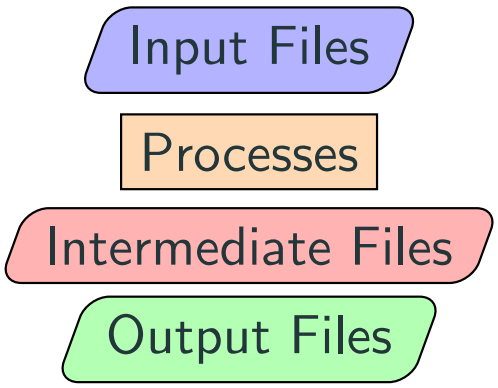
AMRtime structure

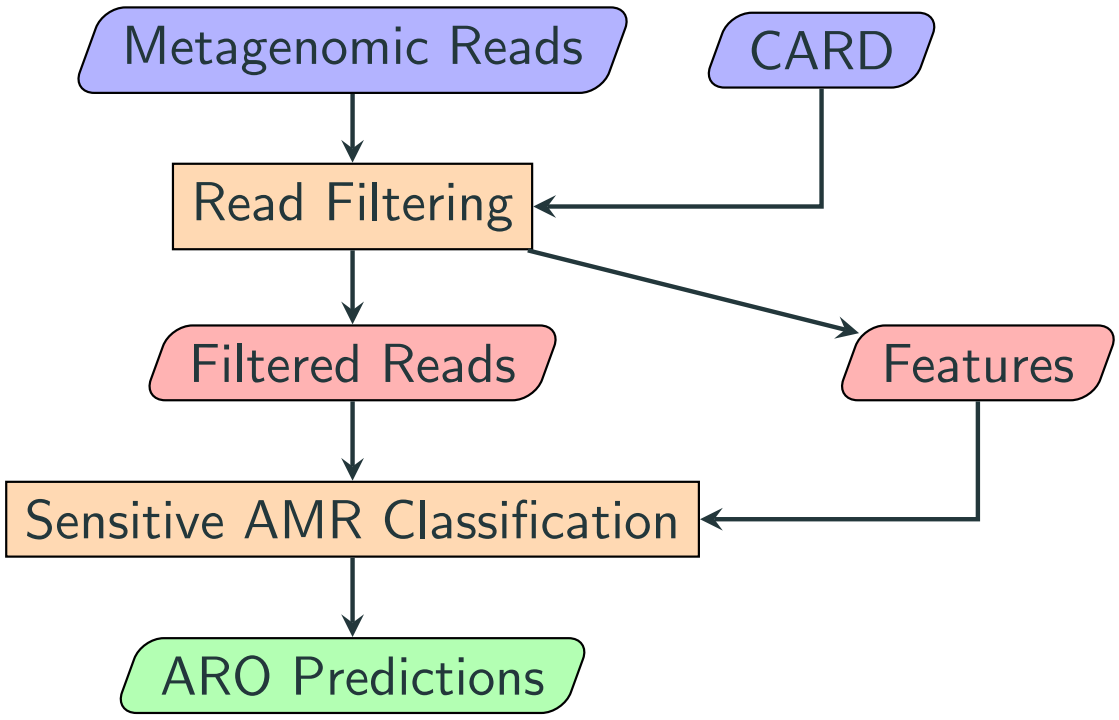


AMRtime structure



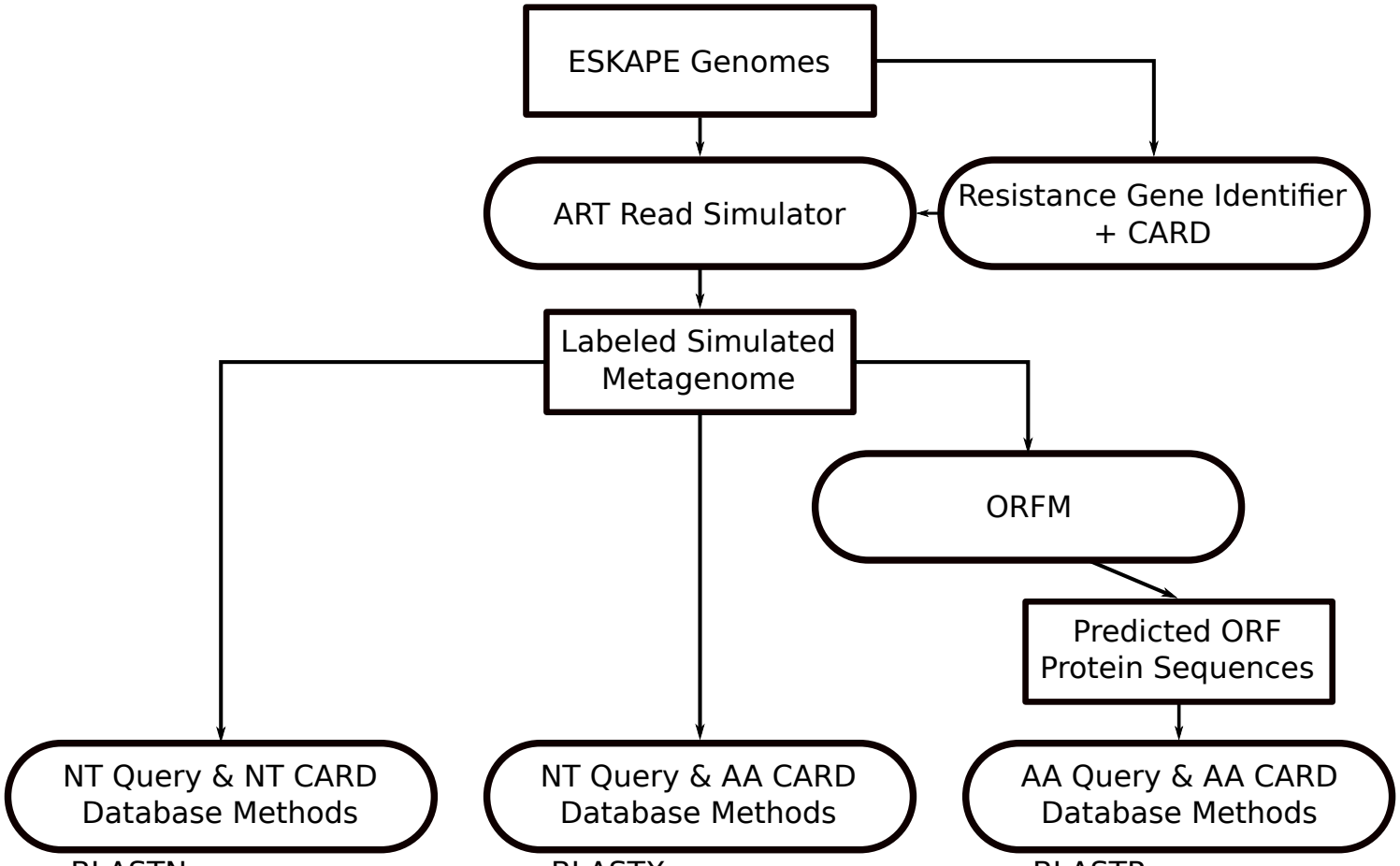
AMRtime structure



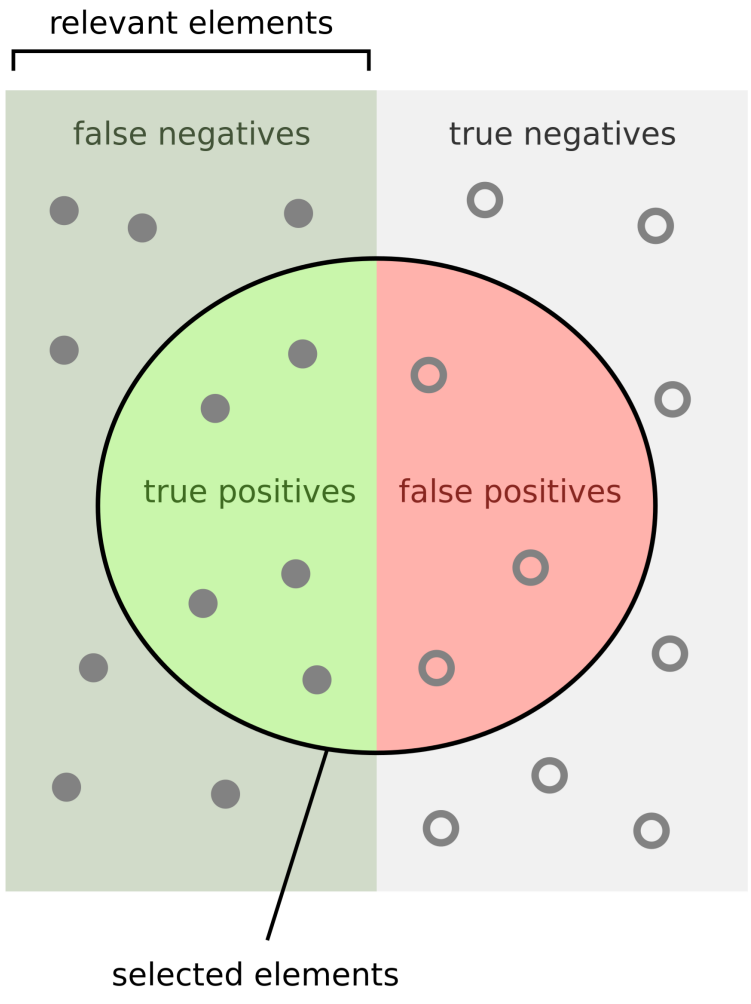


Filtering out non-AMR reads

Testing sequence similarity search tools



Terminology refresher interlude



How many selected items are relevant?

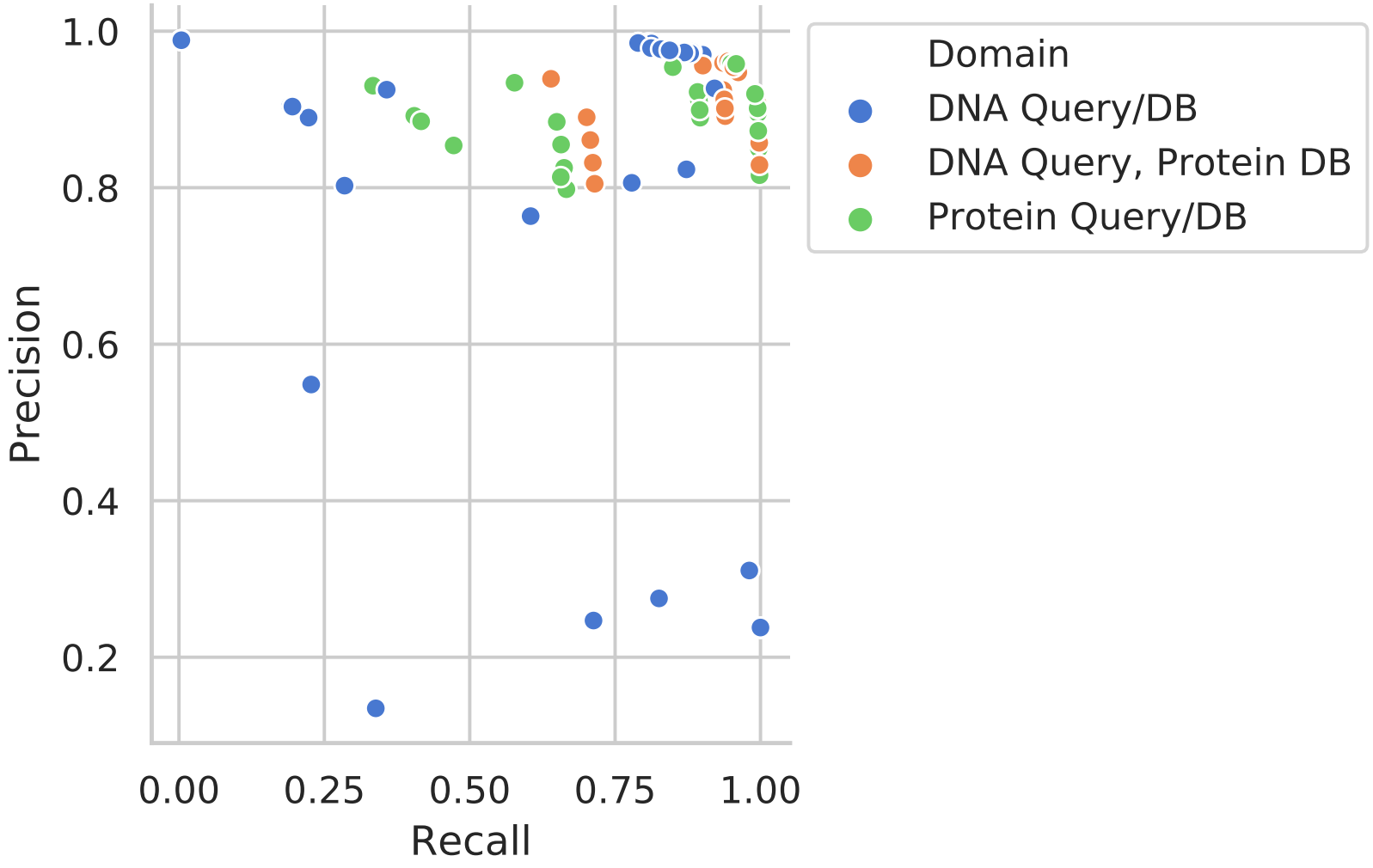
$$\text{Precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

How many relevant items are selected?

$$\text{Recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

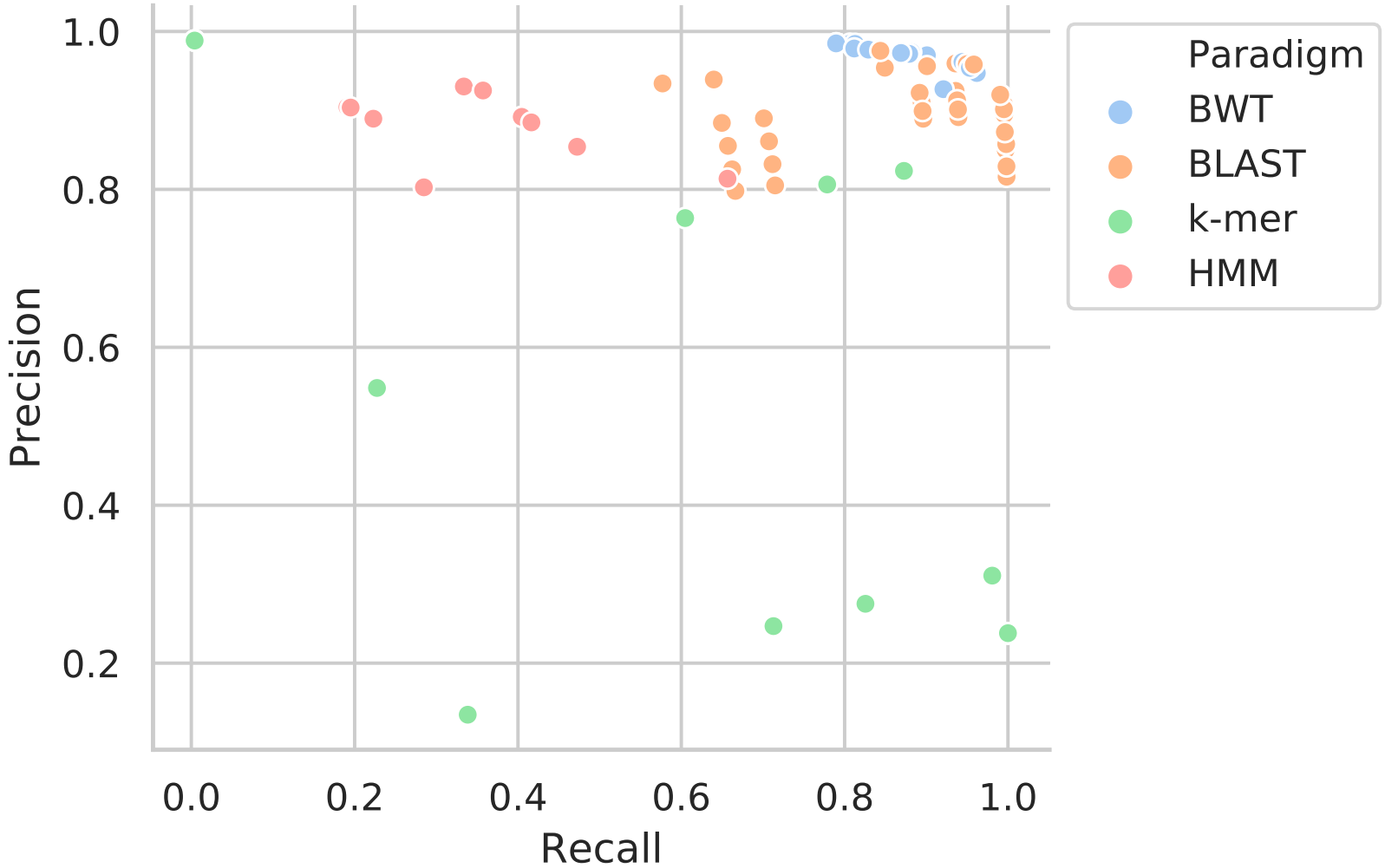
https://commons.wikimedia.org/wiki/File:Precisionrecall.svg

DNA subject best for precision, Protein subject best for recall



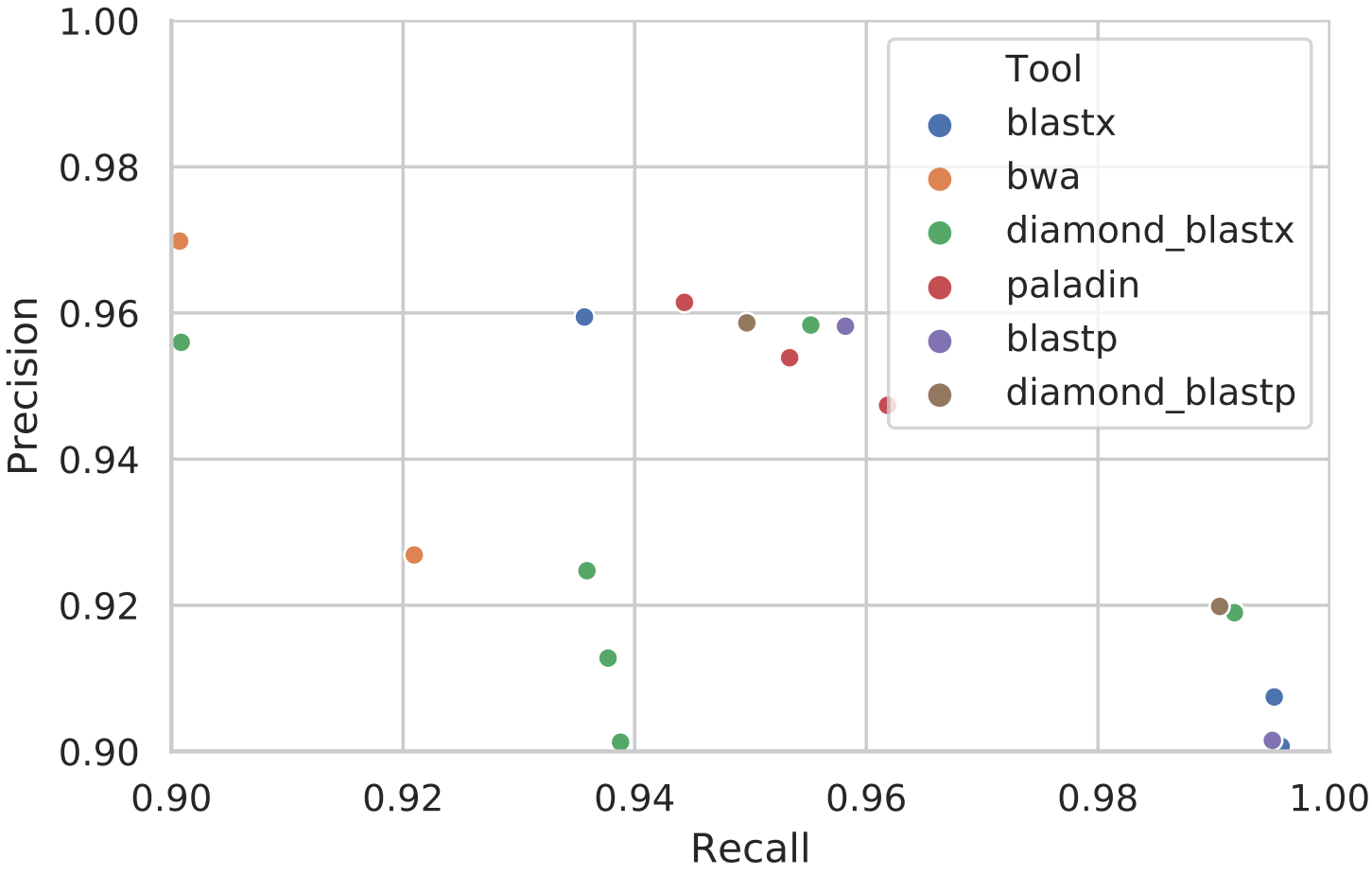
Simulated MiSeq v3 250bp reads, 30.31M reads (7.21M AMR derived)

K-mer methods perform poorly



BWT: bowtie2, bwa-mem, paladin; **BLAST:** blast, diamond; **HMM:** hmmsearch; **K-MER:** biobloom, groot.

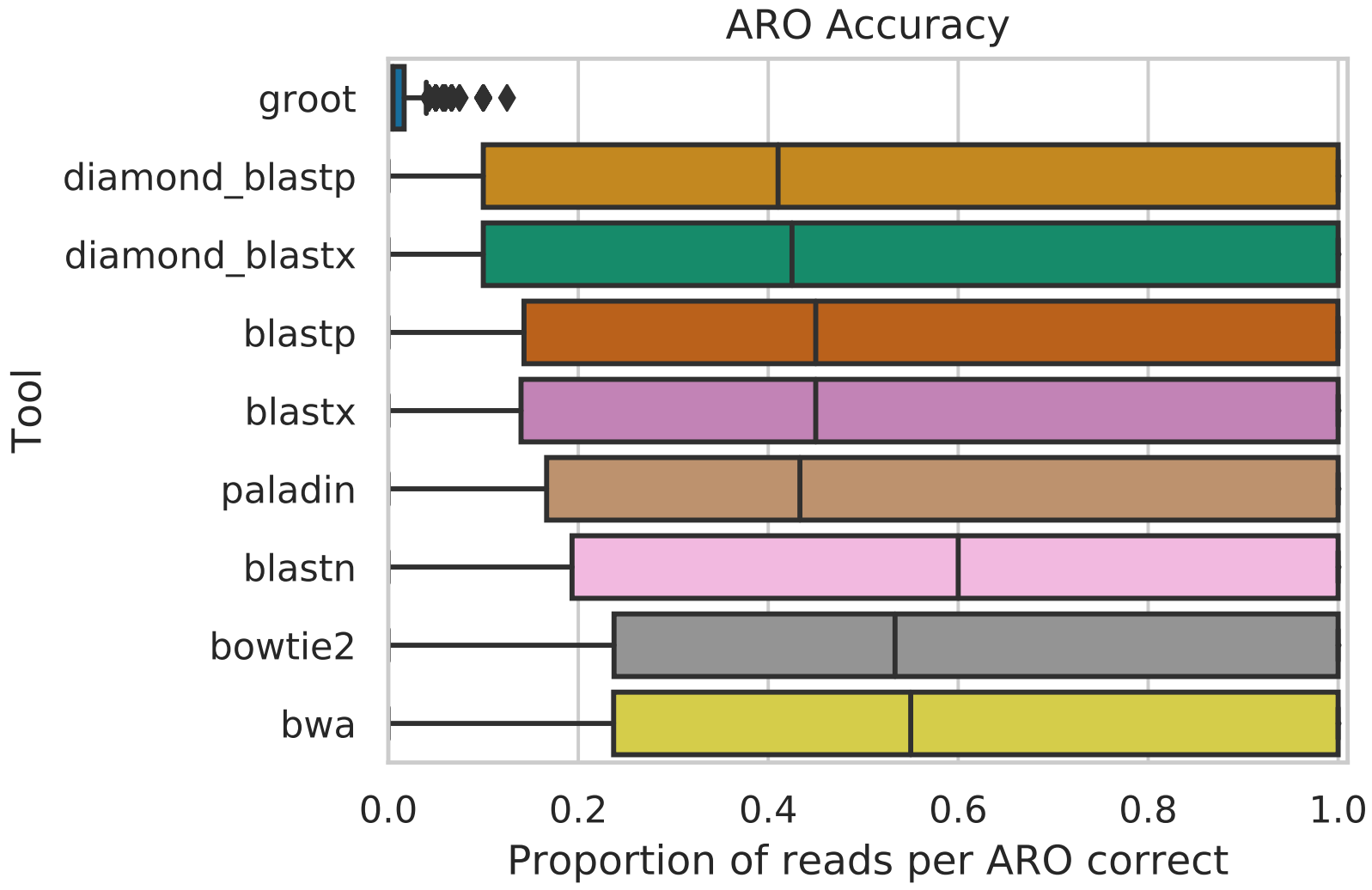
DIAMOND-BLASTX best compromise



DIAMOND-BLASTX 'more sensitive' setting (min $< 1e^{-10}$): 4.926 hours with 2 cores and 8.3Gb of memory. AMR Reads: 7.15M detected, 59.26K missed, 1.87M false positives.

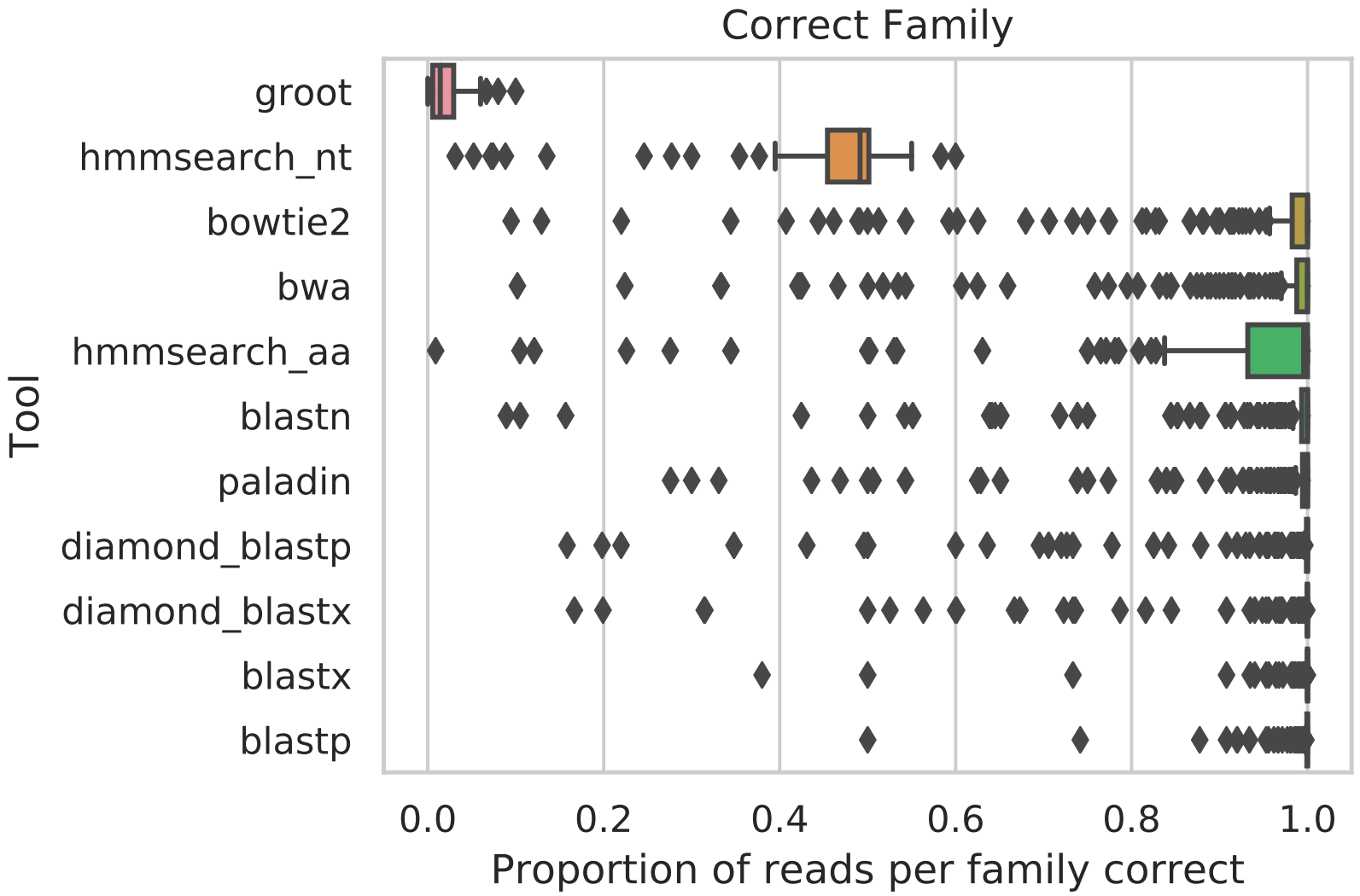
Why not just use these sequence searches?

Poor gene-level accuracy



Performance at optimal settings for ARO accuracy

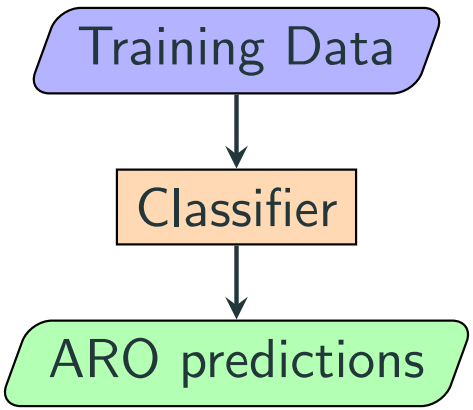
Good family-level accuracy



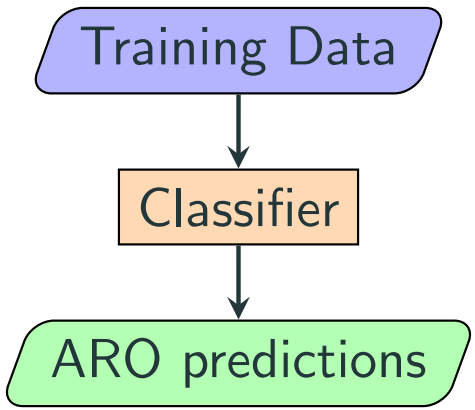
Performance at optimal settings for Family accuracy

Sensitive Homology Classification

Initial classifier

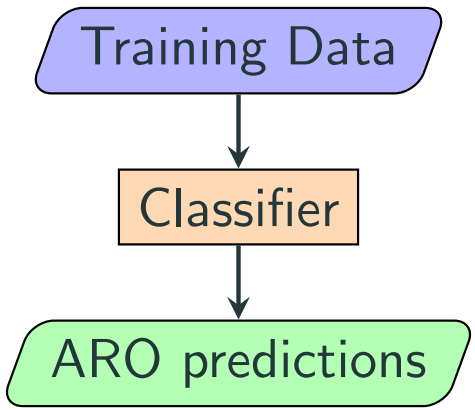


Initial classifier



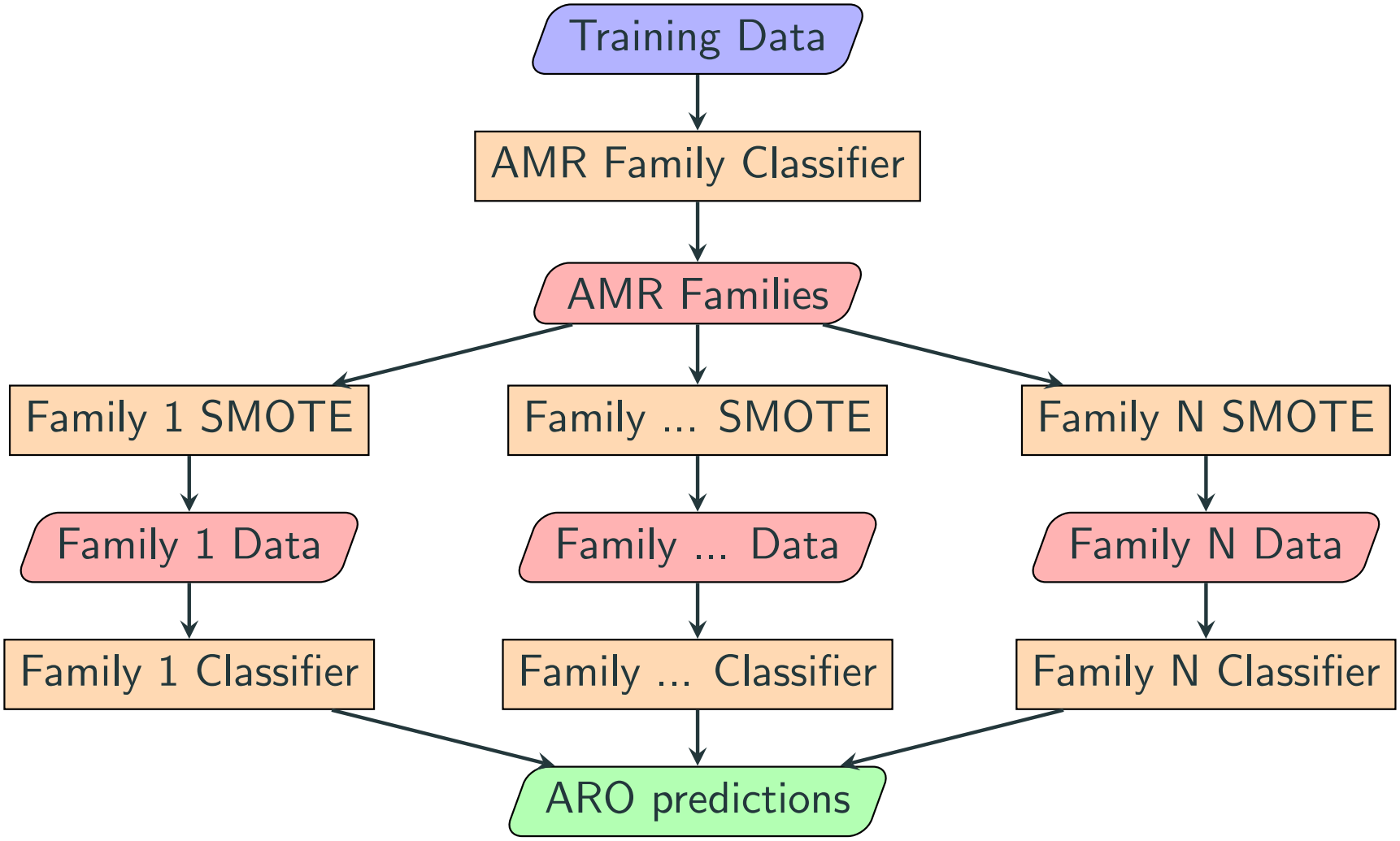
NB 7-mer Average Precision: 0.63

Initial classifier



NB 7-mer Average Precision: 0.63 %

Revised classifier structure: exploiting the ARO

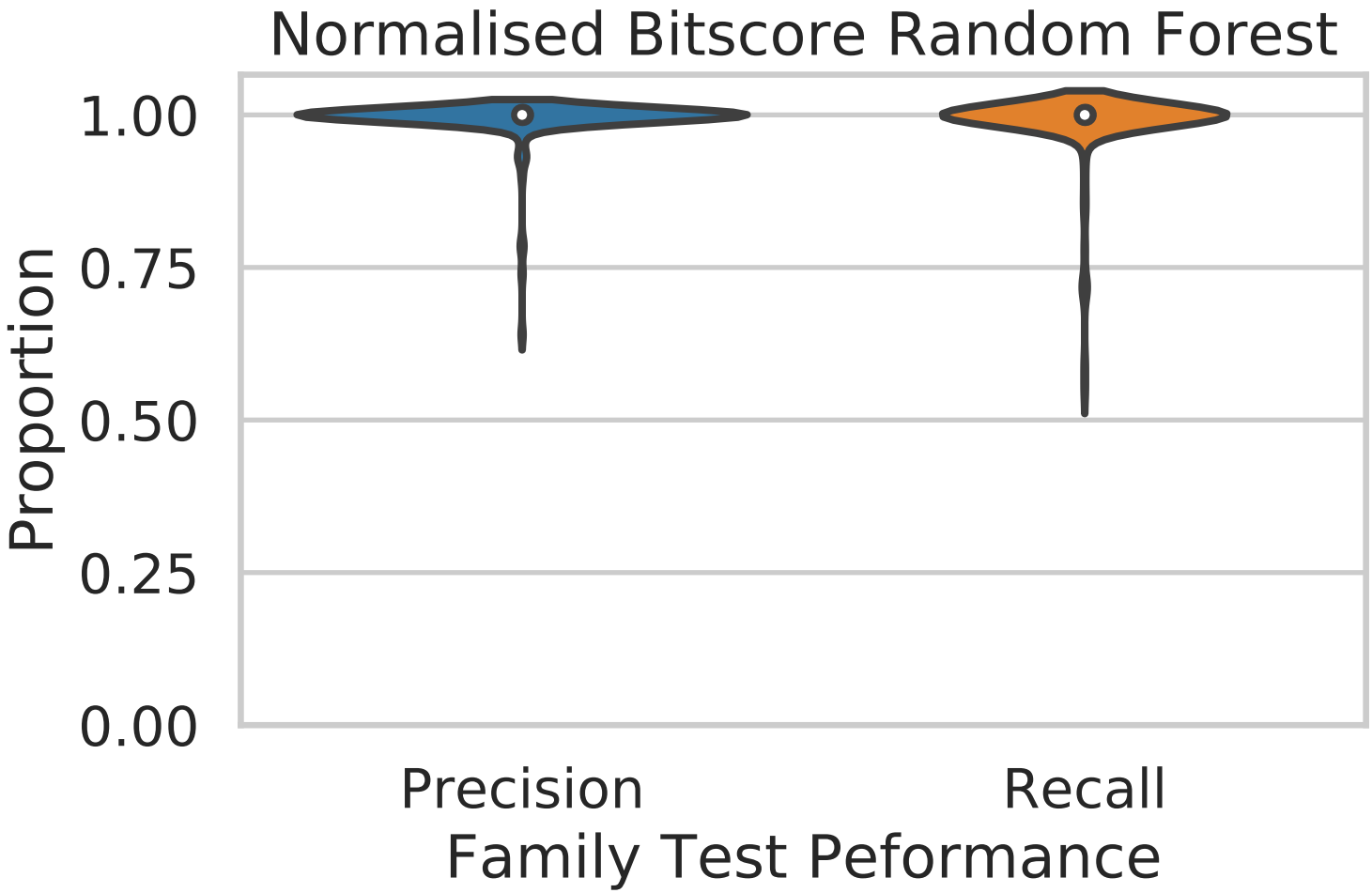


Read encoding

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

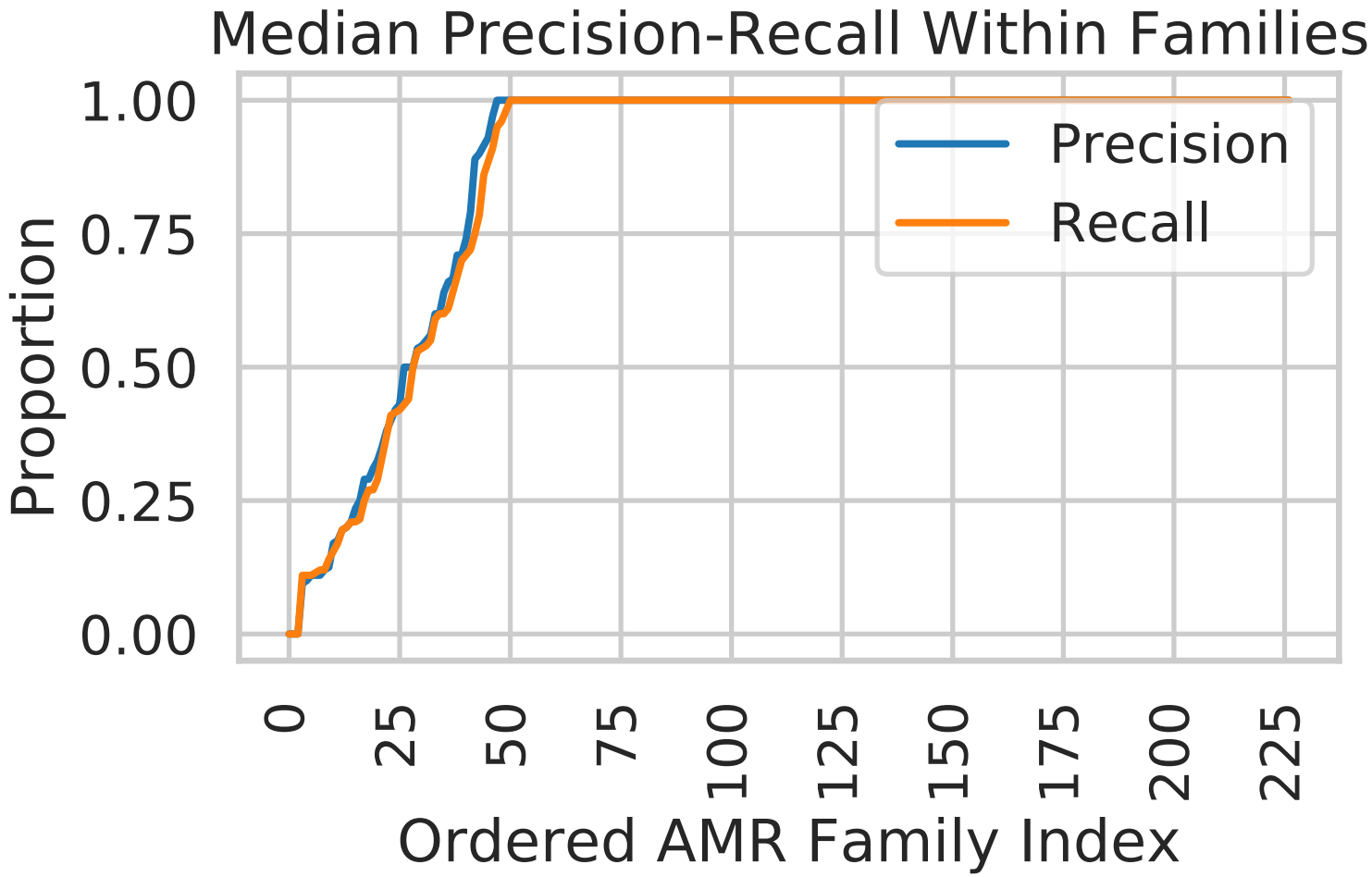
Advantages: read length invariant, low dimensionality, uses filtering data

Held-out test results



Mean Precision: 0.995, Mean Recall: 0.985

ARO level classification more variable



On-going work

- Soft-threshold (i.e. propagating probabilities through layers)
- Multiset labels based on sequence redundancy within families.
- Threshold identification for variant model counts.
- Metamodel rule parsing.
- Galaxy bindings (CARD/IRIDA integration).

Summary

Conclusions

- Direct homology searches are suprisingly poor for AMR metagenomics.

Conclusions

- Direct homology searches are suprisingly poor for AMR metagenomics.
- K-mer based approaches fall flat with sequencing error, low coverage and sparse labels.

Conclusions

- Direct homology searches are suprisingly poor for AMR metagenomics.
- K-mer based approaches fall flat with sequencing error, low coverage and sparse labels.
- Direct homology search results ARE useful when combined with machine learning.

Conclusions

- Direct homology searches are suprisingly poor for AMR metagenomics.
- K-mer based approaches fall flat with sequencing error, low coverage and sparse labels.
- Direct homology search results ARE useful when combined with machine learning.
- The Antibiotic Resistance Ontology provides useful structure to improve predictions.

Conclusions

- Direct homology searches are suprisingly poor for AMR metagenomics.
- K-mer based approaches fall flat with sequencing error, low coverage and sparse labels.
- Direct homology search results ARE useful when combined with machine learning.
- The Antibiotic Resistance Ontology provides useful structure to improve predictions.
- AMRtime: coming soon to CARD and your local government genomic epidemiology platform.

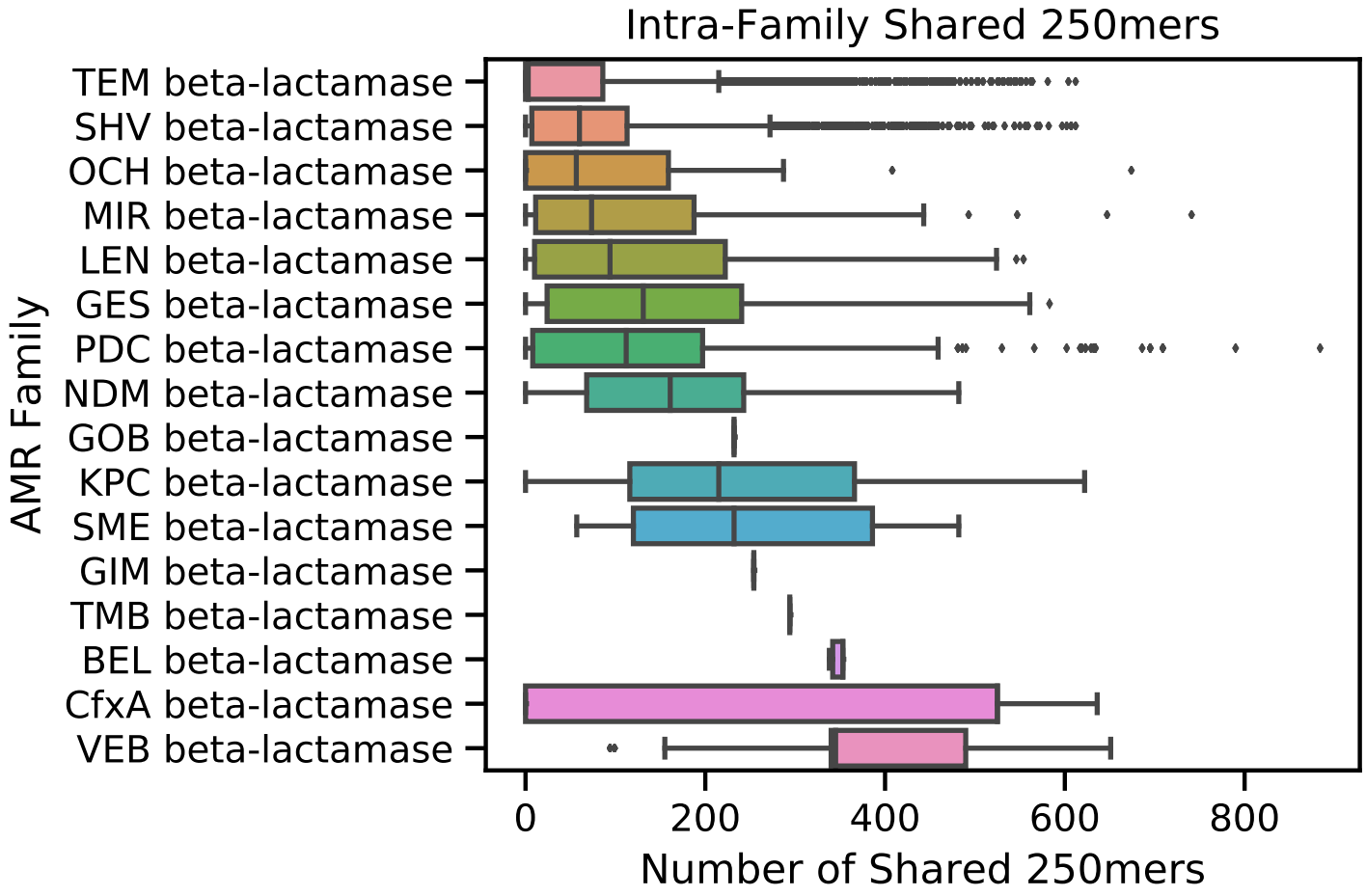
Acknowledgements

Acknowledgements

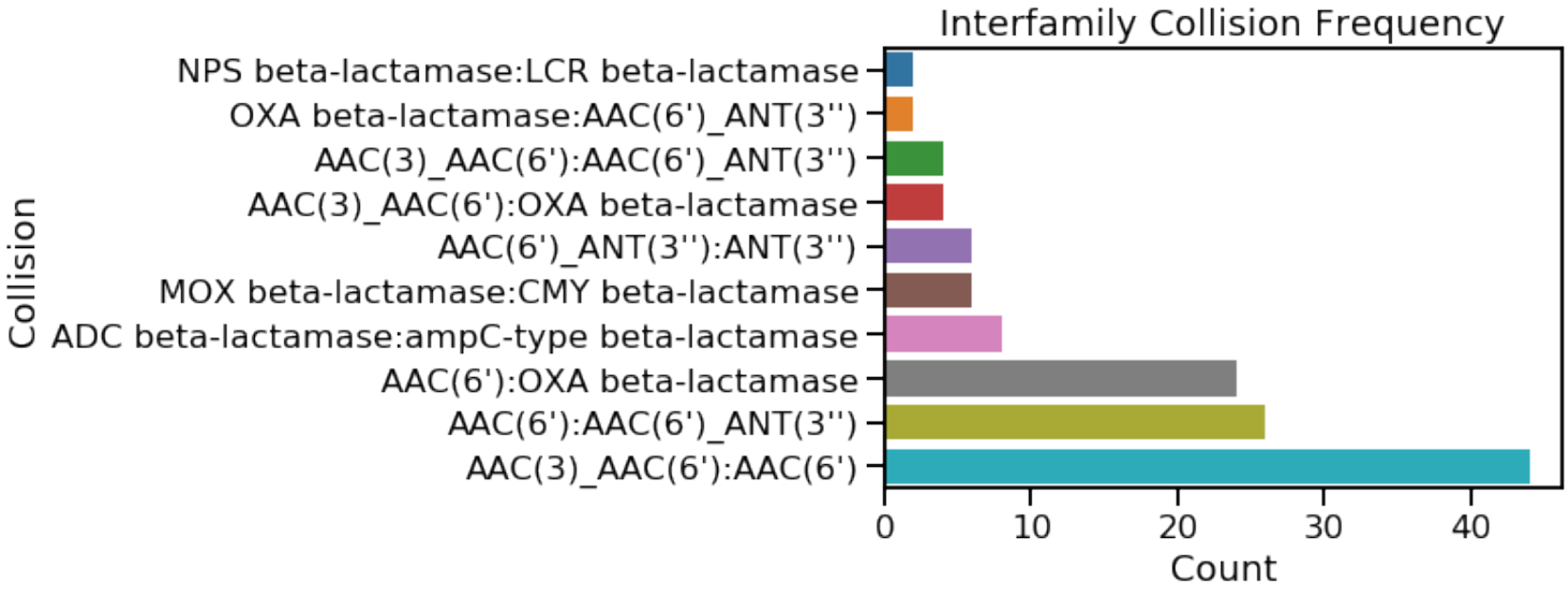
- McMaster University: Brian Alcock and Andrew McArthur
- Simon Fraser University: Fiona Brinkman
- Dalhousie University: Robert Beiko
- Funding: Donald Hill Family Fellowship, Genome Canada Grant.

Questions?

Insufficient Intrafamily Signal



Interfamily Collisions



Interfamily Collisions