The Cost of Speed: Evaluating Systematic Failures in Metagenomic Antimicrobial Resistance Profiling

F Maguire[1], AR Raphenya[2], B Alcock[2], FSL Brinkman[3], AG McArthur[2], RG Beiko[1]

[1]Dalhousie University, Halifax, NS [2]McMcaster University, Hamilton, ON [3]Simon Fraser University, Burnaby, BC

DALHOUSIE UNIVERSITY

IRIDA Integrated Rapid Infectious Disease Analysis

Background

- Metagenomics, the direct sequencing of DNA from biological samples, is emerging as an important method in the study of Antimicrobial Resistance (AMR).
- The key analytic stage in metagenomics analysis is the identification of genes represented within the DNA sequencing reads.
- This is performed via alignment of reads against protein reference databases.
- Protein references are used as they are typically more robust to sequencing error and have more discriminatory power than nucleotide databases.
- The de facto standard tool for conducting this search is BLASTX (1).

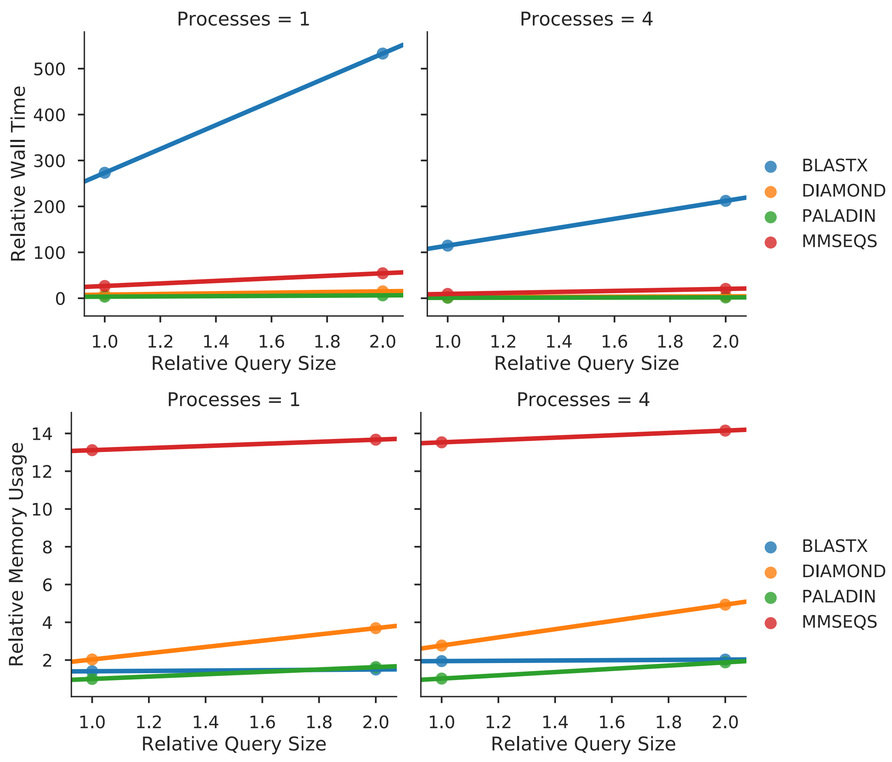

Figure 1: Resource usage analysis relative to minimum time and memory

- Unfortunately, BLASTX v2.2.28 is relatively slow making it infeasible for the millions to billions of reads in a typical metagenome.
- Several tools incorporating additional heuristics have been developed to increase analysis speed: DIAMOND v0.9.19.120 (2), PALADIN v1.31 (3), and MMseqs2 v3-be8f6 (4).
- However, these heuristics, such as reduced alphabets and spaced-seeds, trade-off a certain loss of precision for this speed.

Methods

- In order to assess this error, we generated a labeled ~1 million read synthetic MiSeq metagenome from the Comprehensive Antibiotic Database (CARD) (5) March 2018 release) using ART (6).
- Additionally, an ~800 million read labeled metagenome was simulated from 3,420 ESKAPE, WHO priority and curator selected pathogen genomes underlying the CARD Prevalence database using AMRtime (github.com/beiko-lab/AMRtime).
- Tool performance was then assessed a broad range of parameter settings (BLASTX minimum e-value 1e-3 to match DIAMOND and MMseqs2) against CARD and CARD+CARD Prevalence databases.

Results & Discussion

- BLASTX is by far the most time intensive tool, 250-500X the fastest tool: PALADIN (figure 1).
- MMseqs2 uses the most memory, 12.5-14X more than PALADIN (figure 1).

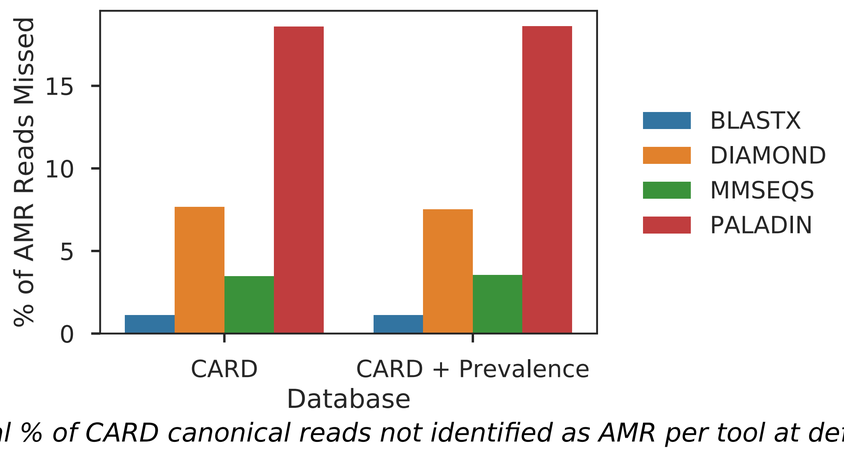

Figure 2: Total % of CARD canonical reads not identified as AMR per tool at default settings

- BLASTX has lowest miss rate (~1%) with default settings while PALADIN performs worst (17%) (figure 2).
- Addition of CARD Prevalence database doesn't decrease missed reads with default settings (figure 2).
- DIAMOND has the most AMR genes that are not identified for >20% of reads simulated from those genes (figure 3).

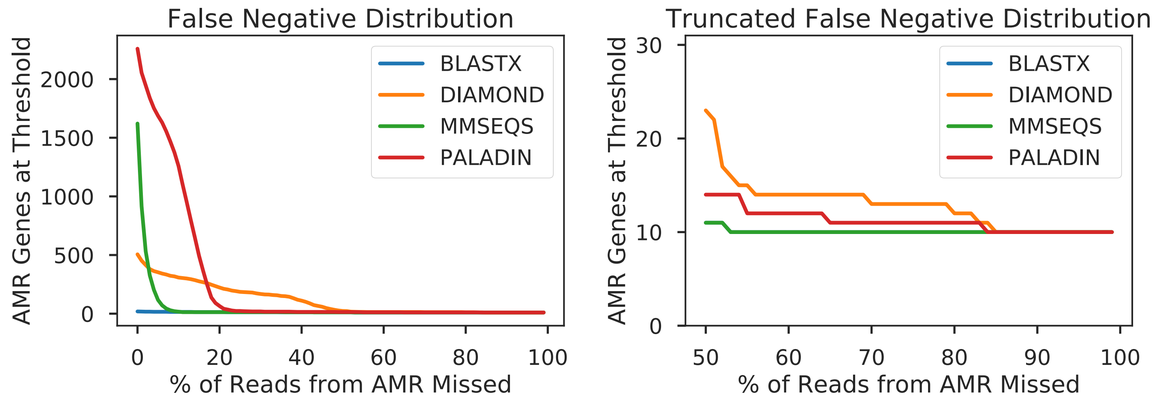

Figure 3: AMR gene specific performance of tools at default settings

- 10 AMR genes were never identified by any tool due to curation error, will be fixed in next CARD release (figure 3).
- MMseqs2 has best profile of the accelerated tools (similar to BLASTX) for >5% AMR gene specific error rate (figure 3).

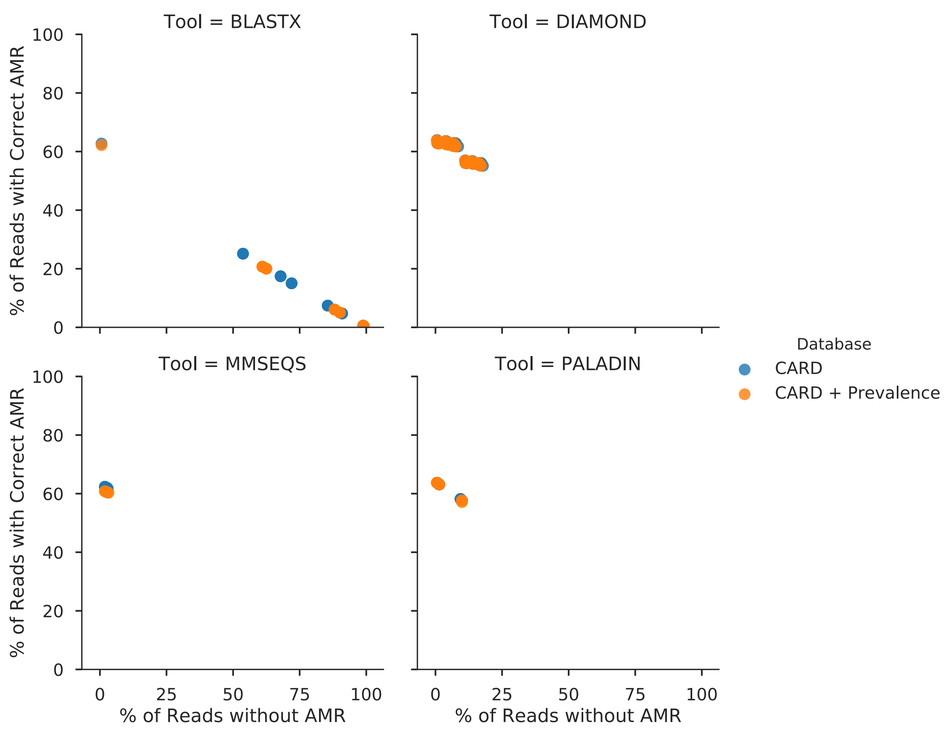

Figure 4: Overall performance across a range of parameter settings

- Parameter sweep supports little improvement from inclusion of CARD + Prevalence database and a ~63% maximum accuracy (figure 4).

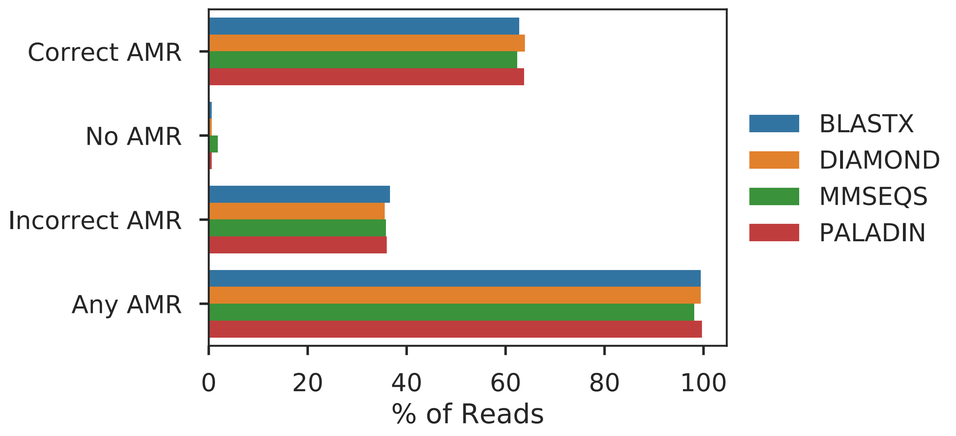

Figure 5: Overall performance of best settings per tool

- Similar overall performance achievable by parameter optimisation for all tools (figure 4, 5).

Future Work

- Assess false positive rates using CARD prevalence genome metagenome as this contains reads that are not derived from AMR genes.
- This is necessary to identify parameter optimisation overfitting issues.

Conclusions

- DIAMOND has worst systematic error profile (figure 3) but maximum possible correct identification of AMR gene 63.9% (CARD database, more sensitive, minimum ORF: 1, minimum e-value: 1e-3) (figure 4, 5).
- MMseqs2 had highest overall miss % at optimal settings (figure 5) and worst memory usage (figure 1) but lowest AMR gene specific systematic failure (figure 3).

References

1. States, David J,J., and Warren Gish. "Combined Use of Sequence Similarity and Codon Bias for Coding Region Identification." (1996)

2. Buchfink, Benjamin, Chao Xie, and Daniel H. Huson. "Fast and sensitive protein alignment using DIAMOND." Nature methods 12.1 (2015): 59.

3. Westbrook, Anthony, et al. "PALADIN: protein alignment for functional profiling whole metagenome shotgun data." Bioinformatics 33.10 (2017): 1473-1478.

4. Steinegger, Martin, and Johannes Söding. "MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets." Nature biotechnology 35.11 (2017): 1026.

5. Jia, Baofeng, et al. "CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database." Nucleic acids research (2016): gkw1004.

6. Huang, Weichun, et al. "ART: a next-generation sequencing read simulator." Bioinformatics 28.4 (2011): 593-594.

This work was funded by Genome Atlantic and forms part of the Public Health Agency of Canada's IRIDA platform AMR metagenomic analysis tool development (AMRtime)