Sarand: Extracting Gene Neighborhood from Complex Metagenomic Assembly Graphs

Somayeh Kafaie, Robert Beiko and Finlay Maguire

Faculty of Computer Science, Dalhousie University

DALHOUSIE UNIVERSITY

Motivation

Antimicrobial Resistance (AMR) is one of the biggest health challenges of our world with ~700,000 annual deaths. Studies show that the genomic context of AMR genes plays a key role in their evolution and transmission via lateral gene transfer. Metagenomic sequencing offers an efficient strategy for AMR profiling as it can comprehensively profile the AMR gene content of a microbial community. Sequence assembly is an important step in metagenomics, but the assembled contigs present an oversimplified view that can miss valid genomic context information. Therefore, we are interested in regaining the context of AMR genes by exploring the assembly graph directly rather than relying on constructed contigs.

Methods

We developed Sarand, a tool to explore the structure of assembly graphs around target AMR genes, and designed efficient algorithms to extract the genomic neighborhood of AMR genes from metagenomic data. Furthermore, we defined gene-coverage to remove invalid paths that arise from assembly errors and only keep valid paths.

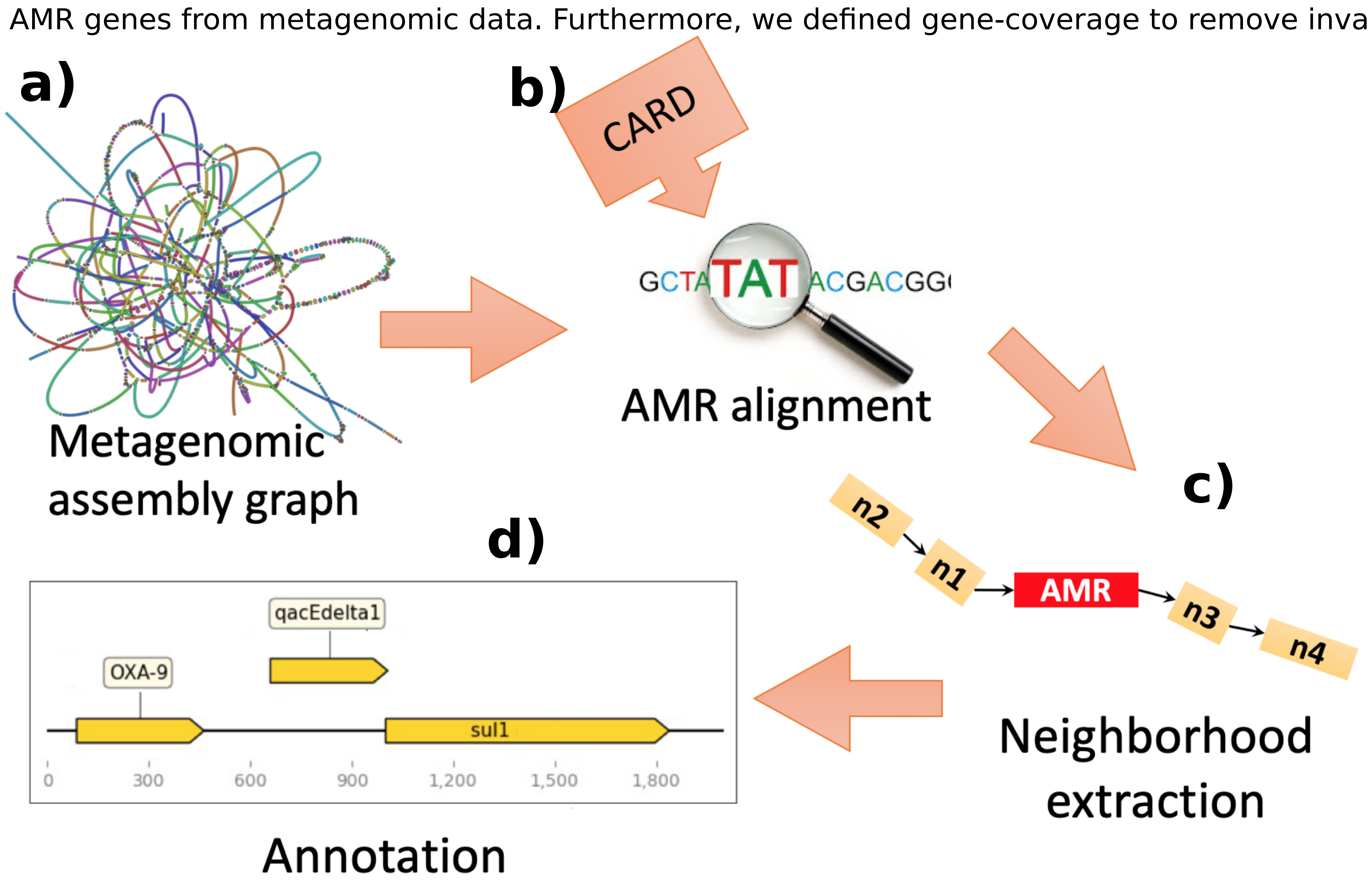
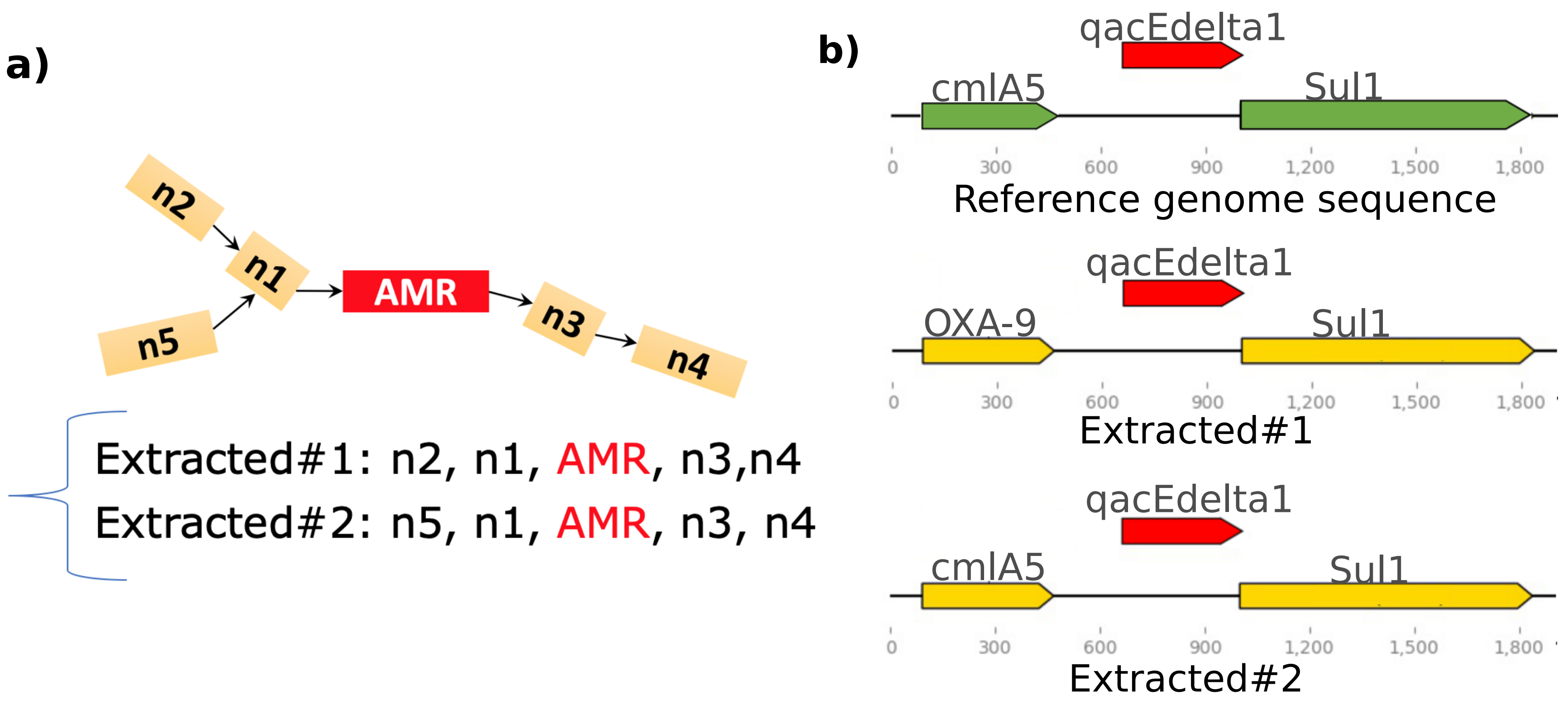


Figure 1 - High-level workflow of Sarand. a) An assembly graph was constructed from metagenomic data. **b)** The BLAST[1] search embedded in Bandage[2] was used to detect AMR genes downloaded from CARD[3] in the assembly graph. **c)** The genomic neighborhood of an AMR gene within a given length of flanking sequences, as a list of graph paths upstream and downstream of AMR, gene was identified. **d)** The extracted sequences were annotated and visualized by running Prokka[4], RGI[3] and dna-features_viewer[5] library.

Extracting Neighborhoods

Recursive functions were used to extract the sequences of all upstream/ downstream nodes of the target AMR gene up to a pre-defined length. Then, all upstream and downstream sequences were concatenated with the AMR sequence.

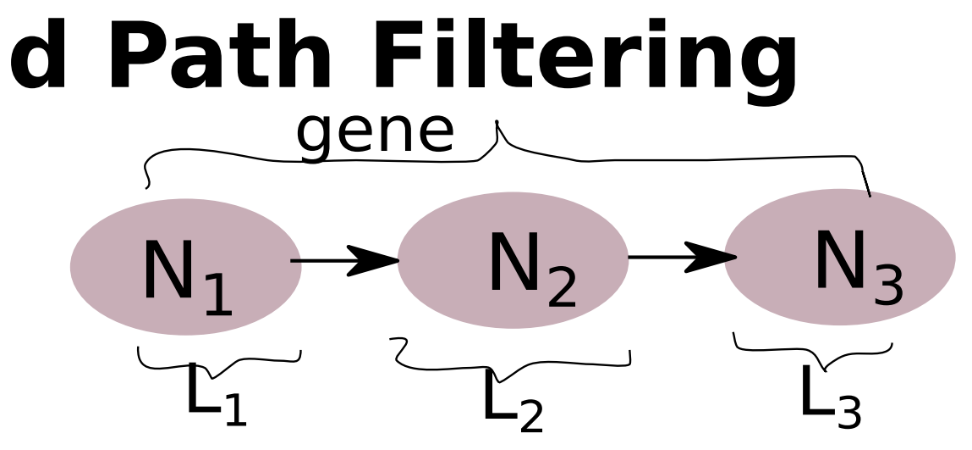


Extracted#1: n2, n1, AMR, n3,n4

Extracted#2: n5, n1, AMR, n3, n4

Figure 3 - Neighborhood extraction. a) Neighborhood of an AMR gene (*qacEdelta1*) was extracted as two separate paths with a maximum length of 1000 bp on each side of the AMR gene. **b)** Two neighborhood sequences were annotated and visualized. Comparing the results with the neighborhood of the same AMR in reference genomes shows that while downstream was constructed successfully, the extracted upstream sequence containing *OXA-9* is not valid. In fact, node n2 is incorrectly linked to n1 as a result of assembly errors leading to the appearance of such a false-positive case.

Gene Coverage and Path Filtering



Gene coverage is calculated as a weighted average over the coverage of all nodes contributing to the gene's sequence where the weight of each node is the proportion of the gene sequence presented in the node sequence. For example, for a gene, represented by nodes N_1, N_2 and N_3, with length $L = L_1 + L_2 + L_3$, the coverage equals:

$$coverage_{gene} = \sum_{i=1}^{3} coverage_{N_i} \times \frac{L_i}{L}$$

The coverage of false-positive genes is often very different from that of the target AMR gene. We implemented a method to prune paths with different coverage scores than the AMR gene to reduce the chance of including false-positive cases.

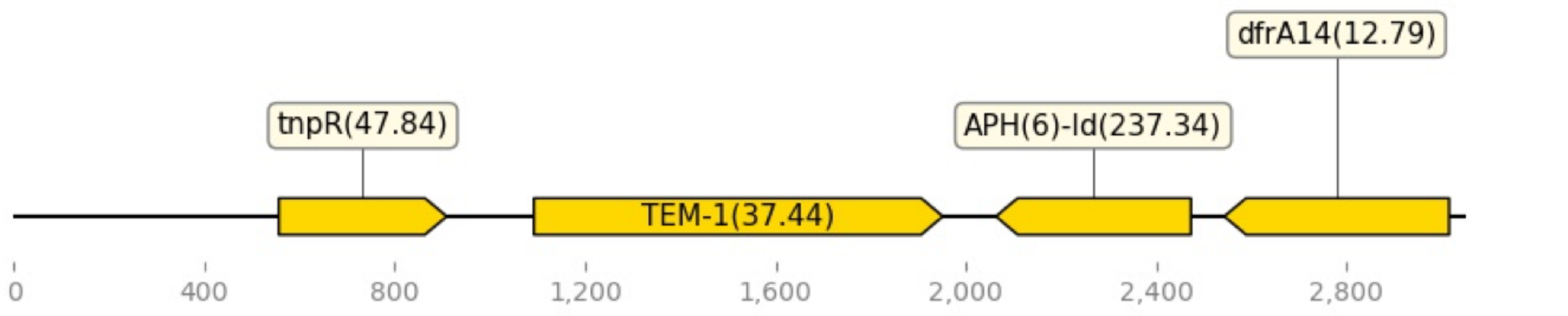


Figure 4 - Path filtering. The neighborhood of *TEM-1* is extracted from the assembly graph with numbers in parentheses presenting gene coverage. Since the coverage of *APH(6)-Id* is very different from that of *TEM-1*, *APH(6)-Id* and any nodes further downstream (e.g., *dfrA14*) will be removed from the extracted path.

Results

We validated Sarand on simulated datasets, and compared our results with contigs results. In fact, for all AMR genes, their neighborhoods were identified and annotated from contigs as well as by our method from the graph. Then, we compared these upstream/downstream annotations with the annotations of neighborhood sequences in the reference genomes, and calculated precision and sensitivity for both contigs and Sarand.

Contacts: somayeh.kafaie@dal.ca, rbeiko@dal.ca and finlay.maguire@dal.ca

Source: https://github.com/beiko-lab/AMR_context

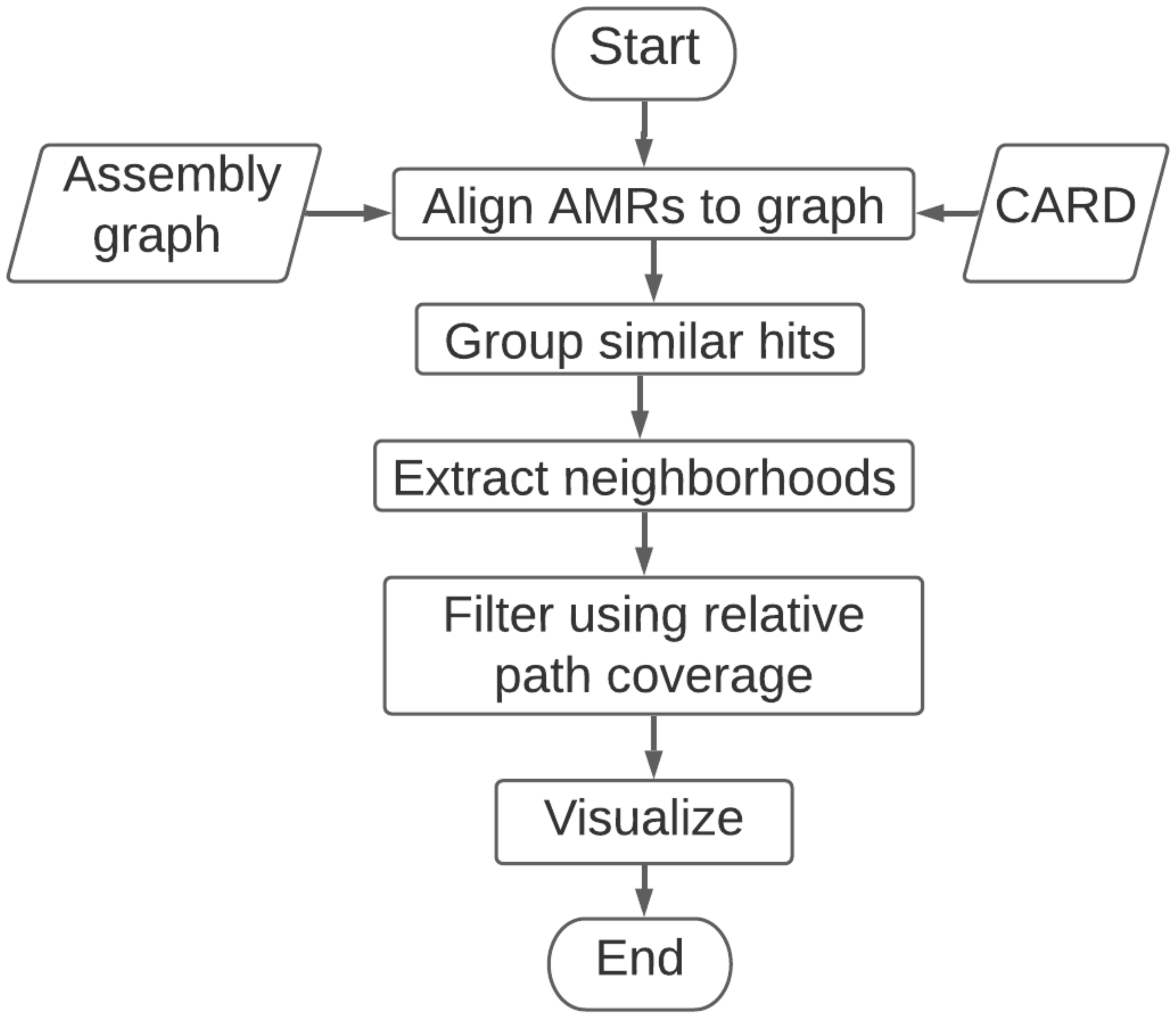


Figure 2 - Flowchart of main steps of Sarand. After AMR alignment, the AMR genes that their sequences are aligned to almost the same location of the graph are grouped. Then, the neighborhood sequences for each group are extracted and stored if no significantly similar sequence already exists. After annotation and storing unique ones, the coverage of each annotated gene is calculated, and only genes that the difference of their coverage from that of the target AMR is less than a given threshold are kept and visualized.

Table 1 - Description of datasets. CAMI[6] refers to Critical Assessment of Metagenome Interpretation project designed for a bias-free evaluation of metagenomics pipelines.

Dataset name	Description	AMR gene count
1_1_1	Includes 3 genomes (*Escherichia coli* SMS-3-5, *Klebsiella pneumoniae* subsp. pneumoniae MGH 78578 and *Staphylococcus aureus* subsp. aureus Mu50*)* with equal abundances	378
2_2_2	Includes 6 genomes (*Escherichia coli* (SMS-3-5 and UMN026), *Klebsiella pneumoniae* subsp. pneumoniae (MGH 78578 and HS11286), *Staphylococcus aureus* subsp. aureus (Mu50 and Mu3)) with equal abundances	451
CAMI_M_1	Includes 132 genomes (32 unique and 100 common strains)	52
CAMI_M_2	Similar to CAMI_M_1	54
CAMI_H_1	Includes 596 genomes (197 unique and 399 common strains)	698

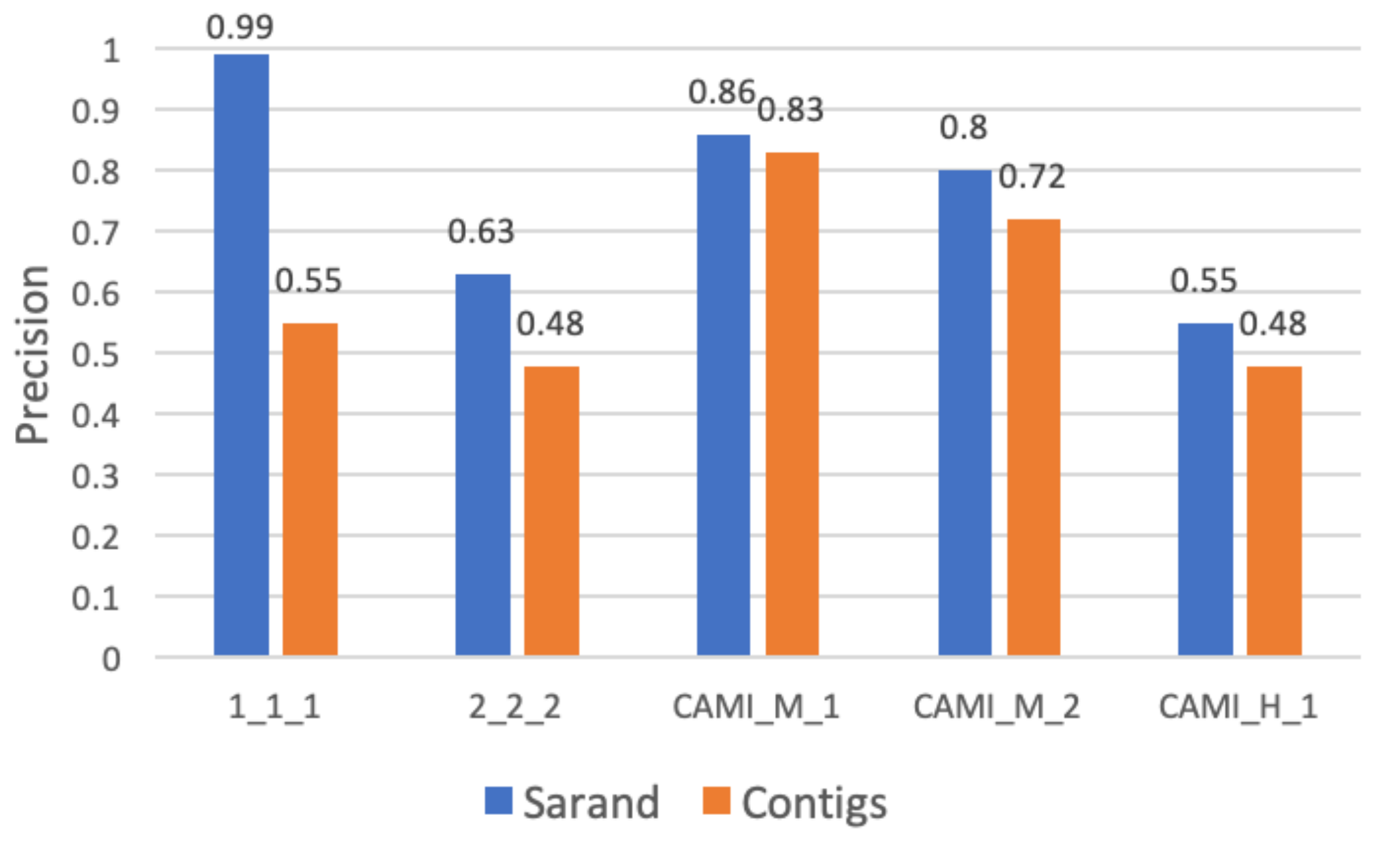


Figure 5 - Comparison of the average precision. Precison measures the fraction of extracted upstream/downstream annotations that are available in the reference genomes.

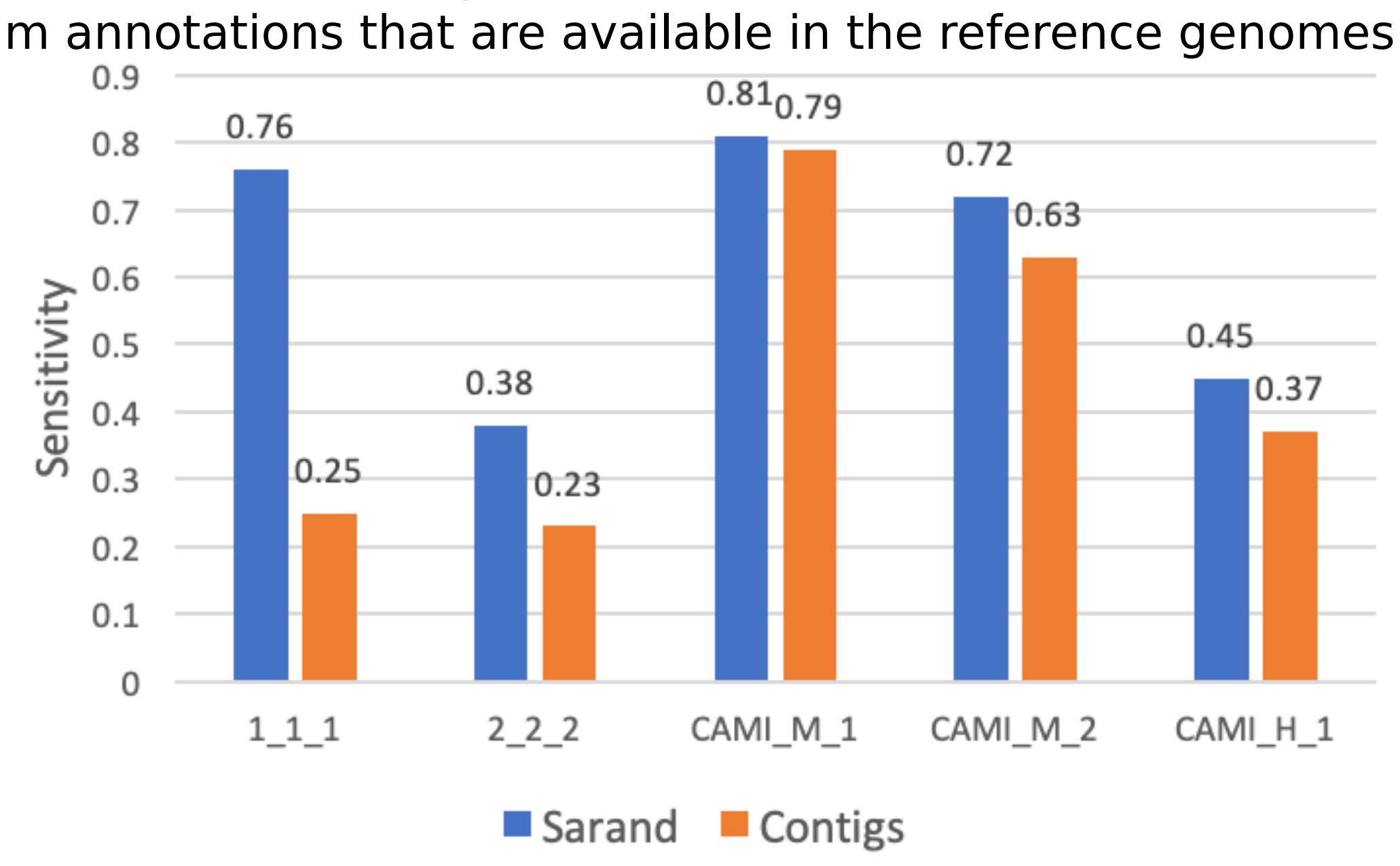


Figure 6 - Comparison of the average sensitivity. Sensitivity measures the fraction of true upstream/downstream annotations that were identified by a method (i.e., Sarand or contigs).

Conclusion and Future Work

Comparing the results of Sarand and contigs across several simulated datasets showed that our method can identify AMR context with higher precision and sensitivity. Sarand is a promising tool to extract the AMR context from metagenomic samples accurately and study the composition of resistomes leading to a better reconstruction of AMR resolution. Since comparison based on annotation and gene names might be misleading in some cases, as the next step, we plan to evaluate our method by comparing the original extracted sequences, rather than their annotation. Currently, we assume that concatenations of all upstream and downstream sequences produce valid sequences which sometimes might not be true. In future, we are interested in defining metrics based on which we can decide that which extracted upstreams and downstreams should be concatenated.

References

1. S. F. Altschul, et al., "Basic local alignment search tool", J. Molecular Biology, 215:403-410, 1990.
2. R. R Wick, et al., "Bandage: interactive visualisation of de novo genome assemblies", Bioinformatics, 31(20): 3350-3352, 2015.
3. B. P Alcock et al., "CARD 2020: Antibiotic Resistome Surveillance with the Comprehensive Antibiotic Resistance Database", Nucleic Acids Research, 48(1): 517-525, 2020.
4. T. Seemann, "Prokka: rapid prokaryotic genome annotation", Bioinformatics, 30(14): 2068–2069, 2014.
5. https://edinburgh-genome-foundry.github.io/DnaFeaturesViewer/
6. A. Sczyrba, et al., "Critical Assessment of Metagenome Interpretation—a benchmark of metagenomics software", Nature methods, 14(11): 1063-1071, 2017.

Acknowledgement - This study was supported by the Canadian Institutes for Health Research, the Donald Hill Family Fellowship, and Genome Canada.