Validation of STEC assignment for *Escherichia coli*

Purpose

Shiga toxin-producing *E. coli* are of critical importance in public health, and determination of STEC strains from genome sequence is considered essential for the PATH-SAFE platform. There are three different tools under consideration:

1. STECFinder - https://github.com/LanLab/STECFinder; Zhang *et al* (2022)
2. AMRFinderPlus - https://github.com/ncbi/amr; Feldgarden *et al* (2021)
3. VirulenceFinder - https://bitbucket.org/genomicepidemiology/ResFinder; Zankari *et al* (2012)

This report aims to provide a comparison of the three tools in order to aid selection of which to include in the PATH-SAFE genome reports.

Evaluation

Data compilation

Complete *E. coli* genomes were downloaded from the NCBI's RefSeq genome assembly database and their species assignment confirmed with the Pathogenwatch speciation tool "Speciator". This resulted in a final set of 4680 genomes. Two attempts were made to assemble datasets with associated Shiga toxin phenotype data, however for unknown technical reasons, these efforts failed. Time constraints meant it was not possible to try other datasets.

Docker images were obtained for each tool and run against two collections of genomes, described below.

Tool versions

Name	Version	Notes
VirulenceFinder	v3.0.0	The Docker image provided by StaPH-B was used: "staphb/virulencefinder:3.0.0". The `-d listeria` option was also set.
STECFinder	v1.1.2	The Docker image provided by StaPH-B was used: "quay.io/biocontainers/stecfinder:1.1.2--pyhdfd78af_0". Default arguments were used.

AMRFinderPlus	4.0.3-2024-10-22.1	The Docker image provided by the NCBI was used: “ncbi/amr:4.0.3-2024-10-22.1” and the “--plus” option used.

Table 1: Versions of software and databases used for this comparison

Comparison datasets

Name	Number of genomes	Description
Complete genomes	4680	Complete genomes downloaded from RefSeq on the 29/7/202

Table 2: Summary of datasets used for this comparison.

Results

Complete genomes

Tool	Genomes with matches	Genes in scheme (alleles)
VirulenceFinder	724	2 (160)
AMRFinderPlus	725	4 (154)
STECFiinder	725	-
NB AMRFinderPlus splits the Shiga-toxin operon into stxA1/2 and stxB1/2 (4 gene families), whereas VirulenceFinder splits it into stx1/2 which appear to correspond to stxA1/2. STECFinder provides a clear “stx type” assignment field, removing any ambiguity.		

Table 3: Summary of number of STEC-encoding genomes found on the “complete genomes” dataset.

All three tools found a similar number of STEC-containing genomes, with STECFinder and AMRFinderPlus exactly agreeing with each other.

Discussion

In terms of the results, there is very little between the tools. The genomics of whether a strain is likely to be Shiga toxin-producing is straightforward, though none of the tools attempts to actually determine that the phenotype is expressed. To examine the concordance with phenotypic data it would be necessary to compare the outputs on a set of genomes with known phenotypes. As discussed above, it was not possible to achieve this. However we can still be assured that all the tools are pretty much equally accurate.

Citations

Zhang, X., Payne, M., Kaur, S., & Lan, R. Improved Genomic Identification, Clustering, and Serotyping of Shiga Toxin-Producing *Escherichia coli* Using Cluster/Serotype-Specific Gene Markers. Frontiers in cellular and infection microbiology, 2022, 11, 772574.

Feldgarden M, Brover V, Gonzalez-Escalona N, Frye JG, Haendiges J, Haft DH, Hoffmann M, Pettengill JB, Prasad AB, Tillman GE, Tyson GH, Klimke W. AMRFinderPlus and the Reference Gene Catalog facilitate examination of the genomic links among antimicrobial resistance, stress response, and virulence. Sci Rep. 2021 Jun 16;11(1):12728. doi: 10.1038/s41598-021-91456-0. PMID: 34135355; PMCID: PMC8208984.

Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, Lund O, Aarestrup FM, Larsen MV. Identification of acquired antimicrobial resistance genes. J Antimicrob Chemother. 2012 Jul 10. PMID: 22782487 doi: 10.1093/jac/dks261

Appendix

AMRFinderPlus

https://github.com/ncbi/amr

VirulenceFinder

https://bitbucket.org/genomicepidemiology/virulencefinder_db/src/master/

STECFinder

https://github.com/LanLab/STECFinder

Appendix B

Google sheet with result details: E. coli validation data

Appendix C

Google drive link to data: Complete genomes 2024-07-29