The Journal of Open Source Software

havic: detect Hepatitis A Virus Infection Clusters from clinical cDNA sequences

Mark B. Schultz*[1], Karolina Mercoulia[1], William Pitchers[1], Patiyan Andersson[1], Marion Easton[3], Joy Gregory[3], Danita Hennessy[3], Linda T. Viberg[1], Michelle Sait[1], Susan A. Ballard[1], Sara Zufan[1], Norelle L Sherry[1], Lilly Yuen[2], Leon Caly[2], Mike G. Catton[2], Julian D.Druce[2], Courtney R. Lane[1], Kristy Horan[1], Torsten Seemann[1], Benjamin P. Howden[1], Anders Gonçalves da Silva[1], and Deborah A. Williamson[1]

1 Microbiological Diagnostic Unit – Public Health Laboratory (MDU-PHL) **2** Victorian Infectious Diseases Reference Laboratory (VIDRL) **3** Department of Health and Human Services (DHHS), Victorian Government, Australia

DOI: 10.21105/joss.0XXXX

Software

- Review
- Repository
- Archive

Editor: Editor Name

Submitted: 01 January XXXX
Published: 01 January XXXX

Introduction

Globally, Hepatitis A Virus (HAV) infects tens of millions of people each year (Blum, 2016). Though mortality is low, morbidity is high resulting in large economic burden due to frequent hospitalisation of cases during the acute phase of infection. Transmission of HAV occurs via the faecal-oral pathway, either directly from person-to-person or indirectly through contaminated food and water (FitzSimons et al., 2010). HAV has its highest prevalence in low-income countries; however, sporadic outbreaks do occur in high-income countries (Blum, 2016), typically arriving via return travellers and import on frozen foods (e.g., see Hu et al., 2020; Shouval, 2020). Lifelong immunity to the virus arises after vaccination or infection (Kroneman et al., 2018).

Molecular epidemiology using DNA sequencing and comparative genomics to characterise virus outbreaks in real-time is now considered an essential public health measure (Hu et al., 2020; Probert et al., 2019; Seemann et al., 2020). For HAV, whole genome sequencing is not yet the normal practice for outbreak surveillance. Instead, the gold-standard approach [i.e., the 'HAVNet' protocol] involves the sequencing of a 460 bp cDNA amplicon spanning the VP1/P2A junction. After sequencing, the amplicon is compared to global databases to make inferences of genotype and to recover putative epidemiological links (HAVNet, 2018). As tiled amplicon approaches to viral sequencing are becoming more common (e.g., Quick et al., 2017), there is and will be a need for software tools that can handle amplicon, whole genome and/or partial genome sequences.

Diversity of HAV has been well characterised and modern nomenclature assigns HAV to one serotype with six genotypes (I-VI). Of the six, three genotypes (i.e., I, II and III) infect humans, and three (i.e., IV, V and VI) infect non-human primates. Phylogenetic analysis conspicuously divides genotypes I, II and III into six subtypes (IA, IB, IIA, IIB, IIIA and IIIB) (Kroneman et al., 2018; Smith & Simmonds, 2018). Pairwise nucleotide divergences between genotypes I to III range from 12.2% to 21.9% (mean 18.2%). Diversity *within* genotypes I to III ranges from 0.3% to 6.5% (mean 4.3%). Between subtypes IA and IB mean nucleotide divergence is estimated at 9.3%; between IIA and IIB divergence is approximately 9.6%; and between IIIA and IIIB divergence is around 11.8% (Smith & Simmonds, 2018).

*Corresponding author

Though the HAVNet protocol is widely adopted, variations to the method are commonplace (e.g., see Probert et al., 2019; Probert & Hacker, 2019). Public databases (e.g., HAVNet and NCBI) are compiled over many years from myriad laboratories (e.g., see Severi et al., 2015). Artefactual nucleotide variations (e.g., low quality, false indels, incompatible orientation) are present. And HAV consensus sequences do not always co-locate within a single genome target, with databases comprised of sequences from whole genome or partial genome sequences from varying regions (refer to Figure 1 in Kroneman et al., 2018). Distance based pairwise nucleotide comparisons used to infer relatedness of samples are naturally sensitive to these sources of variation. In this article, we describe our software tool `havic`, which is written to ease the burden of detecting and characterising HAV outbreak clusters during routine epidemiolgical surveillance in the midst of these challenges.

Statement of utility

`havic` has been developed over a number of years with feedback and feature requests from public health epidemiolgists during routine use of the program for HAV outbreak surveillance. The software aims to provide repeatable analyses that give actionable and objective results, despite inherent imperfections in HAV sequencing data. As the HAV genome comprises a single segment, being a positive-sense single-stranded ribonucleic acid (RNA) of only 7.5 kilobases (kb) (Cohen et al., 1987), `havic` runs can be completed relatively quickly on a standard desktop computer. For larger jobs, the software pipeline has been tested on an HPC system with queing managed by SLURM.

The pipeline is written in python3, implementing a number of R packages, and pipeline control within `havic` is managed using the Ruffus library. Run configurations are defined using a yaml-formatted text file. Installation is performed using `conda`.

The workflow

Briefly, outlining the salient steps in our pipeline, `havic`:

- reads DNA sequences in fasta format from one or many files (BioPython)
- finds and removes duplicates (based on fasta header)
- maps the sequences to a reference genome (reverse complementing the input sequences as required) using `minimap2` (Heng Li, 2017)
- converts the sequence alignment map (SAM) output from `minimap2` to a binary alignment map (BAM) using `samtools` (H. Li et al., 2009) (excluding secondary mappings and unmapped reads)
- extracts the aligned reads from the BAM as a multiple sequence alignment (MSA) using RSamtools ([[Computer Program]], 2017)
- trims sequences in the MSA to a user-defined target region (BioPython)
- infers a nucleotide substitution model from the MSA using ModelFinder (Kalyaanamoorthy et al., 2017)

Under the best-fit substitution model, `havic`:

- runs IQtree2 (Nguyen et al., 2015) to infer the most likely tree using Maximum Likelihood (ML)

- infers support values for branches in the tree using the the ultra-fast bootstrap method (Hoang et al., 2018)

After obtainining the tree, `havic`:

- run `ClusterPicker` (Ragonnet-Cronin et al., 2013) to delineate putative infection clusters within the input sample set
- reports infection clusters in textual, tree and graphical format files

Testing and validation

Example data are pre-packaged with `havic`. Tests are performed by running `havic test <test_suite>`, where `<test_suite>` is any of `hav_amplicon`, `hav_wgs`. Two additional test suites – `measles_wgs` and `hiv_amplicon` – are included for exploration of edge-cases and for future work that will aim to allow iterations of the software to be used for outbreak surveillance in other viruses. Considering the error rate of Q30 Sanger sequences is 1/1000 bases or 0.001, and based on advice from epidemiolgists working in this domain (who have independently validated `havic`-detected clusters using contact tracing metadata over a number of years and multi-jurisdictional outbreaks), we have settled on an optimal maximum within infection cluster divergence of 0.01 for our test analyses. Using the `UFBoot` method in `IQ-Tree2`, we performed our tests using branch supports of greater than or equal to 95%. The software was developed using a test driven ethos, with python unittests forming the basis of all tests. As a `havic` run makes inferences from a single best tree inferred under ML, we strongly recommend running the pipeline multiple times to avoid reliance on a sub-optimal tree.

Visualisation tools

The results of havic are output to a single folder with the option to summarise the results as images. Pairwise nucleotide differences may be summarised as a heatmap Figure 1.

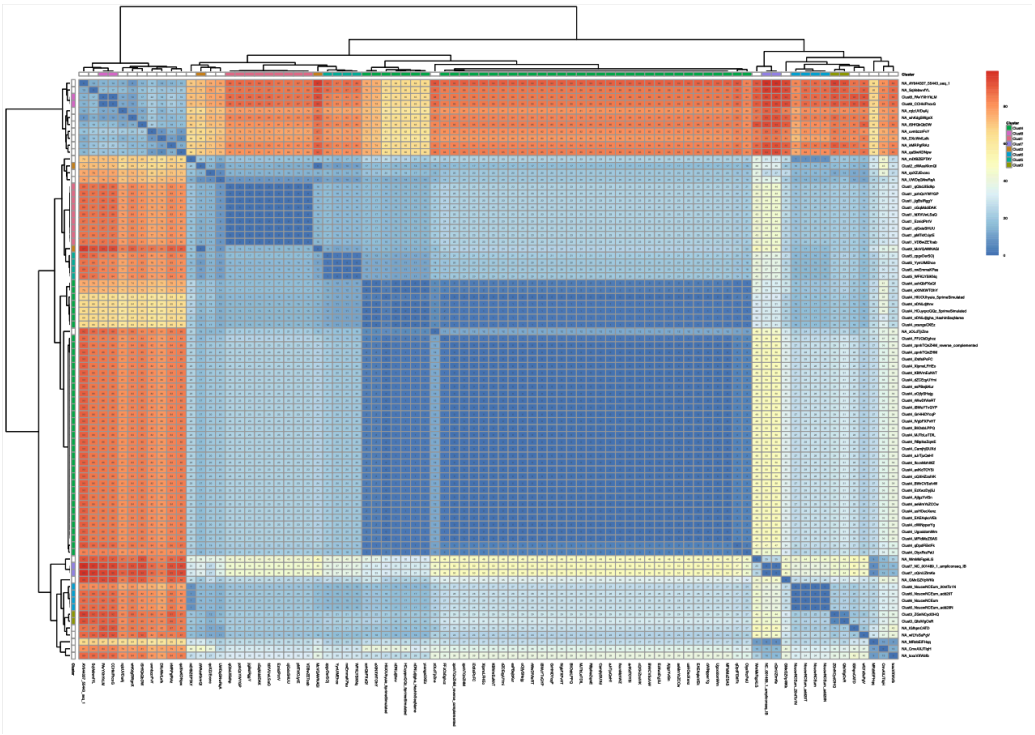

Figure 1: Pairwise genetic distances and ClusterPicker clusters.

And the alignment may be plotted next to the phylogenetic tree with tree tips coloured by infection cluster Figure 2.

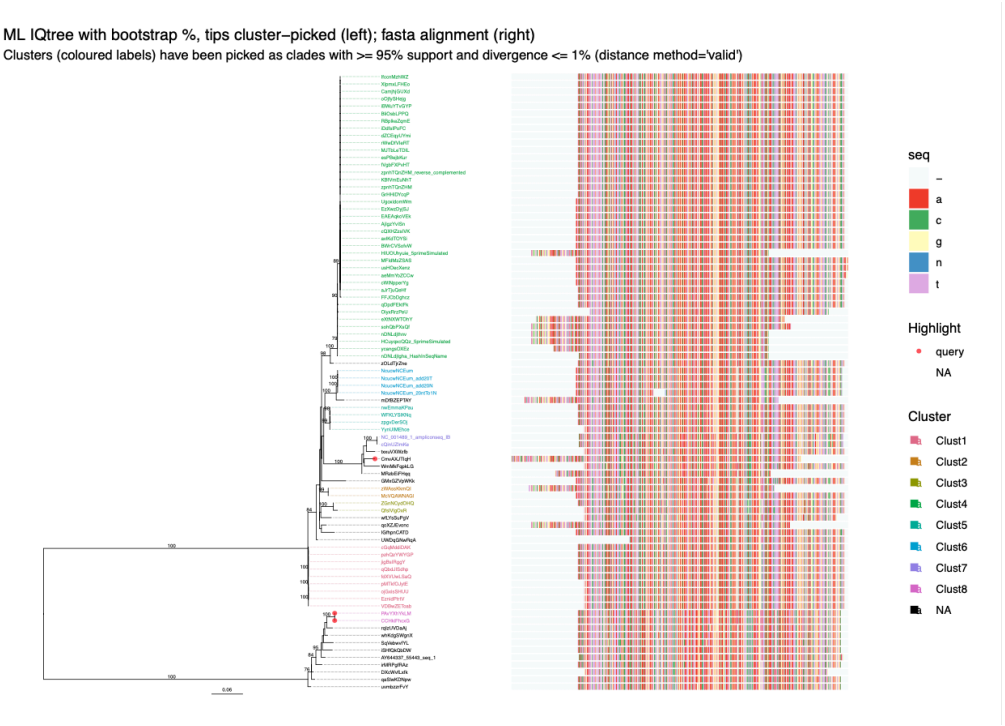

Figure 2: Maximum Likelihood phylogenetic tree with ufboot branch support. The underlying multiple sequence alignment is plotted next to the tree and tips are coloured to highlight the detected infection clusters.

Query isolates may be annotated in the final tree plot using a red dot Figure 3.

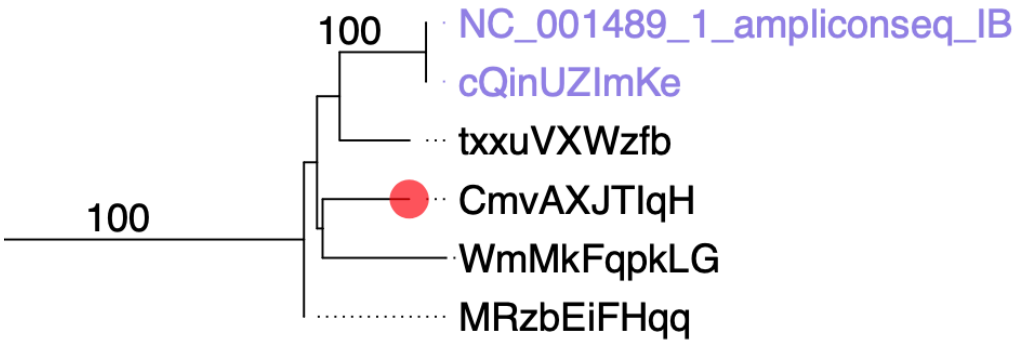

Figure 3: Query sequence highlighted using a red dot.

Acknowledgements

We would like to thank members of the laboratories at Queensland Health Forensic and Scientific Services (QHFSS), DHHS, VIDRL and MDU-PHL.

References

Blum, H. E. (2016). History and Global Burden of Viral Hepatitis [Journal Article]. *Dig Dis*, *34*(4), 293–302. https://doi.org/10.1159/000444466

Cohen, J. I., Ticehurst, J. R., Purcell, R. H., Buckler-White, A., & Baroudy, B. M. (1987). Complete nucleotide sequence of wild-type hepatitis A virus: comparison with different strains of hepatitis A virus and other picornaviruses [Journal Article]. *J Virol*, *61*(1), 50–59. https://www.ncbi.nlm.nih.gov/pubmed/3023706

[Computer Program]. (2017). http://bioconductor.org/packages/release/bioc/html/Rsamtools.html

FitzSimons, D., Hendrickx, G., Vorsters, A., & Van Damme, P. (2010). Hepatitis A and E: update on prevention and epidemiology [Journal Article]. *Vaccine*, *28*(3), 583–588. https://doi.org/10.1016/j.vaccine.2009.10.136

HAVNet. (2018). *Molecular detection and typing of VP1 region of Hepatitis A Virus (HAV)* [Report]. Dutch National Institute of Health; Environment (RIVM).

Hoang, D. T., Chernomor, O., Haeseler, A. von, Minh, B. Q., & Vinh, L. S. (2018). UFBoot2: Improving the Ultrafast Bootstrap Approximation [Journal Article]. *Mol Biol Evol*, *35*(2), 518–522. https://doi.org/10.1093/molbev/msx281

Hu, X., Collier, M. G., & Xu, F. (2020). Hepatitis A Outbreaks in Developed Countries: Detection, Control, and Prevention [Journal Article]. *Foodborne Pathog Dis*, *17*(3), 166–171. https://doi.org/10.1089/fpd.2019.2648

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., Haeseler, A. von, & Jermiin, L. S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates [Journal Article]. *Nat Methods*, *14*(6), 587–589. https://doi.org/10.1038/nmeth.4285

Kroneman, A., Sousa, R. de, Verhoef, L., Koopmans, M. P. G., Vennema, H., & On Behalf Of The, H. N. (2018). Usability of the international HAVNet hepatitis A virus database for geographical annotation, backtracing and outbreak detection [Journal Article]. *Euro Surveill*, *23*(37). https://doi.org/10.2807/1560-7917.ES.2018.23.37.1700802

Li, Heng. (2017). Minimap2: fast pairwise alignment for long nucleotide sequences [Journal Article]. *arXiv:1708.01492*.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools [Journal Article]. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Nguyen, L. T., Schmidt, H. A., Haeseler, A. von, & Minh, B. Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies [Journal Article]. *Mol Biol Evol*, *32*(1), 268–274. https://doi.org/10.1093/molbev/msu300

Probert, W. S., Gonzalez, C., Espinosa, A., & Hacker, J. K. (2019). Molecular Genotyping of Hepatitis A Virus, California, USA, 2017-2018 [Journal Article]. *Emerg Infect Dis*, *25*(8), 1594–1596. https://doi.org/10.3201/eid2508.181489

Probert, W. S., & Hacker, J. K. (2019). New Subgenotyping and Consensus Real-Time Reverse Transcription-PCR Assays for Hepatitis A Outbreak Surveillance [Journal Article]. *J Clin Microbiol*, *57*(9). https://doi.org/10.1128/JCM.00500-19

Quick, J., Grubaugh, N. D., Pullan, S. T., Claro, I. M., Smith, A. D., Gangavarapu, K., Oliveira, G., Robles-Sikisaka, R., Rogers, T. F., Beutler, N. A., Burton, D. R., Lewis-Ximenez, L. L., Jesus, J. G. de, Giovanetti, M., Hill, S. C., Black, A., Bedford, T., Carroll, M. W., Nunes, M., … Loman, N. J. (2017). Multiplex PCR method for MinION and

Illumina sequencing of Zika and other virus genomes directly from clinical samples [Journal Article]. *Nat Protoc*, *12*(6), 1261–1276. https://doi.org/10.1038/nprot.2017.066

Ragonnet-Cronin, M., Hodcroft, E., Hue, S., Fearnhill, E., Delpech, V., Brown, A. J., Lycett, S., & Database, U. H. D. R. (2013). Automated analysis of phylogenetic clusters [Journal Article]. *BMC Bioinformatics*, *14*, 317. https://doi.org/10.1186/1471-2105-14-317

Seemann, T., Lane, C. R., Sherry, N. L., Duchene, S., Goncalves da Silva, A., Caly, L., Sait, M., Ballard, S. A., Horan, K., Schultz, M. B., Hoang, T., Easton, M., Dougall, S., Stinear, T. P., Druce, J., Catton, M., Sutton, B., Diemen, A. van, Alpren, C., … Howden, B. P. (2020). Tracking the COVID-19 pandemic in Australia using genomics [Journal Article]. *Nat Commun*, *11*(1), 4376. https://doi.org/10.1038/s41467-020-18314-x

Severi, E., Verhoef, L., Thornton, L., Guzman-Herrador, B. R., Faber, M., Sundqvist, L., Rimhanen-Finne, R., Roque-Afonso, A. M., Ngui, S. L., Allerberger, F., Baumann-Popczyk, A., Muller, L., Parmakova, K., Alfonsi, V., Tavoschi, L., Vennema, H., Fitzgerald, M., Myrmel, M., Gertler, M., … Rizzo, C. (2015). Large and prolonged food-borne multistate hepatitis A outbreak in Europe associated with consumption of frozen berries, 2013 to 2014 [Journal Article]. *Euro Surveill*, *20*(29), 21192. https://doi.org/10.2807/1560-7917.es2015.20.29.21192

Shouval, D. (2020). The History of Hepatitis A [Journal Article]. *Clin Liver Dis (Hoboken)*, *16*(Suppl 1), 12–23. https://doi.org/10.1002/cld.1018

Smith, D. B., & Simmonds, P. (2018). Classification and Genomic Diversity of Enterically Transmitted Hepatitis Viruses [Journal Article]. *Cold Spring Harb Perspect Med*, *8*(9). https://doi.org/10.1101/cshperspect.a031880