Deconvoluting simulated HPV coinfections using long-read DNA sequencing

Eric T. Dawson[2,3], Erik Garrison[2], Dave Roberson[1], Stephen Chanock[3], Richard Durbin[2] and Sarah Wagner[1]

[1]Cancer Genomics Research Laboratory, Leidos Biomedical Research Inc., Frederick National Laboratory for Cancer Research, Frederick, MD, USA; [2] Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge UK; [3] Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, MD, USA.

Introduction

Human Papilloma Virus (HPV) is a DNA virus with an 8kb genome. Isolates are classified into types (separated by >10% divergence) and subtypes (separated by 2-10% divergence)[1]. Infection with certain types of the virus can lead to cervical cancer, with 78% of all cervical cancer cases in Europe attributable to just two of six carcinogenic types[2]. Infection with multiple types and subtypes of HPV is common[3] and current tests are limited in the number of types/subtypes for which they can test. We describe a novel method for determining the composition of a simulated coinfection. We demonstrate that our approach is accurate on data generated in-silico and on reads produced by the ONT minION sequencer. Finally, we describe our plans for future development.

Method

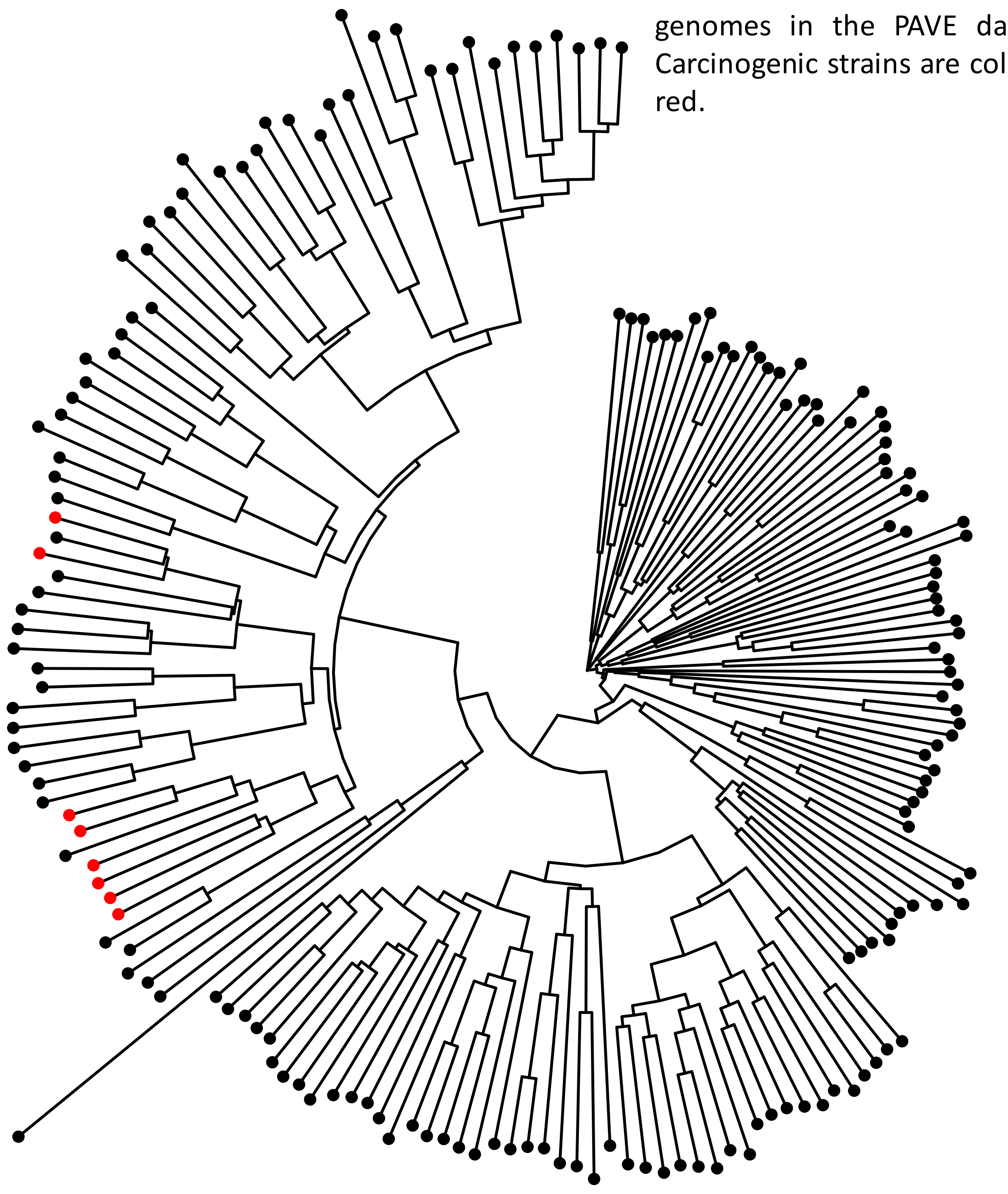
Isolates of HPV16 sublineages 1207 and 1509 were PCR amplified, mixed in a 40/60 mixture of their respective amplicons, and sequenced on the minION using a standard 2D sequencing protocol.

All reference and variant HPV genomes in the PAVE database[4] were downloaded from GenBank. Genomes that were within a Jaccard distance of .10 (as calculated by Mash[5]) were then aligned to each other using vg's msga command[6]. All sequences were made into graphs, circularized and concatenated to form a pangenome graph.

Reads approximating the error profile and length of the minION reads were simulated from the HPV genomes acquired from PAVE. These were mapped to our pangenome structure and the resulting vectors were labeled with their strain of origin. These labeled vectors were used to train an error-correcting tournament model implemented in vowpal-wabbit[7].

We compared our method to Mash, which uses the MinHash algorithm to find common subsequences in samples. We wrote a simple wrapper around Mash to classify reads individually and collate the results.

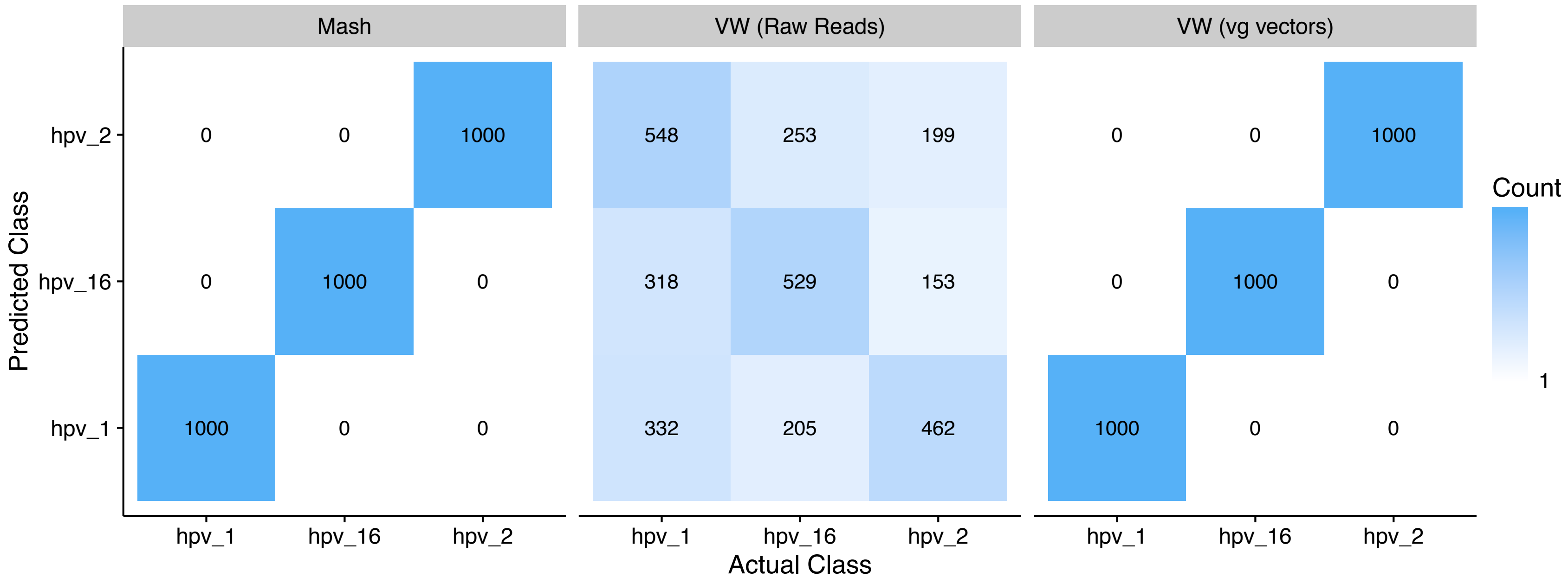
MinION reads from our sequenced coinfection were mapped and vectorized using vg. The read vectors were then fed to our trained model or our modified Mash code for classification. The output was visualized using R.



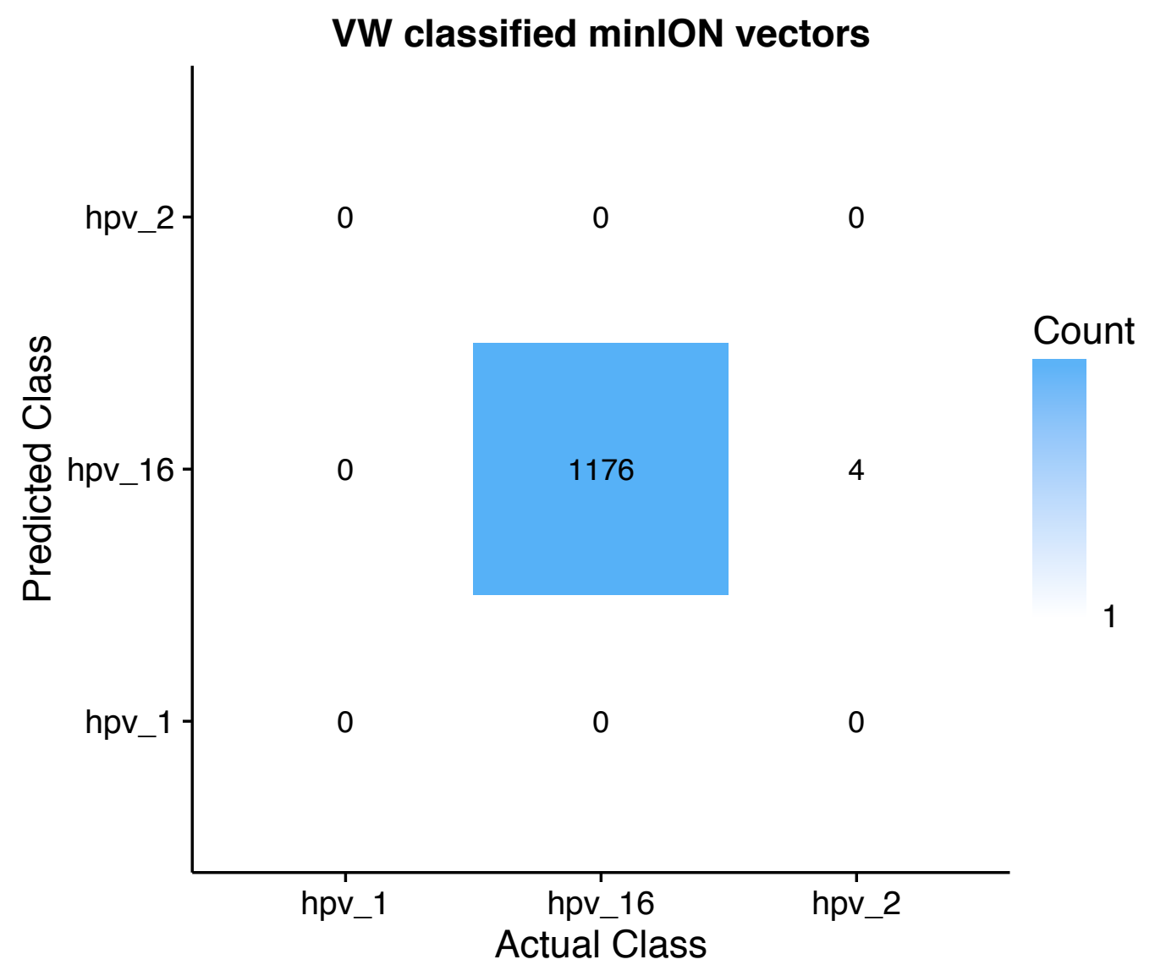
Circular tree of all HPV reference genomes in the PAVE database. Carcinogenic strains are colored in red.

Results

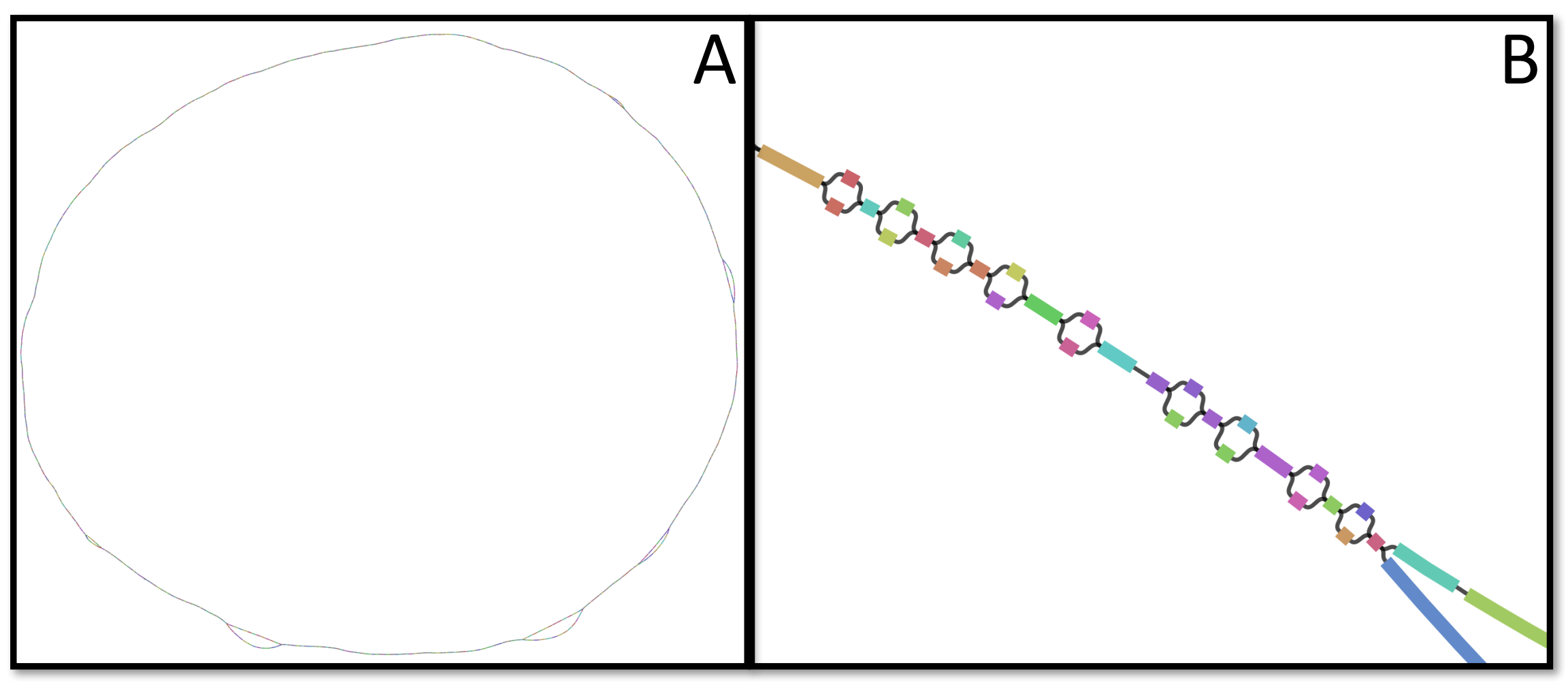
Testing on 3000 7kb simulated reads from three different reference genomes with error rates of 5% and indel rates of 10% showed that our method performs comparably to a modified version of the MinHash algorithm. Both methods deterministically classify reads with 100% accuracy. In comparison, a machine learning model trained on the kmers of reads (without mapping and vectorization) misclassifies the majority of reads.



Our model is also able to accurately classify 2D high-quality minION reads into the correct subtype of origin with only 0.3% error.



Confusion matrix for 1179 minION reads classified using our model trained on simulated reads from three HPV types.

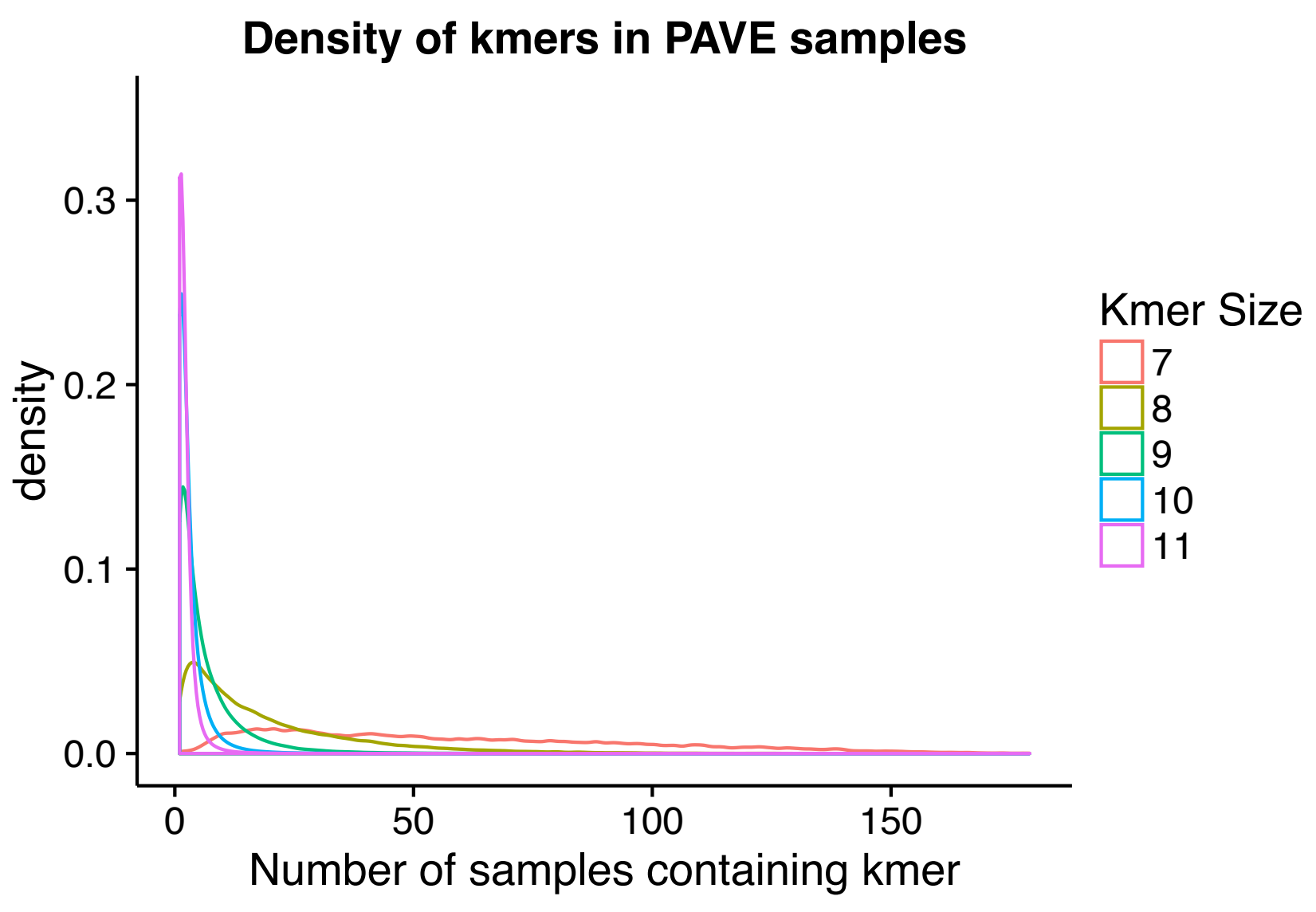


Visualizations of a pangenome graph containing multiple variant genomes aligned to their respective reference genome (A) and a detailed view the structure of the graph with bubbles indicating variation (B).

We next used our approach to generate a mappable HPV pangenome. We then mapped reads and generated vectors on a quad-core desktop. Timings and resulting data sizes for the process are shown to the right.

	Time	Size
Generate a pangenome graph	3m56s	~16MB
Read Mapping (3000 reads)	2m50s	~17MB
Vectorization (3000 reads)	2m23s	~3.3GB

Our model can classify reads roughly as fast as the minION can produce them. Mash can classify reads much faster, approaching several thousand reads a second. However, many of the Mash classifications are based on only a handful of kmer matches, often less than 1% of the sketch size. From the densities of kmers at various sizes, we can see that uniqueness increases rapidly with kmer size. This makes classifying a read difficult, as only the information at the tips of the phylogeny is preserved. We have been investigating a method to use multiple kmer sizes in the MinHash sketch as a way of better preserving phylogenetic information.



Density of kmer sharing for various kmer sizes.

Conclusion and Future Work

Our work demonstrates that it is possible to disentangle complex mixtures of viral sequences from the minION. We demonstrate that the performance of our approach is similar to that of Mash, the current state of the art. Both methods have advantages and drawbacks. Our models are reusable (i.e. need only be trained once) and can be adapted to a variety of input datasets. In addition, the same mappings used for vectorization can be used to call variant sites, and generating them takes almost no extra compute when mapping. Our extended version of Mash, on the other hand, is significantly faster for the same accuracy but loses phylogenetic information since it can only process one kmer size at a time. It is also highly influenced by sequencing error. We plan to continue development of both models with the hope that we can improve the speed and accuracy of our classifier. W hope to extend our approach to other viruses and minION datasets as data becomes available. After implementing multi-kmer MinHash in vg we will incorporate it as a preclustering step. We will also investigate its power in classifying reads without mapping, using the sketches as input to vowpal-wabbit to correct misclassifications. We will also develop a more space-efficient vector format as the current format is sparse and grows linearly with the size of the graph.

Citations

1. Zheng, Z.M. and Baker, C.C. Papillomavirus genome structure, expression, and post-transcriptional regulation. Frontiers in Bioscience 11 2286-2301 (2006).
2. Tjalma W.A., Fiander, A., Reich, O., Powell, N., Nowakowski, A.M., Kirschner, B., *et al*. Differences in human papillomavirus type distribution in high-grade cervical neoplasia and invasive cervical cancer in Europe. Int J Cancer. 2013;132:854-67
3. Chaturvedi, A.K., Katki, H.A., Hildesheim, A., *et al*. Human papillomavirus infection with multiple types: pattern of coinfecton and risk of cervical disease. Journal of Infectious Disease 203(7) 910-920 (2011).
4. Van Doorslaer, K., Tan, Q., Xirasagar, S., Bandaru, S, Gopalan, V., Mohamoud, Y., Huyen, Y., and McBride, A. A. The Papillomavirus Episteme: a central resource for papillomavirus sequence data and analysis. Nucleic Acids Research 41(D1):D571-578. pave.niaid.nih.gov. 2013.
5. Ondov, B.D. *et al*. Fast genome and metagenome distance estimation using MinHash. bioArxiv 029827 (2015).
6. Garrison, E.K. *et al*. vg: tools for manipulating variation graphs. Available at github.com/vgteam/vg. Unpublished.
7. Agarwal, A., Chapelle, O., Dudik, M., and Langford, J. A reliable effective terascale learning system. arXiv:1110.4198v3 (2011, revised 2013)

wellcome trust sanger institute
NIH
NIH Oxford-Cambridge Scholars Program