Fast HPV strain identification and variant calling using MinHash and the Oxford Nanopore minION sequencer.

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

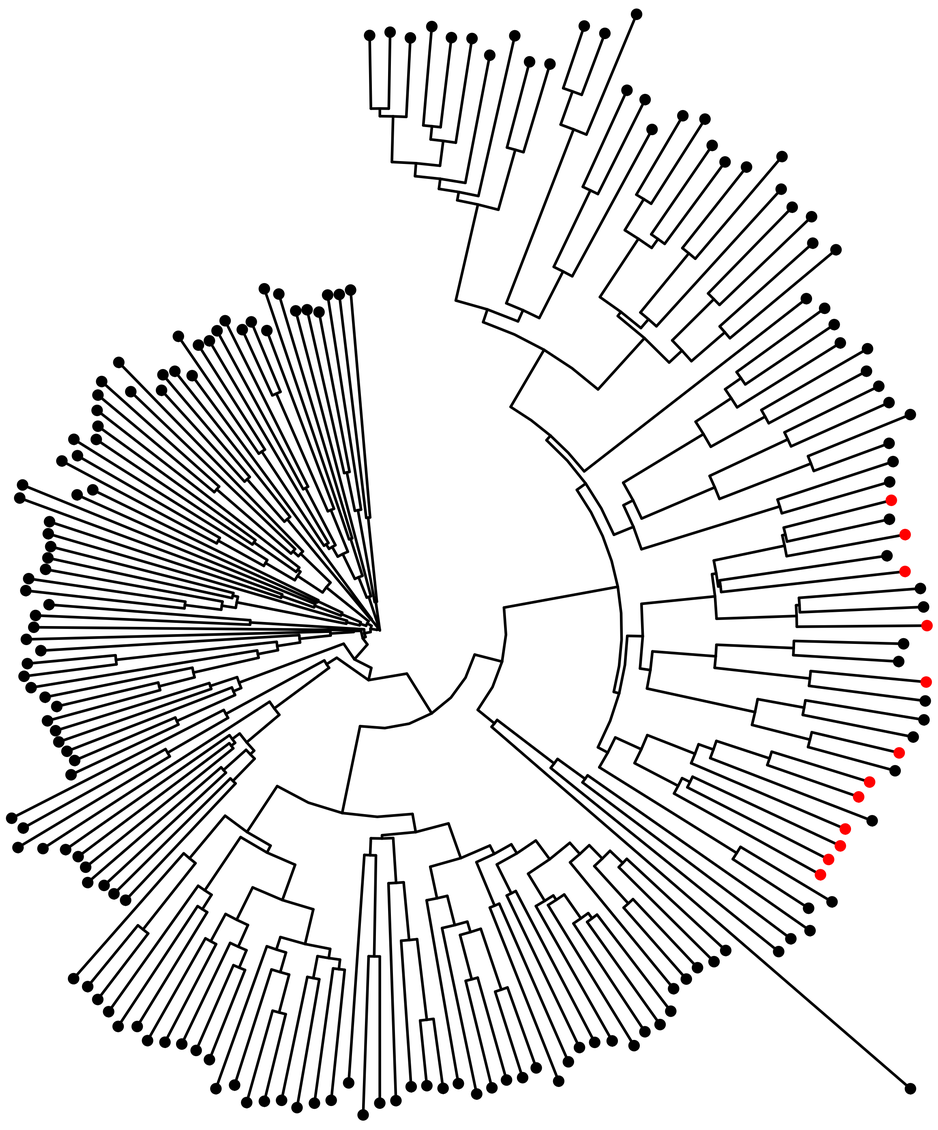
Human Papilloma Virus (HPV) is a DNA virus with an 8kb genome. Isolates are classified into types (separated by >10% divergence) and subtypes (separated by 2-10% divergence)[1]. Infection with certain types of the virus can lead to cervical cancer, with 78% of all cervical cancer cases in Europe attributable 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.

MinION reads were converted to fastq format using `poretools`. We then applied our new method, `rkmh`. `rkmh` generates a MinHash sketch as defined by Mash[4] for the reference sequences and reads provided. It then compares each read signature to the set of references and reports the reference with which a read shares the largest intersection. There are filters for kmer depth in the reads, kmer abundance in references, and minimum sketch intersection size, which operate atomically on the input sequences. We compared each read in our sequenced set to all HPV reference genomes in the PAVE[5] database using a sketch size of 2000 and a minimum sketch size filter of 10.

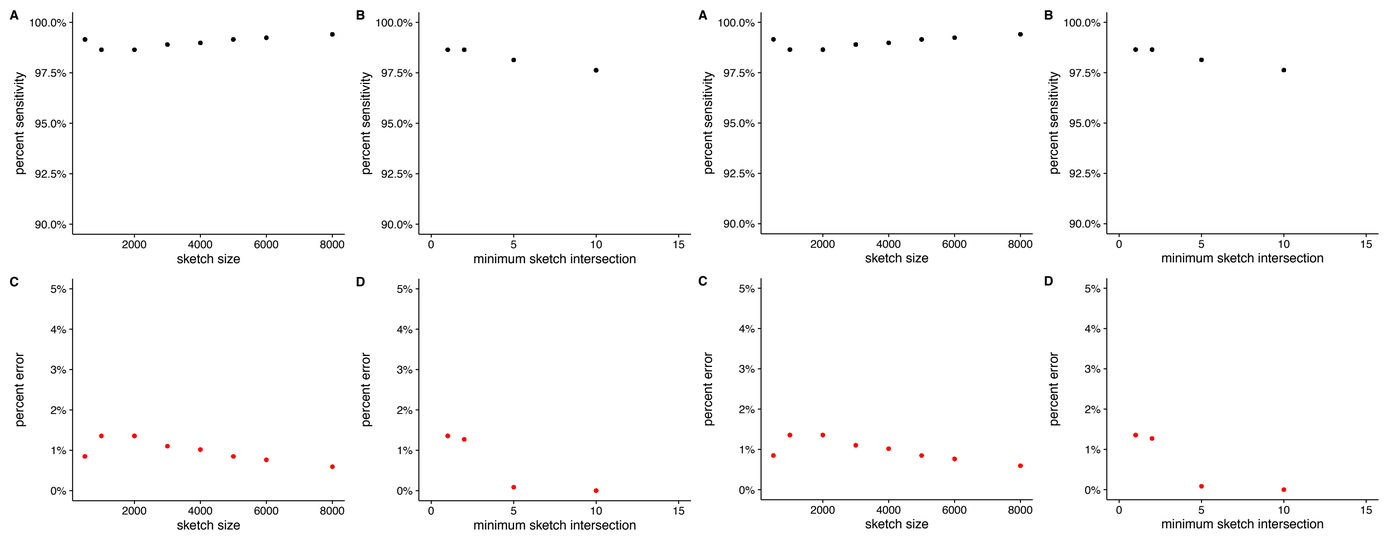
We then simulated 1000 reads from each of 100 reference genomes in the PAVE database using `wgsim`. We classified each read using `rkmh`. We then chose one of these simulated datasets to test the ability of `rkmh` to detect known mutations in the reads relative to the reference genome. We also called mutations in our readset and compared the results to calls from a BLASR-based analysis pipeline.



Circular tree of all HPV reference genomes in the PAVE database. Strains considered carcinogenic are colored red.

Results

`rkmh` identifies the strain of origin of individual reads in our read set with >98% sensitivity and >95% sensitivity. Classifications for 2D high quality reads are 99% accurate and >98% sensitive. `rkmh` can process over 200 7kb reads per second for sustained throughput above 1,500,000bp per second.

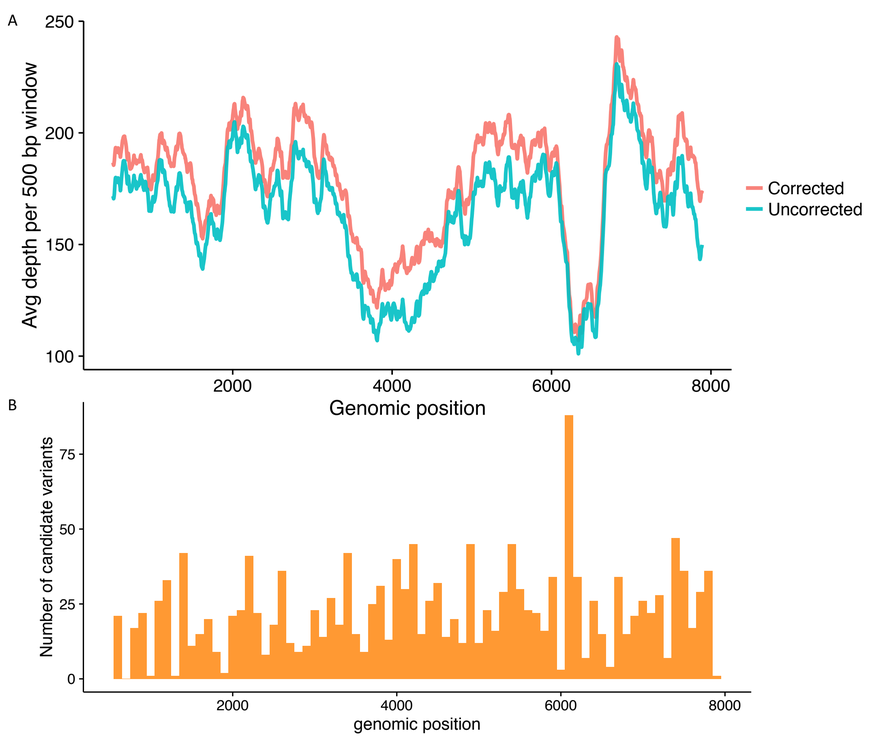


Sensity (A, B) and percent error (C, D) curves for 3,660 minION reads from two known HPV16 subtypes (left). The same plots are reproduced for high quality 2D reads in the same set (right). `rkmh` consistently achieves >98% accuracy and sensitivity, even when sequencing error is above 10%.

`rkmh` also supports alignment-free calling of single base substitutions, insertion and deletions. Kmers from the reference genome at below-average abundance in the reads are modified at each base, and if the resulting abundance is at least 40% higher than the reference kmer a call is reported.

Using this strategy, `rkmh` calls 100% of non-adjacent SNPs in simulated data and more than 80% in real data. Accuracy for INDELs is significantly worse (roughly 20%), with overall accuracy between 60-75% in real and simulated data. Variant calling takes approximately ten seconds for 1000x whole-genome HPV data.

`rkmh` allows exporting hashes in JSON using the schema defined by Ondov *et al*. This is useful for debugging as well as data interchange, although no protocols taking advantage of this feature exist yet.



Sliding window average of kmer depth in reads longer than 6500bp from the mixed read set over the HPV16 reference genome. Average kmer depth varies significantly over the reference genome. Areas of low depth may be due to sequencing error, variation, or the inability to sequence the region.

The number of candidate variant sites in reads longer than 6500bp in our minION run plotted over genomic position in the HPV16 reference genome. The calling process seems susceptible to large losses in precision in regions where average depth drops precipitously. We are continuing to explore different parameters to help minimize this effect.

Conclusions and Future Work

`rkmh` can identify the HPV strain of origin for sequences produced by the ONT minION with high accuracy and sensitivity. It is also roughly ten times faster than alignment-based methods. The program can call single base substitutions, deletions and insertions relative to a reference genome, but INDEL calling performance lags behind standard methods. We plan to expand the calling capabilities to include homopolymer runs as well as to improve INDEL sensitivity. `rkmh` is open-source under the MIT license and is available at `https://github.com/edawson/rkmh`.

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. Ondov, B.D. *et al*. Fast genome and metagenome distance estimation using MinHash. bioArxiv 029827 (2015).
5. 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.

NIH Oxford-Cambridge Scholars Program

wellcome trust sanger institute