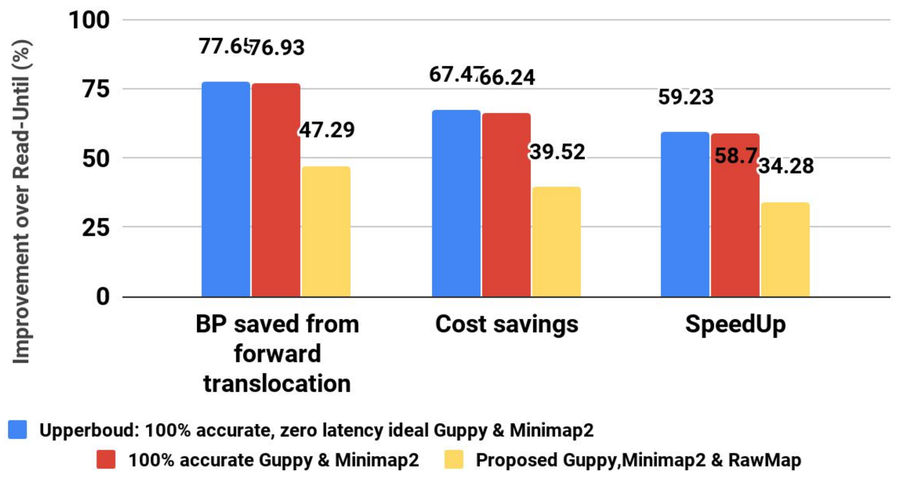
Rapid Real time Squiggle-level Classification for Read-Until Using RawMap

Harisankar Sadasivan[1], Jack Wadden[1], Satish Narayanasamy[1], David Blaauw[1], Reetuparna Das[1], Robert Dickson[2]
[1]Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, 48109, USA
[2]Department of Internal Medicine, University of Michigan Medical School, Ann Arbor, 48109, USA

Introduction

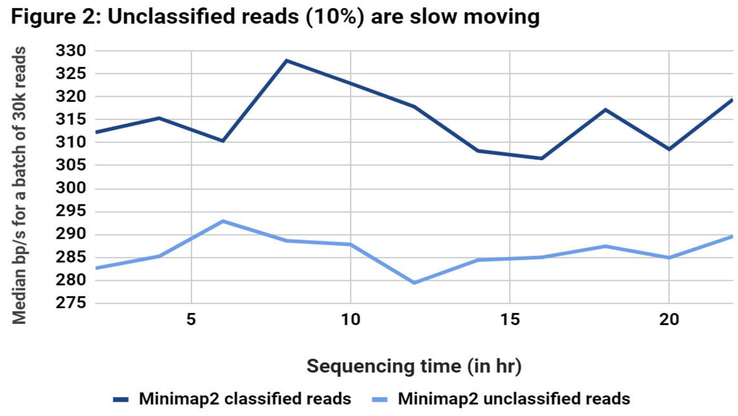
Read-Until enables Oxford Nanopore (ONT) sequencers to selectively sequence target reads of interest in real-time. Read-Until achieves rapid target enrichment for applications such as microbiome abundance estimation where the metagenomic sample has a significant fraction of non-target reads (>99% can be human reads). However, Read-Until requires a fast and accurate classifier that analyzes a short prefix of a read and determines whether the read belongs to one of the target species. The conventional read-until pipeline using a sequence of basecaller (e.g. Guppy), aligner (e.g, Minimap2), and classifier (e.g, Centrifuge) cannot classify 10% of the reads. In figure 1, we show that our proposed pipeline yields significant savings in terms of basepairs saved from forward translocation, cost and speedup on a 99:1 human-zymo DNA mix where the average read length is 20Kbp. Here, cost savings refers to increased lifetime of the flowcell and speedup refers to improvement in end-to-end sequencing time.

Figure 1: Read-Until savings can be improved



The underlying problem with the unaligned 10% of the reads using Minimap2/BLAST is traced back to their low translocation rates as shown in figure 2. This problem necessitates a squiggle-domain classifier for further reducing the non-target reads using Read-Until.

Figure 2: Unclassified reads (10%) are slow moving



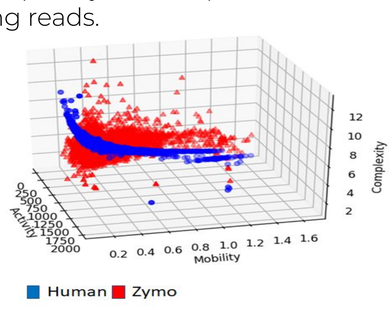
Methodology

We present RawMap, a direct squiggle-space metagenomic classifier for filtering non-target reads. Inspired by brain EEG characterization, RawMap uses a Support Vector Machine (SVM) with an RBF kernel, which is trained to capture the non-linear and non-stationary characteristics of the nanopore squiggles. Each normalized squiggle segment y corresponding to 450 basepairs of a read is mapped to a 3-D feature space. Features are derived from a modified version of Hjorth parameters, where the mean and standard deviation are replaced with median and median absolute deviation respectively. Activity captures the signal power, mobility is the mean frequency and complexity is the change in frequency. RawMap has two SVM models, one trained on fast and other on slow moving reads.

$$Activity = var(y)$$

$$Mobility = \sqrt{\frac{var(y')}{var(y)}}$$

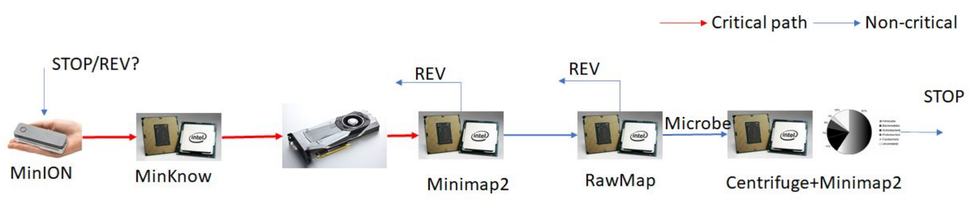
$$Complexity = \frac{mobility(y')}{mobility(y)}$$



Graphical representation of Zymo (target) separation

where y is the normalized raw data segment corresponding to 450basepairs, y' is the first-order difference of the signal and var is the modified variance.

RawMap is combined with Minimap2 for rapid microbiome abundance estimation. Here, Minimap2 acts as the primary classifier for target vs non-target while RawMap acts as secondary classifier that identifies the non-target reads Minimap2 missed.

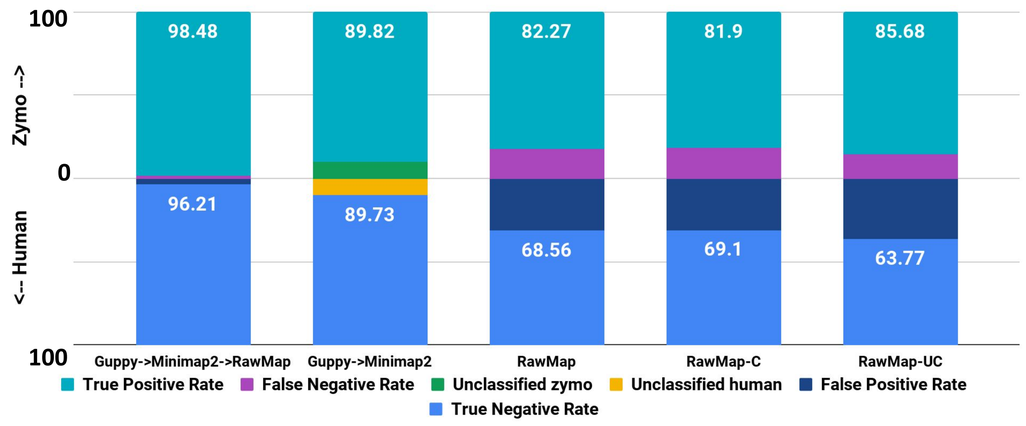


Acknowledgement: I thank Timothy Dunn, my labmate, for peer-review.

Results

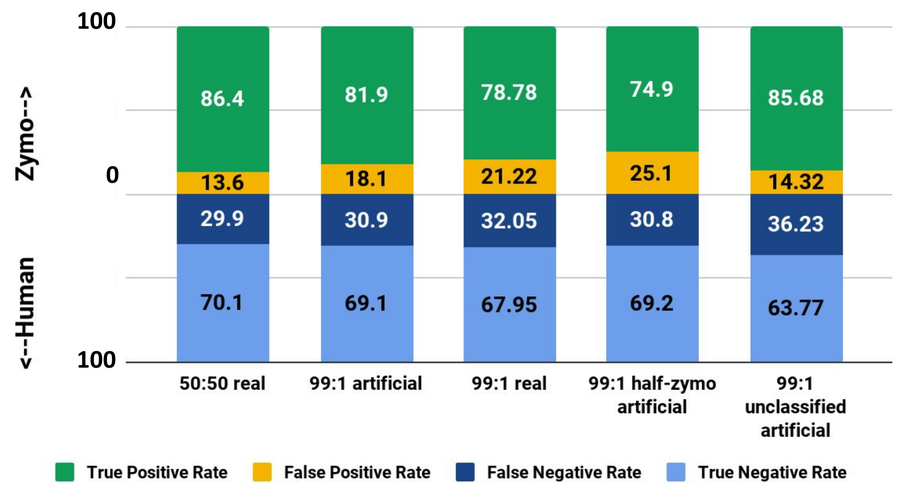
RawMap complements Minimap2 and improves overall pipeline accuracy as depicted in figure 4. RawMap's accuracy on alignable and unalignable reads are also shown (C and UC).

Figure 4: RawMap improves Read-Until accuracy



Extracted DNA of HeLa-human and Zymo-microbial-community-standard are sequenced using SQK-RAD004 on a MinION to form five different datasets. RawMap is tested on these datasets (100K-500K reads), which contain 99:1 and 50:50 HeLa:zymo mixes. As shown in figure 6, datasets mixed in the wetlab are labelled "real" and datasets mixed post-sequencing are labelled "artificial". RawMap is also trained and tested using mutually exclusive Zymo communities.

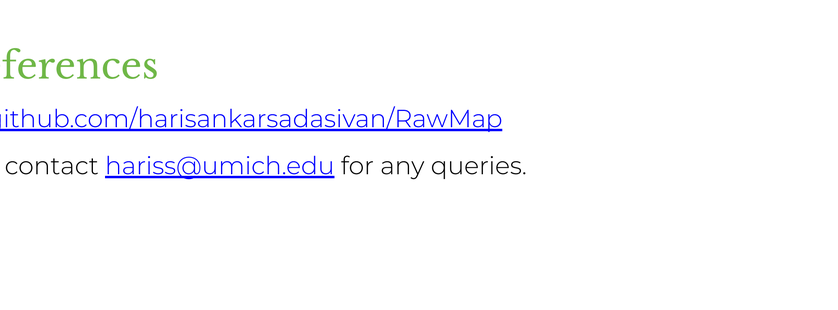
Figure 5: RawMap is good at selecting for Zymo



RamMap, a set of statistical and linear operations, is 900X faster in classifying 450bp than Guppy-followed-by-Minimap2.

The proposed Guppy-followed-by-Minimap2-followed-by-RawMap is strictly an improvement over the conventional pipeline with very less compute overhead. Figure 6 explains the benefits of adding RawMap to the conventional read-until pipeline in an experiment where the read length is 20Kbp and human:zymo is 99:1.

Figure 6: Adding RawMap after Minimap2 yields significant benefits



Conclusion

- RawMap is an efficient squiggle-space classifier that complements Minimap2 and further improves the performance of conventional Read-Until classification pipeline.
- Guppy-followed-by-Minimap2-followed-by-RawMap is recommended for cases where reduced time to answer is critical such as rapid microbiome abundance estimation .

Future work includes integrating RawMap with an event detection algorithm to classify reads into slow/fast moving in real-time. The benefits of using RawMap as the primary classifier followed by Minimap2 is also being explored.

References

www.github.com/harisankarsadasivan/RawMap

Please contact hariss@umich.edu for any queries.