Visualization and analysis of nanopore signal for modification detection and more with Uncalled4

Sam Kovaka[1], Paul W. Hook[2], Vikram Shivakumar[2], Katharine Jenike[3], Luke Morina[2], Roham Razaghi[2], Winston Timp[2,3], Michael C. Schatz[1,3]

[1]Department of Computer Science, Johns Hopkins University, Baltimore, MD. [2] Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD. [3] Department of Genetic Medicine, Johns Hopkins School of Medicine, Baltimore, MD

Nanopore Signal Alignment

Nanopore sequencing works by measuring ionic current as individual DNA or RNA molecules pass through a pore. Many analyses utilize direct alignments of the electrical signal to a nucleotide reference, including detection of nucleotide modifications, polishing, and adaptive sampling. Building off methods developed in Uncalled v1[1], we present Uncalled4: a toolkit for nanopore signal alignment, analysis, and visualization.

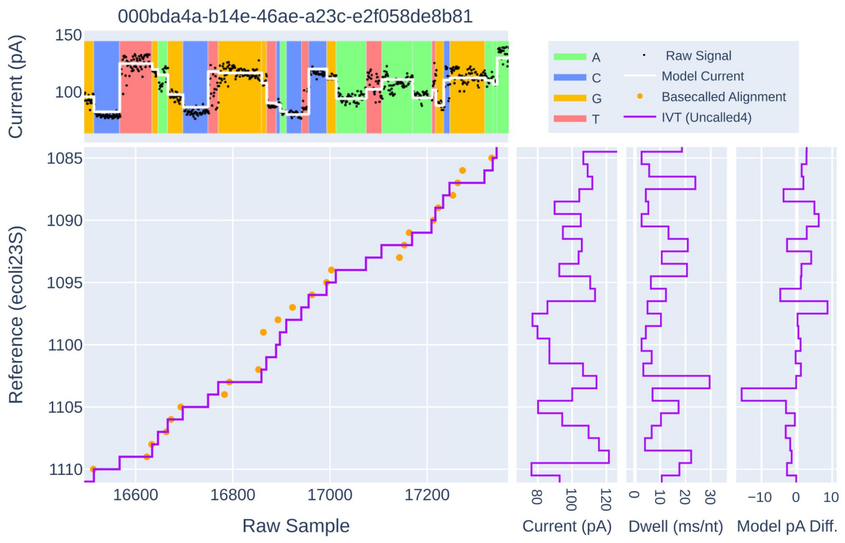

Figure 1. A signal-to-reference dotplot showing an alignment of read signal to its reference sequence. Basecalled alignment (orange) represents a Minimap2[2] alignment projected into signal-space. Top panel shows the expected reference current plotted over raw samples, with the background colored by reference base.

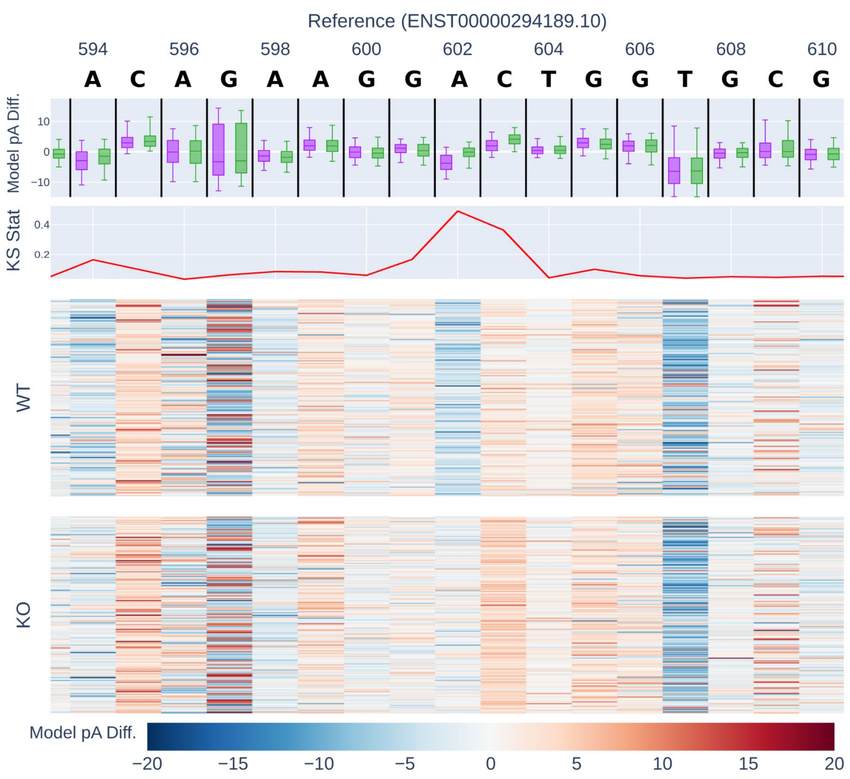

Figure 2. Summary statistics and browser view of the difference between observed and expected current in wildtype (WT) and METTL3 knockout (KO) Uncalled4 direct RNA alignments of HEK293t cell line data. Kolmogorov-Smirnov (KS) test statistics are represented by the red line. A known methylation site is located at position 602.

Modification Detection

Nucleotide modifications can be detected by comparing nanopore signal between two samples: one containing modifications (e.g. wildtype) and one without (e.g. methyltransferase knockout). We demonstrate m6A detection on RNA from wildtype and METTL3 knockout human HEK293t cell lines with m6-ACEseq as ground truth[3], using simple KS statistics with Uncalled4, Nanopolish[4], and Tombo[5] alignments (**Fig. 2, 3a**), and using Uncalled4 and Nanopolish as input to xPore[3] (**Fig. 3b**).

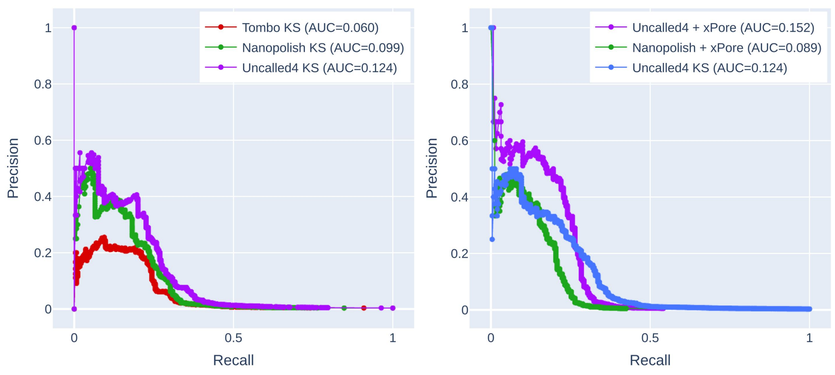

Figure 3. Precision recall curves of identifying m6A in human RNA[3]. **(a)** Using KS statistics based on Tombo, Nanopolish, and Uncalled4 transcriptome alignments, and **(b)** xPore with Uncalled4 and Nanopolish transcriptome alignments collapsed on the gene-level, and gene-level KS statistics on Uncalled4 spliced genome alignments.

Comparing Alignments

Uncalled4 provides statistics and visualizations for comparing alignment methods on the same set of reads. This can be based on two signal-based methods (**Fig. 4a**), or with projected basecalled alignments, which can be considered an approximate ground truth (**Fig. 4b**).

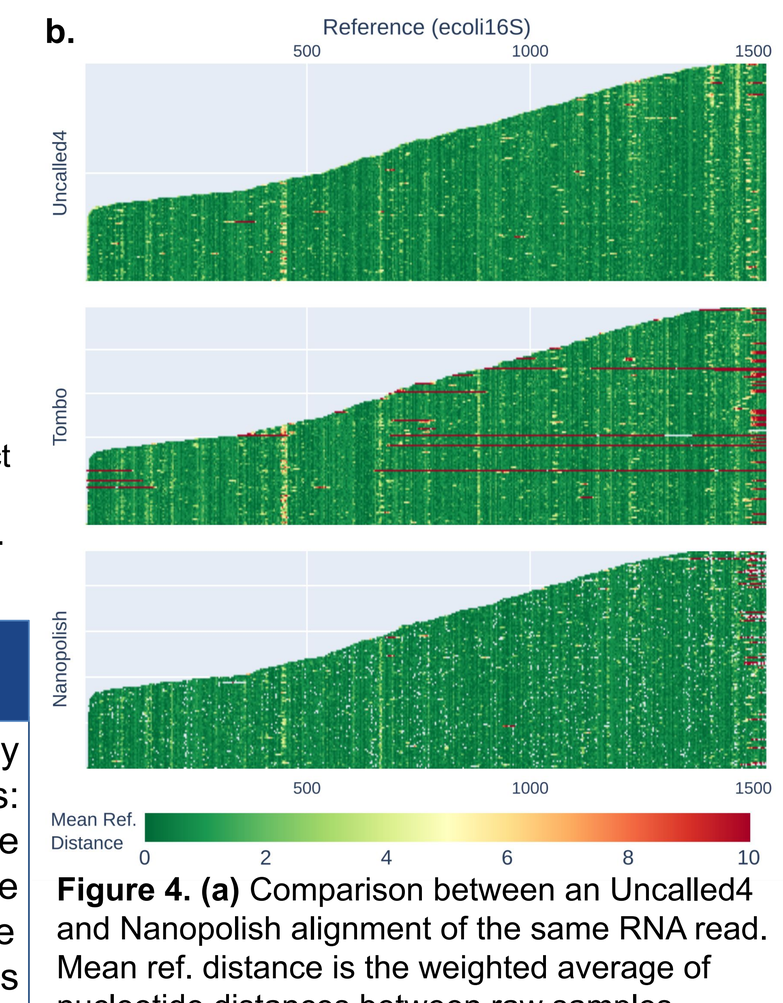

Figure 4. (a) Comparison between an Uncalled4 and Nanopolish alignment of the same RNA read. Mean ref. distance is the weighted average of nucleotide distances between raw samples. **(b)** Distances from projected basecalled alignments of Uncalled4, Nanopolish, and Tombo alignments accross the *E. coli* 16S ribosomal RNA transcript. Data obtained from [6].

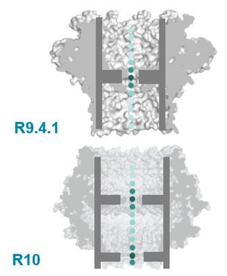

Figure 5. Illustrations of ONT's r9.4 pore, currently the most commonly used sequencing chemistry, and R10.4, a higher accuracy pore which features two "reader heads" (source: [7])

Model Training

Signal alignment requires a pore model to map k-mers to expected current levels. The new R10.4 pore (**Fig. 5**), which covers more bases with two reader heads for higher accuracy than the commonly used R9.4 pore, does not have an established pore model. We use Uncalled4 to train R10.4 DNA models starting with a draft model based on Scrappie reverse basecalling[8], followed by iterative refinment via multiple rounds of signal alignment. We generated models based on unmodified DNA (**Fig. 6**) and methylated DNA with 5mC at CpG sites and m6A in all contexts (**Fig. 7**)

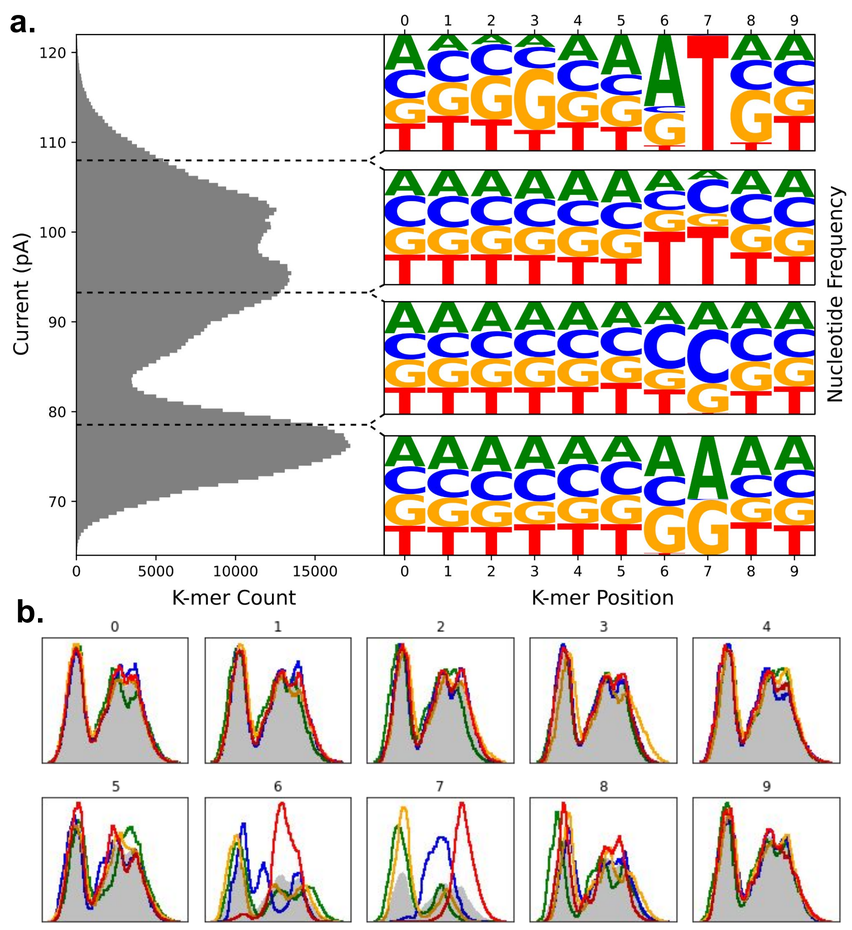

Figure 6. Visualizations of an (unmodified) R10.4 pore model. **(a)** Distribution of current levels for all k-mers in the model, with sequence logos of nucleotide content for different ranges of current (dotted lines). **(b)** Distributions of k-mers with fixed bases at each k-mer position, plotted over the full model distribution in gray.

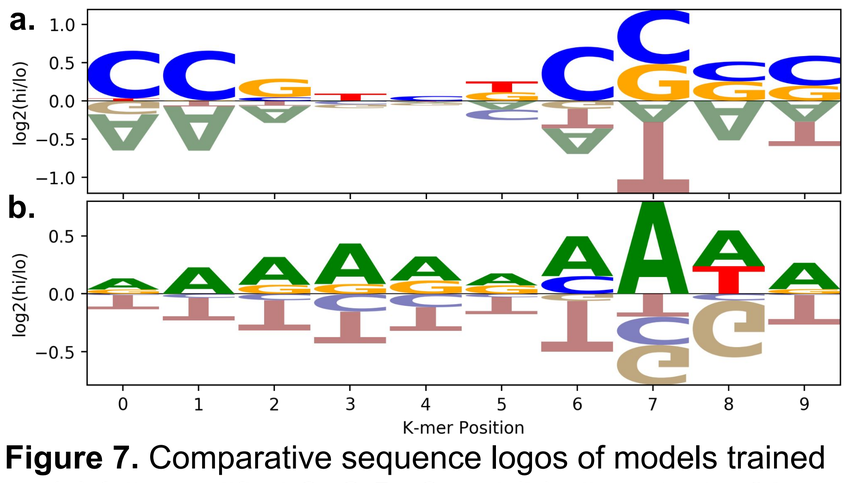

Figure 7. Comparative sequence logos of models trained on **(a)** fully modified CpG 5mC and (**b**) all-context m6A DNA methylation. These models were compared with the unmodified model by measuring the absolute difference in current levels between each k-mer and then computing the log ratios of nucleotide frequencies between k-mers above and below the median absolute difference.

References: **[1]** Kovaka, S., Fan, Y., Ni, B., Timp, W. & Schatz, M. C. Targeted nanopore sequencing by real-time mapping of raw electrical signal with UNCALLED. *Nat. Biotechnol.* **39**, 431–441 (2021) **[2]** Li, H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34, 3094–3100 (2018) **[3]** P. N. Pratanwanich, F. Yao, Y. Chen, C. W. Q. Koh, Y. K. Wan, C. Hendra, P. Poon, Y. T. Goh, P. M. L. Yap, J. Y. Chooi, W. J. Chng, S. B. Ng, A. Thiery, W. S. S. Goh, J. Göke, Nat. Biotechnol. 2021, 39, 1394. **[4]** Simpson, J. T. *et al.* Detecting DNA cytosine methylation using nanopore sequencing. *Nat. Methods* **14**, 407–410 (2017) **[5]** Stoiber, M. *et al.* De novo Identification of DNA Modifications Enabled by Genome-Guided Nanopore Signal Processing. *bioRxiv* 094672 (2017) doi:10.1101/094672 **[6]** Stephenson, W. *et al.* Direct detection of RNA modifications and structure using single molecule nanopore sequencing. *Cold Spring Harbor Laboratory* 2020.05.31.126763 (2020) doi:10.1101/2020.05.31.126763, **[7]** https://github.com/nanoporetech/scrappie **[8]** nanoporetech.com/about-us/news/r103-newest-nanopore-high-accuracy-nanopore-sequencing-now-available-store

skovaka.github.io/UNCALLED