RawHash: Enabling Fast and Accurate Real-Time Analysis of Raw Nanopore Signals for Large Genomes

Can Firtina[1], Nika Mansouri Ghiasi[1], Joel Lindegger[1], Gagandeep Singh[1],
Meryem Banu Cavlak[1], Haiyu Mao[1] and Onur Mutlu[1]

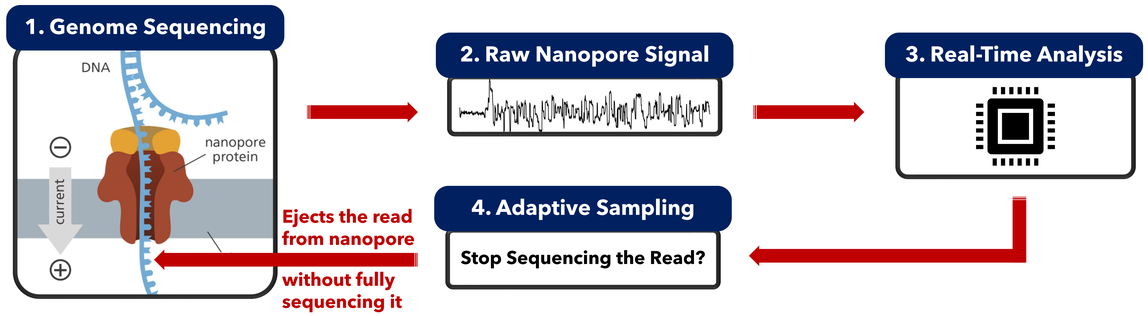
[1] SAFARI ETH zürich

To be presented at ISMB/ECCB 2023

bioRxiv Preprint

Source Code

1: Real-Time Genome Analysis with Adaptive Sampling



2: Challenges in Real-Time Genome Analysis

Fast analysis that can match the throughput of sequencer

Low latency to reduce the sequencing time and cost with effective use of adaptive sampling

Accurate analysis from noisy raw signal data

Efficient analysis to enable portable sequencing on devices with low computational power

3: Problem

Efficient tools (UNCALLED and Sigmap) **cannot provide** either
1. Fast analysis or
2. Accurate analysis for **large genomes**

Accurate tools (e.g., ReadFish) **cannot provide**
1. Efficient analysis

4: Goal

Fast analysis that can scale to large genomes

Low latency to make quick decisions on adaptive sampling

Accurate analysis for large genomes

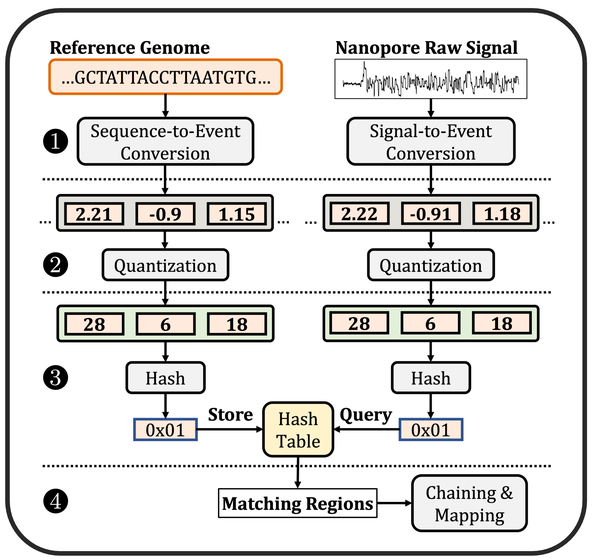
Efficient analysis that can be used with portable devices

5: Key Contributions

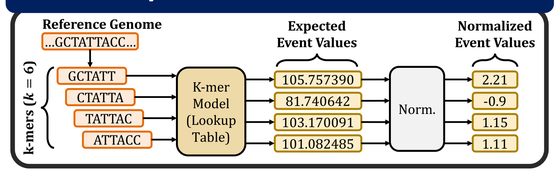
First mechanism that can **efficiently and accurately map raw signals to large reference genomes**

Proposes a novel mechanism, **Sequence Until**, that can dynamically **decide if further sequencing of reads is unnecessary** to **stop the entire sequencing run**

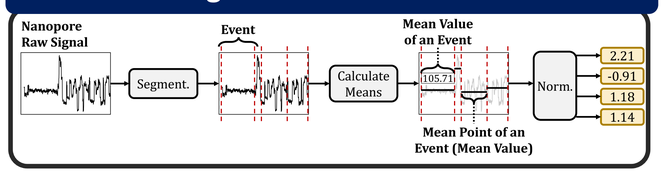
6: RawHash



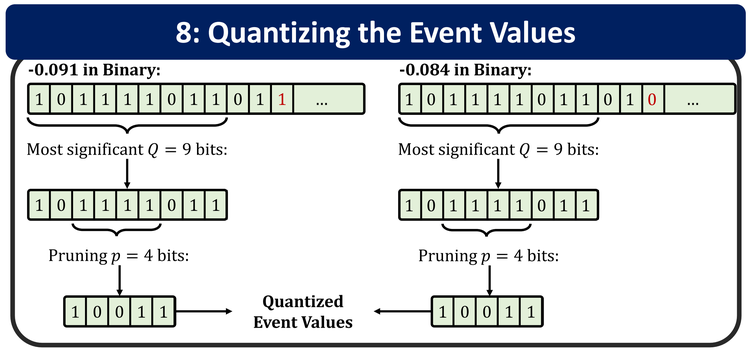
7.1: Sequence-to-Event Conversion



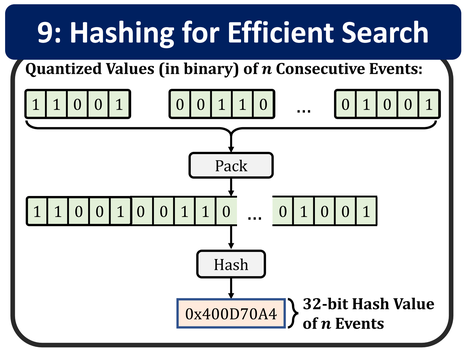
7.2: Signal-to-Event Conversion



8: Quantizing the Event Values



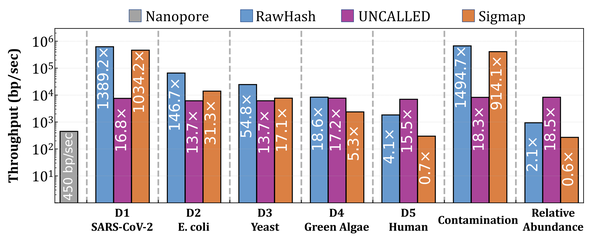
9: Hashing for Efficient Search



10: Evaluation Methodology

- Datasets from very small (viral) to large genomes (human and metagenomics)
- Compared with UNCALLED and Sigmap
- Use cases
  1. Read mapping
  2. Relative abundance estimation
  3. Contamination analysis
- Evaluating
  1. Throughput (bp/sec)
  2. Overall Runtime (sec)
  3. Memory usage (GB)
  4. Number of sequenced bases before ejecting reads (bases)
  5. Accuracy (baseline: minimap2 mappings)
  6. Sequence Until benefits

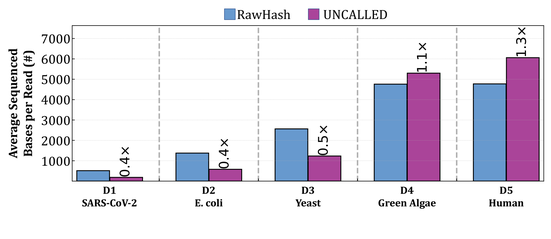
11: Results



Efficient Analysis: **RawHash, UNCALLED, and Sigmap** do not require powerful computational resources (e.g., GPUs)

Fast Analysis: Both **RawHash and UNCALLED** can match the throughput of nanopore

Sigmap falls behind the throughput of nanopores for larger genomes



Low Latency: **RawHash** utilizes adaptive sampling faster for large genomes than UNCALLED

UNCALLED more bases due as it cannot make accurate decisions for large genomes using fewer bases

Dataset		UNCALLED	Sigmap	RawHash
Read Mapping				
D1	Precision	0.9547	**0.9929**	0.9868
SARS-CoV-2	Recall	**0.9910**	0.5540	0.8735
	F_1	**0.9725**	0.7112	0.9267
D2	Precision	0.9816	**0.9842**	0.9573
E. coli	Recall	**0.9647**	0.9504	0.9009
	F_1	**0.9731**	0.9670	0.9282
D3	Precision	0.9459	0.9856	**0.9862**
Yeast	Recall	**0.9366**	0.9123	0.8412
	F_1	0.9412	**0.9475**	0.9079
D4	Precision	0.8836	**0.9741**	0.9691
Green Algae	Recall	0.7778	**0.8987**	0.7015
	F_1	0.8273	**0.9349**	0.8139
D5	Precision	0.4867	0.4287	**0.8959**
Human HG001	Recall	0.2379	0.2641	**0.4054**
	F_1	0.3196	0.3268	**0.5582**
Relative Abundance Estimation				
	Precision	0.7683	0.7928	**0.9484**
D1-D5	Recall	0.1273	0.2739	**0.3076**
	F_1	0.2184	0.4072	**0.4645**
Contamination Analysis				
	Precision	**0.9378**	0.7856	0.8733
D1, D5	Recall	**0.9910**	0.5540	0.8735
	F_1	**0.9637**	0.6498	0.8734

	Estimated Relative Abundance Ratios in 50,000 Random Reads					
Tool	***SARS-CoV-2***	***E. coli***	***Yeast***	***Green Algae***	***Human***	**Distance**
RawHash (100%)	0.0270	0.3636	0.3062	0.1951	0.1081	N/A
RawHash + Sequence Until (7%)	0.0283	0.3539	0.3100	0.1946	0.1133	0.0118

Accurate Analysis: **RawHash** provides the best accuracy for large genomes

Sequence Until dynamically stops the entire sequencing after sequencing **only 7% of the entire sample**

Sequence Until least to almost as accurate relative abundance estimation as using the entire (100%) sample.