RawHash RawHash2 RawAlign

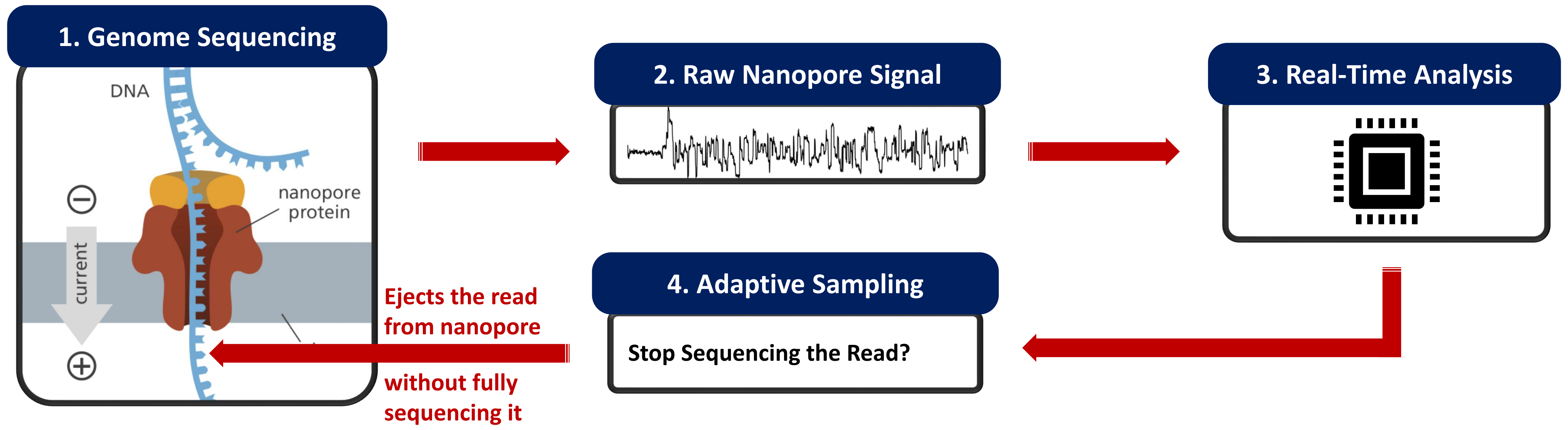
Accurate, Fast, and Scalable Real-Time Analysis of Raw Nanopore Signals

Can Firtina, Joel Lindegger, Nika Mansouri Ghiasi, Melina Soysal, Gagandeep Singh,
Meryem Banu Cavlak, Haiyu Mao, Mohammad Sadrosadati, Mohammed Alser and Onur Mutlu

ETH zürich SAFARI SAFARI Research Group

Source Code

1: Real-Time Genome Analysis with Adaptive Sampling


1. Genome Sequencing
DNA
nanopore protein
current
Ejects the read from nanopore
without fully sequencing it
2. Raw Nanopore Signal
3. Real-Time Analysis
4. Adaptive Sampling
Stop Sequencing the Read?

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

The **first mechanism** to **efficiently and accurately map raw signals to large reference genomes**

A **novel mechanism** "SequenceUntil" to **stop the entire sequencing run** by dynamically **deciding if further sequencing of reads is unnecessary**

The **first mechanism** use **a weighted combination of multiple metrics** for **mapping decisions**

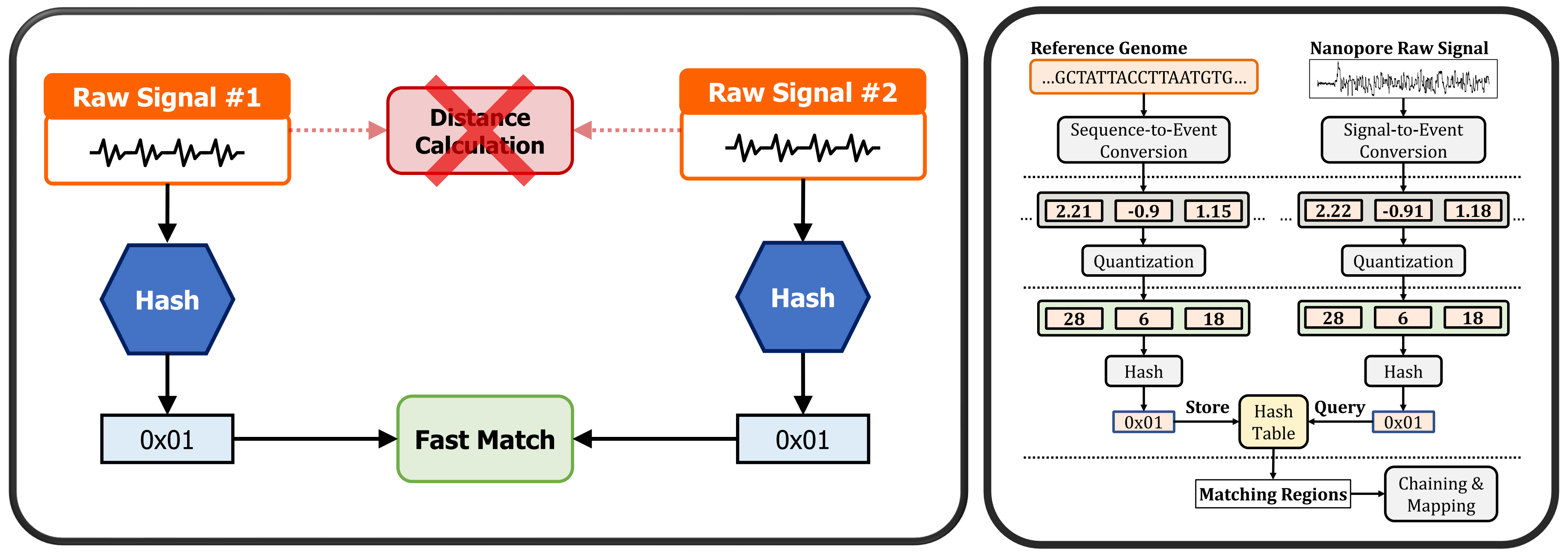
Wide range of **supported file formats**, including **FAST5, POD5, SLOW5, and BLOW5**

Wide range of **supported nanopore chemistries**, including **ONT's R 9.4.1 and R10.4.1**

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
- Evaluation Metrics
  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

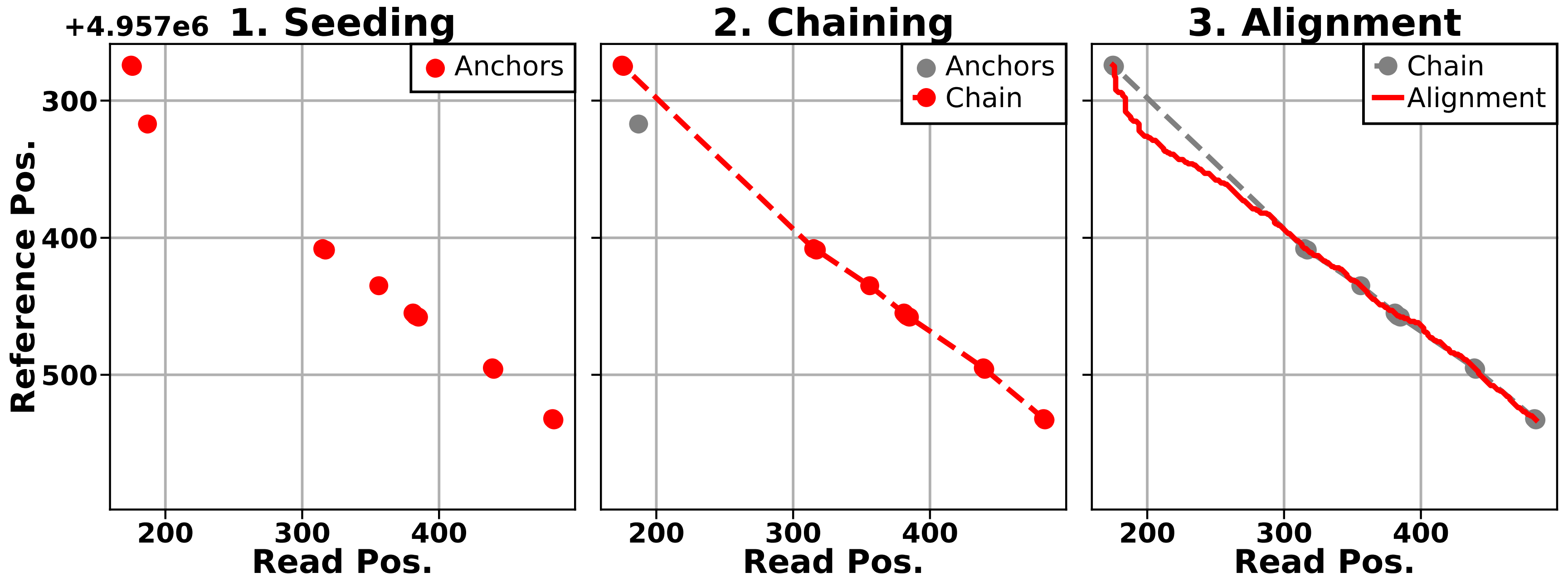
6: RawHash


Raw Signal #1
Raw Signal #2
Distance Calculation
Hash
Hash
0x01
0x01
Fast Match
Reference Genome
...GCTATTACCTTAATGTG...
Nanopore Raw Signal
Sequence-to-Event Conversion
Signal-to-Event Conversion
... 2.21 -0.9 1.15 ...
... 2.22 -0.91 1.18 ...
Quantization
Quantization
28 6 18
28 6 18
Hash
Hash
0x01
Store
Hash Table
Query
0x01
Matching Regions
Chaining & Mapping

Hashing raw signals enables **fast comparison** between long raw signal sequences

RawHash preprocesses the inherently **noisy raw signals** with **Signal-to-Event Conversion** and **Quantization**

7: RawAlign

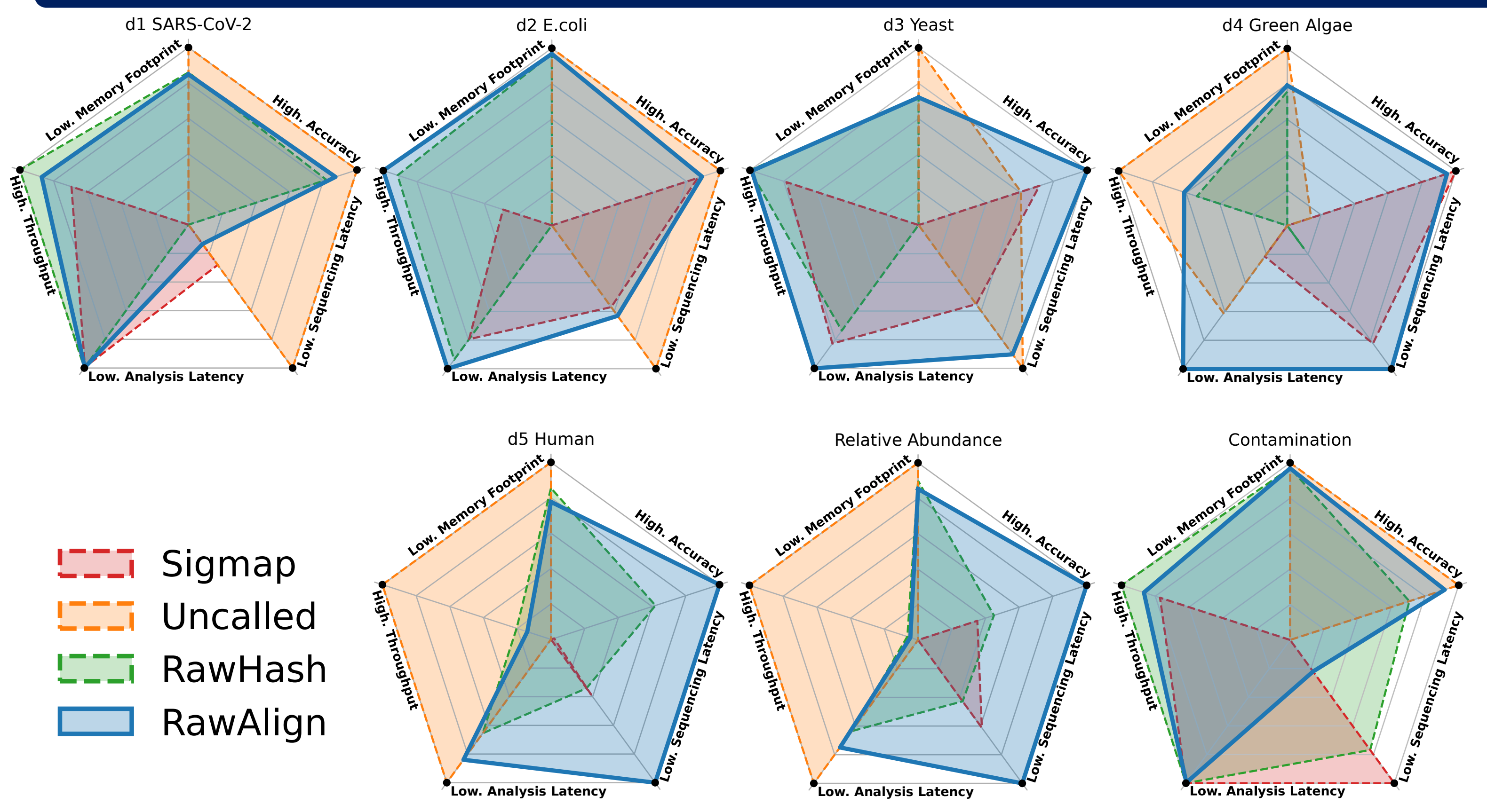


Alignment after seeding and chaining enables **fine-grained and accurate comparison** between subregions of long raw signal sequences

RawAlign uses **dynamic time warping** (DTW) for signal alignment

11: Results

Read Mapping Overview

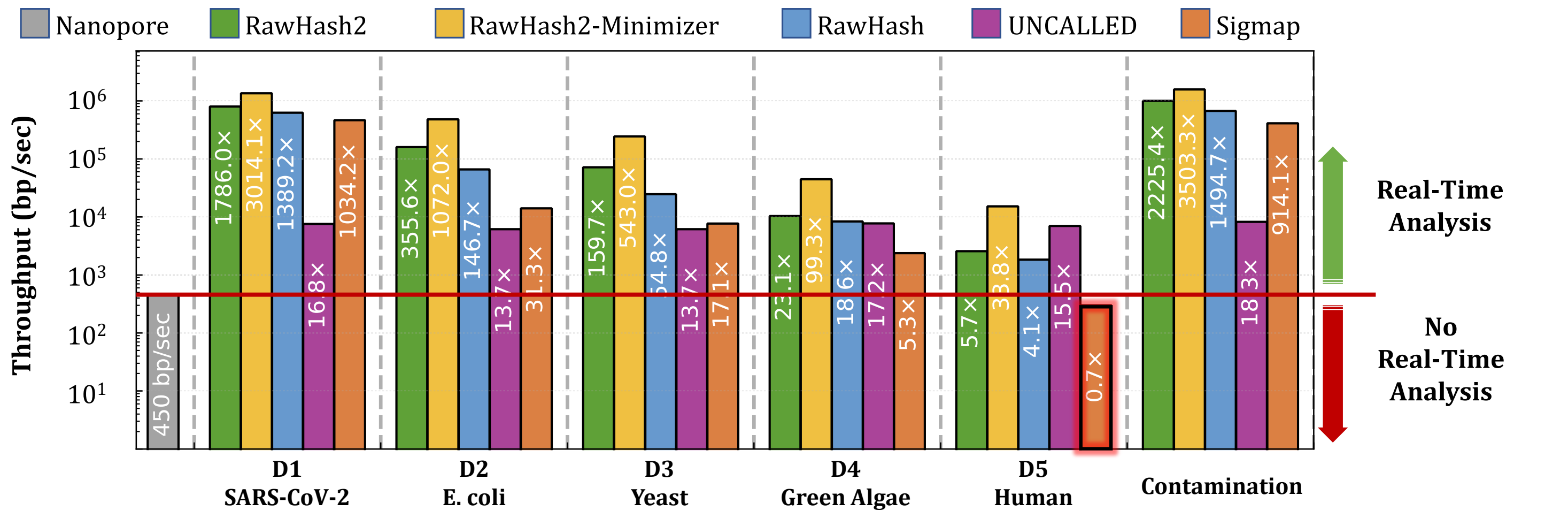


d1 SARS-CoV-2	Memory Footprint (GB)	Throughput (bp/s)	Analysis Latency (ms)	Sequencing Latency (Chunks)	Accuracy (F-1)
Uncalled	**0.280**	6,575.310	29.244	**0.410**	**0.972**
Sigmap	28.250	350,565.180	1.111	1.005	0.711
RawHash	4.210	**502,043.190**	**0.942**	1.238	0.925
RawAlign	4.520	438,089.990	1.070	1.126	0.939
d2 E.coli					
Uncalled	0.800	5,174.050	115.787	**1.290**	**0.973**
Sigmap	111.170	19,215.930	34.441	2.111	0.967
RawHash	4.270	49,559.740	19.754	3.200	0.928
RawAlign	**0.000**	**53,693.170**	**13.323**	1.995	0.968
d3 Yeast					
Uncalled	**0.580**	5,151.670	159.304	**2.773**	0.941
Sigmap	14.710	15,217.010	67.602	4.139	0.947
RawHash	4.530	**17,996.930**	77.586	5.826	0.906
RawAlign	4.530	17,854.670	**48.394**	3.071	**0.963**
d4 Green Algae					
Uncalled	**1.260**	**8,174.320**	440.815	11.790	0.840
Sigmap	53.710	2,251.370	608.898	5.804	**0.938**
RawHash	14.060	5,429.580	700.304	10.646	0.824
RawAlign	12.200	5,871.450	**276.094**	**4.514**	0.932
d5 Human					
Uncalled	**13.170**	**5,612.920**	**1,077.536**	12.959	0.320
Sigmap	313.400	195.180	16,296.435	10.401	0.327
RawHash	56.940	1,298.520	6,318.984	10.695	0.557
RawAlign	80.350	956.310	3,510.682	**6.321**	**0.703**
Contamination					
Uncalled	**1.060**	6,607.850	199.283	3.557	**0.964**
Sigmap	111.650	405,956.490	1.206	**2.062**	0.650
RawHash	4.280	**524,042.570**	**1.139**	2.409	0.872
RawAlign	4.500	455,376.380	2.004	3.227	0.938
Relative Abundance					
Uncalled	**10.870**	**6,721.770**	**309.079**	4.921	0.218
Sigmap	506.340	181.880	5,670.365	3.338	0.406
RawHash	60.760	596.740	2,264.014	3.816	0.459
RawAlign	83.760	480.050	1,652.162	**2.336**	**0.754**

RawHash's seeding and chaining effect **high throughput and accuracy**

RawAlign's alignment further **improves accuracy on top of RawHash across all datasets**

Read Mapping Throughput



RawHash2's seed filtering further **improves throughput on top of RawHash**

Relative Abundance

Tool	SARS-CoV-2	E.coli	Yeast	Green Algae	Human	Distance
Ground Truth	0.652	0.167	0.024	0.030	0.127	-
minimap2	0.613	0.163	0.025	0.053	0.147	**0.050**
Uncalled	0.072	0.466	0.001	0.150	0.312	0.689
Sigmap	0.201	0.446	0.002	0.123	0.229	0.549
RawHash	0.309	0.440	0.000	0.073	0.178	0.445
RawAlign	0.565	0.248	0.002	0.050	0.136	0.123

RawAlign can calculate **relative abundances with high accuracy, similar to** the **state-of-the-art** basecalling-based analysis tool **minimap2**