

# Computer Architecture

## Lecture 26c: RawHash2

Can Firtina

ETH Zurich

Fall 2024

13 December 2024

# Key Contributions in RawHash2

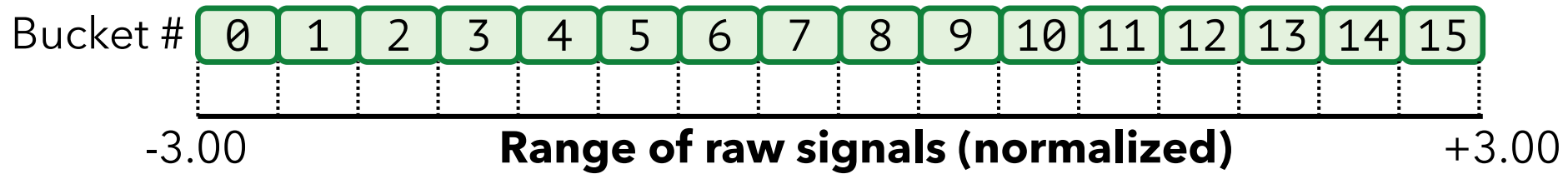
**A new adaptive quantization** that better fits the expected nanopore signal pattern to achieve **high accuracy**

**Weighted decision making** for more robust mapping

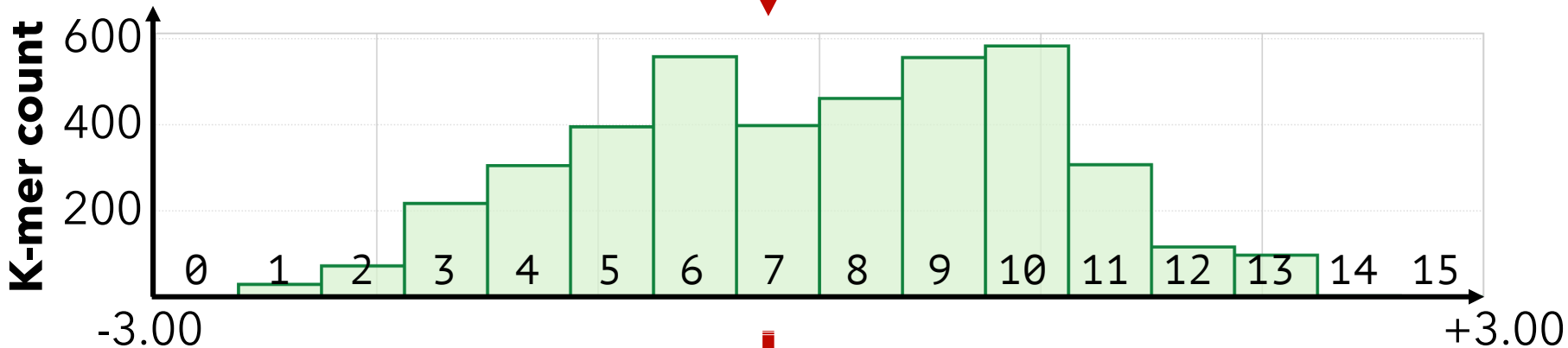
**Frequency filter** and **minimizer sketching** to reduce seed matches for faster and space-efficient mapping

**Improved chaining algorithm** with sensitive penalty scores

# Better Understanding of Noise



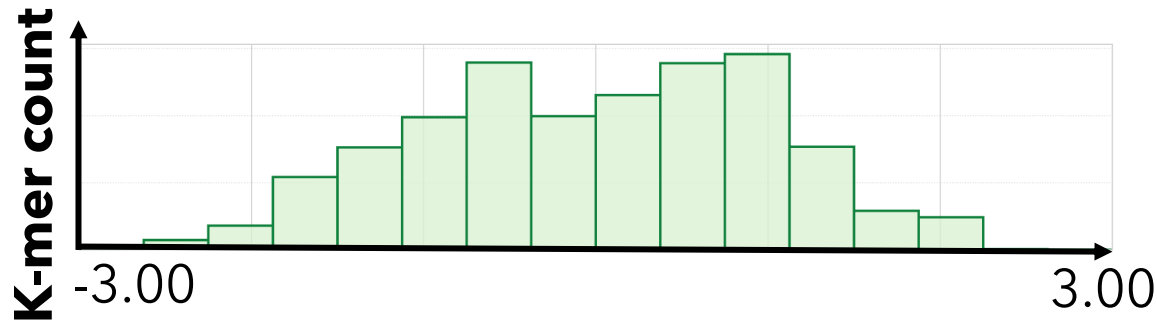
✗ Equal-width buckets leads to **unbalanced loading**



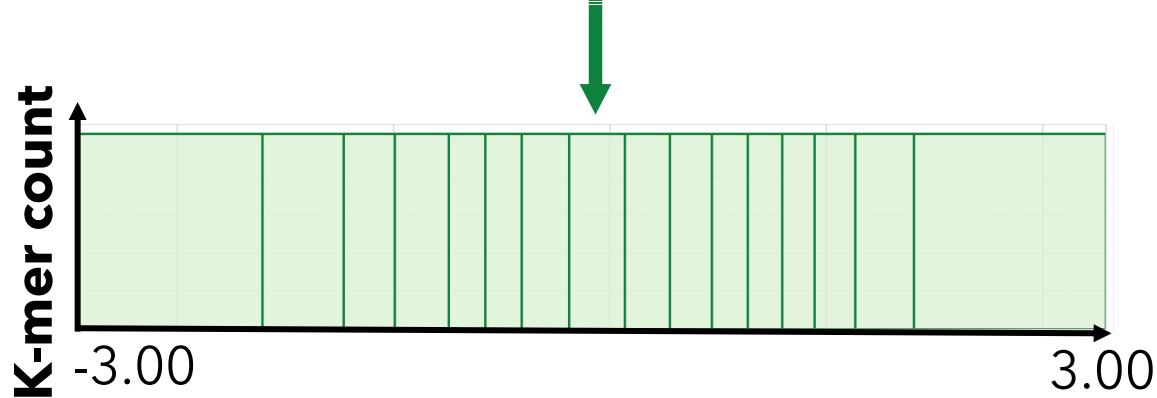
✗ More **false hash matches** due to reduced uniqueness

# Adaptive Quantization

- **Key Idea:** Quantizing raw signals with **non-equal bucket widths to maximize load balancing**



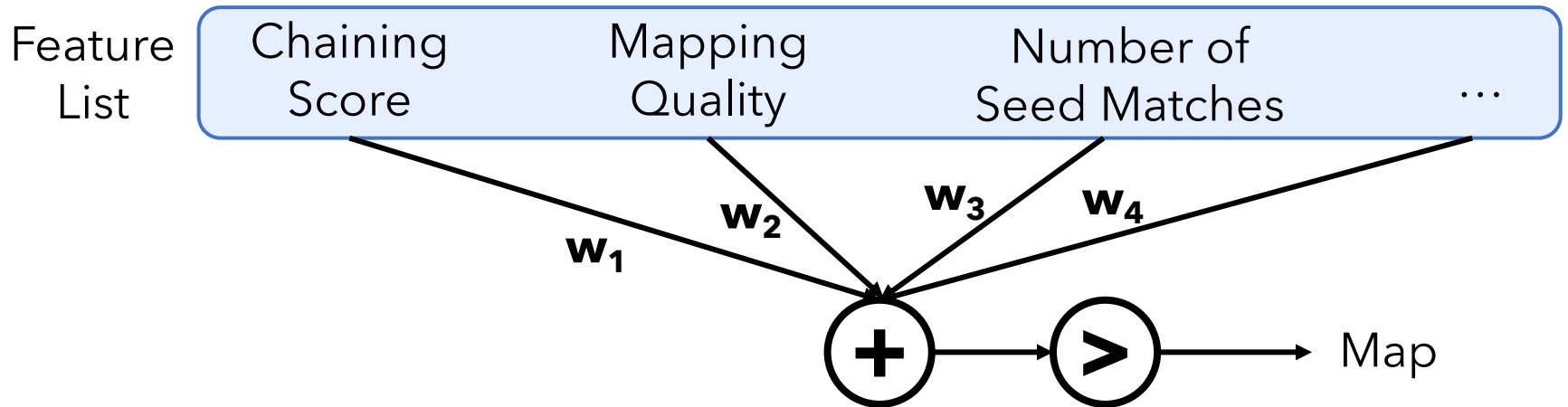
**Goal: Ideal Load Balance**



**Adaptive quantization reduces collisions** caused due to skewed raw signal distributions

# Other Key Improvements in RawHash2

- **Weighted mapping decisions**



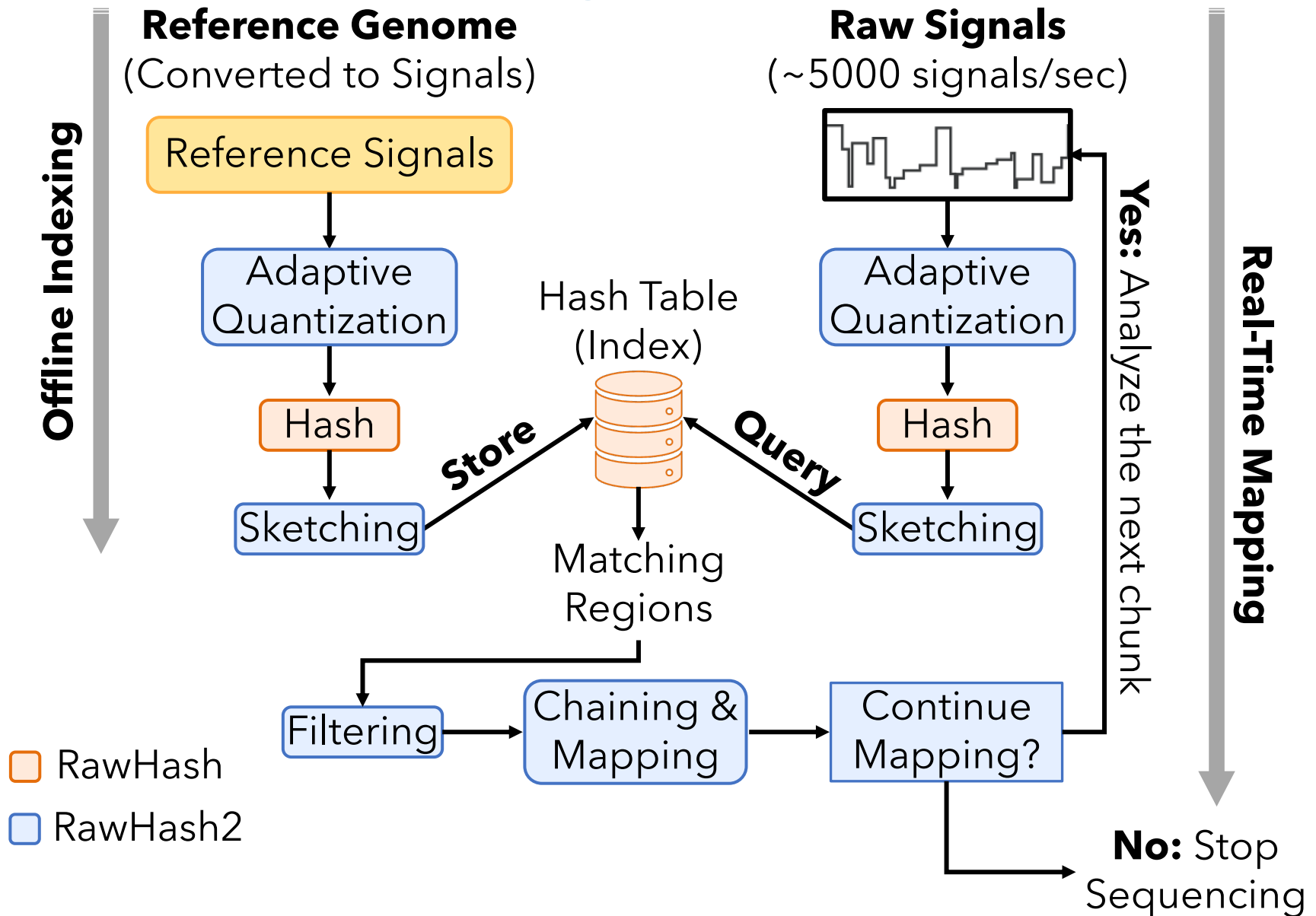
- **Sampling strategies** for reduced storage and computation overheads
  - Frequency filter and minimizer sketching



- **Improved chaining algorithm**

- More sensitive scoring functions

# Real-Time Mapping with RawHash2



# Outline

Background

RawHash

RawHash2

Evaluation

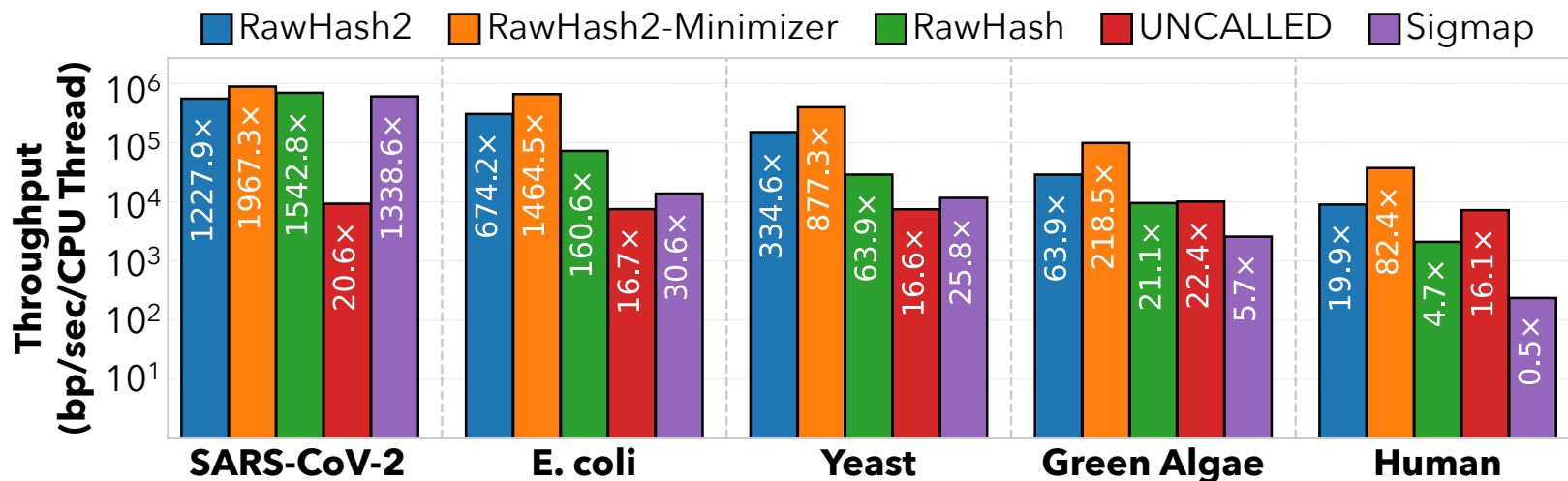
Conclusion

# Evaluation Methodology

- Two settings for RawHash2:
  - **RawHash2:** All hash values without sampling
  - **RawHash2-Minimizer:** Minimizer sketching
- Compared to **UNCALLED** [Kovaka+, Nat. Biotech.'21], **Sigmap** [Zhang+, ISMB/ECCB'21], and **RawHash** [Firtina+, ISMB/ECCB'23]
- **Use cases** for real-time genome analysis:
  1. Read mapping
  2. Relative abundance estimation
  3. Contamination analysis

# Key Results – Throughput

- Data generation throughput of **a single nanopore**: **~450 bp/sec**
  - **A single nanopore device** contains roughly 512 to 2500 nanopores
- Computation throughput of a **single CPU thread**: **bases processed/sec**
  - **Scalability**: The number of nanopores that a single CPU thread can process

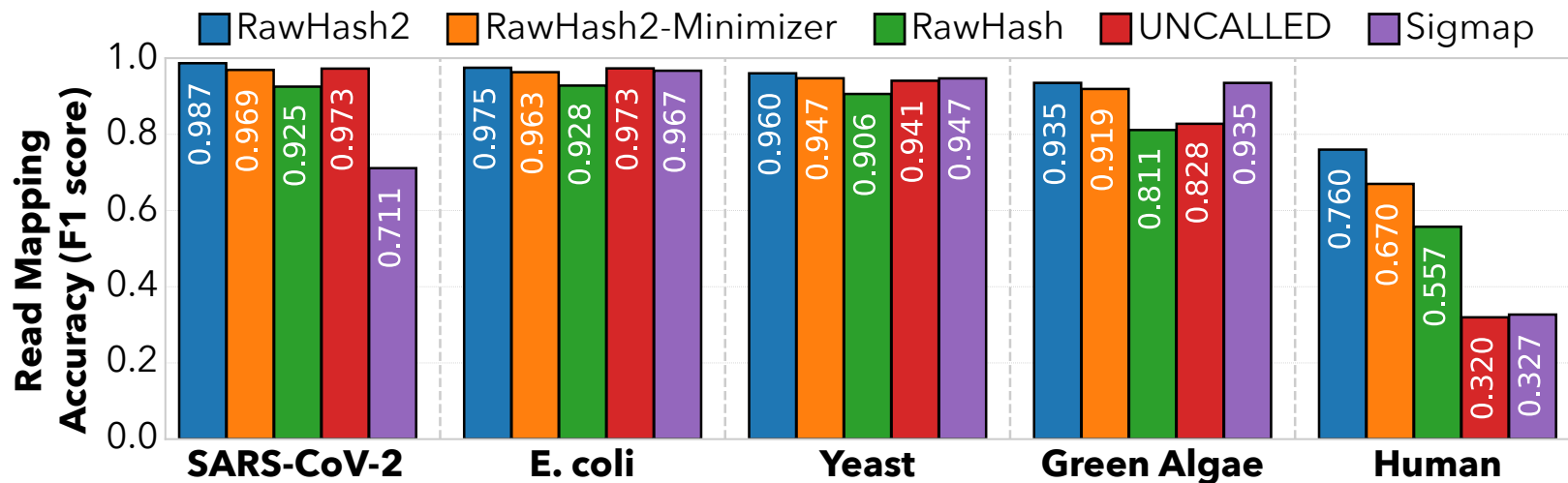


**RawHash2 average speedup:**  
**26.5x** (UNCALLED), **19.2x** (Sigmap), and **4x** (RawHash)

**RawHash2-Minimizer average speedup: 2.5x** (RawHash2)

# Key Results – Mapping Accuracy

- Accuracy of **mapping positions** (F1 score)
  - Ground truth: Mapping positions of **basecalled sequences** using minimap2

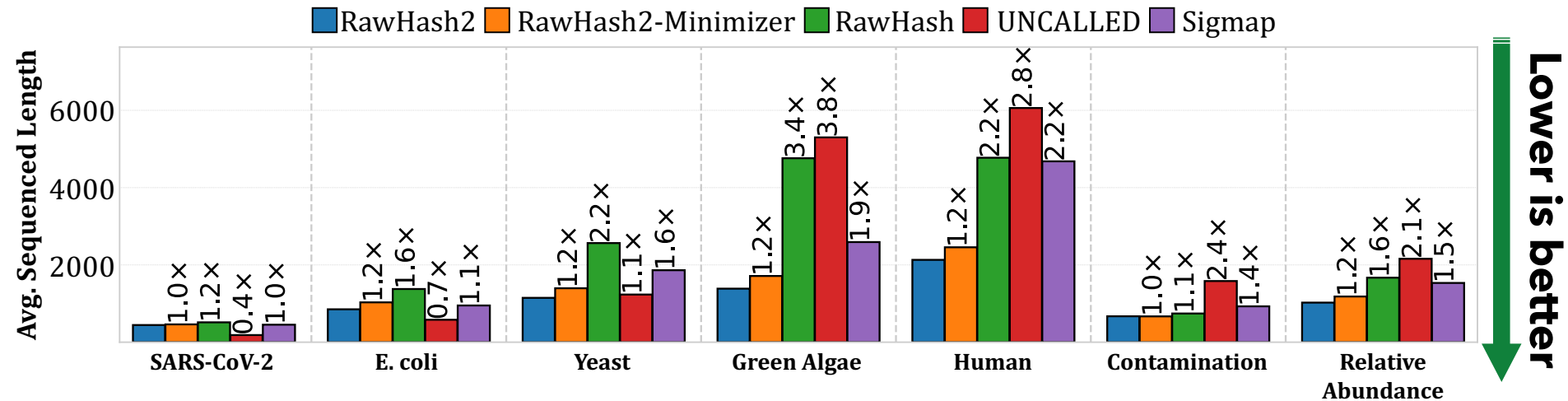


**RawHash2 provides the best accuracy in all datasets**  
(up to ~2.4x for large genomes)

**RawHash2-Minimizer** provides mapping accuracy  
**comparable to RawHash2**

# Average Sequenced Length

- Fewer bases to sequence → Less unnecessary sequencing



**RawHash2 can reduce sequencing time and cost: on average by 1.9x compared to UNCALLED and RawHash**

# Benefits of Sequence Until

- Running RawHash with and without Sequence Until

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

**Sequence Until** enables **sequencing only 7%** of the entire sample **while providing high accuracy**

**UNCALLED and RawHash** benefit from Sequence Until significantly by enabling up to **100× reductions in sequencing**

# Outline

Background

RawHash

RawHash2

Evaluation

Conclusion

# Conclusion

## Key Contributions:

- 1) The **first hash-based mechanism** for mapping raw nanopore signals
- 2) The novel **Sequence Until** technique can accurately and **dynamically stop the entire sequencing of all reads at once** if further sequencing is not necessary

## Key Results: Across 3 use cases and 5 genomes of varying sizes

- **27× 19×, and 4× better average throughput** compared to the state-of-the-art works
- **Most accurate raw signal mapper for all datasets**
- Sequence Until **reduces the sequencing time and cost by 15×**

## Many opportunities for analyzing raw nanopore signals in real-time:

- Many hash-based **sketching techniques** can now be used for raw signals
- **Indexing is very cheap:** Many future use cases with the on-the-fly index construction
- We should rethink the algorithms to fully perform downstream analysis with raw signals



# RawHash

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

Can Firtina

Nika Mansouri Ghiasi

Joel Lindegger

Gagandeep Singh

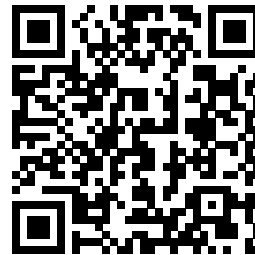
Meryem Banu Cavlak

Haiyu Mao

Onur Mutlu



[RawHash](#)



[RawHash2](#)



[Code](#)

**SAFARI**

**ETH** zürich

# Computer Architecture

## Lecture 26c: RawHash2

Can Firtina

ETH Zurich

Fall 2024

13 December 2024

# Sequencing Data Analysis

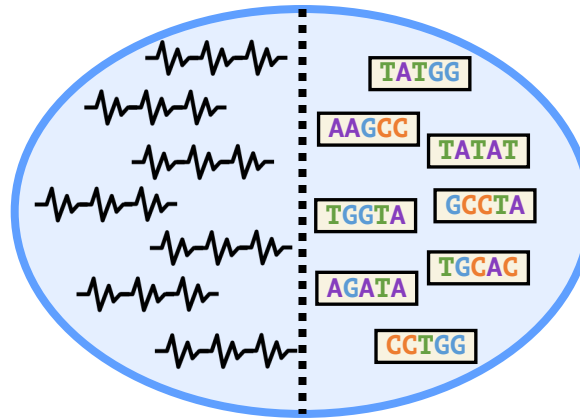
Heuristic Algorithms



Data Structures



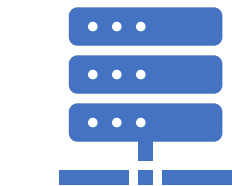
Filters



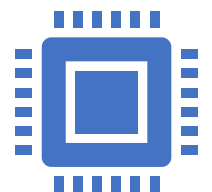
**Quick, accurate, and energy-efficient** analysis



**Imperfections in sequencing data**  
impacts design choices

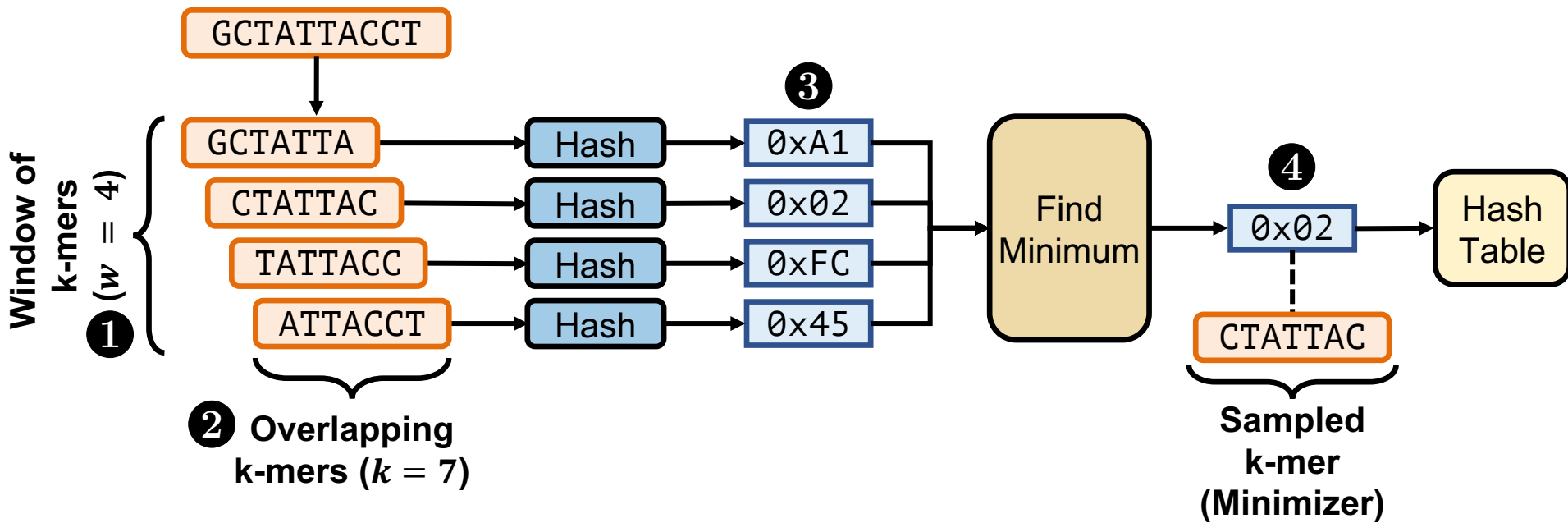


Distributed Computing

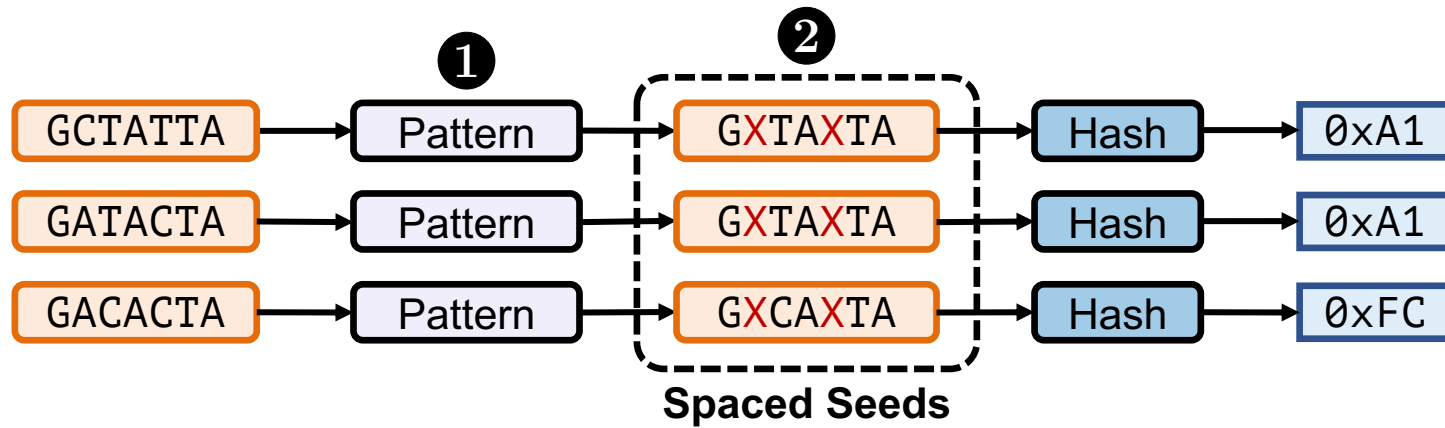


Hardware Accelerators

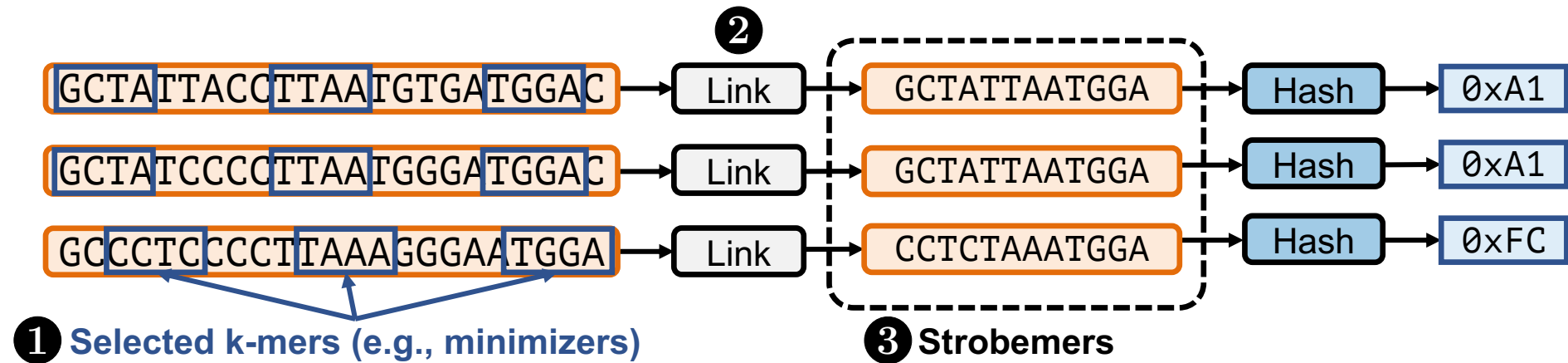
# Minimizer Sketching



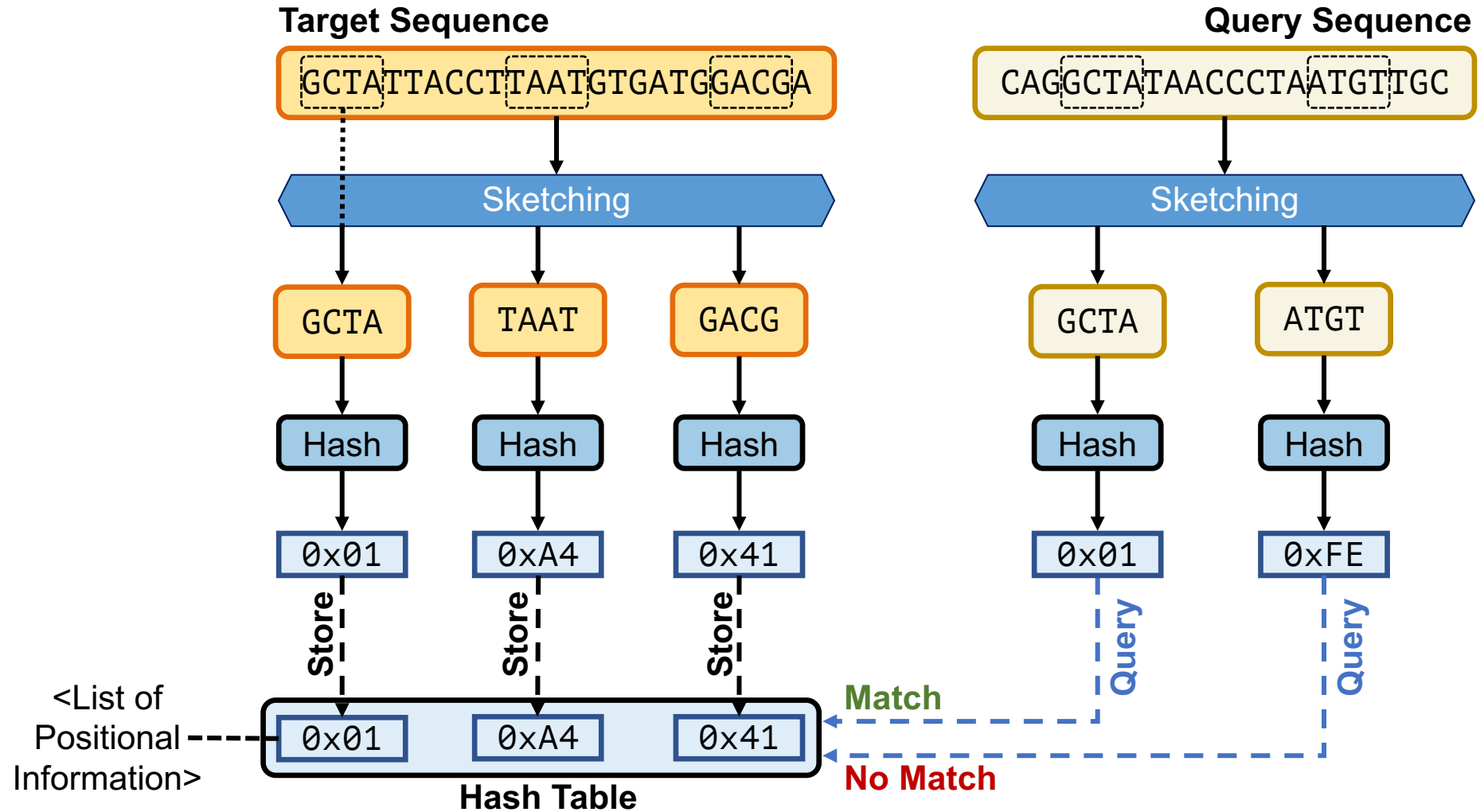
# Spaced Seeding



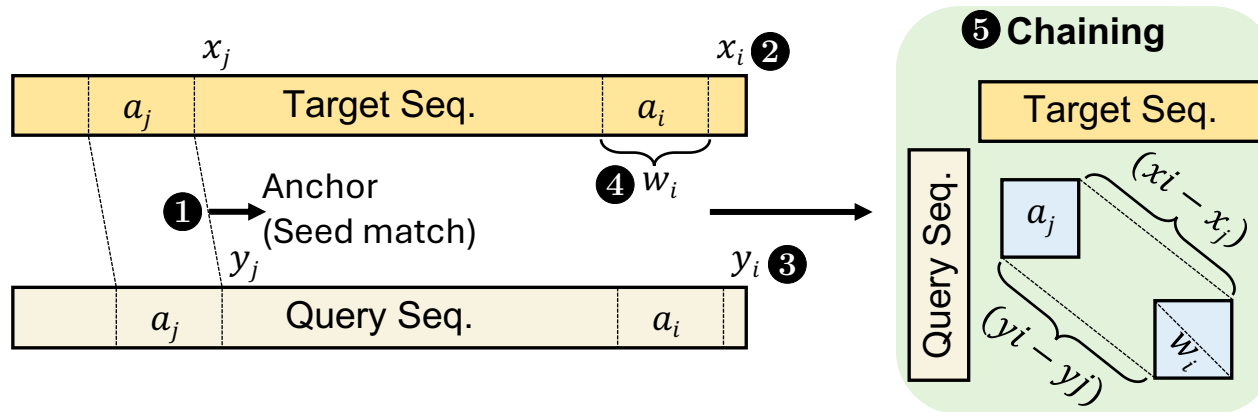
# Strobemer Sketches



# Hash-Based Sketching and Seed Matching

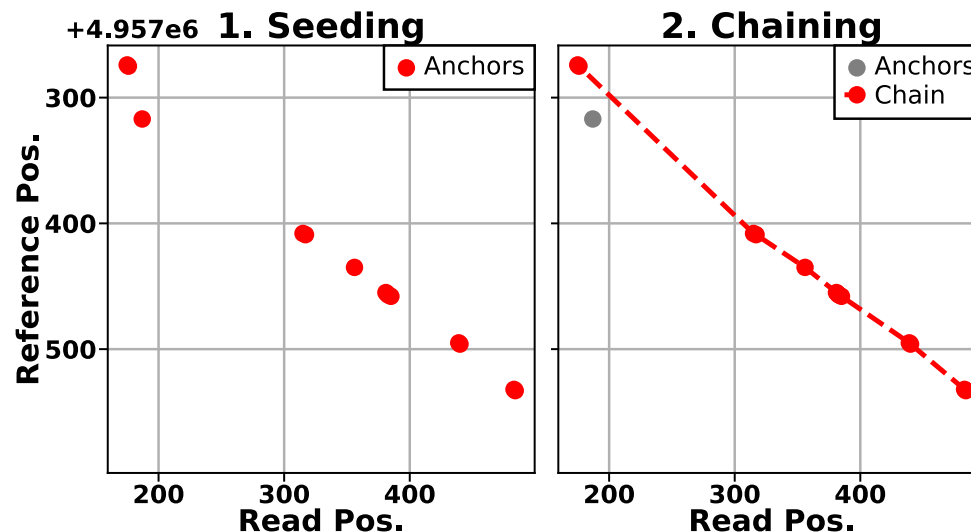


# Chaining (Two Points)

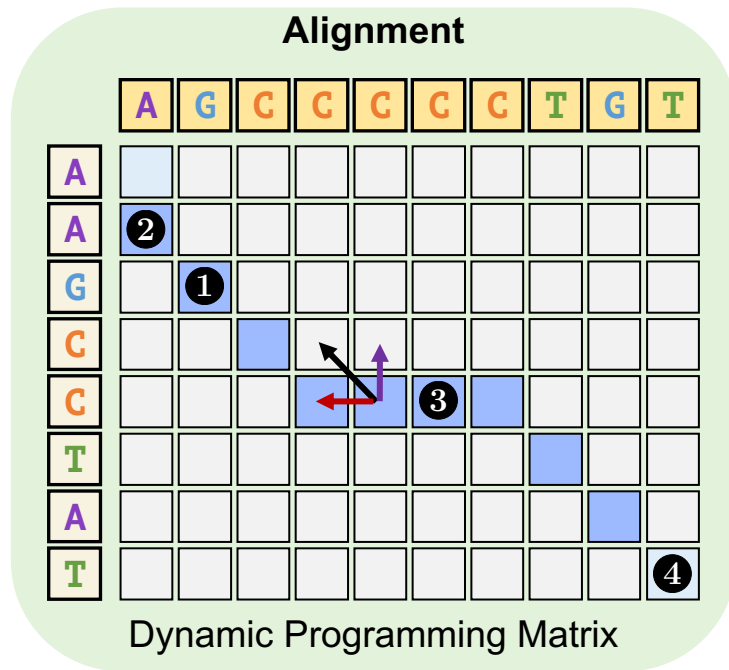


# Chaining (Multiple Points)

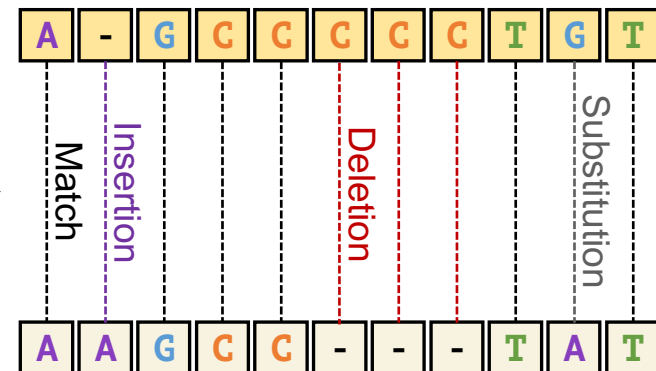
- **Exact hash value matches:** Needed for finding matching regions between a reference genome and a read
- What if there are mutations or errors?
  - **No hash (seed) match** will occur in such positions
- The chaining algorithm links **exact matches in a proximity** even though there are gaps (no seed matches) between them



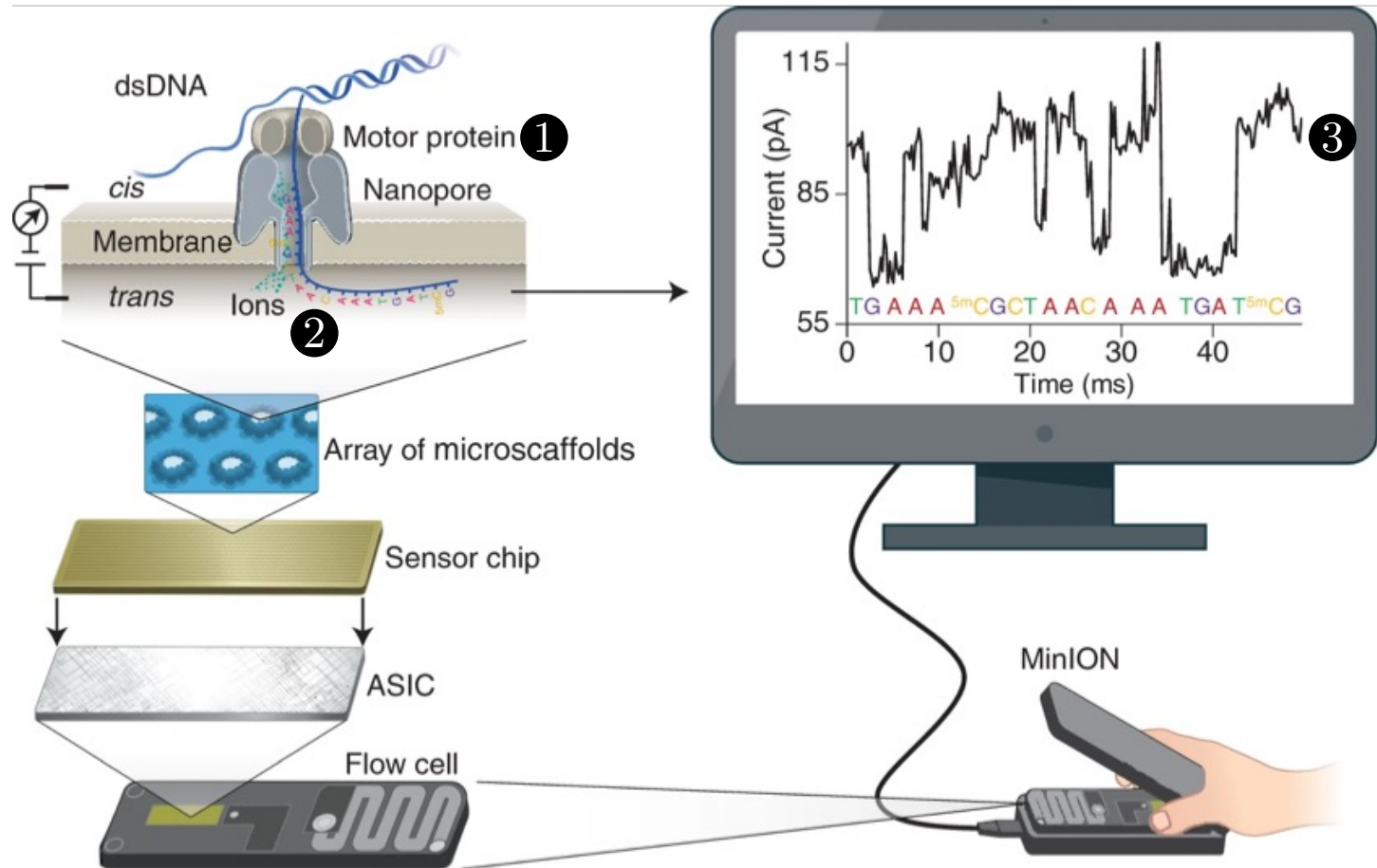
# Sequence Alignment



5



# Nanopore Sequencing

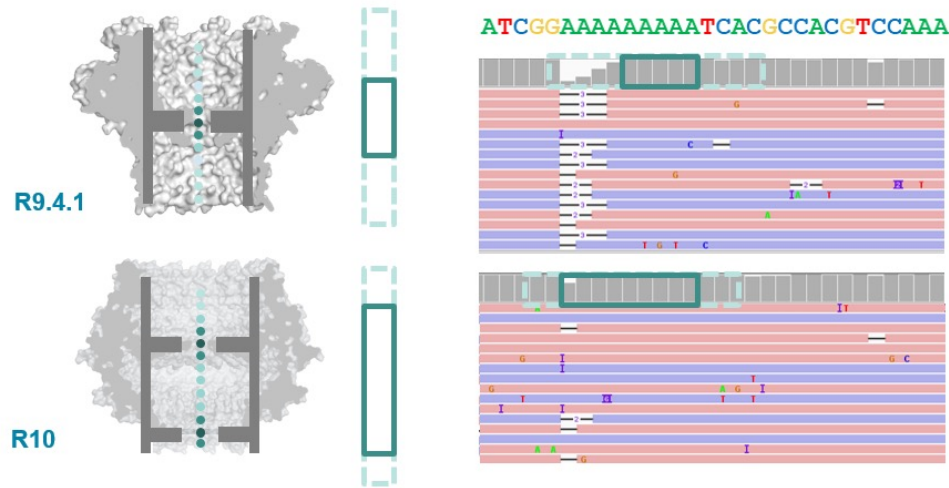


# Source of Noise in Nanopore Sequencing

- **Stochastic thermal fluctuations in the ionic current**
  - Random ionic movement due to inherent thermal energy (Brownian motion)
- **Variations in the translocation speed**
  - Mainly due to the motor protein
- **Environmental factors**
  - **Temperature:** Affecting enzymes including the motor protein
  - **pH levels:** Affecting charge and the shape of molecules
- **Maybe: Aging & material-related noise between nanopores**
  - Their effects potentially can be minimized with normalization techniques

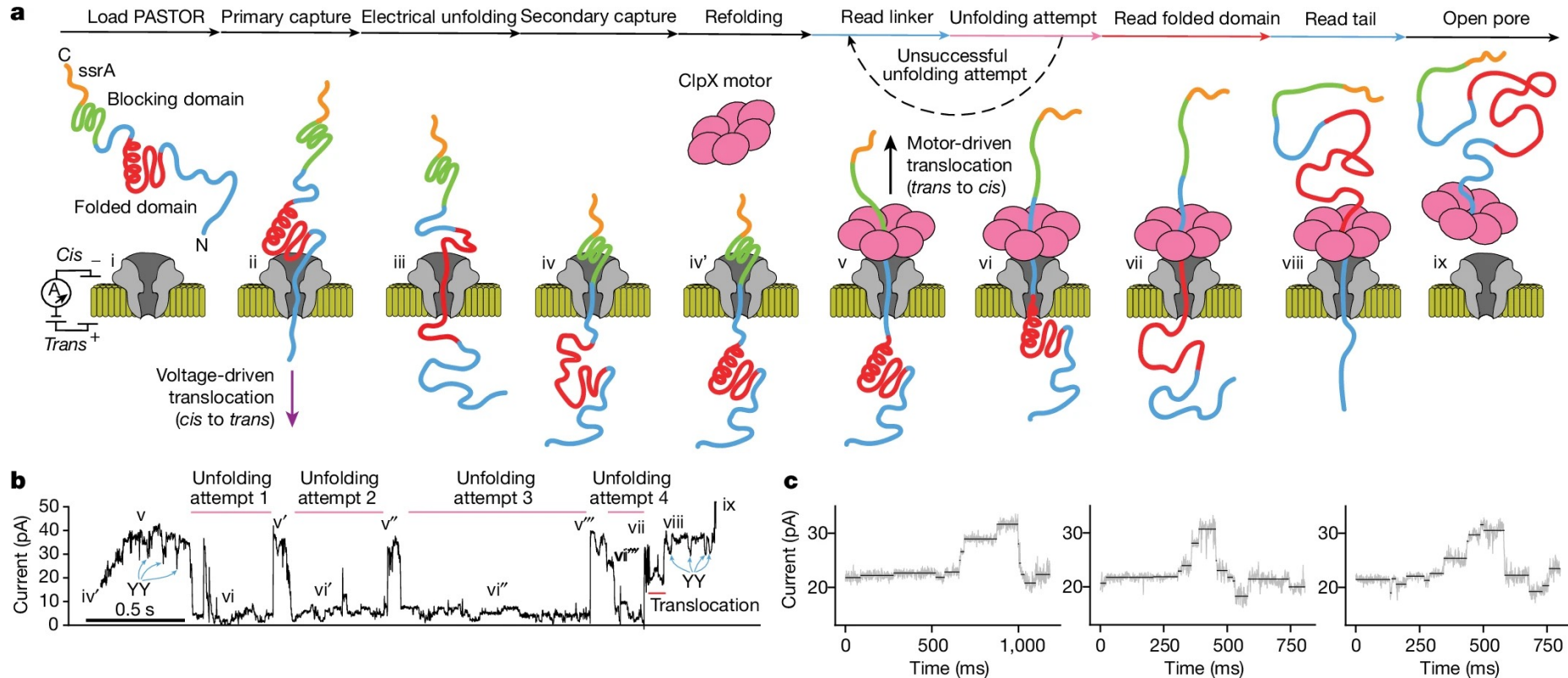
# R9 vs. R10 Chemistries

- **Dual reader head**



- **Motor protein** with more consistent translocation speed in R10
- **Duplex sequencing** in R10

# Proteomics with Nanopores



# Applications of Read Until

**Depletion:** Reads mapping to a particular reference genome is ejected

- Microbiome studies by removing host DNA
- Eliminating known residual DNA or RNA (e.g., mitochondrial DNA)
- High abundance genome removal

**Enrichment:** Reads **not** mapping to a particular reference genome is ejected

- Removing contaminated organisms
- Targeted sequencing (e.g., to a particular region of interest in the genome)
- Low abundance genome enrichment

# Applications of Run Until & Sequence Until

**Run Until:** Stopping the entire sequencing run

- Stopping when reads reach to a particular depth of coverage
- Stopping when the abundance of all genomes reach a particular threshold

**Sequence Until:** Run Until with accuracy-aware decision making

- Stopping when relative abundance estimations do not change substantially (for high-abundance genomes)
- Stopping when finding that the sample is contaminated with a particular set of genomes
- ...

# In Vitro (e.g., PCR) vs. In Silico

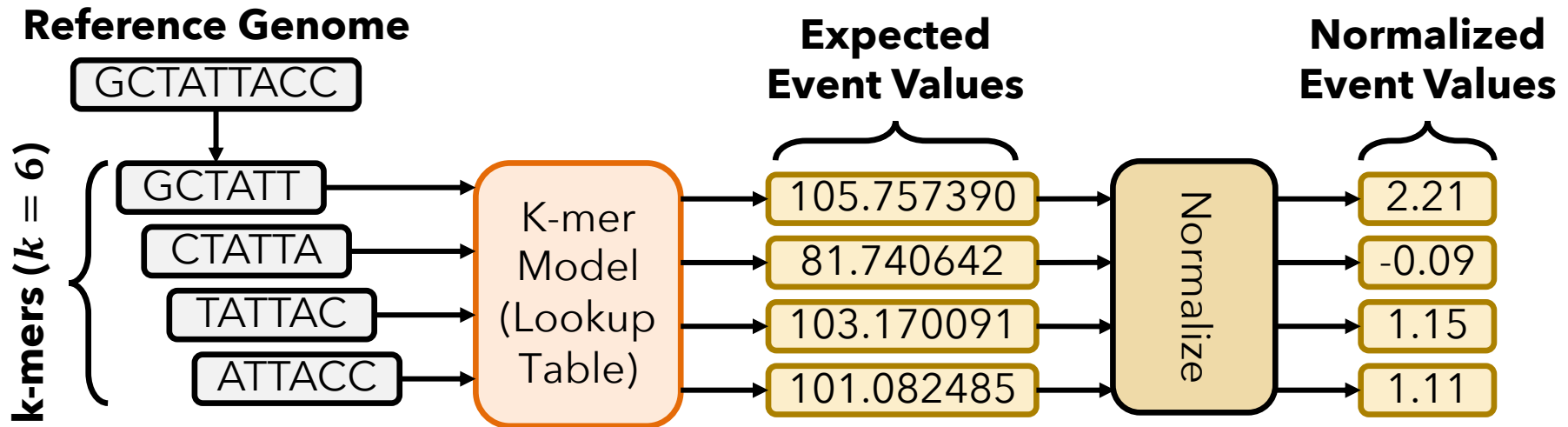
- **Polymerase Chain Reaction (PCR)** as a way of in vitro “analysis”
  - Can increase the quantity of DNA in a sample
  - **Non-dynamic** targeted sequencing (e.g., low abundance *known* targets)
  - **Requires additional resources:** Time and money for preparation and execution of PCR
- **Adaptive sampling** as a way of in silico (i.e., computational) analysis
  - **Cannot** increase the existing quantity of DNA in a sample
  - **Dynamic targeted sequencing:** Decisions can be made based on real-time analysis (e.g., Sequence Until)
  - Minimal additional resources
    - **Almost no additional resources** for preparation and execution
    - **Simultaneous** enrichment and depletion is possible
    - Better suited for rapid whole genome sequencing
  - *Beauty* of computational analysis (e.g., high flexibility – no need for primers)
- PCR and adaptive sampling can be combined depending on the analysis type

# Finding Mapping Positions

- Useful for **any application** that requires exact genomic position
  - Variant calling in downstream analysis
  - Specifically: Identifying rare variants in cancer genomics
  - Methylation profiling
- Accurate and flexible **depth of coverage estimation**
  - **Alternative: DNA quantification** (without computational analysis)
    - DNA quantification is challenging for metagenomics analysis
  - **Computational method:** We can map to almost entire set of known reference genomes to accurately estimate the coverage of a metagenomics sample
- **Transcriptome analysis**
  - Accurately quantifying expression levels & alternative splicing
- **Better resolution** (i.e., more sensitive analysis) for any other application that does not specifically require mapping positions

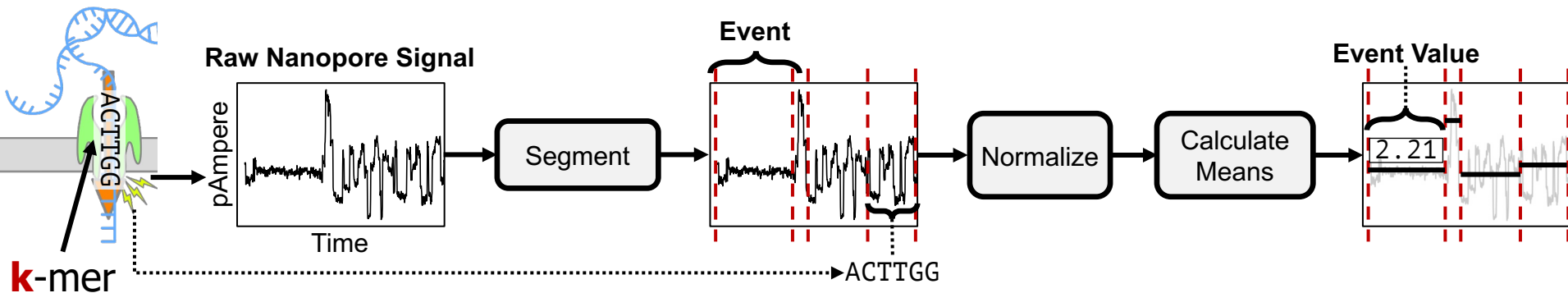
# Reference-to-Event Conversion

- **K-mer model:** Provides **expected** event values **for each k-mer**
  - Preconstructed based on nanopore sequencer characteristics
- Use the **k-mer model** to convert **all k-mers** of a reference genome to their **expected** event values



# Enabling Analysis From Electrical Signals

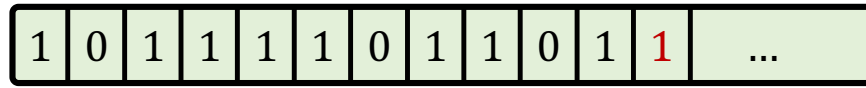
- **K** many nucleotides (**k**-mers) sequenced at a time
- **Event**: A **segment** of the raw signal
  - Corresponds to a **particular k**-mer



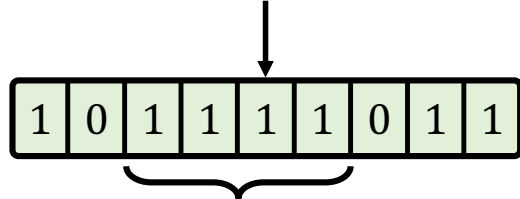
- **Observation**: Event values generated after sequencing **the same k-mer** are **similar** in value (not necessarily the same)

# Quantization -- RawHash

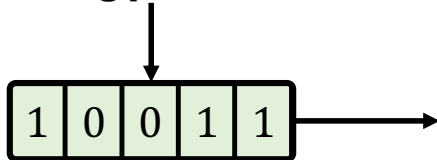
**-0.091 in Binary:**



Most significant  $Q = 9$  bits:

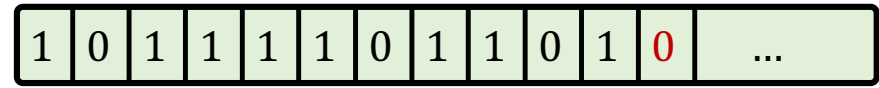


Pruning  $p = 4$  bits:

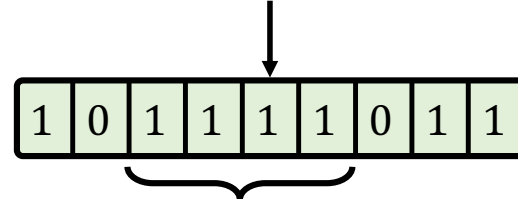


**Quantized  
Event Values**

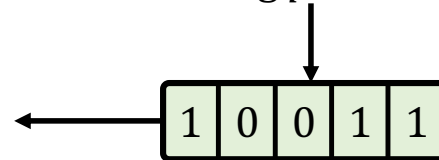
**-0.084 in Binary:**



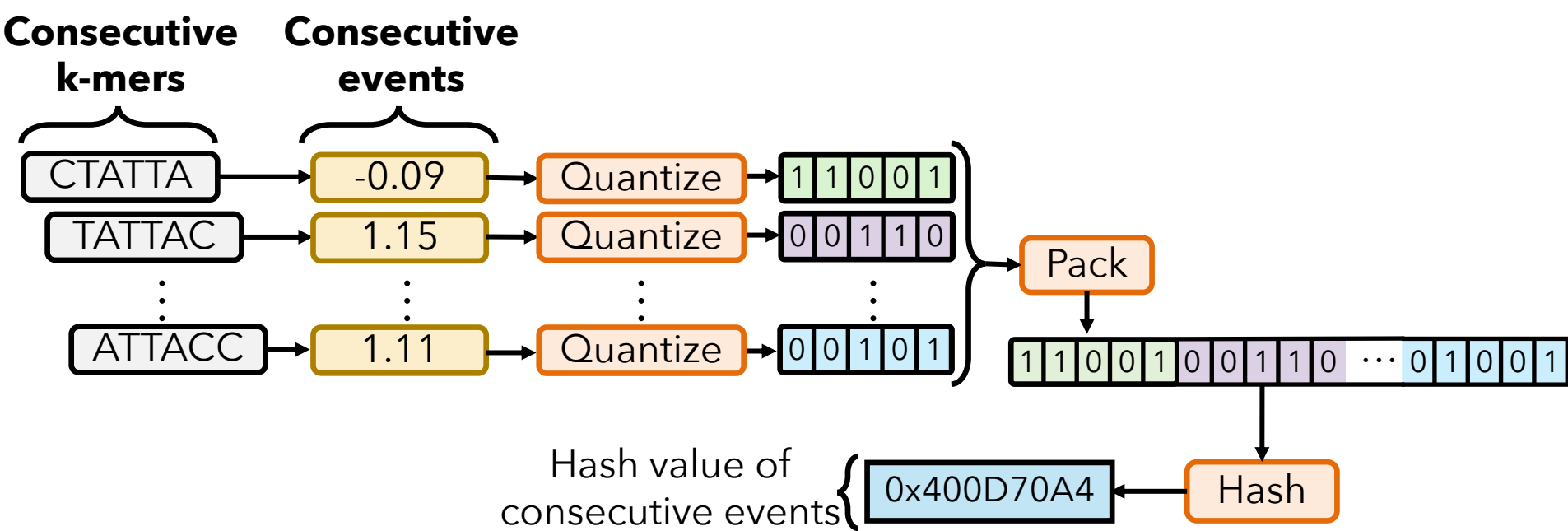
Most significant  $Q = 9$  bits:



Pruning  $p = 4$  bits:



# Packing and Hashing



# Sequence Until – RawHash & UNCALLED

| Tool             | Estimated Relative Abundance Ratios |                |              |                    |              | Distance      |
|------------------|-------------------------------------|----------------|--------------|--------------------|--------------|---------------|
|                  | <i>SARS-CoV-2</i>                   | <i>E. coli</i> | <i>Yeast</i> | <i>Green Algae</i> | <i>Human</i> |               |
| Ground Truth     | 0.0929                              | 0.4365         | 0.0698       | 0.1179             | 0.2828       | N/A           |
| UNCALLED (25%)   | 0.0026                              | 0.5890         | 0.0613       | 0.1332             | 0.2139       | 0.1910        |
| RawHash (25%)    | 0.0271                              | 0.4853         | 0.0920       | 0.0786             | 0.3170       | <b>0.0995</b> |
| UNCALLED (10%)   | 0.0026                              | 0.5906         | 0.0611       | 0.1316             | 0.2141       | 0.1920        |
| RawHash (10%)    | 0.0273                              | 0.4869         | 0.0963       | 0.0772             | 0.3124       | <b>0.1004</b> |
| UNCALLED (1%)    | 0.0026                              | 0.5750         | 0.0616       | 0.1506             | 0.2103       | 0.1836        |
| RawHash (1%)     | 0.0259                              | 0.4783         | 0.0987       | 0.0882             | 0.3088       | <b>0.0928</b> |
| UNCALLED (0.1%)  | 0.0040                              | 0.4565         | 0.0380       | 0.1910             | 0.3105       | 0.1242        |
| RawHash (0.1%)   | 0.0212                              | 0.5045         | 0.1120       | 0.0810             | 0.2814       | <b>0.1136</b> |
| UNCALLED (0.01%) | 0.0000                              | 0.5551         | 0.0000       | 0.0000             | 0.4449       | 0.2602        |
| RawHash (0.01%)  | 0.0906                              | 0.6122         | 0.0000       | 0.0000             | 0.2972       | <b>0.2232</b> |

# Sequence Until – RawHash

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

# Presets

| Preset (-x) | Corresponding parameters | Usage                             |
|-------------|--------------------------|-----------------------------------|
| viral       | -e 5 -q 9 -l 3           | Viral genomes                     |
| sensitive   | -e 6 -q 9 -l 3           | Small genomes (i.e., < 50M bases) |
| fast        | -e 7 -q 9 -l 3           | Large genomes (i.e., > 50M bases) |

# Versions – RawHash

| Tool     | Version |
|----------|---------|
| RawHash  | 0.9     |
| UNCALLED | 2.2     |
| Sigmap   | 0.1     |
| Minimap2 | 2.24    |

# Related Works

- **Basecalled real-time analysis**

- ReadFish, ReadBouncer, RUBRIC: Basecalled read mapping
- SPUMONI, SPUMONI 2: Basecalled binary classification using r-index
- Coriolis: Basecalled metagenomics classification
- baseLess: k-mer calling for classification

- **Raw signal analysis without basecalling**

- SquiggleNet, DeepSelectNet, RawMap: Target/non-target classification
- Sigmoni: Target/non-target classification using r-index
- UNCALLED, Sigmap, RawHash: Read mapping

# Adaptive Quantization

$$q(s) = \begin{cases} \lfloor n \times (f_r \times \frac{(s-f_{min})}{f_{max}-f_{min}}) \rfloor & \text{if } f_{min} \leq s \leq f_{max} \\ \lfloor n \times (f_r + c_r \times s) \rfloor & \text{if } s < f_{min} \\ \lfloor n \times (f_r + c_r + c_r \times s) \rfloor & \text{if } s > f_{max} \end{cases}$$

# Chaining Scores – RawHash vs RawHash2

- **RawHash Chaining**

$$f(i) = \max \left\{ \max_{i>j\geq 1} \{f(j) + \alpha(j, i)\}, w_i \right\}$$

$$\alpha(j, i) = \min \left\{ \min\{y_i - y_j, x_i - x_j\}, w_i \right\}$$

- **RawHash2 Chaining**

$$f(i) = \max \left\{ \max_{i>j\geq 1} \{f(j) + \alpha(j, i) - \beta(j, i)\}, w_i \right\}$$

$$\beta(j, i) = \gamma_c((y_i - y_j) - (x_i - x_j))$$

$$\gamma_c(l) = \begin{cases} 0.01 \cdot \bar{w} \cdot |l| + 0.5 \log_2 |l| & (l \neq 0) \\ 0 & (l = 0) \end{cases}$$

# Datasets

| Organism                      | Device Type        | Flow Cell Type | Transloc. Speed         | Sampling Frequency | Basecaller Model | Reads (#)              | Bases (#) | SRA Accession | Reference Genome | Genome Size     |        |
|-------------------------------|--------------------|----------------|-------------------------|--------------------|------------------|------------------------|-----------|---------------|------------------|-----------------|--------|
| Read Mapping                  |                    |                |                         |                    |                  |                        |           |               |                  |                 |        |
| D1                            | <i>SARS-CoV-2</i>  | MinION         | R9.4.1 e8 (FLO-MIN106)  | 450                | 4000             | Guppy HAC v3.2.6       | 1,382,016 | 594M          | CADDE Centre     | GCF_009858895.2 | 29,903 |
| D2                            | <i>E. coli</i>     | GridION        | R9.4.1 e8 (FLO-MIN106)  | 450                | 4000             | Guppy HAC v5.0.12      | 353,317   | 2,365M        | ERR9127551       | GCA_000007445.1 | 5M     |
| D3                            | <i>Yeast</i>       | MinION         | R9.4.1 e8 (FLO-MIN106)  | 450                | 4000             | Albacore v2.1.7        | 49,989    | 380M          | SRR8648503       | GCA_000146045.2 | 12M    |
| D4                            | <i>Green Algae</i> | PromethION     | R9.4.1 e8 (FLO-PRO002)  | 450                | 4000             | Albacore v2.3.1        | 29,933    | 609M          | ERR3237140       | GCF_000002595.2 | 111M   |
| D5                            | <i>Human</i>       | MinION         | R9.4.1 e8 (FLO-MIN106)  | 450                | 4000             | Guppy Flip-Flop v2.3.8 | 269,507   | 1,584M        | FAB42260         | T2T-CHM13 (v2)  | 3,117M |
| D6                            | <i>E. coli</i>     | GridION        | R10.4 e8.1 (FLO-MIN112) | 450                | 4000             | Guppy HAC v5.0.16      | 1,172,775 | 6,123M        | ERR9127552       | GCA_000007445.1 | 5M     |
| D7                            | <i>S. aureus</i>   | GridION        | R10.4 e8.1 (FLO-MIN112) | 450                | 4000             | Dorado SUP v0.5.3      | 407,727   | 1,281M        | SRR21386013      | GCF_000144955.2 | 2.8M   |
| Contamination Analysis        |                    |                |                         |                    |                  |                        |           |               |                  |                 |        |
| D1 and D5                     |                    |                |                         |                    |                  |                        | 1,651,523 | 2,178M        | D1 and D5        | D1              | 29,903 |
| Relative Abundance Estimation |                    |                |                         |                    |                  |                        |           |               |                  |                 |        |
| D1-D5                         |                    |                |                         |                    |                  |                        | 2,084,762 | 5,531M        | D1-D5            | D1-D5           | 3,246M |

# Accuracy

| Dataset        | Metric    | RH2           | RH2-Min. | RH     | UNCALLED | Sigmap |
|----------------|-----------|---------------|----------|--------|----------|--------|
| SARS-CoV-2     | F1        | <b>0.9867</b> | 0.9691   | 0.9252 | 0.9725   | 0.7112 |
| E. coli        | F1        | <b>0.9748</b> | 0.9631   | 0.9280 | 0.9731   | 0.9670 |
| Yeast          | F1        | <b>0.9602</b> | 0.9472   | 0.9060 | 0.9407   | 0.9469 |
| Green Algae    | F1        | <b>0.9351</b> | 0.9191   | 0.8114 | 0.8277   | 0.9350 |
| Human          | F1        | <b>0.7599</b> | 0.6699   | 0.5574 | 0.3197   | 0.3269 |
| Contamination  | Precision | <b>0.9595</b> | 0.9424   | 0.8702 | 0.9378   | 0.7856 |
| Rel. Abundance | Distance  | <b>0.2678</b> | 0.4243   | 0.4385 | 0.6812   | 0.5430 |

# Mapping Accuracy – Radar

RawHash2

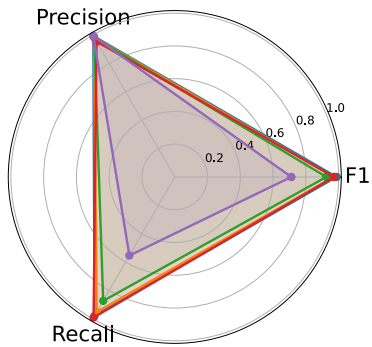
RawHash2-Minimizer

RawHash

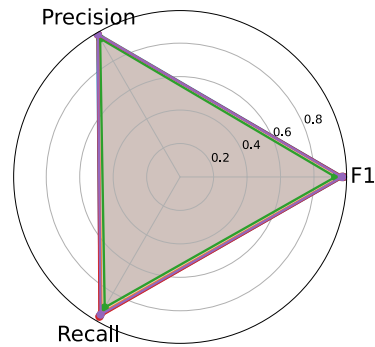
UNCALLED

Sigmap

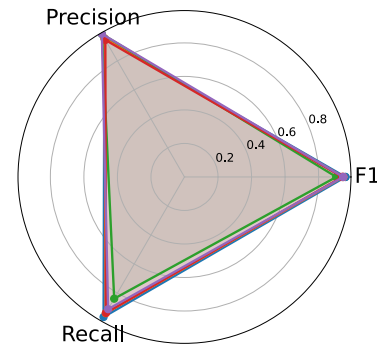
**SARS-CoV-2**



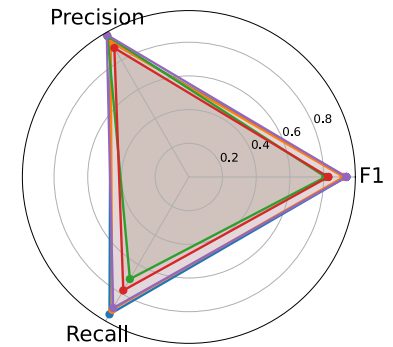
**E. coli**



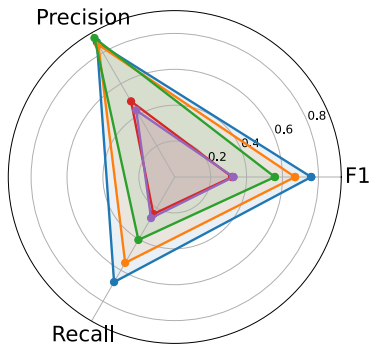
**Yeast**



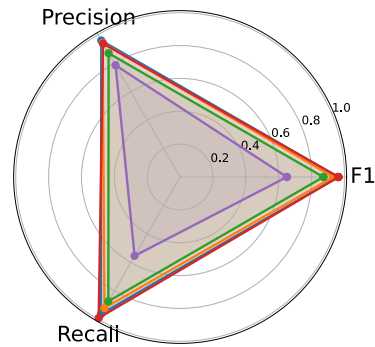
**Green Algae**



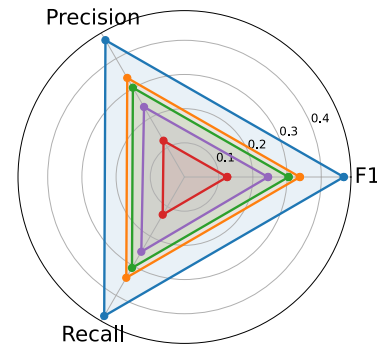
**Human**



**Contamination**



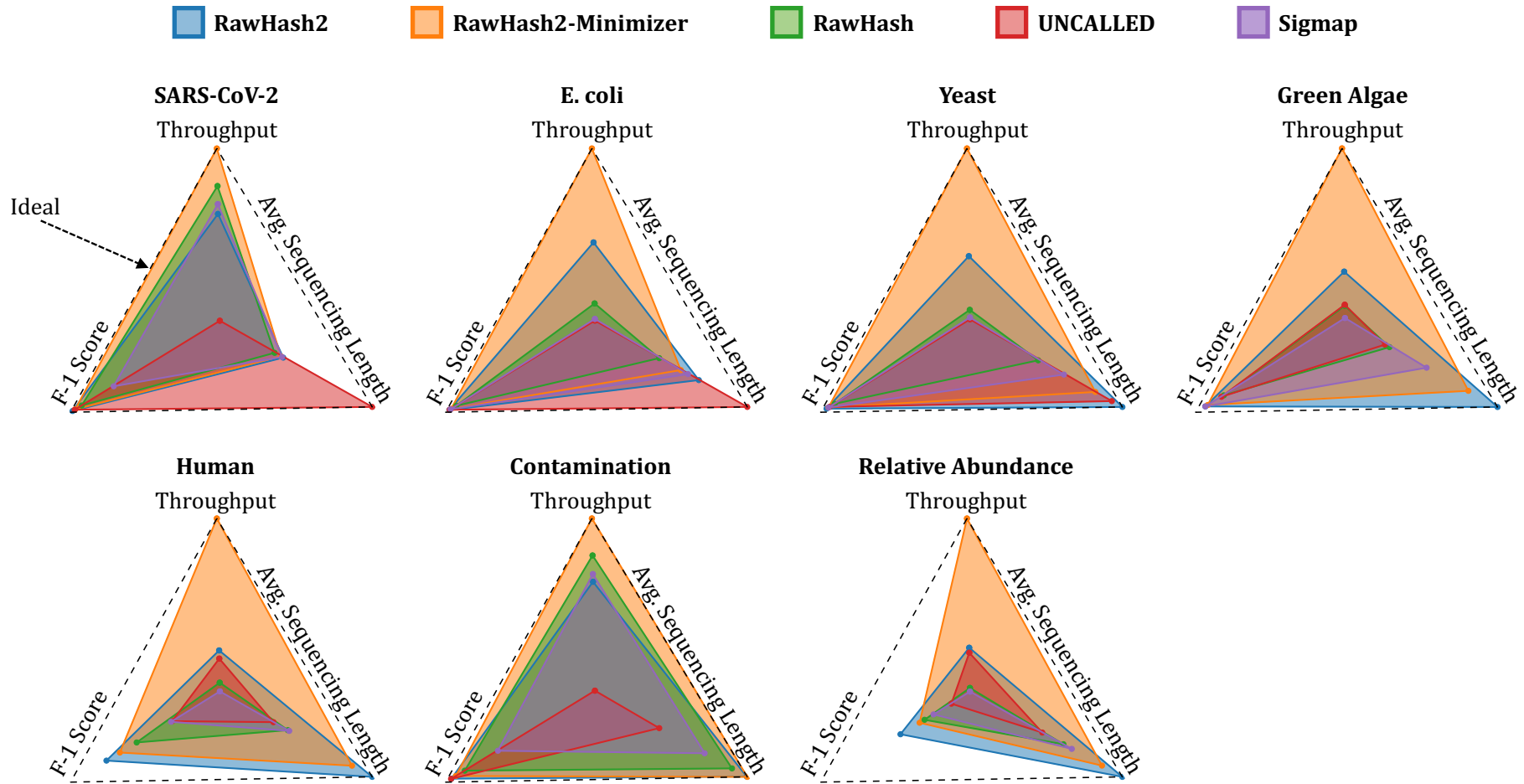
**Relative Abundance**



# Mapping Accuracy – All Metrics

| Dataset        | Metric    | RH2           | RH2-Min. | RH            | UNCALLED      | Sigmap        |
|----------------|-----------|---------------|----------|---------------|---------------|---------------|
| SARS-CoV-2     | F1        | <b>0.9867</b> | 0.9691   | 0.9252        | 0.9725        | 0.7112        |
|                | Precision | <b>0.9939</b> | 0.9868   | 0.9832        | 0.9547        | 0.9929        |
|                | Recall    | 0.9796        | 0.9521   | 0.8736        | <b>0.9910</b> | 0.5540        |
| E. coli        | F1        | <b>0.9748</b> | 0.9631   | 0.9280        | 0.9731        | 0.9670        |
|                | Precision | <b>0.9904</b> | 0.9865   | 0.9563        | 0.9817        | 0.9842        |
|                | Recall    | 0.9597        | 0.9408   | 0.9014        | <b>0.9647</b> | 0.9504        |
| Yeast          | F1        | <b>0.9602</b> | 0.9472   | 0.9060        | 0.9407        | 0.9469        |
|                | Precision | 0.9553        | 0.9561   | 0.9852        | 0.9442        | <b>0.9857</b> |
|                | Recall    | <b>0.9652</b> | 0.9385   | 0.8387        | 0.9372        | 0.9111        |
| Green Algae    | F1        | <b>0.9351</b> | 0.9191   | 0.8114        | 0.8277        | 0.9350        |
|                | Precision | 0.9284        | 0.9280   | 0.9652        | 0.8843        | <b>0.9743</b> |
|                | Recall    | <b>0.9418</b> | 0.9104   | 0.6999        | 0.7779        | 0.8987        |
| Human          | F1        | <b>0.7599</b> | 0.6699   | 0.5574        | 0.3197        | 0.3269        |
|                | Precision | 0.8675        | 0.8511   | <b>0.8943</b> | 0.4868        | 0.4288        |
|                | Recall    | <b>0.6760</b> | 0.5523   | 0.4049        | 0.2380        | 0.2642        |
| Contamination  | F1        | 0.9614        | 0.9317   | 0.8718        | <b>0.9637</b> | 0.6498        |
|                | Precision | <b>0.9595</b> | 0.9424   | 0.8702        | 0.9378        | 0.7856        |
|                | Recall    | 0.9632        | 0.9212   | 0.8736        | <b>0.9910</b> | 0.5540        |
| Rel. Abundance | F1        | <b>0.4659</b> | 0.3375   | 0.3045        | 0.1249        | 0.2443        |
|                | Precision | <b>0.4623</b> | 0.3347   | 0.3018        | 0.1226        | 0.2366        |
|                | Recall    | <b>0.4695</b> | 0.3404   | 0.3071        | 0.1273        | 0.2525        |

# Combined Benefits – Radar



# Sequenced Length

| Dataset        | RH2             | RH2-Min.      | RH       | UNCALLED      | Sigmap   |
|----------------|-----------------|---------------|----------|---------------|----------|
| SARS-CoV-2     | 443.92          | 460.85        | 513.95   | <b>184.51</b> | 452.38   |
| E. coli        | 851.31          | 1,030.74      | 1,376.14 | <b>580.52</b> | 950.03   |
| Yeast          | <b>1,147.66</b> | 1,395.87      | 2,565.09 | 1,233.20      | 1,862.69 |
| Green Algae    | <b>1,385.59</b> | 1,713.46      | 4,760.59 | 5,300.15      | 2,591.16 |
| Human          | <b>2,130.59</b> | 2,455.99      | 4,773.58 | 6,060.23      | 4,680.50 |
| Contamination  | 670.69          | <b>667.89</b> | 742.56   | 1,582.63      | 927.82   |
| Rel. Abundance | <b>1,024.28</b> | 1,182.04      | 1,669.46 | 2,158.50      | 1,533.04 |

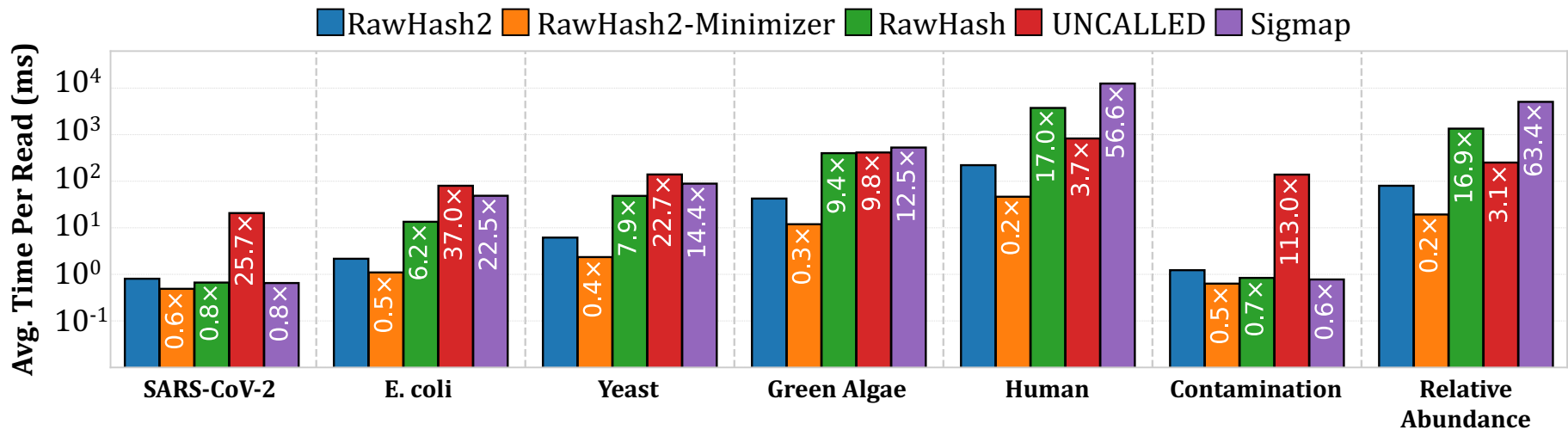
# Computational Resources #1

| Dataset                   | RH2         | RH2-Min.         | RH           | UNCALLED     | Sigmap        |
|---------------------------|-------------|------------------|--------------|--------------|---------------|
| Indexing CPU Time (sec)   |             |                  |              |              |               |
| SARS-CoV-2                | 0.12        | 0.06             | 0.16         | 8.40         | <b>0.02</b>   |
| E. coli                   | 2.48        | <b>1.61</b>      | 2.56         | 10.57        | 8.86          |
| Yeast                     | 4.56        | <b>3.02</b>      | 4.44         | 16.40        | 25.29         |
| Green Algae               | 27.60       | <b>17.73</b>     | 24.51        | 213.13       | 420.25        |
| Human                     | 1,093.56    | <b>588.30</b>    | 809.08       | 3,496.76     | 41,993.26     |
| Contamination             | 0.13        | 0.06             | 0.15         | 8.38         | <b>0.03</b>   |
| Rel. Abundance            | 747.74      | <b>468.14</b>    | 751.67       | 3,666.14     | 36,216.87     |
| Indexing Peak Memory (GB) |             |                  |              |              |               |
| SARS-CoV-2                | <b>0.01</b> | <b>0.01</b>      | <b>0.01</b>  | 0.06         | <b>0.01</b>   |
| E. coli                   | 0.35        | 0.19             | 0.35         | <b>0.11</b>  | 0.40          |
| Yeast                     | 0.75        | 0.39             | 0.76         | <b>0.30</b>  | 1.04          |
| Green Algae               | 5.11        | <b>2.60</b>      | 5.33         | 11.94        | 8.63          |
| Human                     | 80.75       | <b>40.59</b>     | 83.09        | 48.43        | 227.77        |
| Contamination             | <b>0.01</b> | <b>0.01</b>      | <b>0.01</b>  | 0.06         | <b>0.01</b>   |
| Rel. Abundance            | 152.59      | 75.62            | 152.84       | <b>47.80</b> | 238.32        |
| Mapping CPU Time (sec)    |             |                  |              |              |               |
| SARS-CoV-2                | 1,705.43    | <b>1,227.05</b>  | 1,539.64     | 29,282.90    | 1,413.32      |
| E. coli                   | 1,296.34    | <b>787.49</b>    | 7,453.21     | 28,767.58    | 22,923.09     |
| Yeast                     | 545.77      | <b>246.37</b>    | 4,145.38     | 7,181.44     | 7,146.32      |
| Green Algae               | 2,135.83    | <b>657.63</b>    | 22,103.03    | 12,593.01    | 26,778.44     |
| Human                     | 100,947.58  | <b>21,860.05</b> | 1,825,061.23 | 245,128.15   | 6,101,179.89  |
| Contamination             | 3,783.69    | <b>2,332.28</b>  | 3,480.43     | 234,199.60   | 3,011.78      |
| Rel. Abundance            | 250,076.90  | <b>62,477.76</b> | 4,551,349.79 | 569,824.13   | 15,178,633.11 |

# Computational Resources #2

| Mapping Peak Memory (GB)                                      |            |                   |            |              |            |
|---------------------------------------------------------------|------------|-------------------|------------|--------------|------------|
| SARS-CoV-2                                                    | 4.15       | 4.16              | 4.20       | <b>0.17</b>  | 28.26      |
| E. coli                                                       | 4.13       | 4.03              | 4.18       | <b>0.50</b>  | 111.12     |
| Yeast                                                         | 4.38       | 4.12              | 4.37       | <b>0.36</b>  | 14.66      |
| Green Algae                                                   | 6.11       | 4.98              | 11.77      | <b>0.78</b>  | 29.18      |
| Human                                                         | 48.75      | 25.04             | 52.43      | <b>10.62</b> | 311.94     |
| Contamination                                                 | 4.16       | 4.14              | 4.17       | <b>0.62</b>  | 111.70     |
| Rel. Abundance                                                | 49.14      | 25.82             | 54.89      | <b>8.99</b>  | 486.63     |
| Mapping Throughput (bp/sec)                                   |            |                   |            |              |            |
| SARS-CoV-2                                                    | 552,561.25 | <b>885,263.48</b> | 694,274.92 | 9,260.31     | 602,380.96 |
| E. coli                                                       | 303,382.45 | <b>659,013.57</b> | 72,281.32  | 7,515.76     | 13,750.97  |
| Yeast                                                         | 150,547.61 | <b>394,766.80</b> | 28,757.15  | 7,471.48     | 11,624.82  |
| Green Algae                                                   | 28,742.46  | <b>98,323.70</b>  | 9,488.79   | 10,069.41    | 2,569.89   |
| Human                                                         | 8,968.78   | <b>37,086.38</b>  | 2,099.35   | 7,225.67     | 236.45     |
| Contamination                                                 | 563,129.81 | <b>884,929.30</b> | 696,873.20 | 9,343.95     | 601,936.49 |
| Rel. Abundance                                                | 9,501.37   | <b>36,919.79</b>  | 962.79     | 8,437.70     | 196.48     |
| CPU Threads Needed for the entire MinION Flowcell (512 pores) |            |                   |            |              |            |
| SARS-CoV-2                                                    | <b>1</b>   | <b>1</b>          | <b>1</b>   | 25           | <b>1</b>   |
| E. coli                                                       | <b>1</b>   | <b>1</b>          | 4          | 31           | 17         |
| Yeast                                                         | 2          | <b>1</b>          | 9          | 31           | 20         |
| Green Algae                                                   | 9          | <b>3</b>          | 25         | 23           | 90         |
| Human                                                         | 26         | <b>7</b>          | 110        | 32           | 975        |
| Contamination                                                 | <b>1</b>   | <b>1</b>          | <b>1</b>   | 25           | <b>1</b>   |
| Rel. Abundance                                                | 25         | <b>7</b>          | 240        | 28           | 1173       |

# Average Time Spent per Read



# FAST5 vs. POD5. vs S/BLOW5

| Tool                 | <i>E. coli</i> | <i>Yeast</i> |
|----------------------|----------------|--------------|
| Elapsed Time (mm:ss) |                |              |
| RH2-FAST5            | 19:27          | 08:35        |
| RH2-POD5             | 16:55          | 07:33        |
| RH2-BLOW5            | 17:32          | 07:38        |
| RH2-Min.-FAST5       | 12:13          | 03:56        |
| RH2-Min.-POD5        | 09:42          | 02:56        |
| RH2-Min.-BLOW5       | 10:16          | 03:02        |

# Flow Cell Types R9 vs R10.4

| Flow Cell                                |                         | RH2        | RH2-Min.   |
|------------------------------------------|-------------------------|------------|------------|
| <b>Read Mapping Accuracy (E. coli)</b>   |                         |            |            |
| R9.4                                     | F1                      | 0.9748     | 0.9631     |
|                                          | Precision               | 0.9904     | 0.9865     |
|                                          | Recall                  | 0.9597     | 0.9408     |
| R10.4                                    | F1                      | 0.8960     | 0.8389     |
|                                          | Precision               | 0.9506     | 0.9325     |
|                                          | Recall                  | 0.8473     | 0.7623     |
| <b>Read Mapping Accuracy (S. aureus)</b> |                         |            |            |
| R10.4                                    | F1                      | 0.7749     | 0.6778     |
|                                          | Precision               | 0.8649     | 0.8167     |
|                                          | Recall                  | 0.7018     | 0.5793     |
| <b>Performance (E. coli)</b>             |                         |            |            |
| R9.4                                     | Throughput [bp/sec]     | 303,382.45 | 659,013.57 |
|                                          | Mean time per read [ms] | 2.161      | 1.099      |
| R10.4                                    | Throughput [bp/sec]     | 175,351.94 | 480,471.75 |
|                                          | Mean time per read [ms] | 6.598      | 2.505      |
| <b>Performance (S. aureus)</b>           |                         |            |            |
| R10.4                                    | Throughput [bp/sec]     | 256,680.4  | 617,308.7  |
|                                          | Mean time per read [ms] | 5.478      | 2.243      |

# Ratio of Filtered Seed Hits

| <b>Dataset</b>    | <b>Average Filtered Ratio</b> |
|-------------------|-------------------------------|
| SARS-CoV-2        | 0.0627                        |
| E. coli           | 0.5505                        |
| Yeast             | 0.5356                        |
| Green Algae       | 0.8106                        |
| Human             | 0.5104                        |
| E. coli (R10.4)   | 0.6895                        |
| S. aureus (R10.4) | 0.6003                        |

# Presets

| Preset                  | Corresponding parameters                                                                                                            | Usage                              |
|-------------------------|-------------------------------------------------------------------------------------------------------------------------------------|------------------------------------|
| viral                   | -e 6 -q 4 -max-chunks 5 -bw 100 -max-target-gap 500<br>-max-target-gap 500 -min-score 10 -chain-gap-scale 1.2 -chain-skip-scale 0.3 | Viral genomes                      |
| sensitive               | -e 8 -q 4 -fine-range 0.4                                                                                                           | Small genomes (i.e., < 500M bases) |
| fast                    | -e 8 -q 4 -max-chunks 20                                                                                                            | Large genomes (i.e., > 500M bases) |
| Other helper parameters |                                                                                                                                     |                                    |
| depletion               | -best-chains 5 -min-mapq 10 -w-threshold 0.5<br>-min-anchors 2 -min-score 15 -chain-skip-scale 0                                    | Contamination analysis             |
| r10                     | -k9 -seg-window-length1 3 -seg-window-length2 6 -seg-threshold1 6.5<br>-seg-threshold2 4 -seg-peak-height 0.2 -chain-gap-scale 1.2  | For R10.4 Flow Cells               |

# Versions

| Tool         | Version    |
|--------------|------------|
| RawHash2     | 2.1        |
| RawHash      | 1.0        |
| UNCALLED     | 2.3        |
| Sigmap       | 0.1        |
| Minimap2     | 2.24       |
| FAST5 (HDF5) | 1.10       |
| POD5         | 0.2.2      |
| S/BLOW5      | 1.2.0-beta |