



Rawsamble

Overlapping and Assembling
Raw Nanopore Signals

Using a Hash-based Seeding Mechanism

Can Firtina

Maximilian Mordig

Joel Lindegger

Harun Mustafa

Sayan Goswami

Stefano Mercogliano

Yan Zhu

Andre Kahles

Onur Mutlu

SAFARI

ETH zürich



**BIOMEDICAL
INFORMATICS**



Executive Summary

Problem: Existing solutions **cannot** interpret raw signals directly **if a reference genome is unknown or does not exist**

Goal: Enable raw signal analysis **without a reference genome**

Key Contributions:

1. **Rawsamble: The first mechanism** that can find **all-vs-all overlapping** pairs between raw nanopore signals
2. **The first *de novo* assembly** constructed directly from raw signal overlaps **without basecalling**
3. **A new assembler** to build and output the assemblies of signals

Key Results: Across 3 genomes of varying sizes, Rawsamble provides

- **Throughput: 139× - 1031×** faster with one thread compared to a single pore
- **Overlap statistics: 37%** of overlapping pairs **shared with minimap2**
- **Assembly:** Unitigs of length **up to one million nucleotides** from overlapping raw signals **without basecalling**

Outline

Background

Rawsamble Mechanism

Evaluation

Conclusion

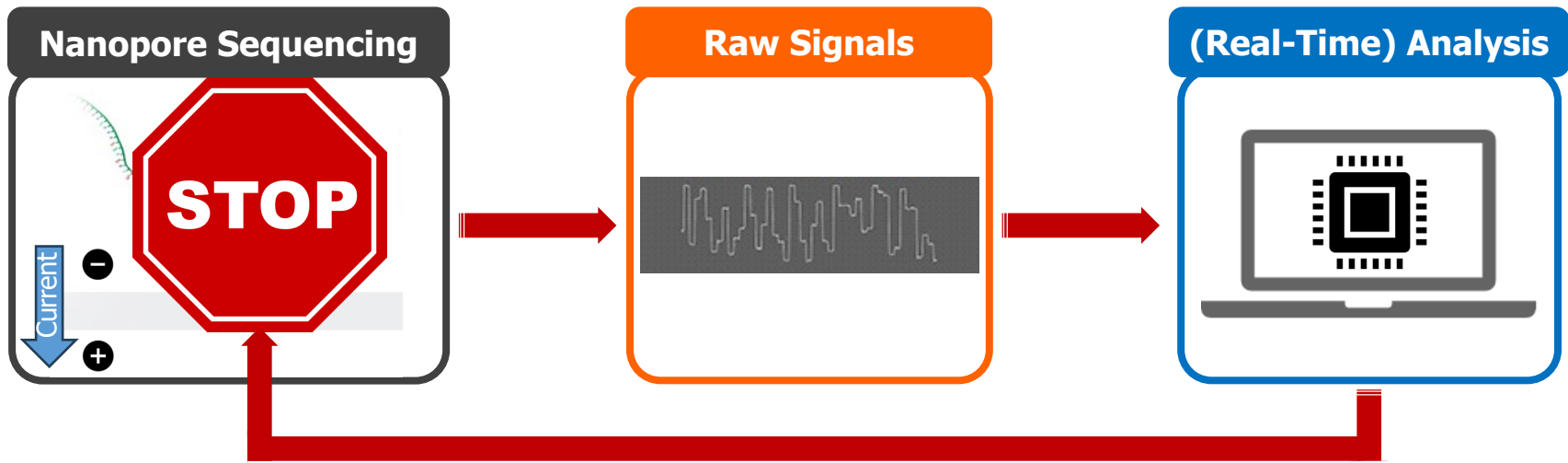
Nanopore Sequencing

Nanopore Sequencing: a widely used sequencing technology

- Can sequence large fragments of nucleic acid molecules (up to ~4Mbp)
- Offers high throughput
- Cost-effective
- Enables **real-time and portable genome analysis**



Nanopore Sequencing



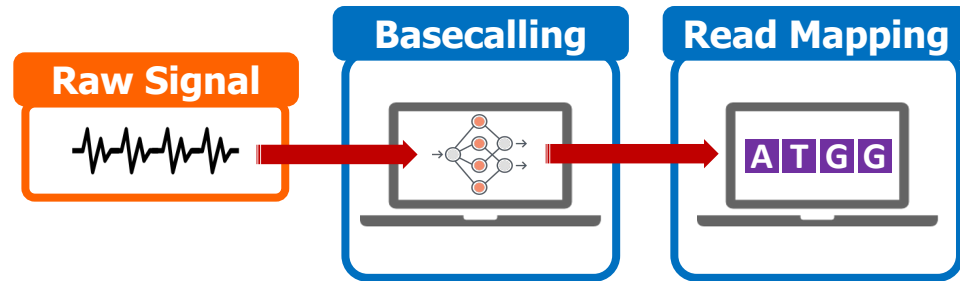
Raw Signals: Ionic current measurements generated at a certain **throughput**

(Real-Time) Analysis: Signals can be analyzed while they are generated

Real-Time Decisions: Stopping sequencing **early** based on real-time analysis

Analyzing Raw Nanopore Signals

Traditional: Translating (**basecalling**) signals to bases **before** analysis

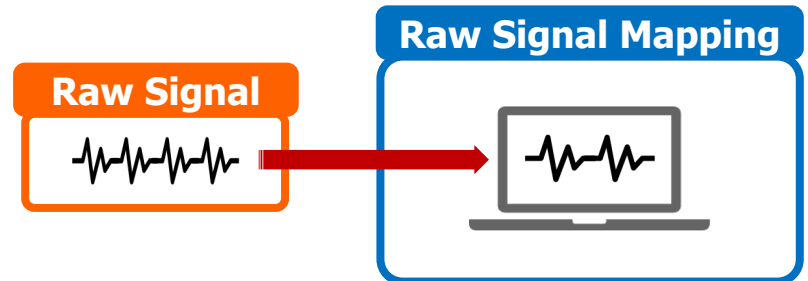


✓ Basecalled sequences are less noisy than raw signals

✓ Many analysis tools use basecalled sequences

✗ Costly and power-hungry computational requirements

Recent Work: Directly analyzing signals **without basecalling**

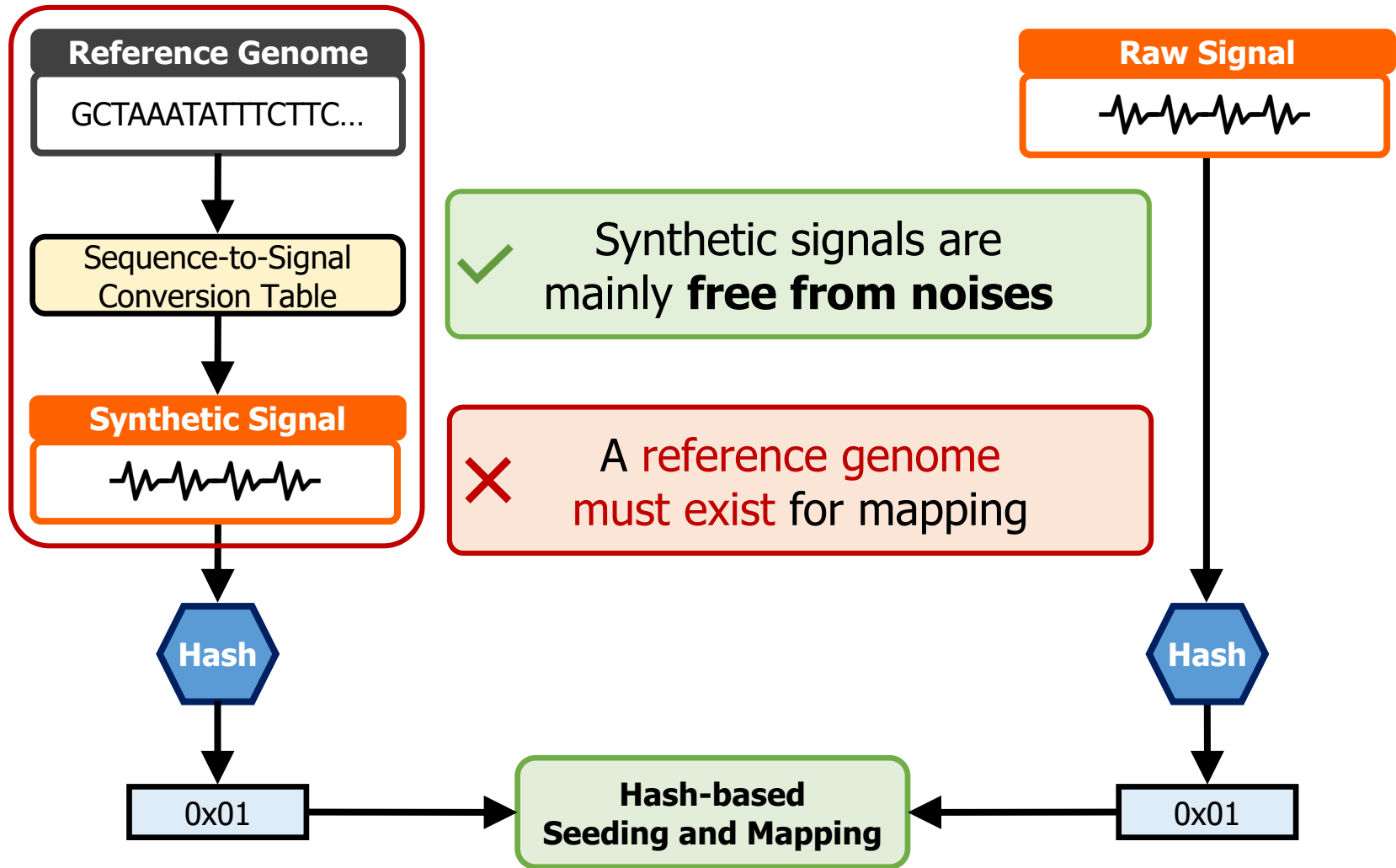


✓ Efficient analysis with better scalability and portability

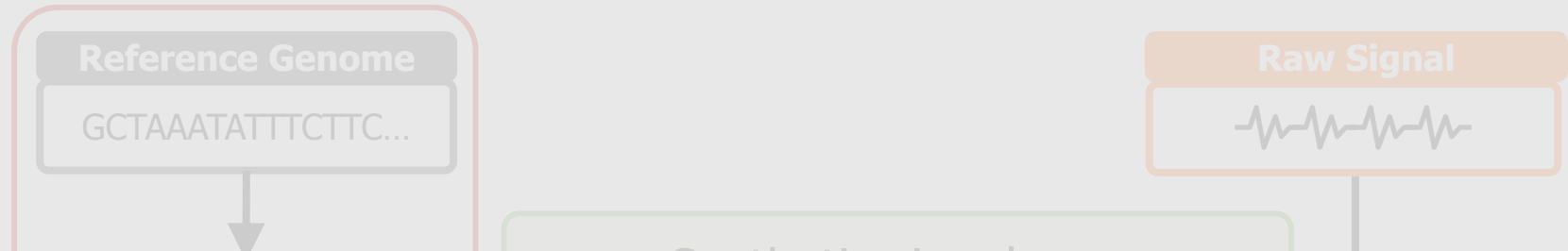
✓ Raw signals retain more information than just bases

✗ Lack of established tools for downstream analysis

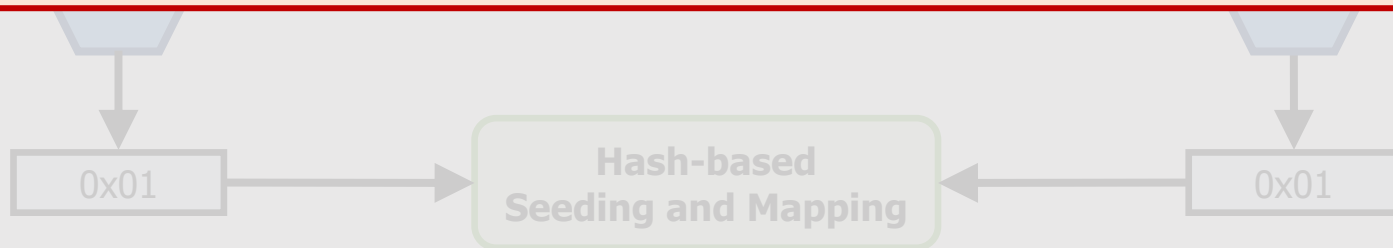
The State-of-the-Art Raw Signal Mapper



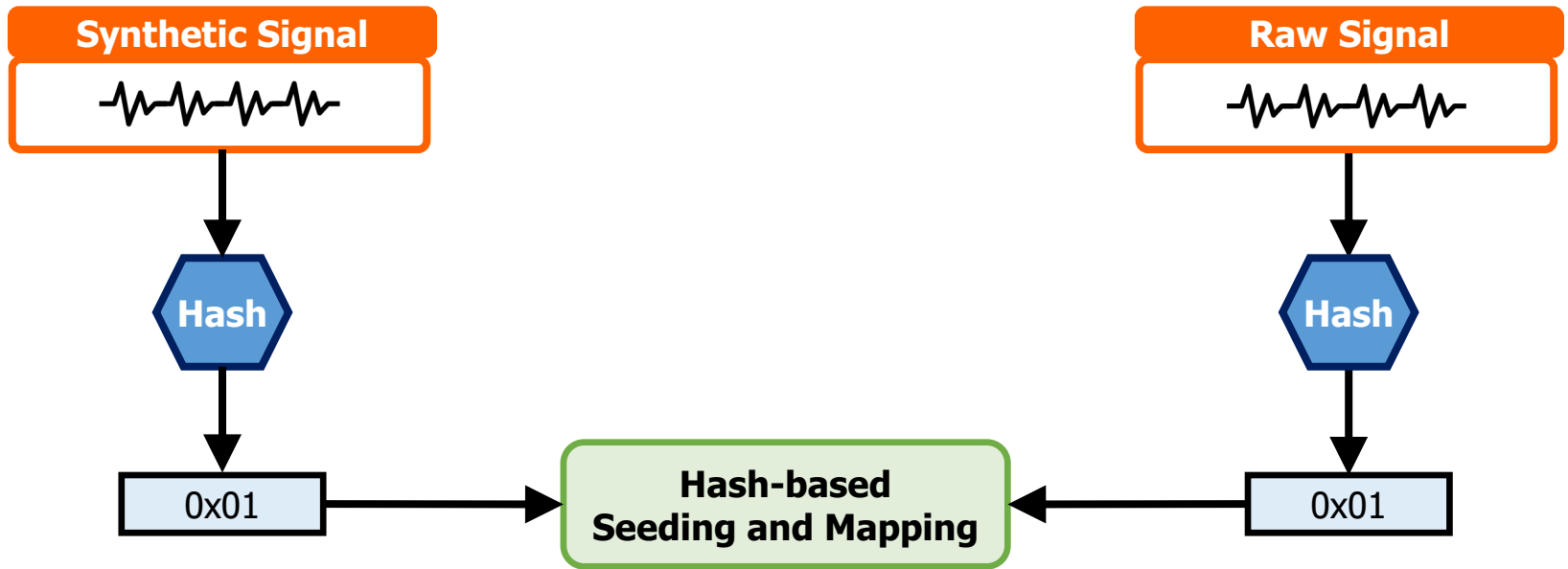
The State-of-the-Art Raw Signal Mapper



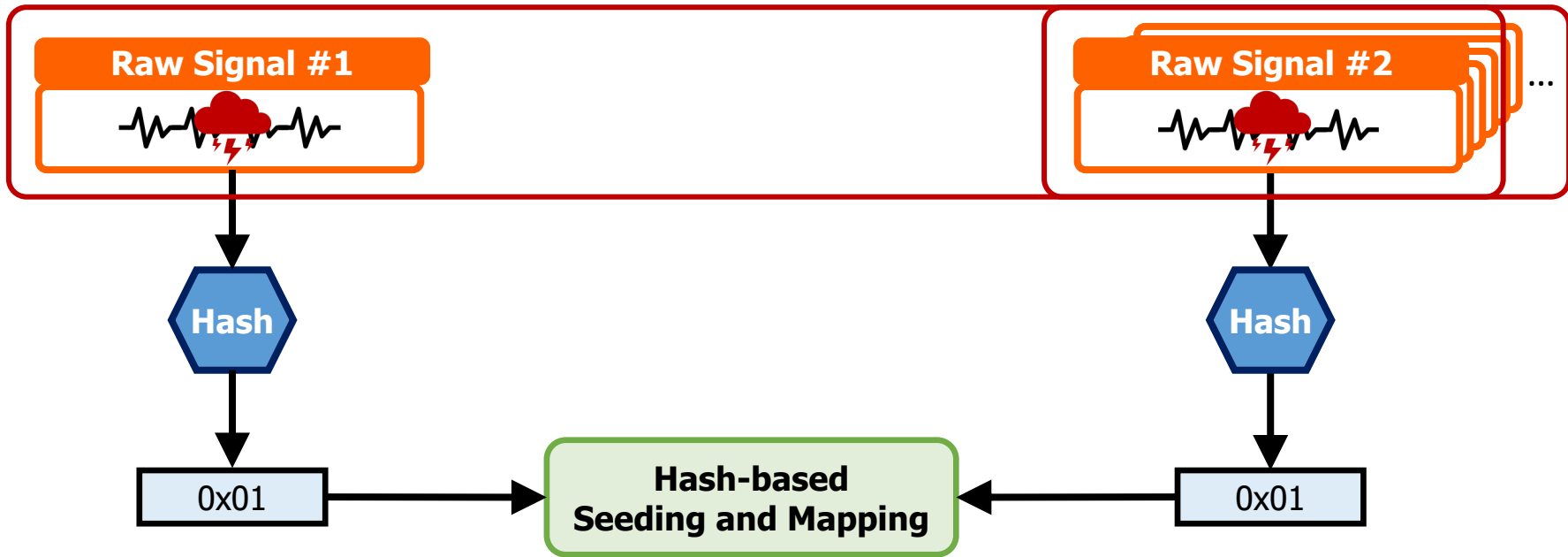
Existing solutions **cannot** analyze raw signals directly **without a reference genome**



Beyond Reference Mapping: Overlapping



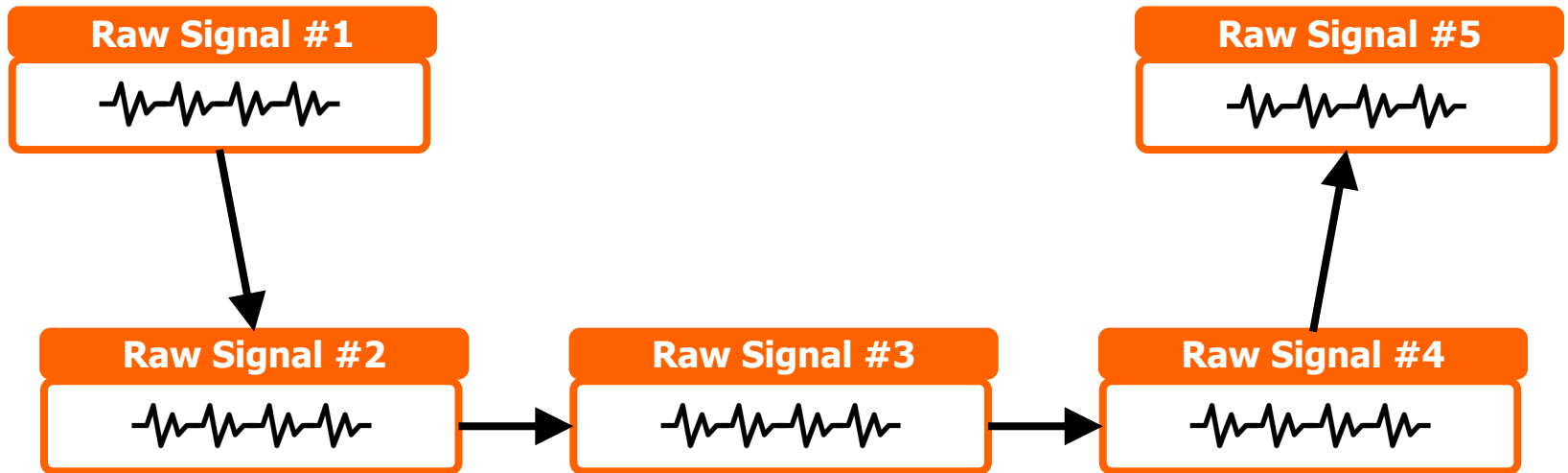
Beyond Reference Mapping: Overlapping



Challenge #1: Identifying accurate matches when **both** signals are noisy

Challenge #2: Finding **all** overlapping read pairs

Beyond Reference Mapping: Overlapping



Challenge #1: Identifying accurate matches when **both** signals are noisy

Challenge #2: Finding **all** overlapping read pairs

Challenge #3: Generating long paths from many overlaps to build assemblies

Outline

Background

Rawsamble Mechanism

Evaluation

Conclusion

Goal

Enable raw signal analysis
without a reference genome



Rawsamble

The first mechanism that can quickly and accurately find **all-vs-all overlapping of raw signals**

The first *de novo* assembly constructed directly from raw signal overlaps **without basecalling**

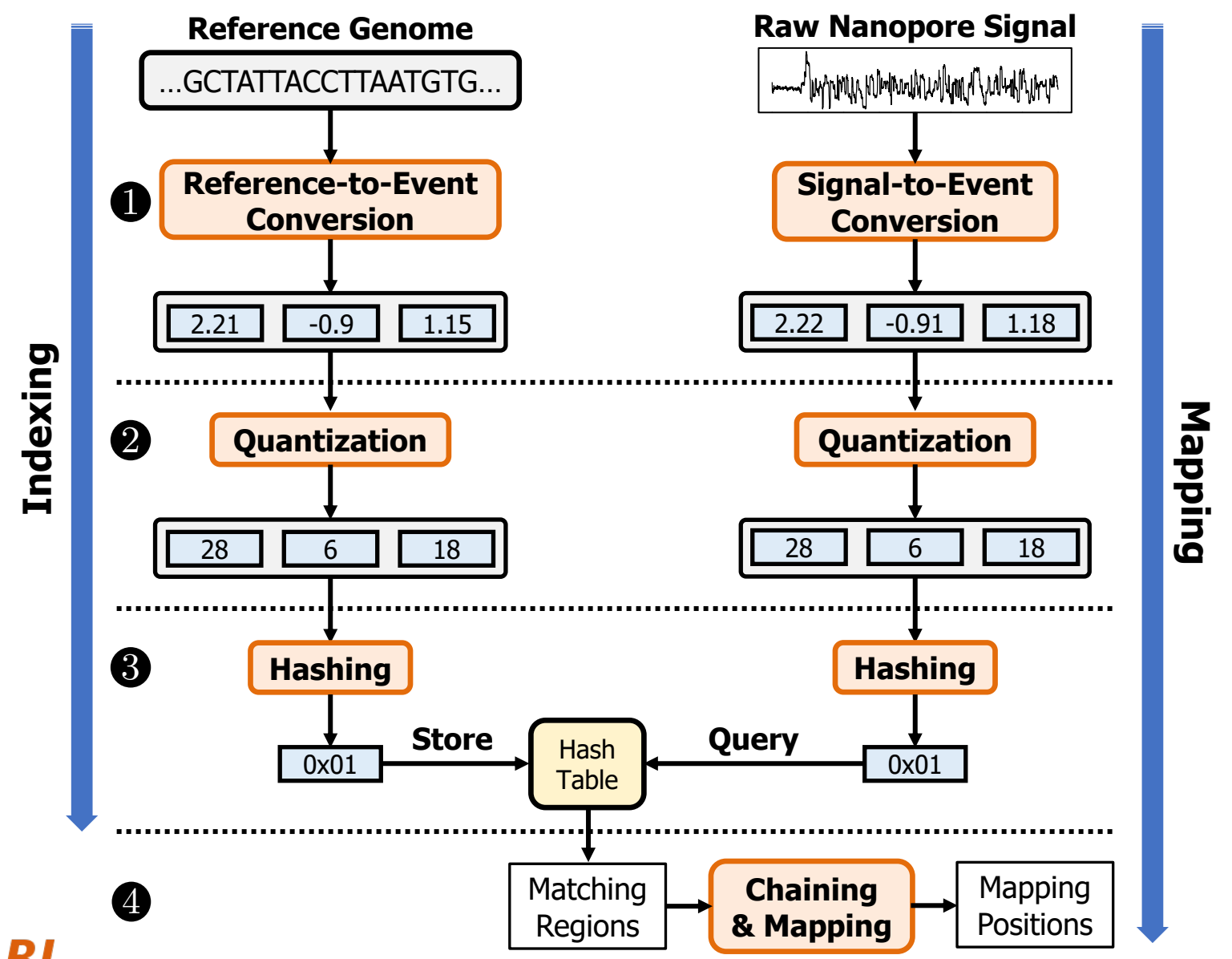
A new assembler to build and output the assemblies from raw signals

Rawsamble Key Ideas

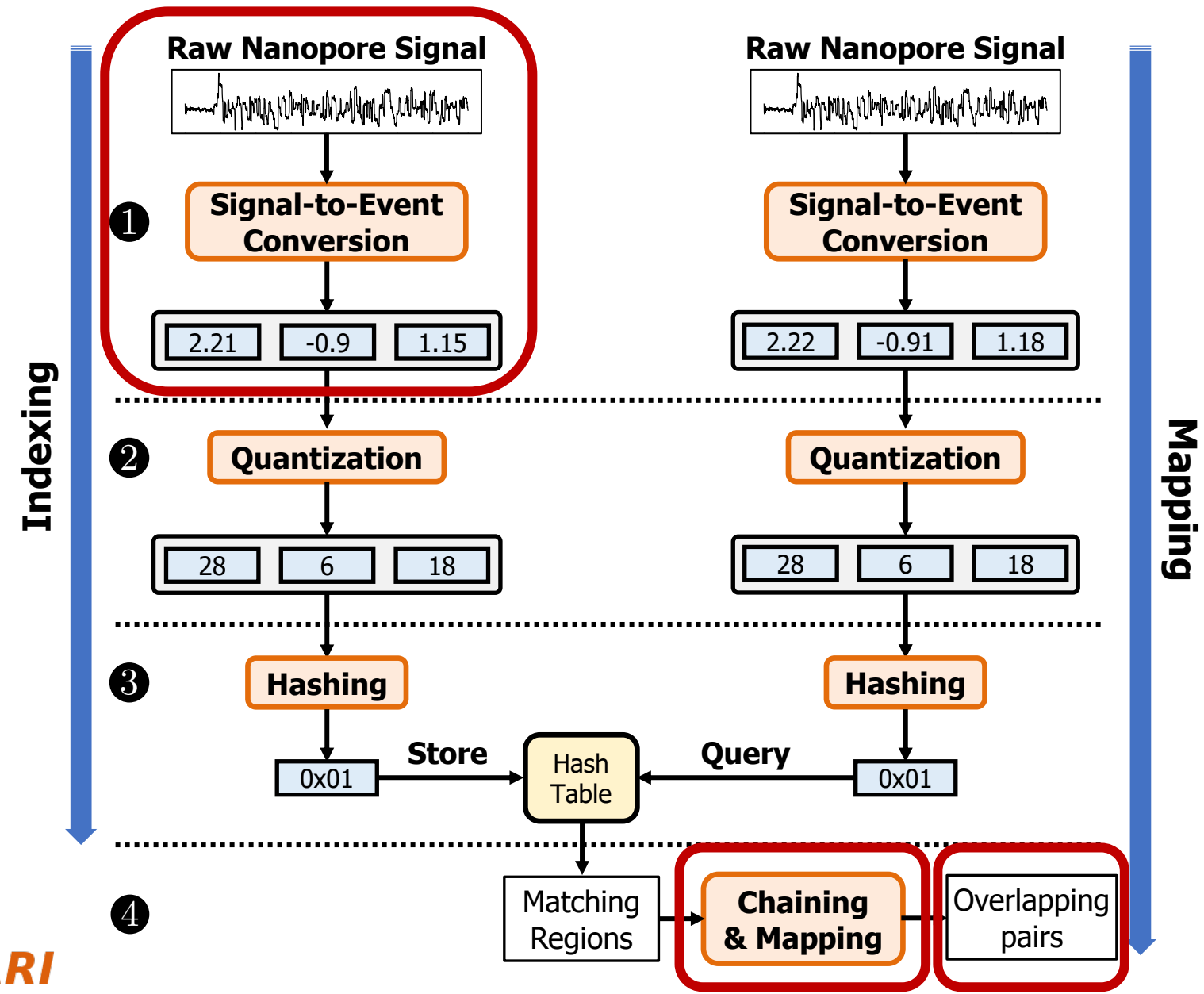
Build on the existing state-of-the-art raw signal mapper: **RawHash**

Extend RawHash to **support overlapping**

RawHash Overview

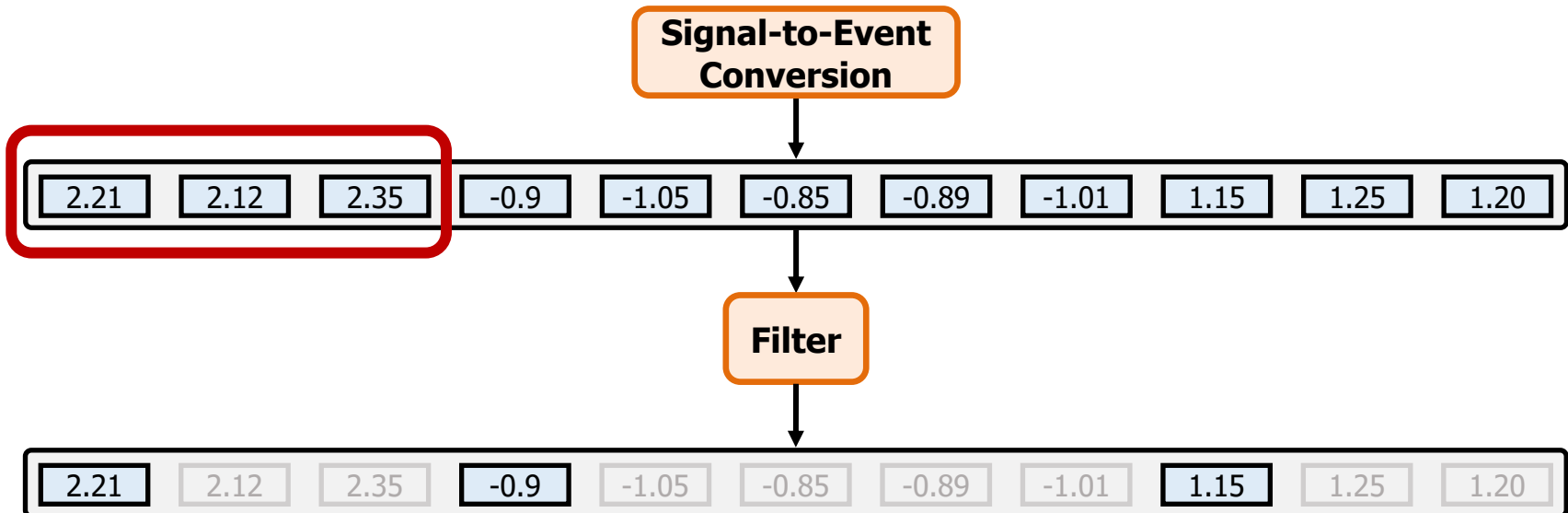
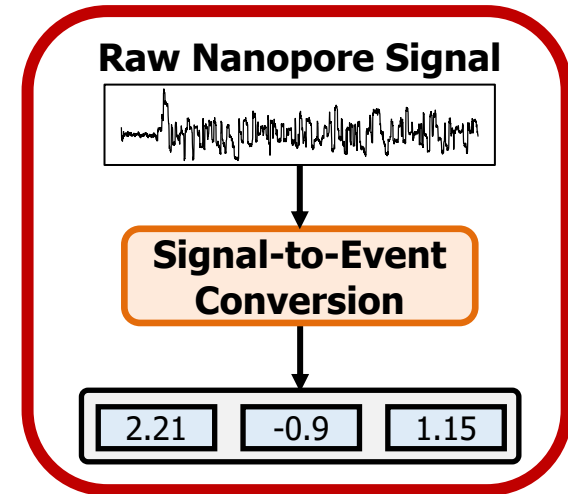


Rawsamblе Overview



Indexing using Raw Signals

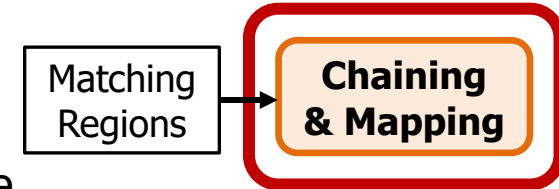
- Building the index **directly from raw signals**
 - **Free from the conversion table** used while converting the sequences to signals
- Converted signals are **filtered aggressively**:
 - To **avoid nanopore-related errors** better
 - Based on the similarity between adjacent signals



Chaining and Outputting Overlaps

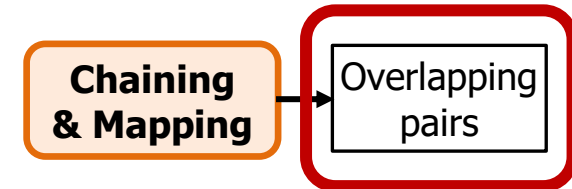
- **Adjusting the minimum chaining score** to avoid false chains

- All-vs-all overlapping tends to find a larger number of seed hits than mapping to a reference genome
- Minimum score for a chain during overlapping is set to be **$\sim 5\times$ larger than mapping**



- **Adjusting the outputting strategy** for accurate assembly

- **All chains are reported** to enable a raw signal overlapping with many raw signals (all-vs-all)
- **Cyclic overlaps are avoided** with simple comparisons between read names (minimap2 strategy)



Outline

Background

Rawsamble Mechanism

Evaluation

Conclusion

Evaluation Methodology

- Compared to **minimap2** overlaps (forward strand only) [Li, Bioinformatics'18]
 - Rawsambl is integrated into RawHash2 [Firtina+, ISMB/ECCB'23, Firtina+, arXiv]
- **Use case(s)** for raw signal overlapping
 1. *De novo* assembly construction using miniasm [Li, Bioinformatics'16]
 2. More new directions to be discussed
- **Evaluation metrics:**
 - **Throughput** (bases processed per second per CPU thread) and **overall time**
 - **Percentage of shared and unique overlapping pairs** between tools
 - **Assembly statistics**

	Organism	Flow Cell	Reads (#)	Bases (#)	SRA Acc.
D1	<i>E. coli</i>	R9.4	353,317	2,365M	ERR9127551
D2	<i>Yeast</i>	R9.4	49,989	380M	SRR8648503
D3	<i>Human</i>	R9.4	269,507	1,584M	FAB42260 (ONT)

- **Datasets:**

Performance & Peak Memory

Organism	Tool	CPU time (hh:mm:ss)	Peak Memory (GB)	Throughput (bp/sec)
D1	Rawsamble	4:06:44	14.98	95,626
<i>E. coli</i>	minimap2	0:20:32	30.66	NA
D2	Rawsamble	0:05:39	9.87	62,548
<i>Yeast</i>	minimap2	0:00:35	5.74	NA
D3	Rawsamble	0:15:14	18.04	463,973
<i>Human</i>	minimap2	0:02:05	18.68	NA

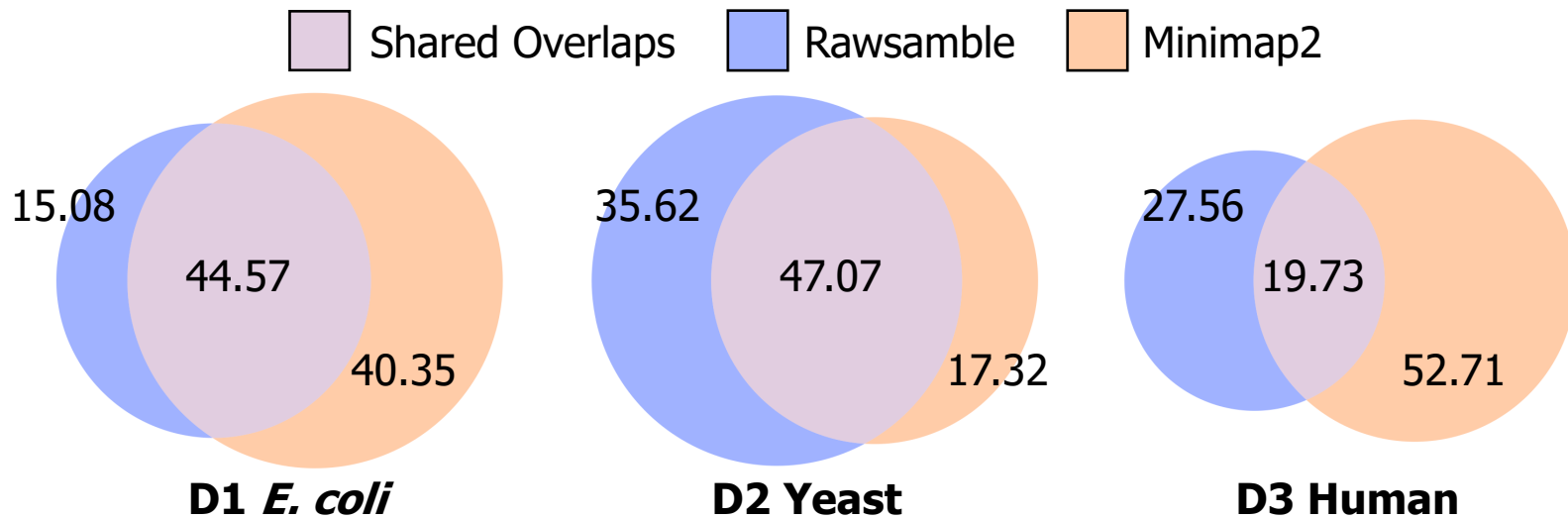
Although **minimap2 is substantially faster**,
Rawsamble avoids the basecalling step

- **Real-time analysis requires** faster throughput than sequencer
 - Throughput of a single nanopore: **~450 bp/sec (data generation speed)**

139× - 1031× faster throughput with a **single CPU thread**
compared to a single pore

All-vs-All Overlapping Statistics

- Percentage of overlapping pairs
 - Shared between Rawsamble and minimap2
 - Unique to either Rawsamble or minimap2



On average, **37.12%** of overlapping pairs is **shared** with minimap2

How can we evaluate the impact of these ratios?

De novo Assembly From Overlaps

- **Goal:** To build long *de novo* assemblies from raw signal overlaps
 - Miniasm can be used off-the-shelf as both tools provide PAF outputs

Organism	Tool	N50 Unitig Length	Avg. Unitig Length	Max. Unitig Length	No. of unitigs
D1	Rawsamble	543,505	373,594	1,431,572	39
<i>E. coli</i>	minimap2	5,210,589	2,611,044	5,210,938	4
D2	Rawsamble	60,605	47,250	256,116	431
<i>Yeast</i>	minimap2	122,735	82,757	386,005	278
D3	Rawsamble	23,717	16,376	66,163	59
<i>Human</i>	minimap2	18,128	10,572	42,654	53

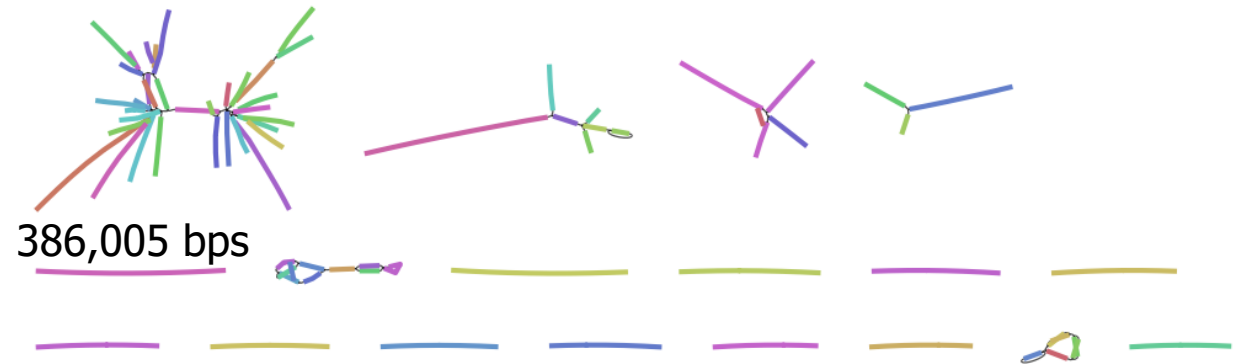
Raw signal overlaps can be used for constructing *de novo* assemblies **without basecalling**

Overlaps from minimap2 lead to longer unitigs
mainly due to using less noisy sequencing data

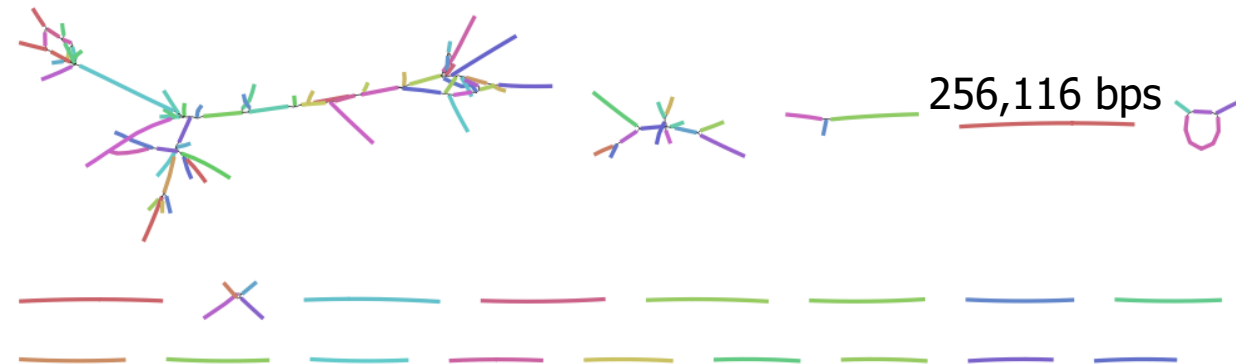
De novo Assembly From Overlaps

- Visualizing the miniasm outputs with Bandage [Wick+, Bioinformatics'15]

Minimap2 (Yeast):



Rawsambl (Yeast):

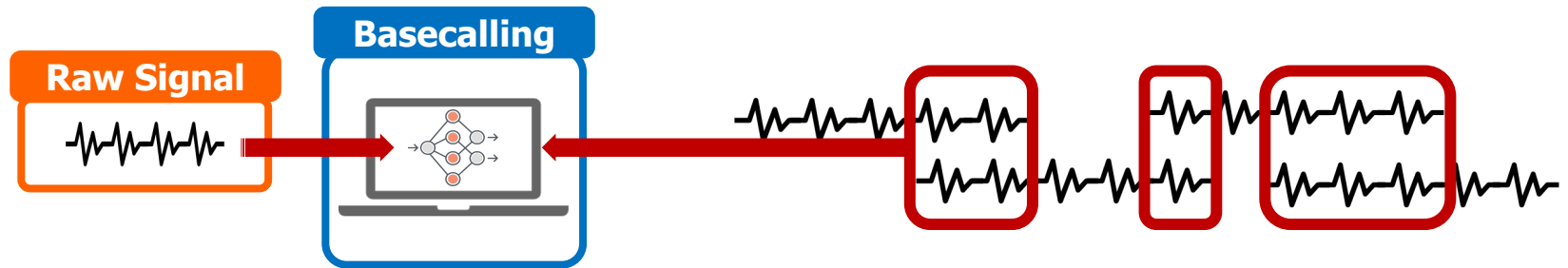


Unitigs can be assembled into long components using **raw signals**

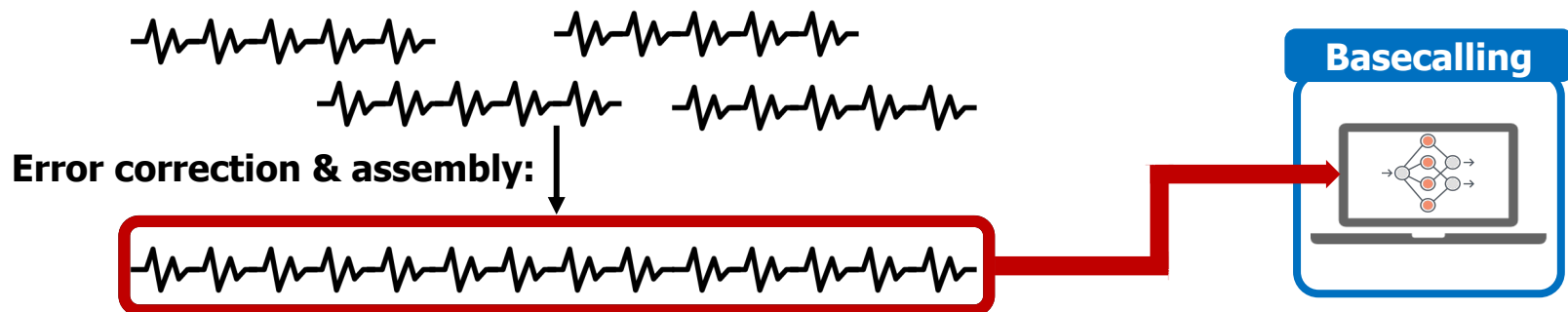
New Directions in Raw Signal Analysis

✓ Constructing (and analyzing) *de novo* assemblies

Utilizing the overlap information for more accurate basecalling



Utilizing the constructed assembly for basecalling



Rawsamble



- Can Firtina, Maximilian Mordig, Joël Lindegger, Harun Mustafa, Sayan Goswami, Stefano Mercogliano, Yan Zhu, Andre Kahles, and Onur Mutlu,

"Rawsamble: Overlapping and Assembling Raw Nanopore Signals using a Hash-based Seeding Mechanism"

[32nd Annual Conference on Intelligent Systems for Molecular Biology \(ISMB\)](#), Jul 2024

[\[Source Code\]](#)

[Preprint to be available soon]

Rawsamble: Overlapping and Assembling Raw Nanopore Signals using a Hash-based Seeding Mechanism

Can Firtina¹ Maximilian Mordig^{1,2} Joël Lindegger¹ Harun Mustafa^{1,3,4} Sayan Goswami¹

Stefano Mercogliano¹ Yan Zhu^{1,5} Andre Kahles^{1,3,4} Onur Mutlu¹

¹ETH Zurich

²Max Planck Institute for Intelligent Systems

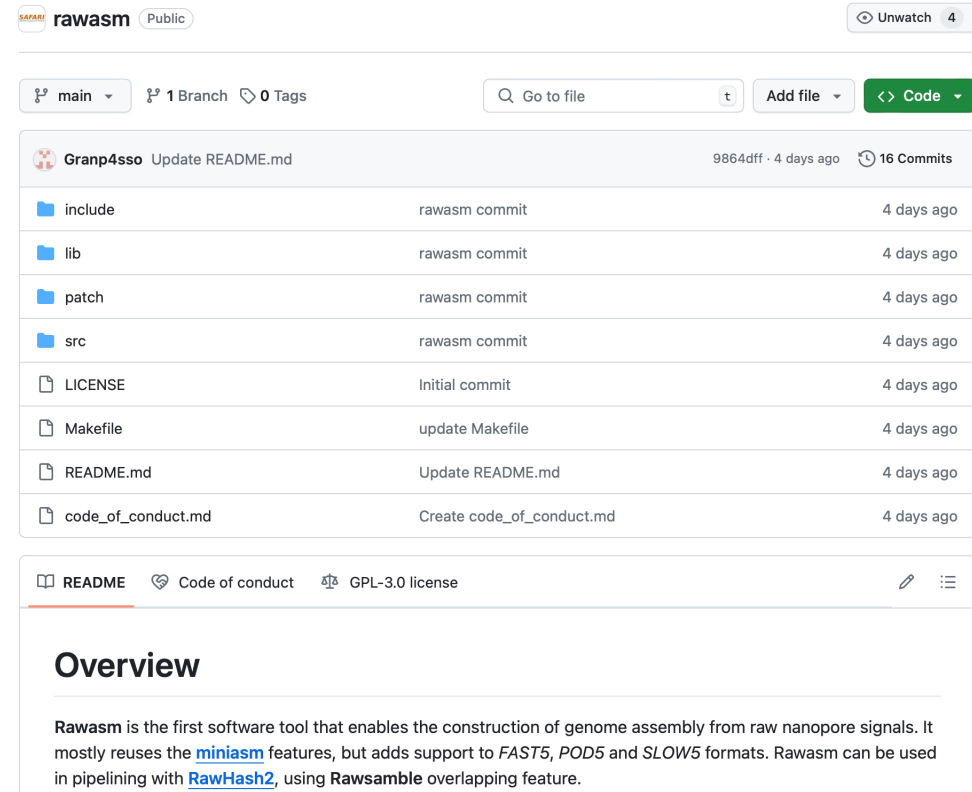
³University Hospital Zurich

⁴Swiss Institute of Bioinformatics

⁵University of Toronto

Rawasm: Raw Signal Assembler [Beta]

- **Slightly modified version of miniasm**
 - To output assembled raw signals instead of basecalled sequences
- Supports **all major raw signal file formats**
 - FAST5, POD5, S/BLOW5 file formats
- Still in a testing phase:
Feedback is appreciated!



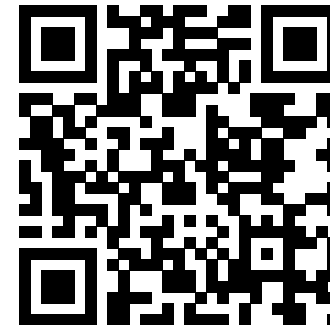
The screenshot shows the GitHub repository for 'rawasm' by CMU-SAFARI. The repository is public and has 1 branch and 0 tags. The commit history shows a recent update to the README.md file by Grnp4sso. The overview section provides a brief description of the tool.

File	Commit	Time
include	rawasm commit	4 days ago
lib	rawasm commit	4 days ago
patch	rawasm commit	4 days ago
src	rawasm commit	4 days ago
LICENSE	Initial commit	4 days ago
Makefile	update Makefile	4 days ago
README.md	Update README.md	4 days ago
code_of_conduct.md	Create code_of_conduct.md	4 days ago

Overview

Rawasm is the first software tool that enables the construction of genome assembly from raw nanopore signals. It mostly reuses the [miniasm](#) features, but adds support to *FAST5*, *POD5* and *SLOW5* formats. Rawasm can be used in pipelining with [RawHash2](#), using [Rawsamble](#) overlapping feature.

<https://github.com/CMU-SAFARI/rawasm>



Outline

Background

Rawsamble Mechanism

Evaluation

Conclusion

Conclusion

Key Contributions:

1. **Rawsample: The first mechanism** that can find **all-vs-all overlapping** pairs between raw nanopore signals
2. **The first *de novo* assembly** constructed directly from raw signal overlaps **without basecalling**
3. **A new assembler** to build and output the assemblies of signals

- Key Results:** Across 3 genomes of varying sizes, Rawsample provides
- **Throughput: 139× - 1031×** faster with one thread compared to a single pore
 - **Overlap statistics: 37%** of overlapping pairs **shared with minimap2**
 - **Assembly:** Unitigs of length **up to one million nucleotides** from overlapping raw signals **without basecalling**

Many opportunities for analyzing raw nanopore signals:

- **Indexing is very cheap:** Many future use cases with the on-the-fly index construction
- We should rethink the algorithms to perform downstream analysis **fully using raw signals**
- We should rethink the basecalling approaches to integrate information from raw signal analysis



Rawsamble

Overlapping and Assembling
Raw Nanopore Signals

Using a Hash-based Seeding Mechanism

Can Firtina

Maximilian Mordig

Joel Lindegger

Harun Mustafa

Sayan Goswami

Stefano Mercogliano

Yan Zhu

Andre Kahles

Onur Mutlu

SAFARI

ETH zürich



**BIOMEDICAL
INFORMATICS**



Backup Slides

Future Work

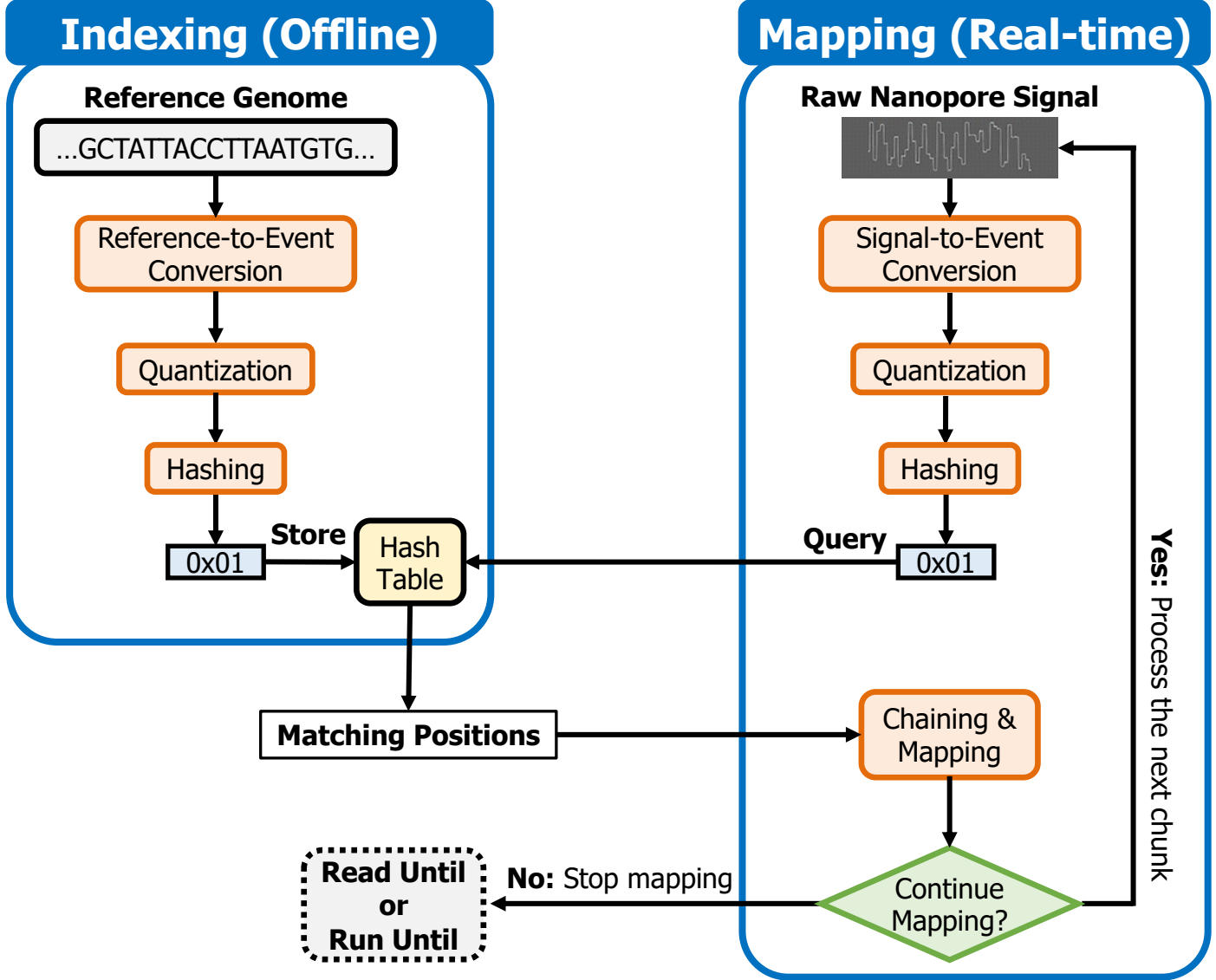
Reverse Complementing Raw Nanopore Signals

- Without reverse complementing, we are missing half of the useful information

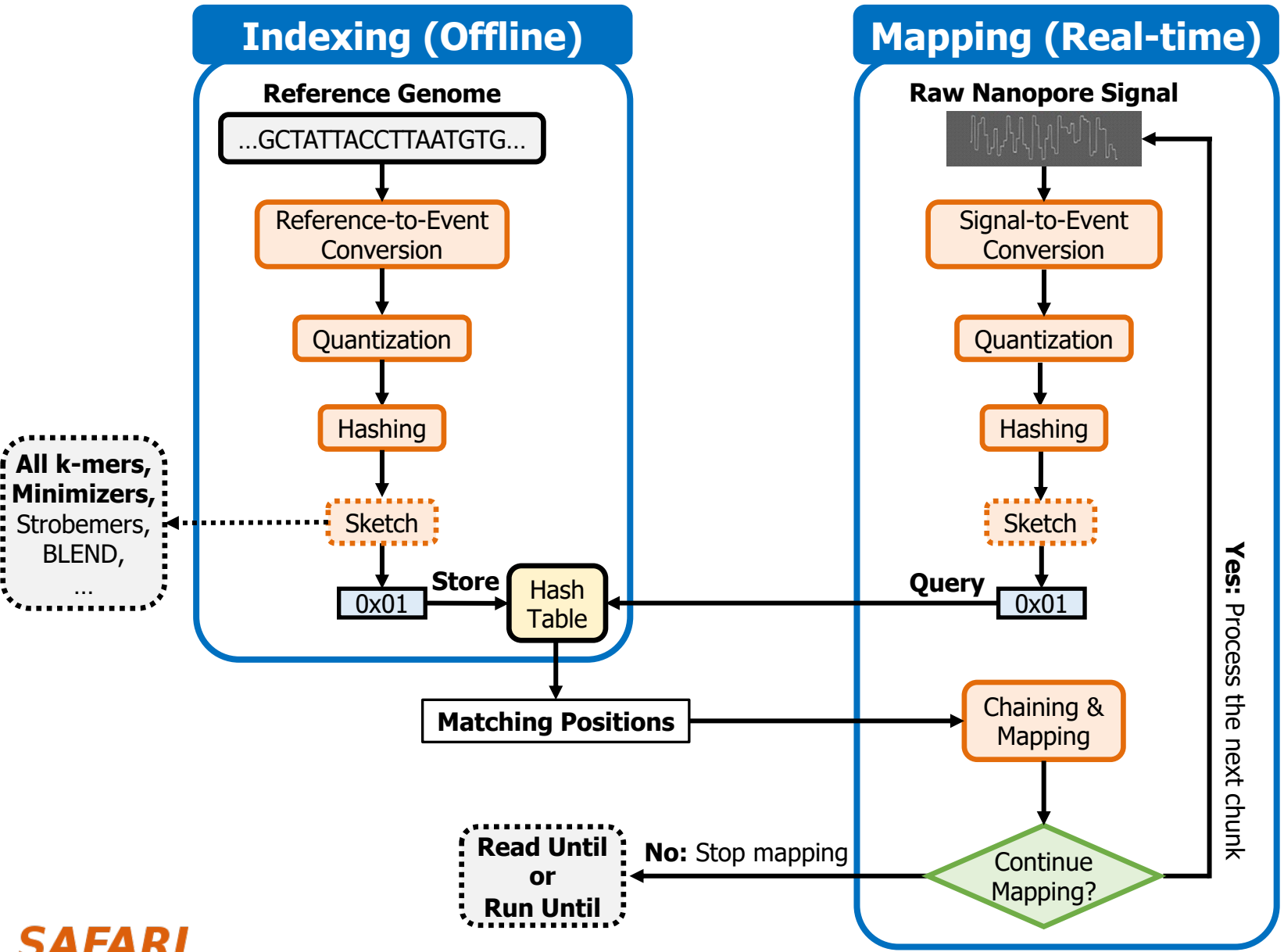
Dynamically Building the Hash Table in Real-Time

- Needed for real-time *de novo* assembly construction
- What are the useful applications for real-time *de novo* assembly construction?

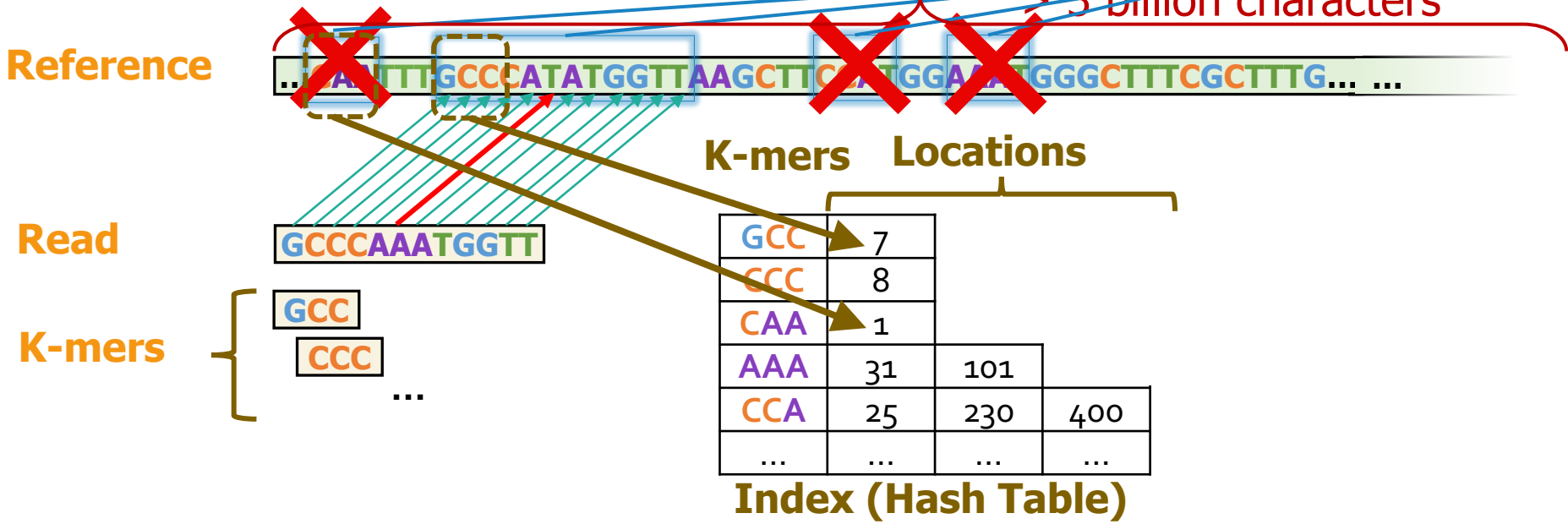
Real-Time Mapping using Hash-based Indexing



Sketching with Hash-based Indexing



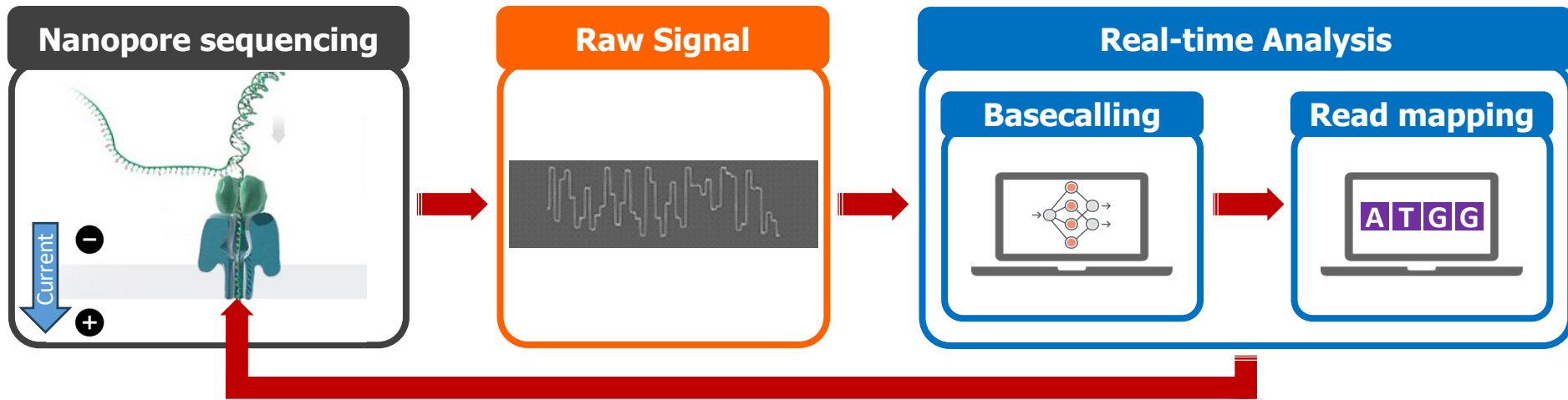
Practical Similarity Identification Seeds



- Seeding**: Determine potential matching regions (seeds) in the reference genome
- Seed Filtering (e.g., Chaining)**: Prune some seeds in the reference genome
- Alignment**: Determine the exact differences between the read and the reference genome

Existing Solutions – Real-time Basecalling

Deep neural networks (**DNNs**) for translating **signals** to **bases**



DNNs provide **less noisy analysis** from basecalled sequences

Costly and power-hungry computational requirements

Applications of Read Until

Depletion: Reads mapping to a particular reference genome is ejected

- Removing contaminated reads from a sample
- Relative abundance estimation
- Controlling low/high-abundance genomes in a sample
- Controlling the sequencing of depth of a genome

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

- Purifying the sample to ensure it contains only the selected genomes
- Removing the host genome (e.g., human) in contamination analysis



Rawsamble

Overlapping and Assembling
Raw Nanopore Signals

Using a Hash-based Seeding Mechanism

Can Firtina

Maximilian Mordig

Joel Lindegger

Harun Mustafa

Sayan Goswami

Stefano Mercogliano

Yan Zhu

Andre Kahles

Onur Mutlu

SAFARI

ETH zürich



**BIOMEDICAL
INFORMATICS**

