Pan-genomic indexes for robust classification of nanopore and metagenomic reads

Omar Ahmed[1], Massimiliano Rossi[2], Sam Kovaka[1], Michael C. Schatz[1], Travis Gagie[3], Christina Boucher[2], and Ben Langmead[1]

Genome Informatics
November 3-5, 2021

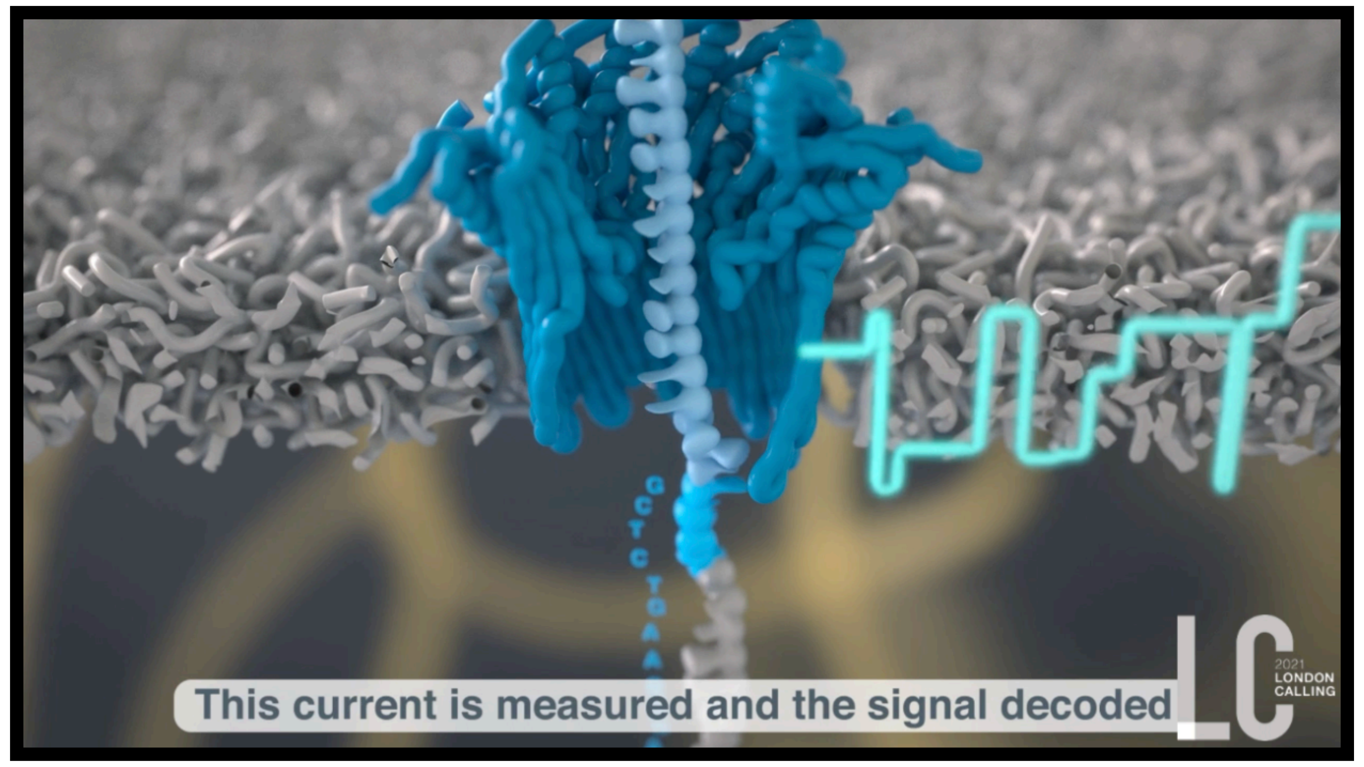
[1] Department of Computer Science, Johns Hopkins University, Baltimore, MD, USA

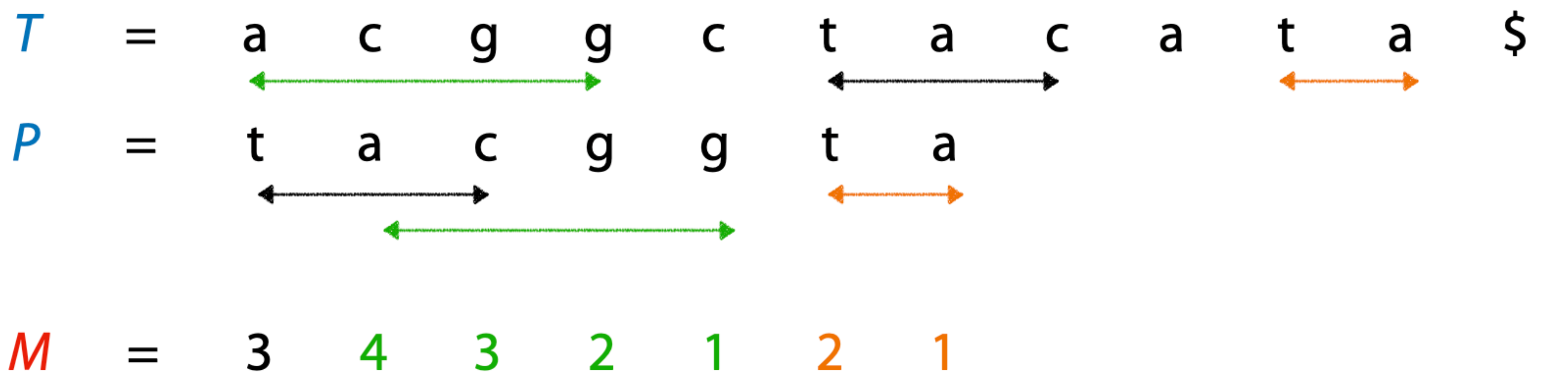
[2] Department of Computer and Information Science and Engineering, University of Florida, Gainesville, FL, USA

[3] Faculty of Computer Science, Dalhousie University, Halifax, NS, CAN

Overview of Presentation

- Development of SPUMONI for classification of nanopore reads





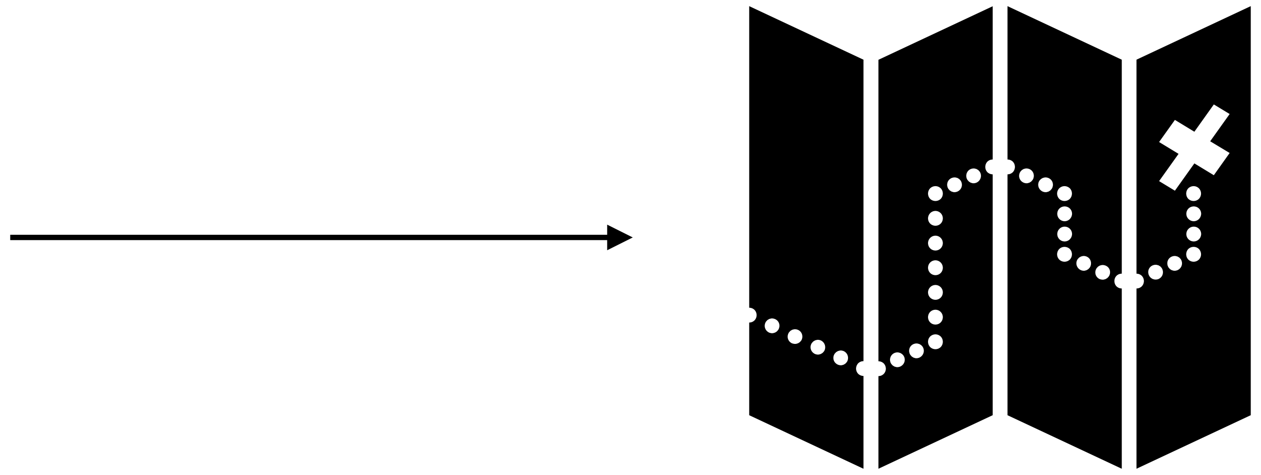
T = a c g g c t a c a t a $

P = t a c g g t a

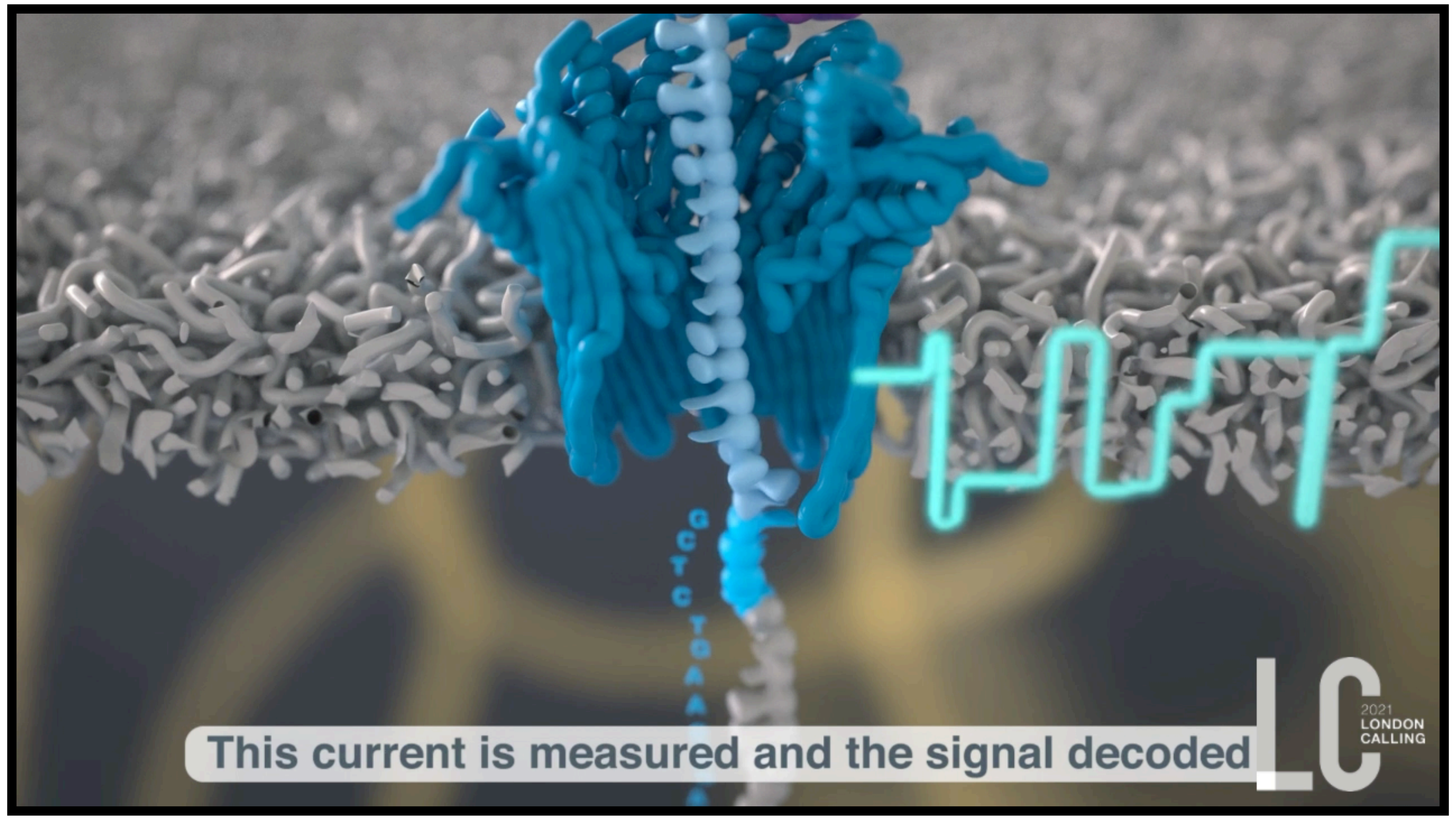
M = 3 4 3 2 1 2 1

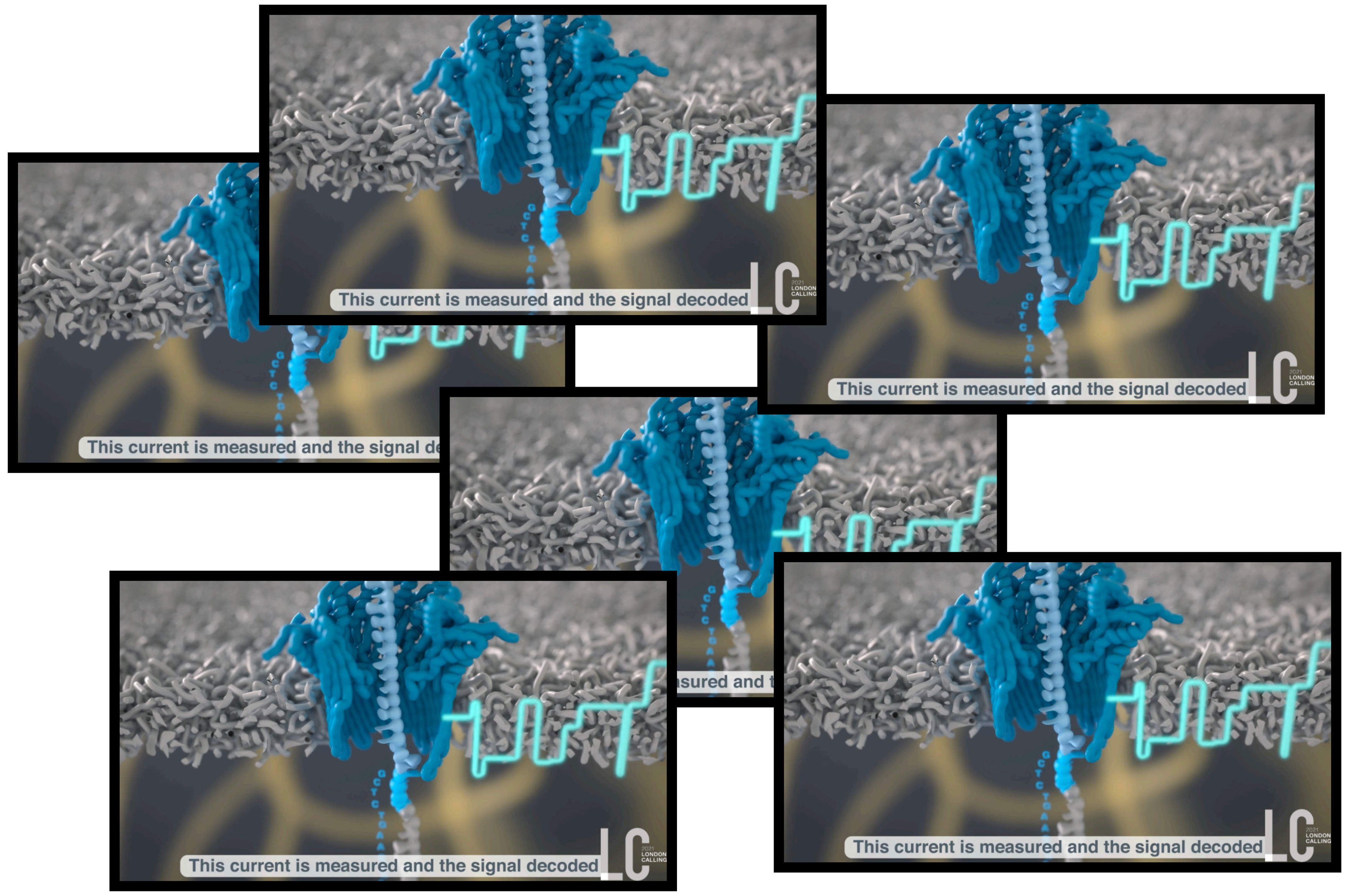
- Scaling SPUMONI to handle larger pan-genomes more efficiently

```
>seq0
ACGATAGA
TACTACTAT
>seq1
ATAAAGCTA
GATATCATG
```



Nanopore Sequencing





- Allows users to perform targeted sequencing using software

- UNCALLED & Readfish allow users to target sequences but not optimized for large, and repetitive databases

Motivation: A need for faster methods to classify reads against large, repetitive databases

Video From: https://nanoporetech.com/about-us/news/towards-real-time-targeting-enrichment-or-other-sampling-nanopore-sequencing-devices

Method Overview

Check Max Rossi's Poster 150

SPUMONI - Streaming PseUdo MONI[1]

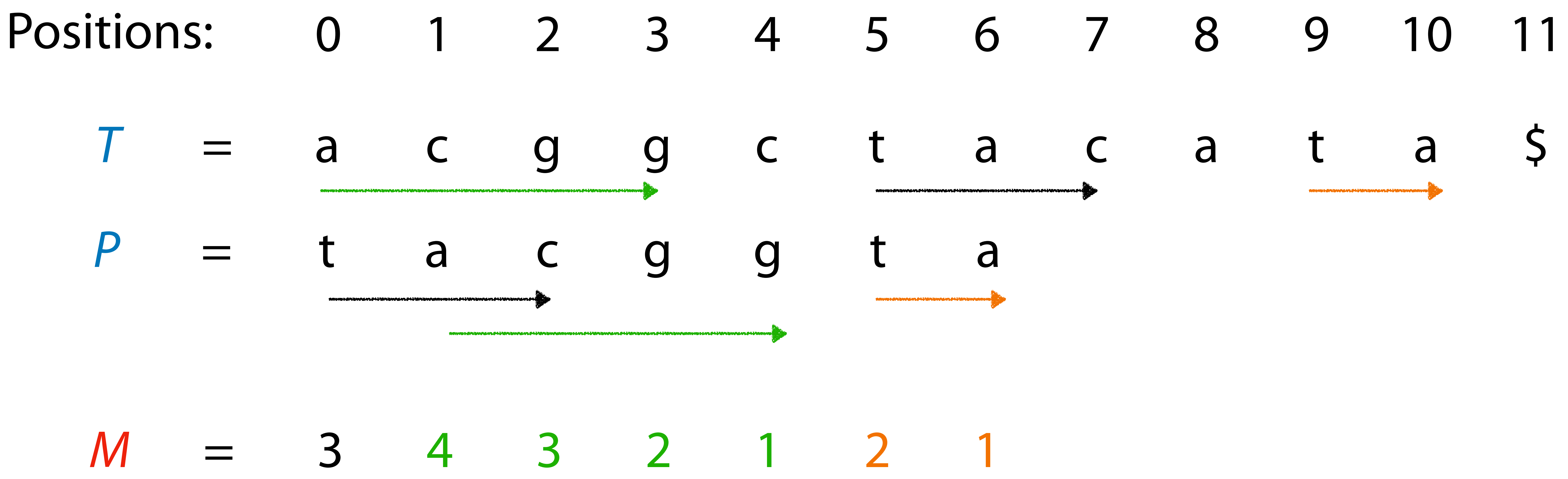
- Makes rapid targeting decisions based on input database
    - Key Intuition: A read's MS/PMLs with respect to a reference can reveal if there appears to be "good" approximate match to reference
- Uses the r-index[2] to enable efficient indexing of large, repetitive collections
    - Number of runs in BWT, r, typically grows sub-linearly w.r.t to length of input sequence, n.
- Extends MONI[1] in two key areas
    - Adds a "null distribution" and hypothesis testing framework for finding "significant" matches
    - Replaces MONI's "batch" matching statistic (MS) algorithm with a faster, streaming algorithm
        - Calculates new quantity called pseudo-matching lengths (PMLs)

[1] Rossi, M., Oliva, M., Langmead, B., Gagie, T., & Boucher, C. (2021). MONI: A pangenomics index for finding MEMs. *Proc. RECOMB.*

[2] Mun, T., Kuhnle, A., Boucher, C., Gagie, T.,Langmead, B., and Manzini, G. (2020). Matching reads to many genomes with the r-index. J. Comput. Biol.27, 514–518

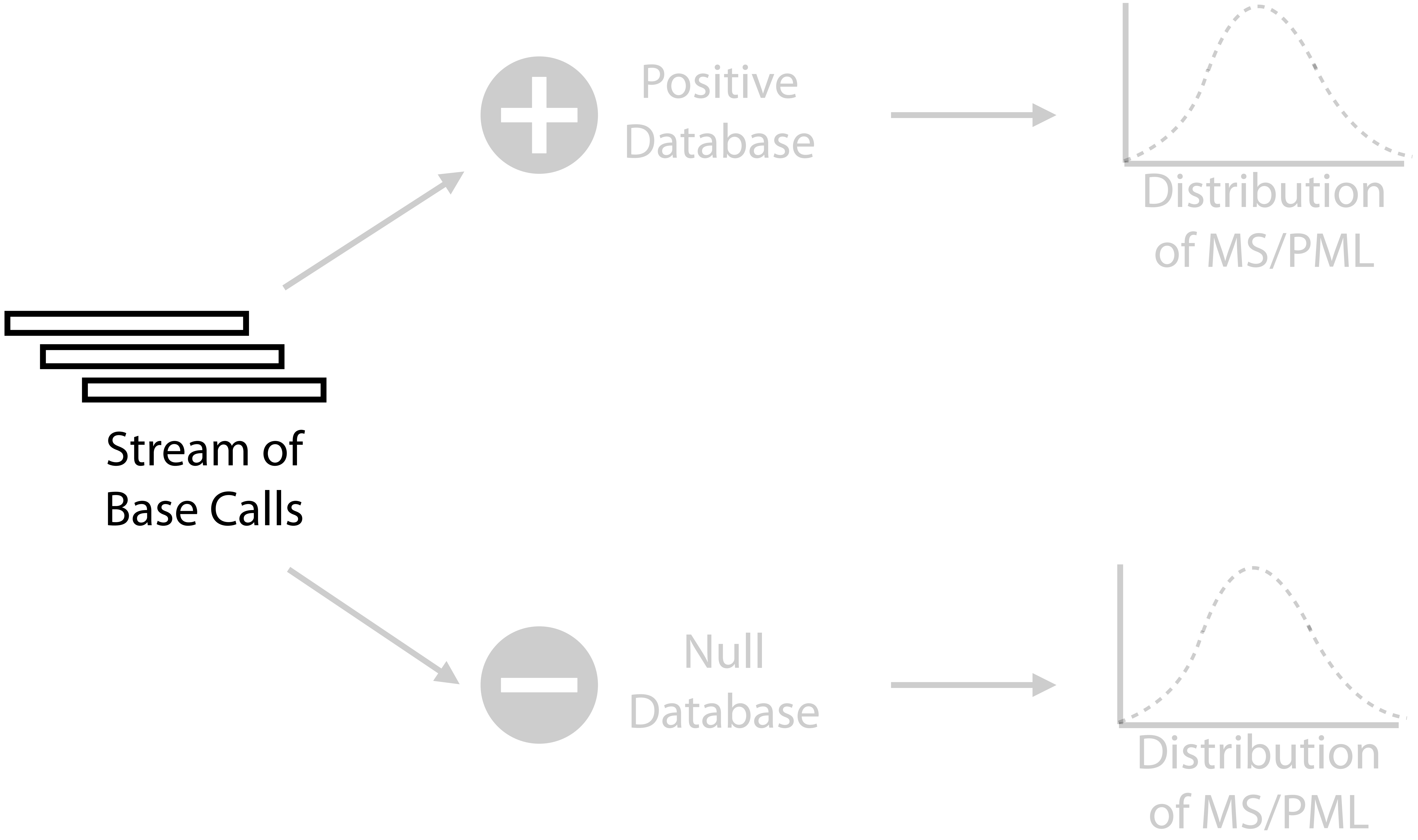
Matching Statistics

- The matching statistics of *P* w.r.t to *T* is an array *M* of length m where *M*[*i*] is the length of the longest prefix of the pattern *P[i..m]* that occurs in text *T*
  - Let *T* be a text of length n, and *P* be a pattern of length m
- Think of matching statistics like they are half MEMs (Maximally Exact Matches)
  - Example:



Positions:	0	1	2	3	4	5	6	7	8	9	10	11
T =	a	c	g	g	c	t	a	c	a	t	a	$
P =	t	a	c	g	g	t	a					
M =	3	4	3	2	1	2	1					

SPUMONI Approach



Stream of Base Calls

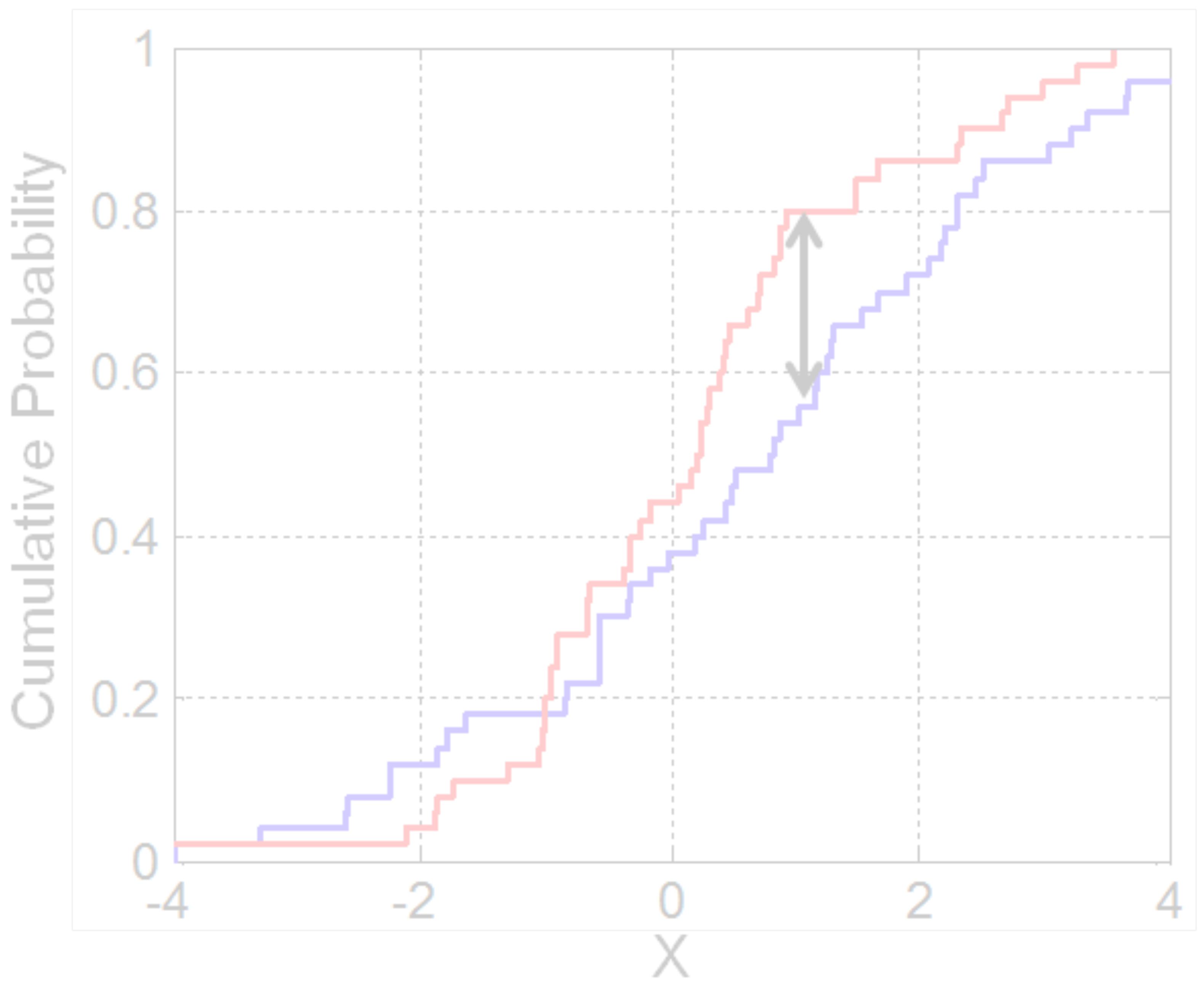
Positive Database

Null Database

Distribution of MS/PML

Distribution of MS/PML

Statistical Tests:



KS-Test

Image: https://en.wikipedia.org/wiki/Kolmogorov%E2%80%93Smirnov_test

Results - Real Mock Community Experiment

- Question: Can using a pan-genome reference allow us to target a particular strain that is not present in the reference? and how does it compare on time and memory?
  - Using **real** mock community reads[1] where we want to "target" the yeast reads, and eject all the microbial species

Reference:	One genome ref (**7** genomes from **7** species)		Pan-genome ref (**3537** genomes from **7** species)	
Approach:	SPUMONI	minimap2	SPUMONI	minimap2
Accuracy:	86.72	87.82	96.02	97.52

- **SPUMONI is …**
  - 12X faster than minimap2
  - Uses 4X less memory than minimap2

- Answer: ① Yes, using a pan-genome reference, allowed us to target the ZymoMC strains

  ② Faster and uses less memory than minimap2 with similar classification metrics

[1] Kovaka, S., Fan, Y., Ni, B. *et al.* Targeted nanopore sequencing by real-time mapping of raw electrical signal with UNCALLED. *Nat Biotechnol* **39,** 431–441 (2021).

Pillars of SPUMONI

- What are the key ideas for why SPUMONI outperforms alignment in classifying reads?

Key Ideas	Why?
SPUMONI is **faster.**	Non-alignment method
SPUMONI is **scalable to large databases.**	Use of r-index[1]
SPUMONI's classification is **robust.**	Non-parametric, Non-kmer method

Motivated us to **push SPUMONI into other domains** like metagenomic classification!

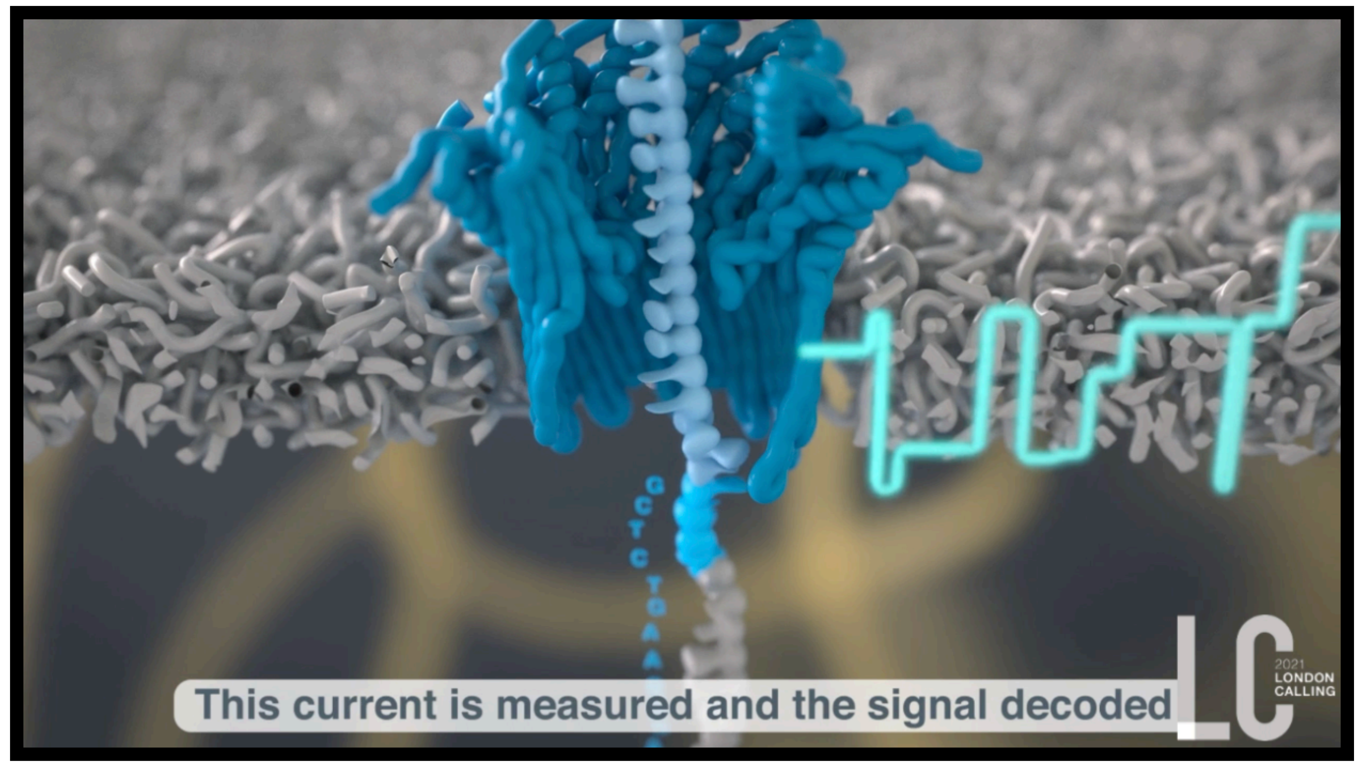
- For details on additional specifics of method and results:

Ahmed, O., Rossi, M., Kovaka, S., Schatz, M. C., Gagie, T., Boucher, C., & Langmead, B. (2021). Pan-genomic Matching Statistics for Targeted Nanopore Sequencing. iScience, 102696.

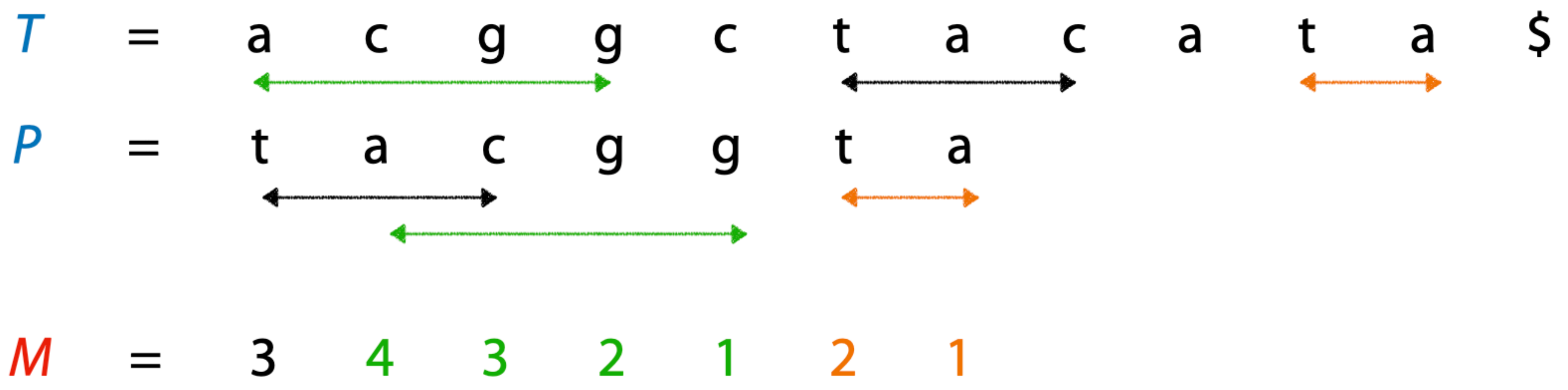
[1] Mun, T., Kuhnle, A., Boucher, C., Gagie, T.,Langmead, B., and Manzini, G. (2020). Matching reads to many genomes with the r-index. J. Comput. Biol.27, 514–518

Overview of Presentation

- Development of SPUMONI for classification of nanopore reads



This current is measured and the signal decoded



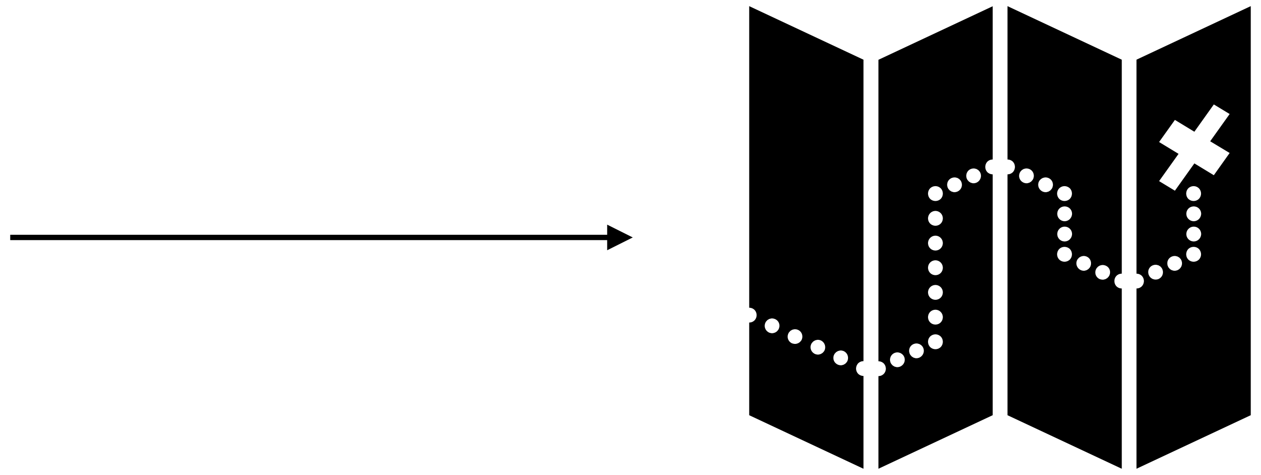
T = a c g g c t a c a t a $

P = t a c g g t a

M = 3 4 3 2 1 2 1

- Scaling SPUMONI to handle larger pan-genomes more efficiently

```
>seq0
ACGATAGA
TACACTAT
>seq1
ATAAAGCTA
GATATCATG
```

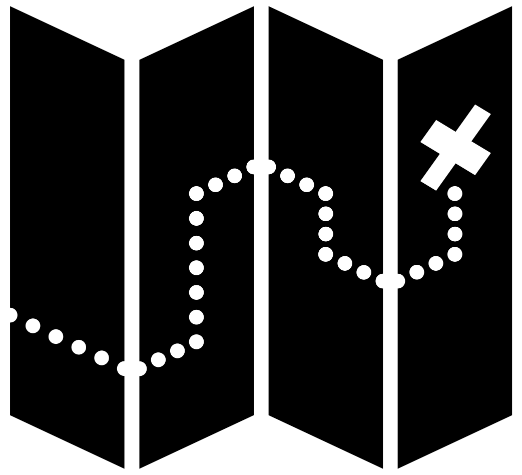


Speeding Up SPUMONI

- How can we speed up SPUMONI?

```
>seq0
ACGATAGA
TACACTAT
>seq1
ATAAAGCTA
GATATCATG
```

FASTA file



SPUMONI Index

Space: O(r)

Key Idea: Extract and concatenate the minimizers[1] to generate a smaller reference.

- How did we use the minimizer concept?

Small Window Size (k)

Reference ACTAGATAGACCAATCAGGTAATACGCGTAGGCTACTAGGATA

Large Window Size (w)

Minimizers AGAT ATCA GTAA GTAG TAGG

```
>original
ACTAGATAG
ACCAATCA
GGTAATACG
CGTAGGCTA
CTAGGATA
```

apply the minimizer scheme

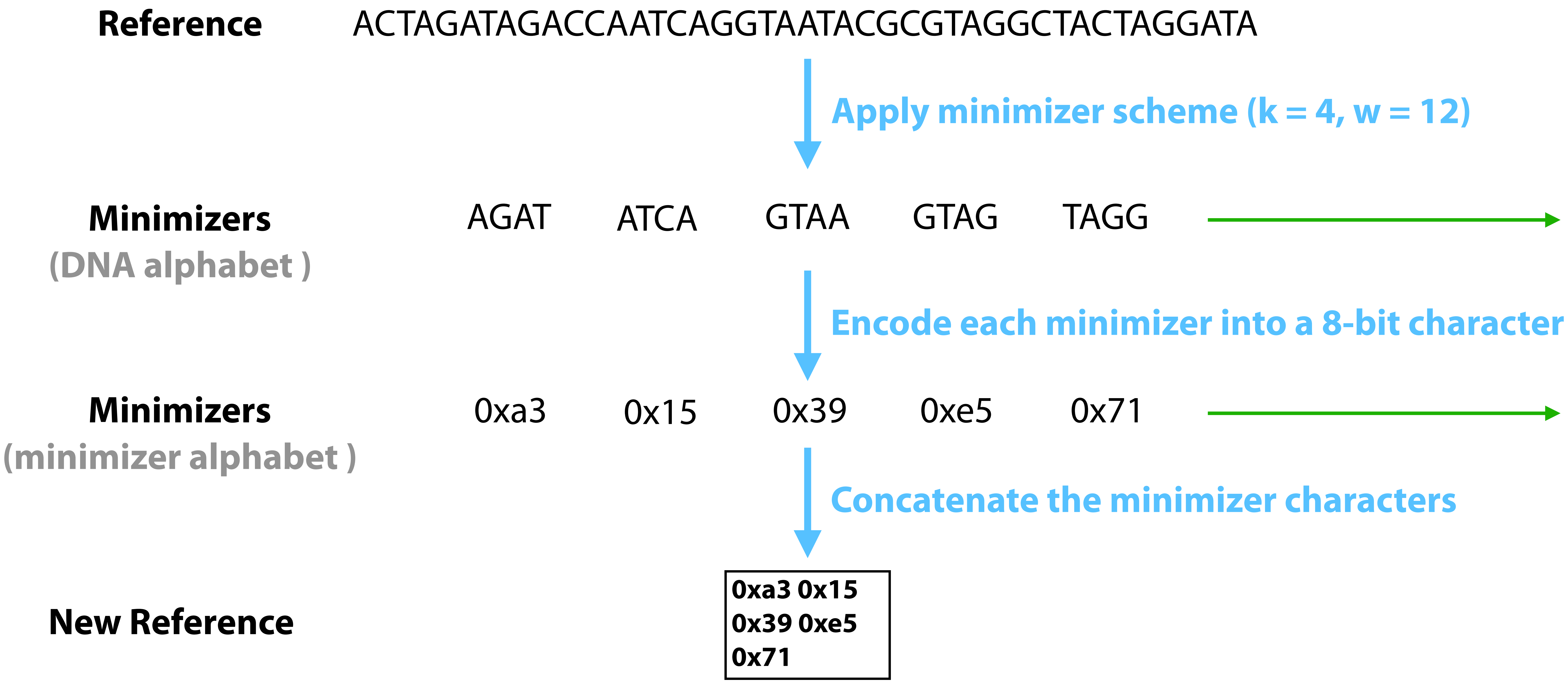
```
>new_ref
AGATATCAG
TAAGTAGTA
GG
```

We can apply the same scheme to reads, and thereby index this **smaller** reference.

[1] Roberts, M., Hayes, W., Hunt, B. R., Mount, S. M., & Yorke, J. A. (2004). Reducing storage requirements for biological sequence comparison. *Bioinformatics*, *20*(18), 3363-3369.

Speeding Up SPUMONI

- And we can take it even further …
  - Ekim et al. (2021)[1] use minimizer-alphabet for de Bruijn graphs
- Let’s take a look at how we use the minimizer-alphabet in SPUMONI …



Reference ACTAGATAGACCAATCAGGTAATACGCGTAGGCTACTAGGATA

Apply minimizer scheme (k = 4, w = 12)

Minimizers **(DNA alphabet)** AGAT ATCA GTAA GTAG TAGG

Encode each minimizer into a 8-bit character

Minimizers **(minimizer alphabet)** 0xa3 0x15 0x39 0xe5 0x71

Concatenate the minimizer characters

New Reference **0xa3 0x15 0x39 0xe5 0x71**

Impact of the Alphabet Promotion

```
seq.length = 5 * 4
for i = 1 to i = seq.length {
    // Compute MS at position i
}
```

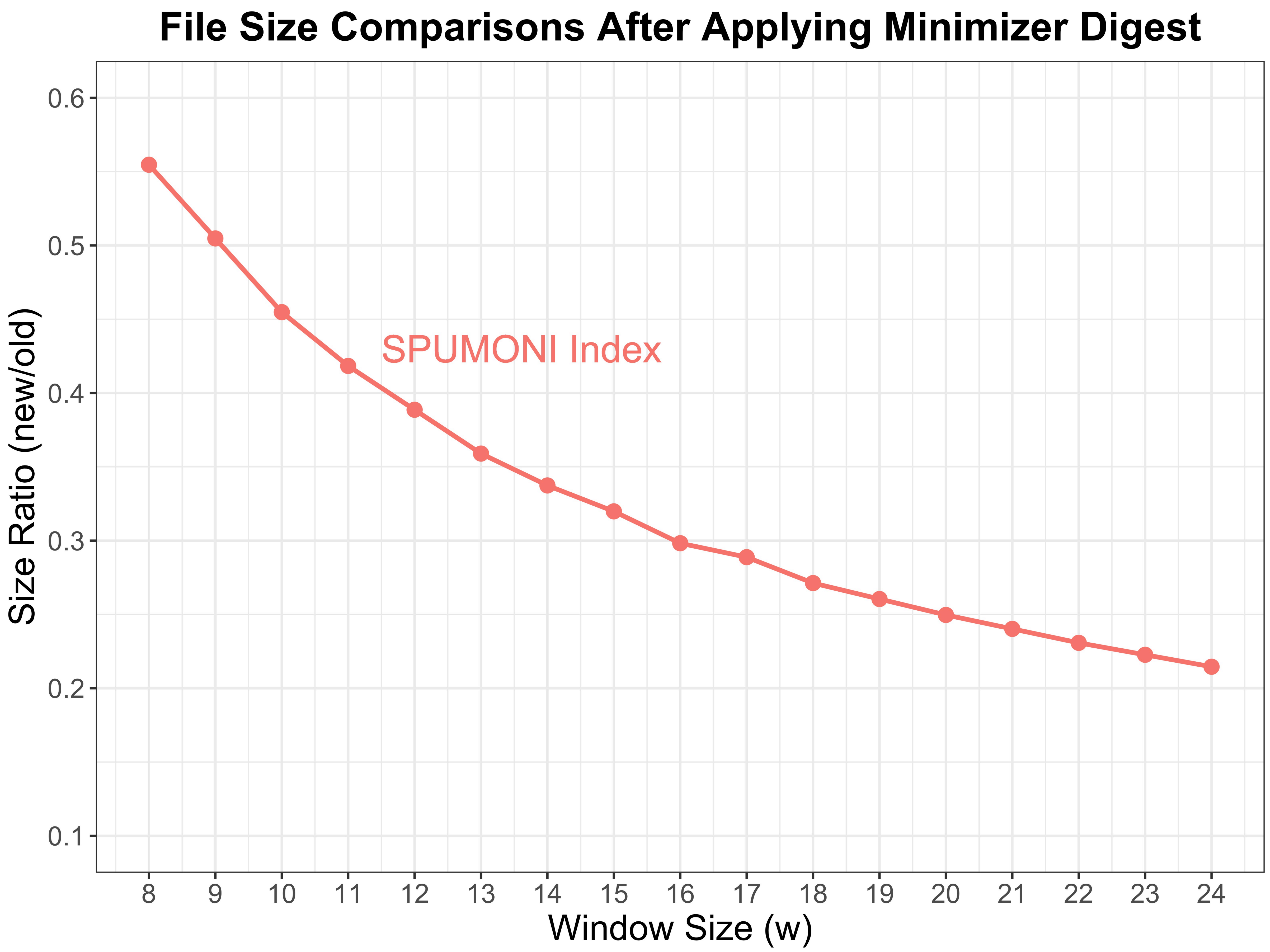
```
seq.length = 5
for i = 1 to i = seq.length {
    // Each iteration covers 4
    // characters
}
```

Combining the minimizer scheme with alphabet promotion leads to even smaller references.

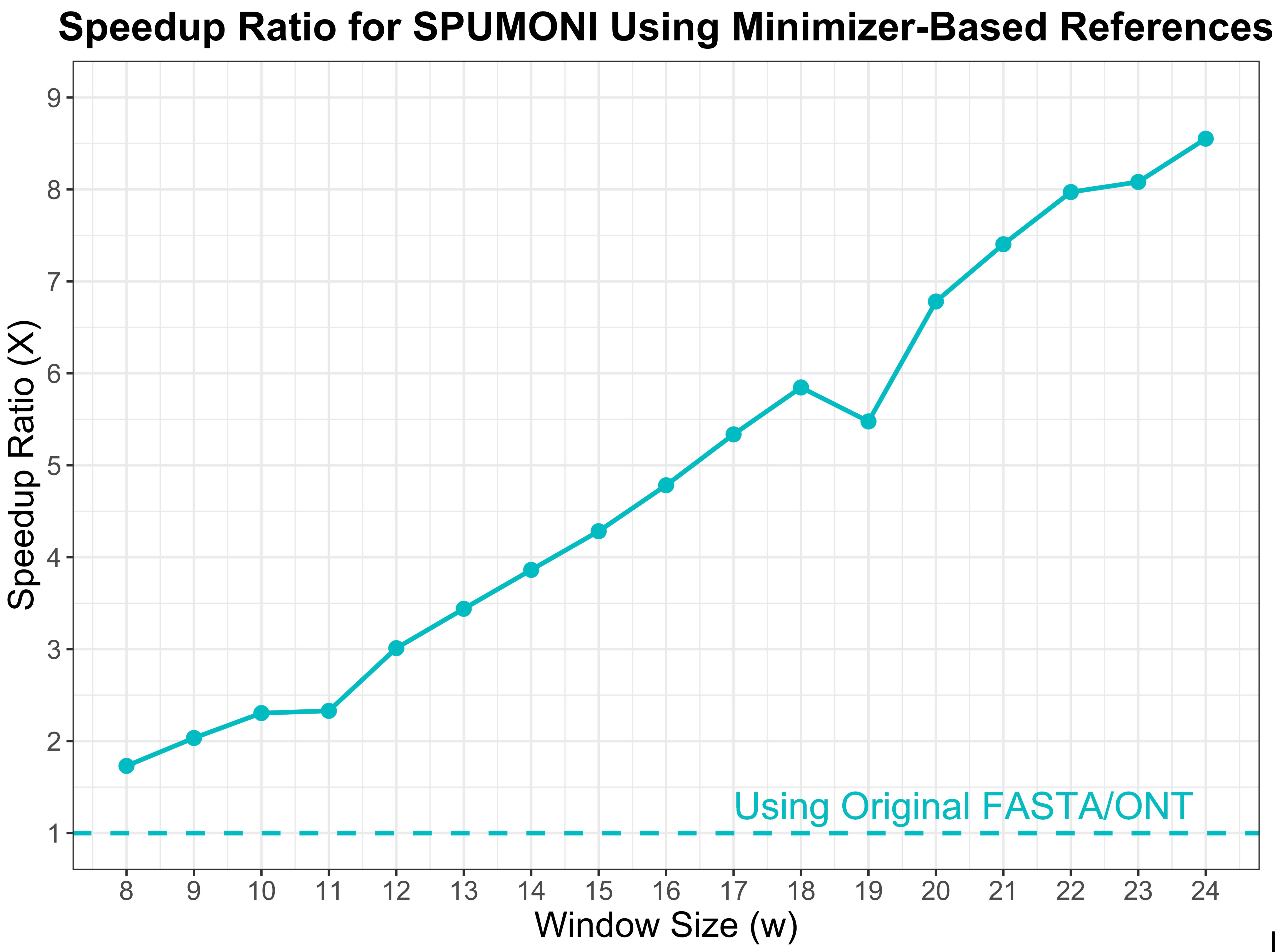
[1] Ekim, B., Berger, B., & Chikhi, R. (2021). Minimizer-space de Bruijn graphs: Whole-genome assembly of long reads in minutes on a personal computer. *Cell Systems*.

Results - Using Minimizer-Based References

- **Question:** How much smaller will the index be when applying this minimizer scheme where k = 4 to the reference?
- **Answer:** *Built an index over 1833 E. coli genomes[1] (~9 GB) to see …*



File Size Comparisons After Applying Minimizer Digest



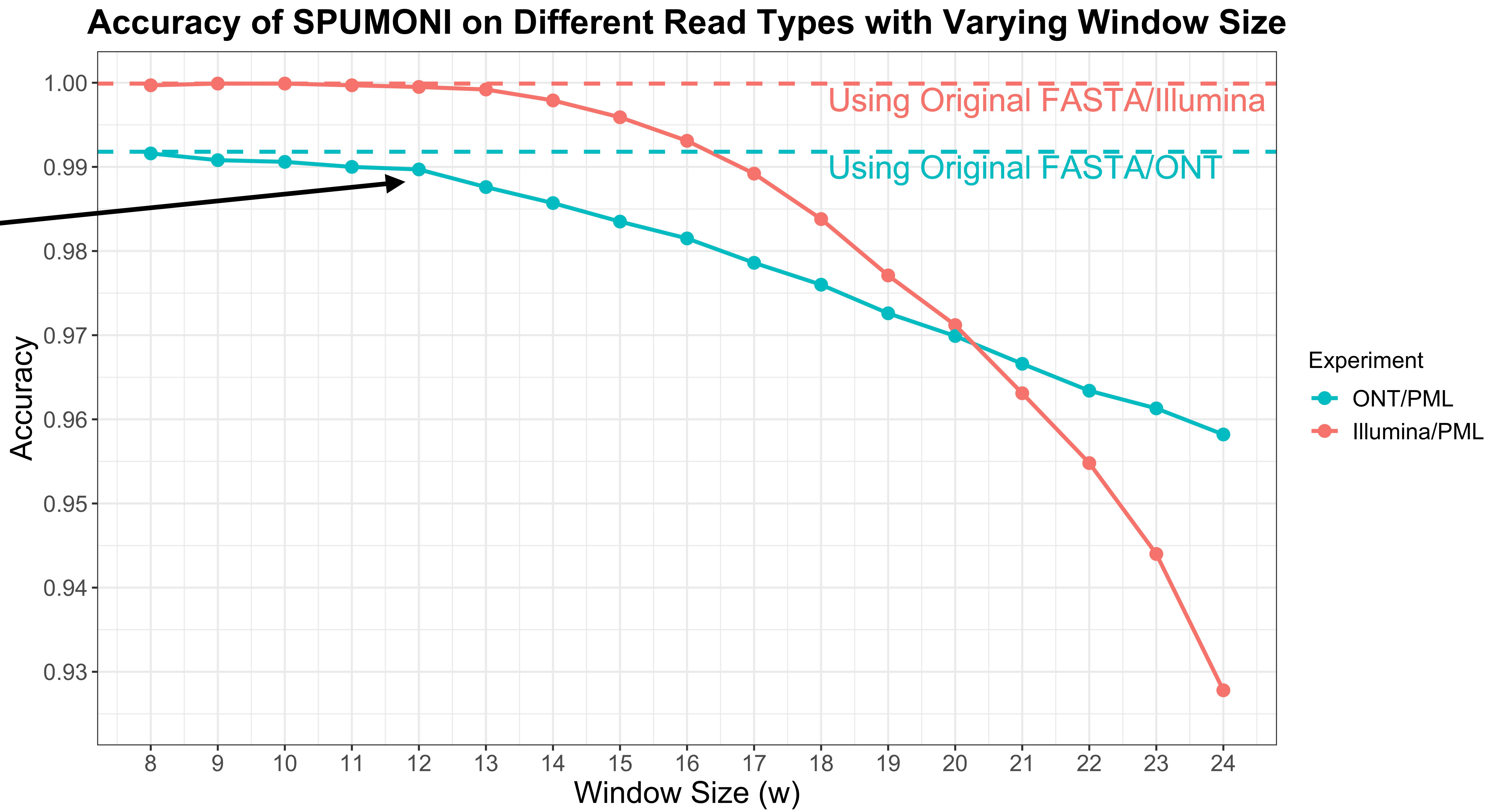
Speedup Ratio for SPUMONI Using Minimizer-Based References

[1] O'Leary NA et al. **Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation.** *Nucleic Acids Res.* 2016.

Results - Using Minimizer-Based References

- Question: How will this reduction in reference size affect our binary classification ability?
- Answer: *Simulated Human[1] and E. coli reads[2], and used our E. coli index to binary classify the reads …*

At w = 12, we are indexing a file **1/5** the original!

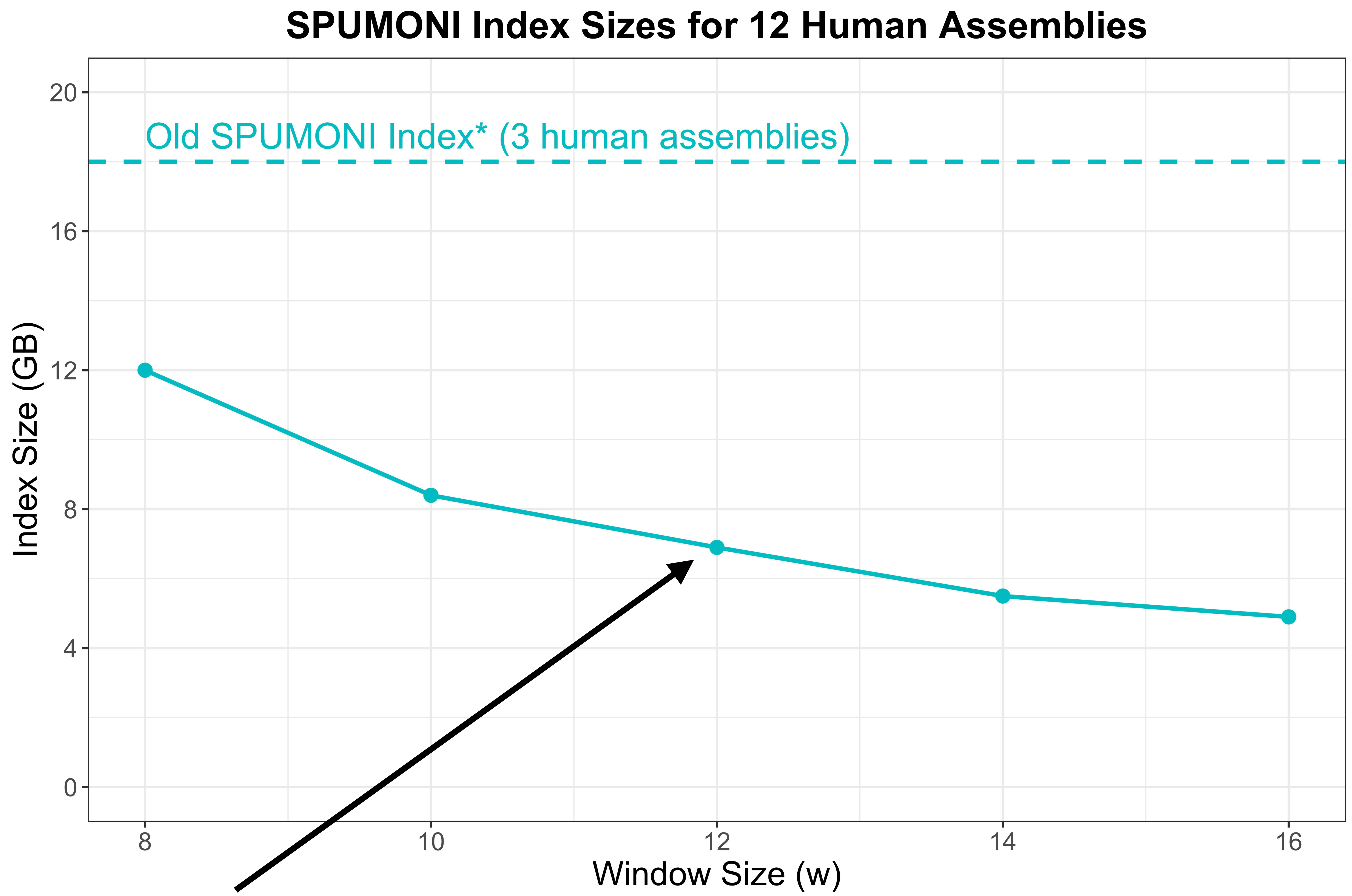


[1] Nurk S, Koren S, Rhie A, Rautiainen M, et al. **The complete sequence of a human genome.** bioRxiv, 2021.

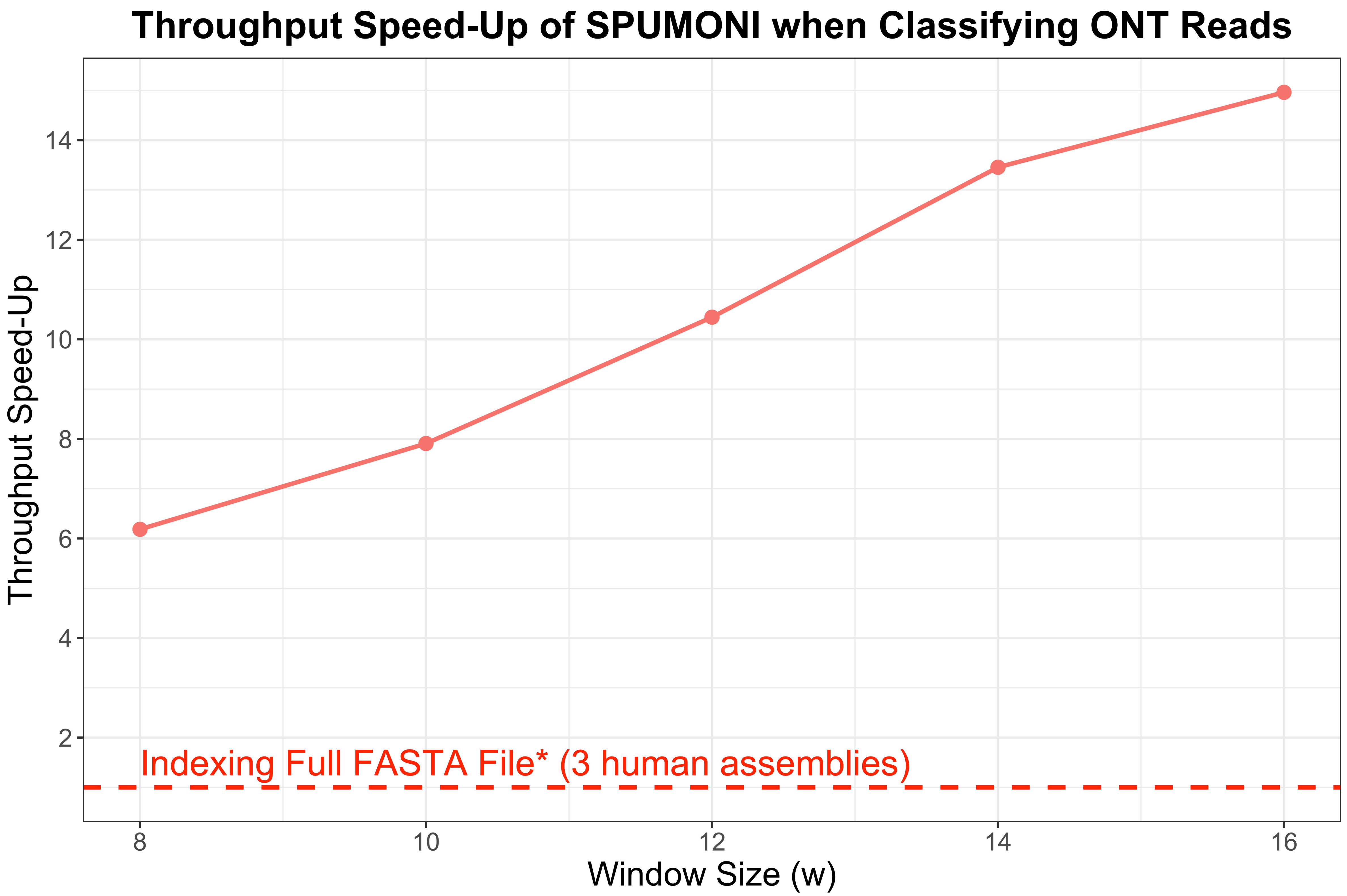
[2] O'Leary NA, et al. **Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation.** *Nucleic Acids Res.* 2016.

Results - Indexing 12 Human Assemblies

- Question: Can we index larger databases more efficiently than previously?
  - For reference previously, SPUMONI could **index 3 human genomes in 18 GB**
  - SPUMONI was **~2x faster than minimap2** at classifying ONT reads using the 3 human genome index.
- Answer: *Built an index over 12 human assemblies, and compared to our previous throughput …*


SPUMONI Index Sizes for 12 Human Assemblies

Index is ~7 GB, and **it has 4X as many genomes!**


Throughput Speed-Up of SPUMONI when Classifying ONT Reads

Results - Extending to Multiple Classes

- Question: Can we use matching statistics to distinguish multiple classes?
  - In this experiment, we simulated E. coli[2], Salmonella[2], and Human Reads[1]
  - Built three separate indexes …
    - 3 human genomes (~9 GB)
    - 1833 E. coli genomes (~9 GB)
    - 988 Salmonella genomes (~4.6 GB)

    } *Relatively similar sized so we can just test it with simple classification rule.*
  - Used a simple test of largest mean to classify read however …
- Answer: *Yes, we can. Here are the confusion matrices for classifying short and long reads …*

Classifying Short Reads

		Predicted Class		
		E. coli	Human	Salmonella
True Class	E. coli	24,588	39	373
	Human	14	24,981	5
	Salmonella	484	0	24,516

Classifying Long Reads

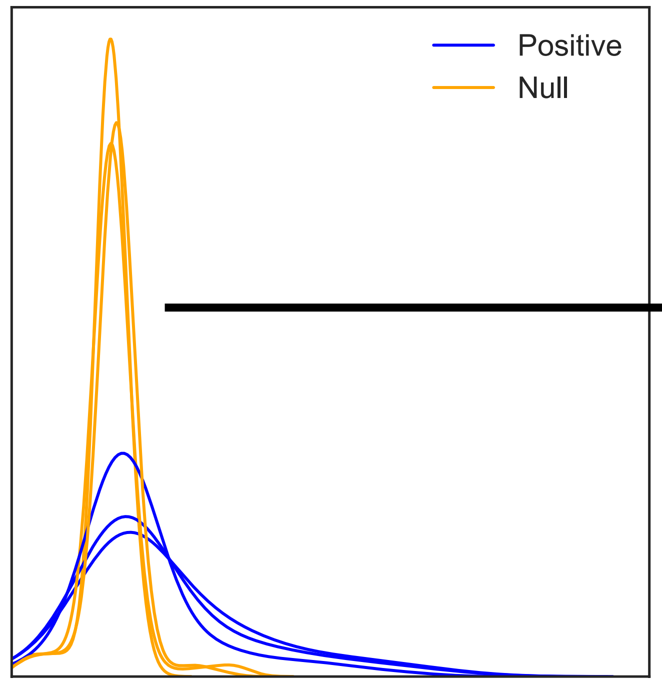
		Predicted Class		
		E. coli	Human	Salmonella
True Class	E. coli	24,459	46	495
	Human	30	24,946	24
	Salmonella	272	17	24,711

[1] Nurk S, Koren S, Rhie A, Rautiainen M, et al. **The complete sequence of a human genome.** bioRxiv, 2021.

[2] O'Leary NA, et al. **Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation.** *Nucleic Acids Res.* 2016.

Key Takeaways

- SPUMONI is a rapid tool for binary read classification that uses a read's MS or PMLs to classify it.
  - Use of the **minimizers** and a **promoted alphabet** allows SPUMONI to index large databases more efficiently.
- In a host-depletion scenario, where SPUMONI indexed human genomes …
  - Previously, SPUMONI was ~2X faster at classifying than minimap2 (3 human genomes).
  - Using the new indexing approach and **4x as many human genomes**, SPUMONI can classify reads **6-15X faster than previously with only 3 human genomes.**
- We envision in the future we can classify metagenomic reads robustly **using distributions of matching statistics** to account for database growth and variable-sized classes.



"Null" distribution allows the **notion of significance to be a function of database sequences**, and the **error rate of the query read.**

Thank you!

Contact: oahmed6@jhu.edu

Twitter: @oyfahmed

GitHub: https://github.com/oma219/spumoni

Acknowledgements:

- Special thanks to Massimiliano Rossi, Daniel N. Baker, & Ben Langmead for assistance on project.
- Thanks to Sam Kovaka, Michael C. Schatz, Travis Gagie, Christina Boucher for help & assistance on the project
- Thanks to Nae-Chyun Chen, Taher Mun, Kathleen Newcomer, Anna Liebhoff, Dominik Kempa, Mao-Jan Lin and Kavya Vaddadi

Funding: NIH NSF NSERC CRSNG