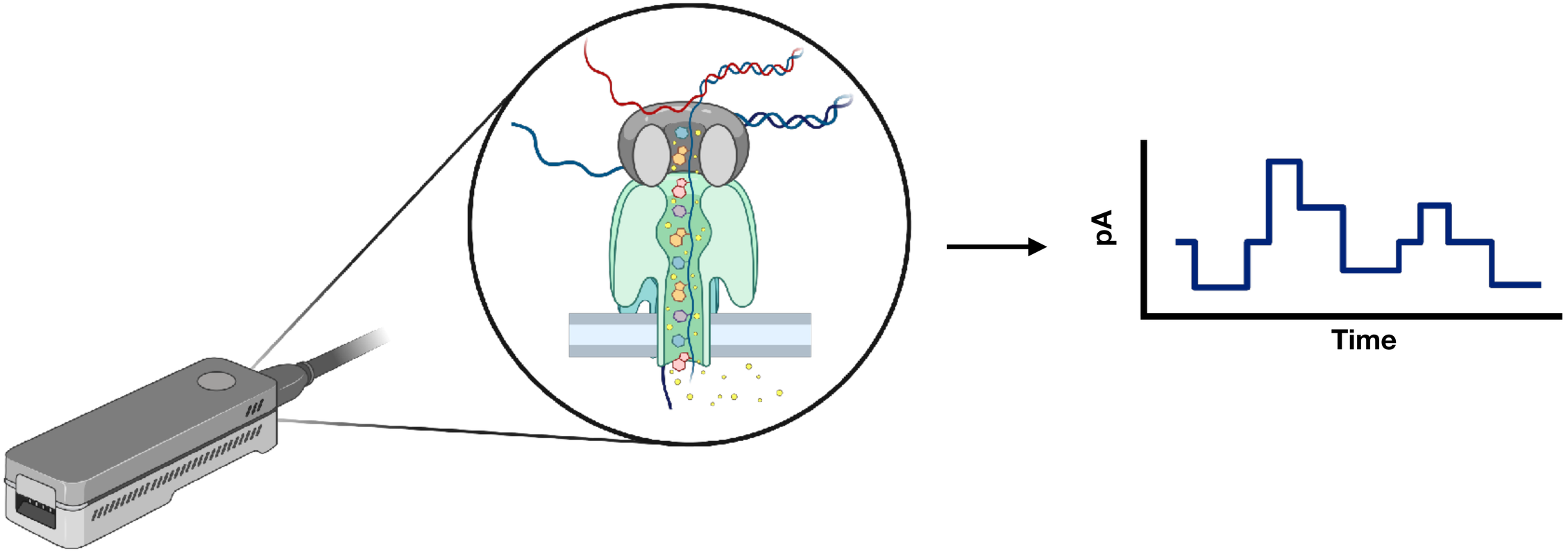
Sigmoni: efficient pangenome multi-classification of nanopore signal

Vikram Shivakumar

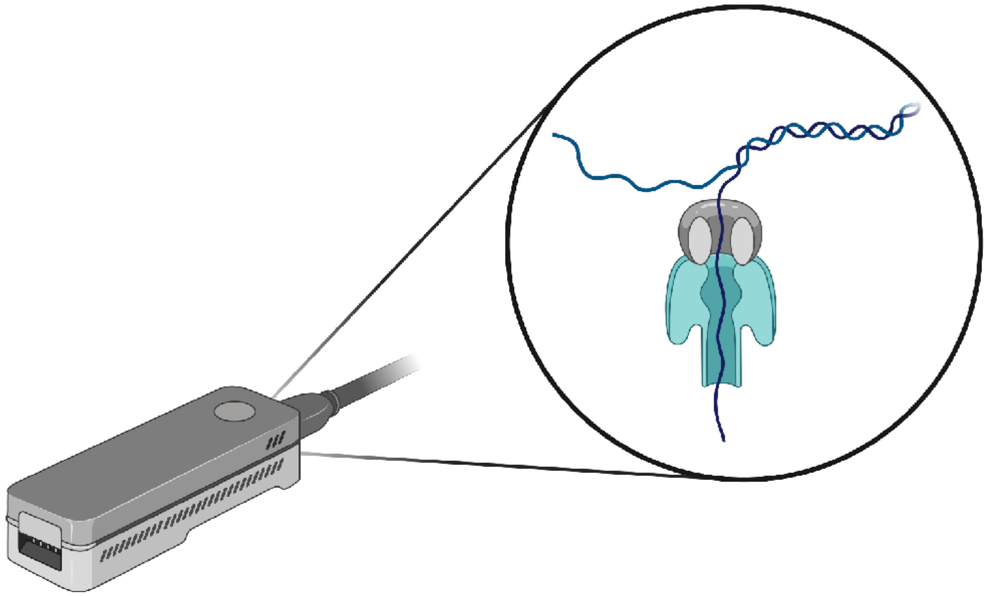
7/13/24
ISMB
Montréal, Québec

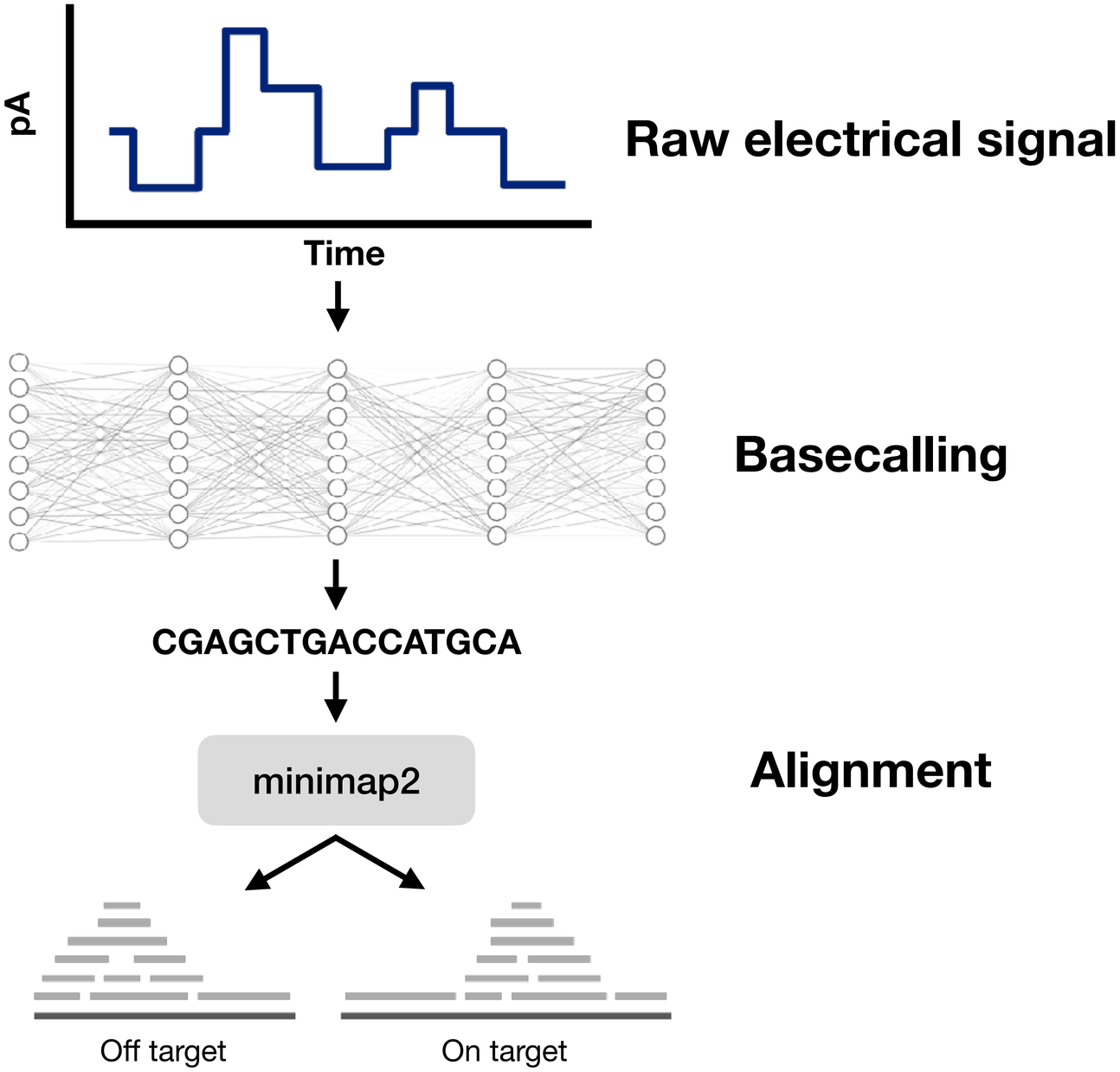
JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Nanopore sequencing

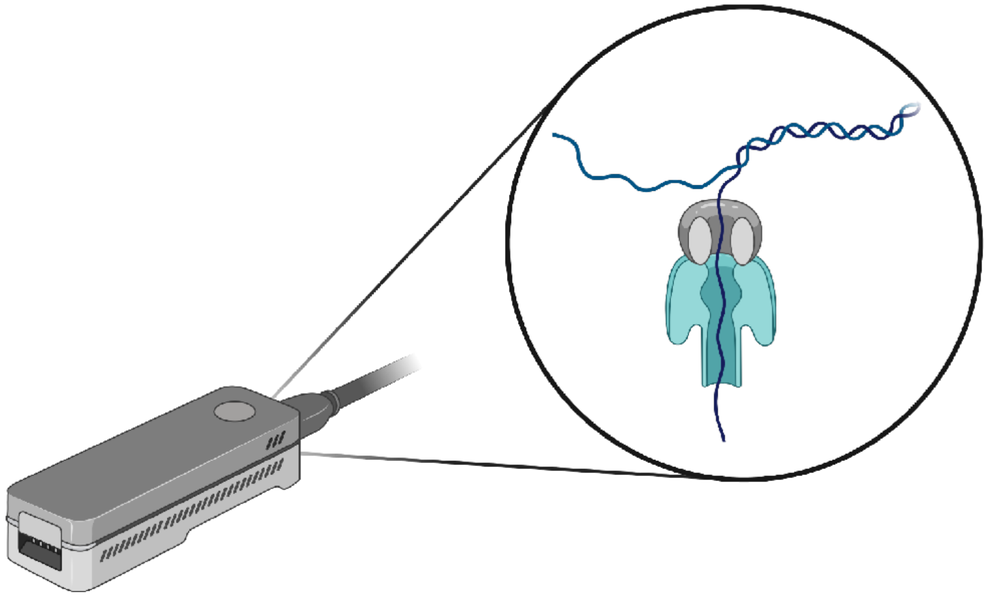


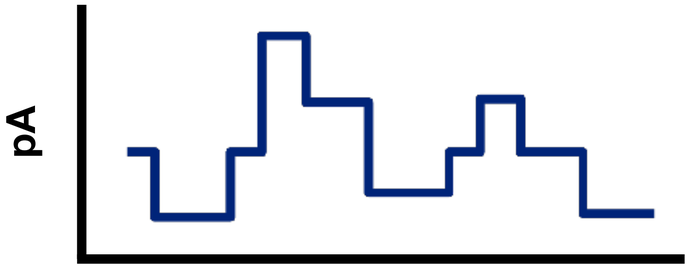
Nanopore sequencing



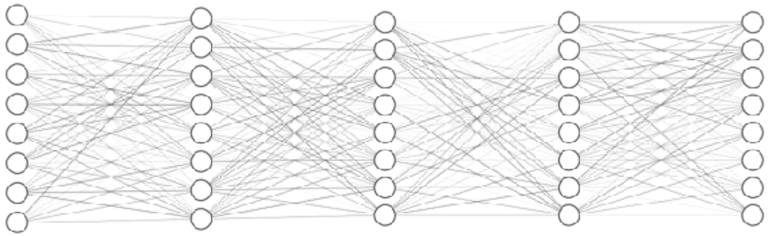


Nanopore sequencing





Raw electrical signal

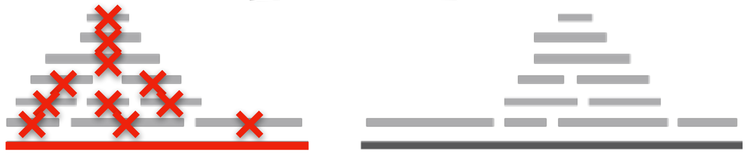


Basecalling

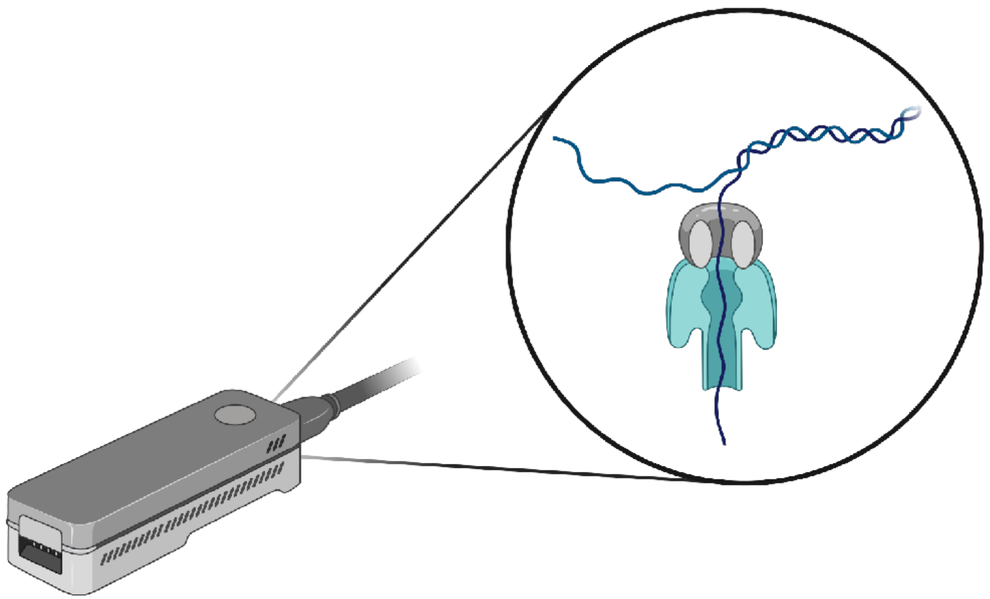
CGAGCTGACCATGCA

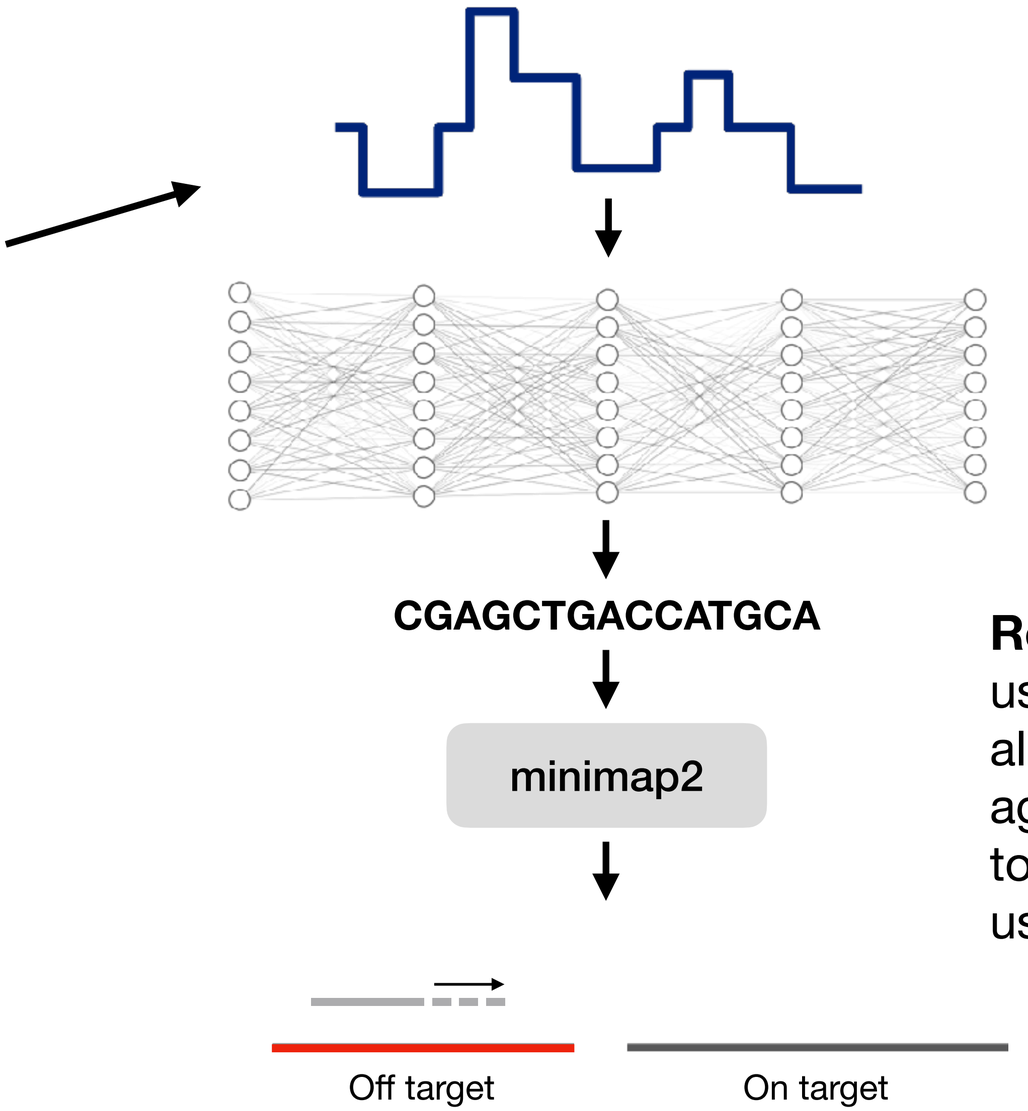
minimap2

Alignment



Adaptive sampling

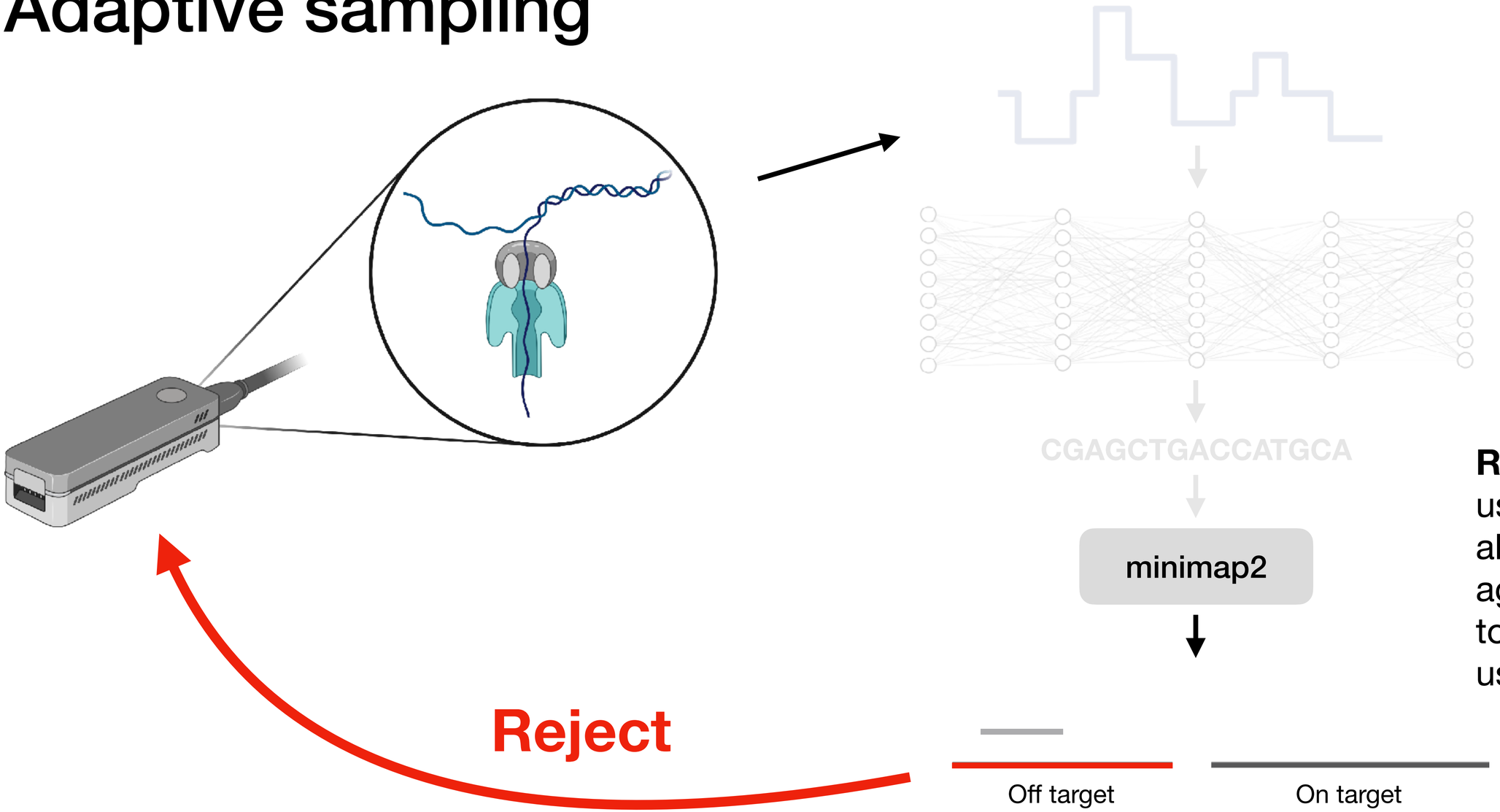




Readfish
uses minimap2, aligning read chunks against a reference to classify reads, but uses the basecaller

Payne, et al. *Nature Biotechnology* (2021)

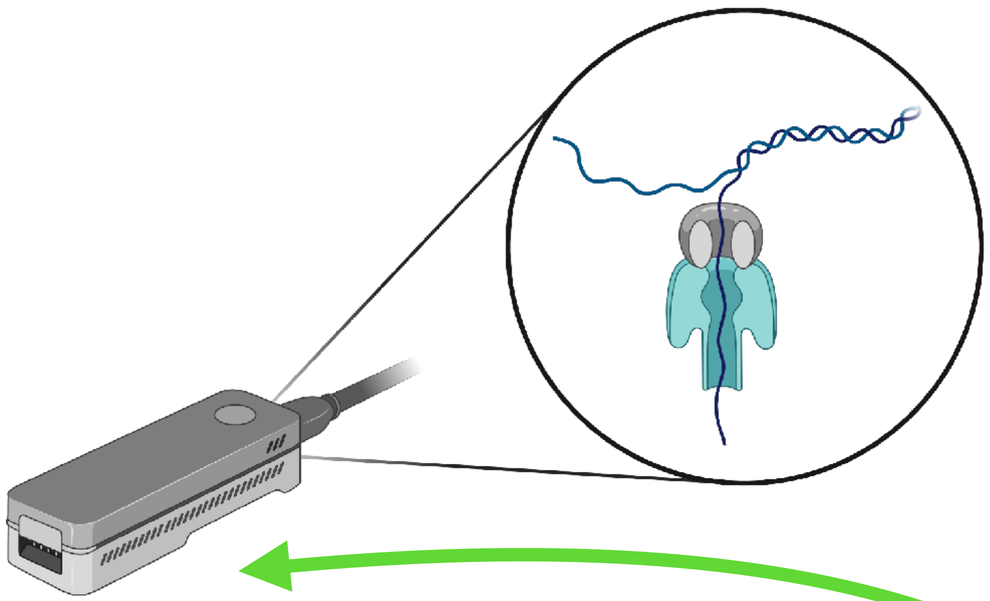
Adaptive sampling



Readfish
uses minimap2, aligning read chunks against a reference to classify reads, but uses the basecaller

Payne, et al. *Nature Biotechnology* (2021)

Adaptive sampling



CGAGCTGACCATGCA

minimap2

Off target

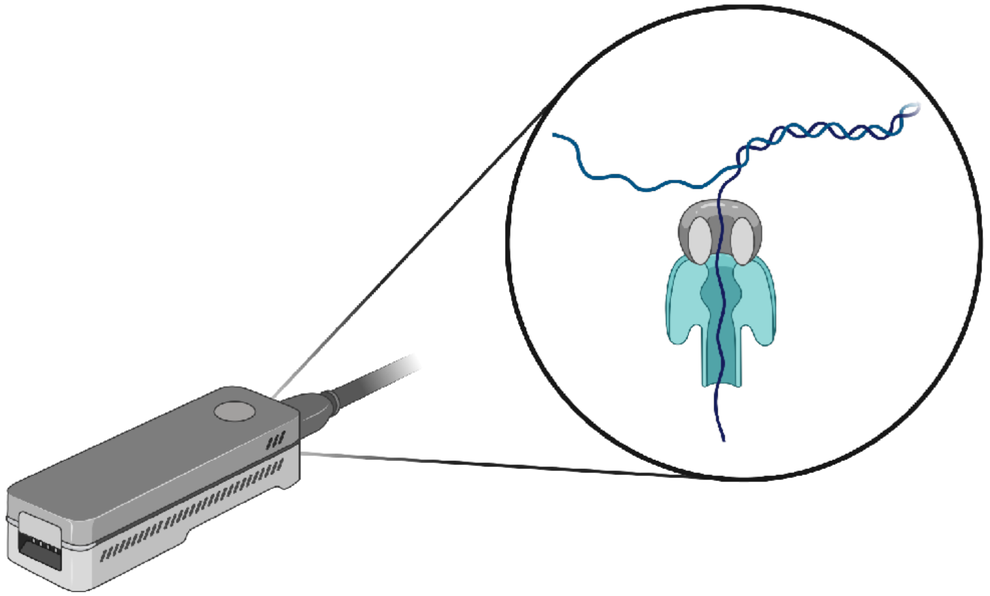
On target

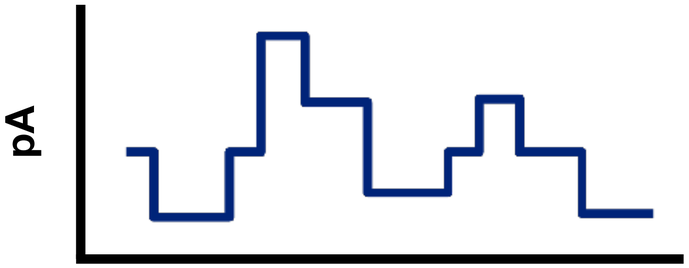
Readfish
uses minimap2, aligning read chunks against a reference to classify reads, but uses the basecaller

Continue sequencing

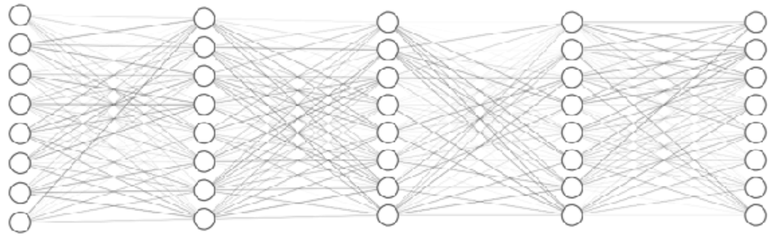
Payne, et al. *Nature Biotechnology* (2021)

Nanopore sequencing





Raw electrical signal

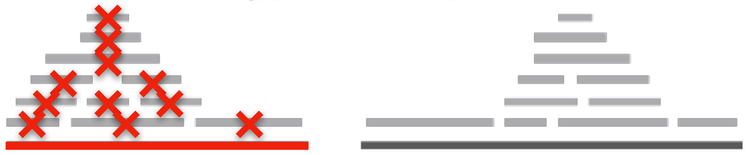


Basecalling

CGAGCTGACCATGCA

minimap2

Alignment



Nanopore sequencing

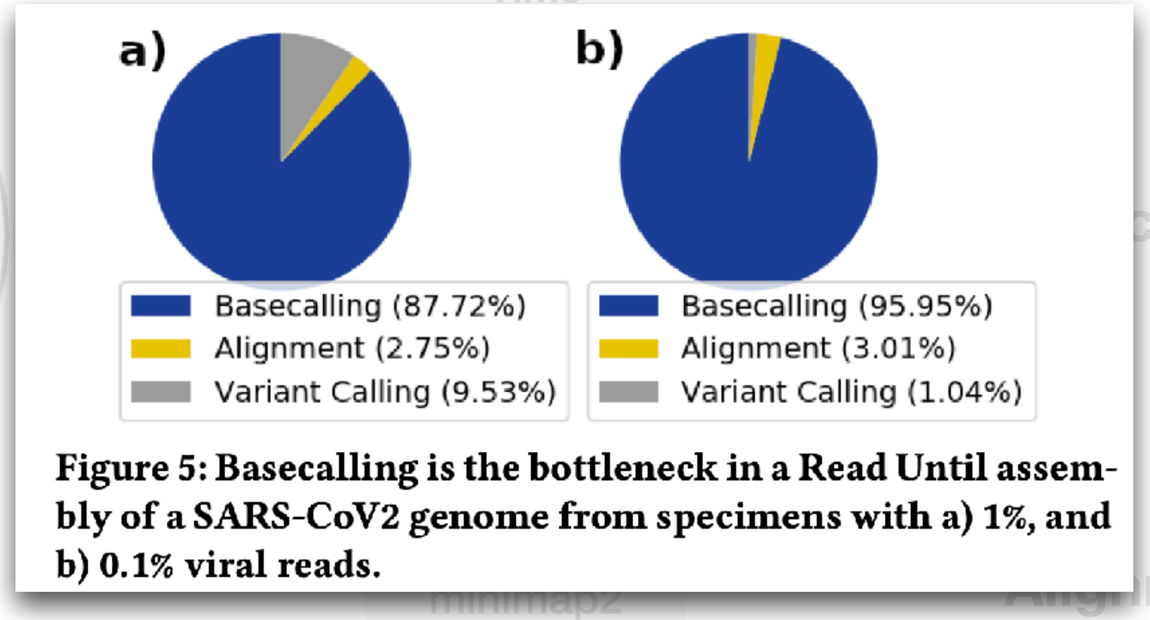
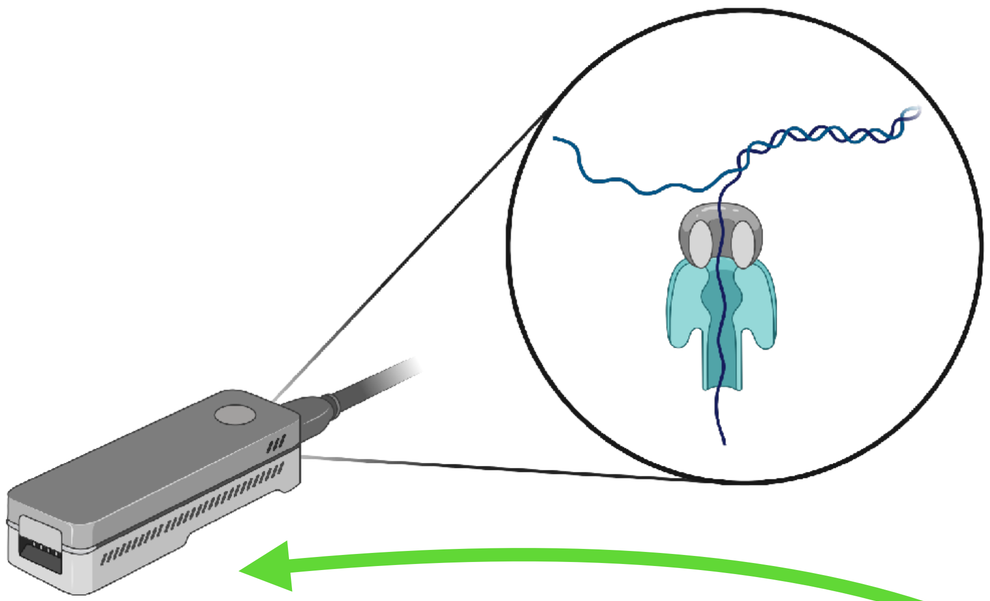
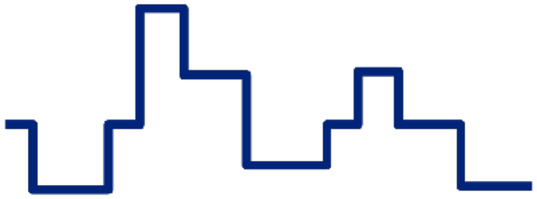


Figure 5: Basecalling is the bottleneck in a Read Until assembly of a SARS-CoV2 genome from specimens with a) 1%, and b) 0.1% viral reads.

Dunn, T., *et al*. (2021). Squigglefilter: An accelerator for portable virus detection. In *MICRO-54: 54th Annual IEEE/ACM International Symposium on Microarchitecture* (pp. 535-549).

Adaptive sampling

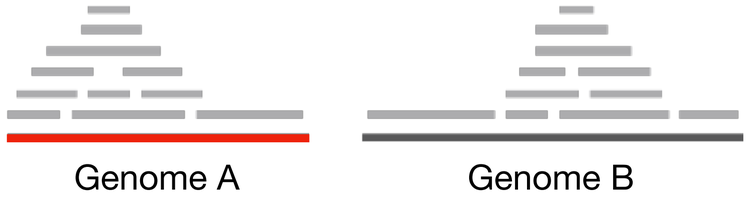




UNCALLED
identifies likely k-mer sequences from the signal and finds collinear matches using an FM-index

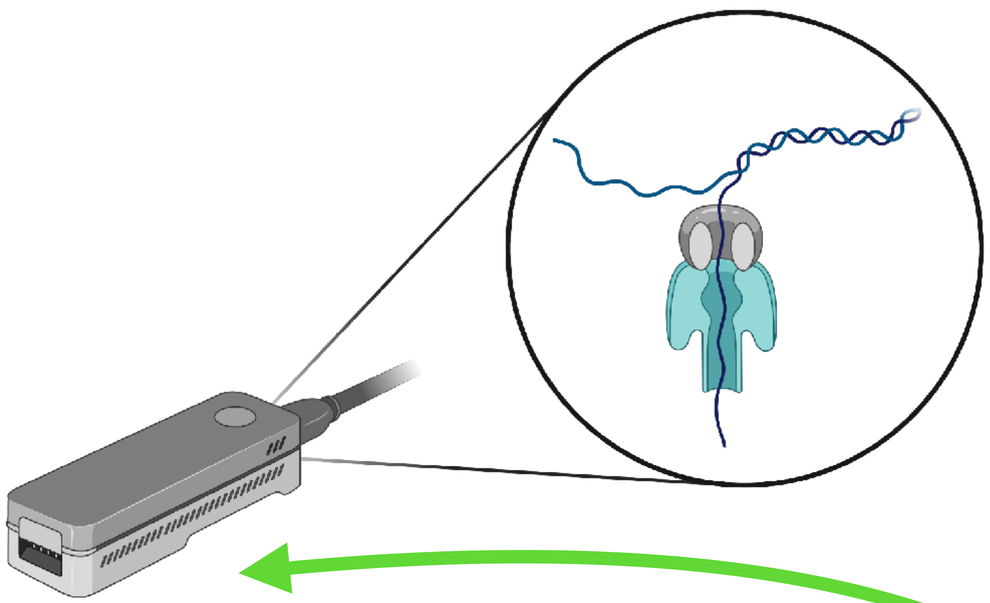
Continue sequencing

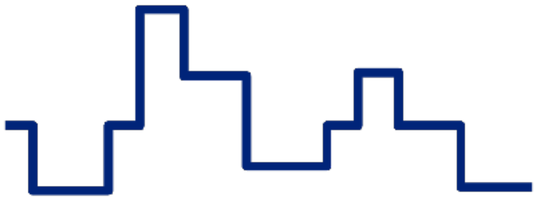
Reject



Kovaka, et al. *Nature Biotechnology* (2021)

Adaptive sampling

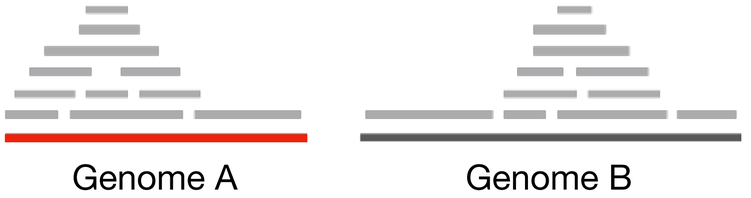




Sigmap
uses a *k-d* tree
to query signal
seeds against
the reference

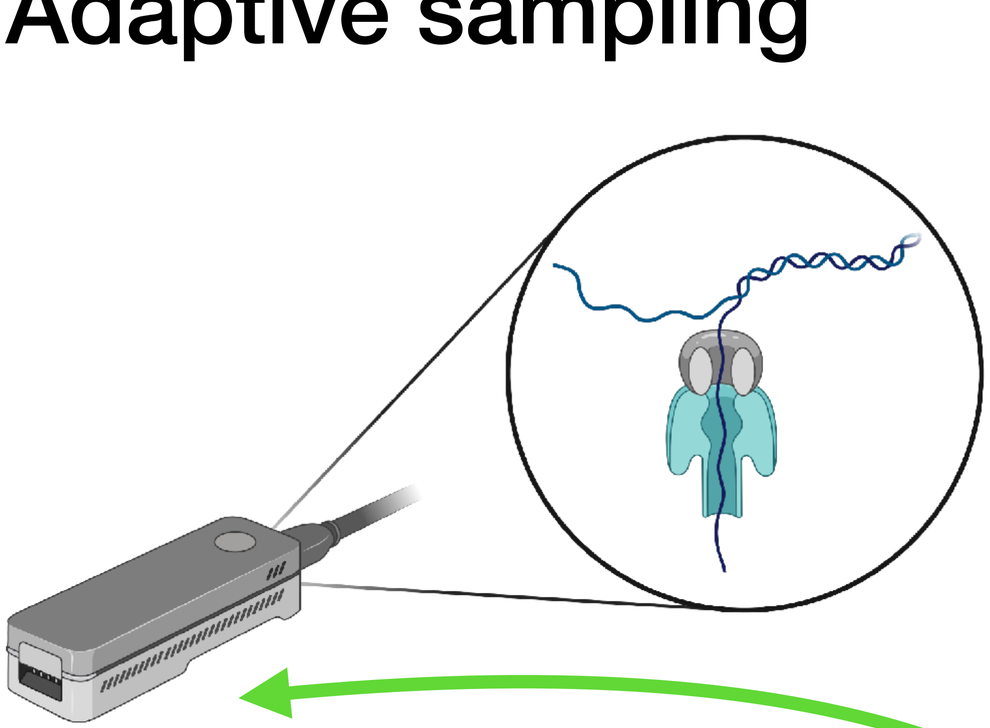
Continue sequencing

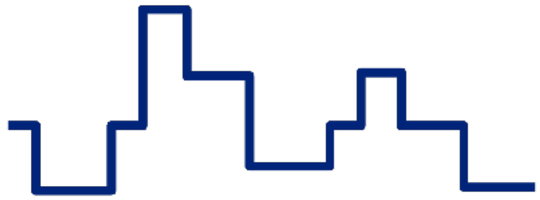
Reject



Zhang, et al. *Bioinformatics* (2021)

Adaptive sampling



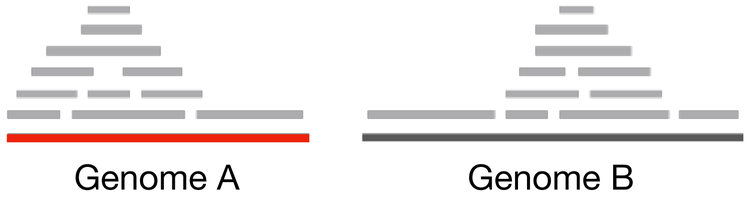


?

pangenome index?

Continue sequencing

Reject



r-index for exact matching

- Previous work (**SPUMONI**) uses the *r*-index to classify *basecalled* nanopore reads against a reference

- *r*-index
  (1) run-length encoded BWT +
  (2) sampled suffix array +
  (3) auxiliary data-structures

$O(r)$ in size

- SPUMONI framework enables linear time exact match computation against a reference index

Home > Genome Biology > Article

SPUMONI 2: improved classification using a pangenome index of minimizer digests

Software | Open access | Published: 18 May 2023

Volume 24, article number 122, (2023) Cite this article

Omar Y. Ahmed, Massimiliano Rossi, Travis Gagie, Christina Boucher & Ben Langmead

$O(r)$ data structures scale well

De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes

MATTHEW B. HUFFORD, ARUN S. SEETHARAM, MARGARET R. WOODHOUSE, KAPEEL M. CHOUGULE, SHUJUN OU, JIANING LIU, WILLIAM A. RICCI, TINGTING GUO, ANDREW OLSON, [...], AND R. KELLY DAWE +36 authors

SCIENCE · 6 Aug 2021 · Vol 373, Issue 6555 · pp. 655-662 · DOI: 10.1126/science.abg5289

42.33

Article | Open access | Published: 10 May 2023

A draft human pangenome reference

Wen-Wei Liao, Mobin Asri, Jana Ebler, Daniel Doerr, Marina Haukness, Glenn Hickey, Shuangjia Lu, Julian K. Lucas, Jean Monlong, Haley J. Abel, Silvia Buonaiuto, Xian H. Chang, Haoyu Cheng, Justin Chu, Vincenza Colonna, Jordan M. Eizenga, Xiaowen Feng, Christian Fischer, Robert S. Fulton, Shilpa Garg, Cristian Groza, Andrea Guarracino, William T. Harvey, Simon Heumos, ... Benedict Paten

+ Show authors

Nature **617**, 312–324 (2023) | Cite this article

135.16

Article | Open access | Published: 11 April 2024

A pan-genome of 69 *Arabidopsis thaliana* accessions reveals a conserved genome structure throughout the global species range

Qichao Lian, Bruno Huettel, Birgit Walkemeier, Baptiste Mayjonade, Céline Lopez-Roques, Lisa Gil, Fabrice Roux, Korbinian Schneeberger & Raphael Mercier

Nature Genetics (2024) | Cite this article

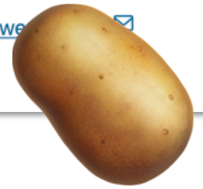
42.03

Article | Open access | Published: 08 June 2022

Genome evolution and diversity of wild and cultivated potatoes

Dié Tang, Yuxin Jia, Jinzhe Zhang, Hongbo Li, Lin Cheng, Pei Wang, Zhigui Bao, Zhihong Liu, Shuangshuang Feng, Xijian Zhu, Dawei Li, Guangtao Zhu, Hongru Wang, Yao Zhou, Yongfeng Zhou, Glenn J. Bryan, C. Robin Buell, Chunzhi Zhang & Sanwe

Nature **606**, 535–541 (2022) | Cite this article



30.20

Article | Open access | Published: 31 July 2023

Telomere-to-telomere assemblies of 142 strains characterize the genome structural landscape in *Saccharomyces cerevisiae*

Samuel O'Donnell, Jia-Xing Yue, Omar Abou Saada, Nicolas Agier, Claudia Caradec, Thomas Cokelaer, Matteo De Chiara, Stéphane Delmas, Fabien Dutreux, Téo Fournier, Anne Friedrich, Etienne Kornobis, Jing Li, Zepu Miao, Lorenzo Tattini, Joseph Schacherer, [...] & Gilles Fischer

Nature Genetics **55**, 1390–1399 (2023) | Cite this article

65.30

compression ratio (n/r)

r-index for exact matching

- SPUMONI framework enables **linear time** exact match computation against a reference index, in $O(r)$ space

Ref : GATTACATACATAAT

Query: ATGAATTACTAA

Matching statistic: half maximal exact matches (half MEM)

r-index for exact matching

- SPUMONI framework enables **linear time** exact match computation against a reference index, in $O(r)$ space

Matching statistic: half maximal exact matches (half MEM)

Ref : GATTACATACATAAT

Query: ATGAATTACTAA

matching statistics:

2,

r-index for exact matching

- SPUMONI framework enables **linear time** exact match computation against a reference index, in $O(r)$ space

Matching statistic: half maximal exact matches (half MEM)

Ref : GATTACATACATAAT

Query: ATGAATTACTAA

matching statistics:

2, 1,

r-index for exact matching

- SPUMONI framework enables **linear time** exact match computation against a reference index, in $O(r)$ space

Matching statistic: half maximal exact matches (half MEM)

Ref : GATTACATACATAAT

Query: ATGAATTACTAA

matching statistics:

2, 1, 2,

r-index for exact matching

- SPUMONI framework enables **linear time** exact match computation against a reference index, in $O(r)$ space

Matching statistic: half maximal exact matches (half MEM)

Ref : GATTACATACATAAT

Query: ATGAATTACTAA

matching statistics:

2, 1, 2, 3,

r-index for exact matching

- SPUMONI framework enables **linear time** exact match computation against a reference index, in $O(r)$ space

Matching statistic: half maximal exact matches (half MEM)

Ref : GATTACATACATAAT

Query: ATGAATTACTAA

matching statistics:

2, 1, 2, 3, 5,

r-index for exact matching

- SPUMONI framework enables **linear time** exact match computation against a reference index, in $O(r)$ space

Matching statistic: half maximal exact matches (half MEM)

Ref : GATTACATACATAAT

Query: ATGAATTACTAA

matching statistics:

2, 1, 2, 3, 5, 4, 3, 2, 1, 3, 2, 1

r-index for exact matching

- SPUMONI framework enables **linear time** exact match computation against a reference index, in $O(r)$ space

Matching statistic: half maximal exact matches (half MEM)

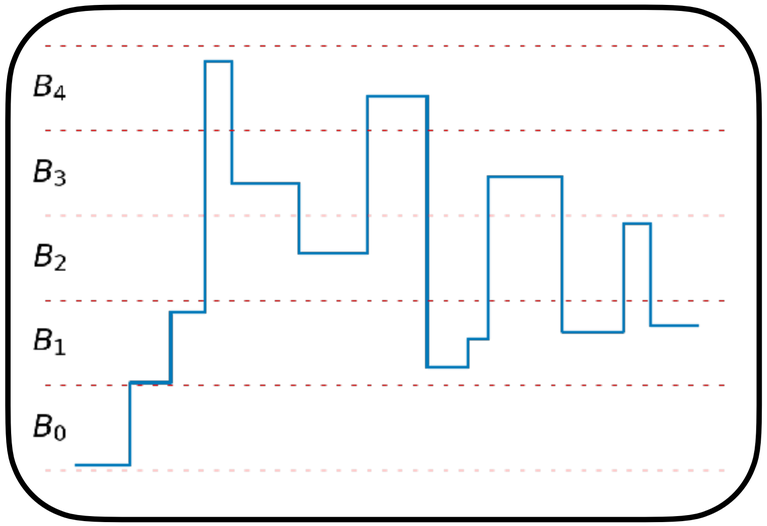
We actually use **pseudomatching lengths** (similar to matching statistics, but faster to compute in a streaming manner)

Sigmoni

- Signal-based read classification using the *r*-index
- Idea: quantize signal into a discrete alphabet, project reference to the same alphabet, and compute exact matches
- Similar to SPUMONI, but avoiding basecalling reads by using a rapid signal binning procedure

Methods

Query



$B_0\ B_1\ B_4\ B_3\ \ldots\ B_1\ B_2\ B_1$

Reference

GATTACACATACAAG…

GATTACA

ATTACAC

TTACACA

TACACAT

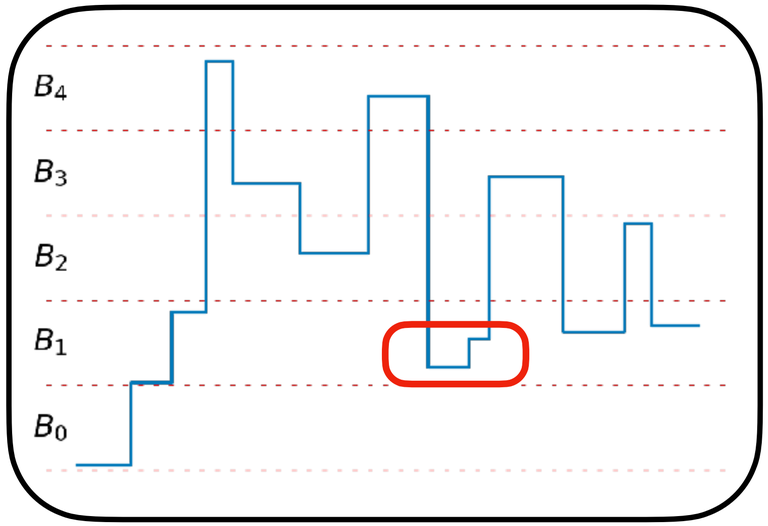
$\bar{C}_0\ \bar{C}_1\ \bar{C}_1\ \bar{C}_3\ \bar{C}_2$ Expected Current

$B_0\ B_1\ B_3\ B_2\ \ldots\ B_0\ B_2\ B_3$ Bin sequence (w/ Homopolymer Compression)

***r*-index**

Methods

Query



$B_0\ B_1\ B_4\ B_3\ \ldots\ B_1\ B_2\ B_1$

Reference

GATTACACATACAAG…

GATTACA

ATTACAC

TTACACA

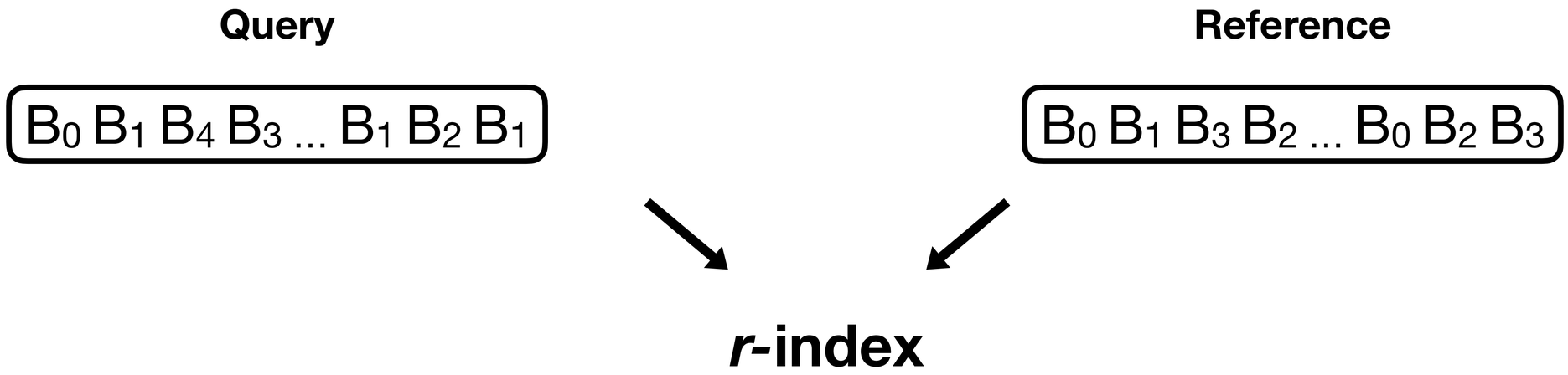
TACACAT

$\bar{C}_0\ \bar{C}_1\ \bar{C}_1\ \bar{C}_3\ \bar{C}_2$ Expected Current

$B_0\ B_1\ B_3\ B_2\ \ldots\ B_0\ B_2\ B_3$ Bin sequence (w/ Homopolymer Compression)

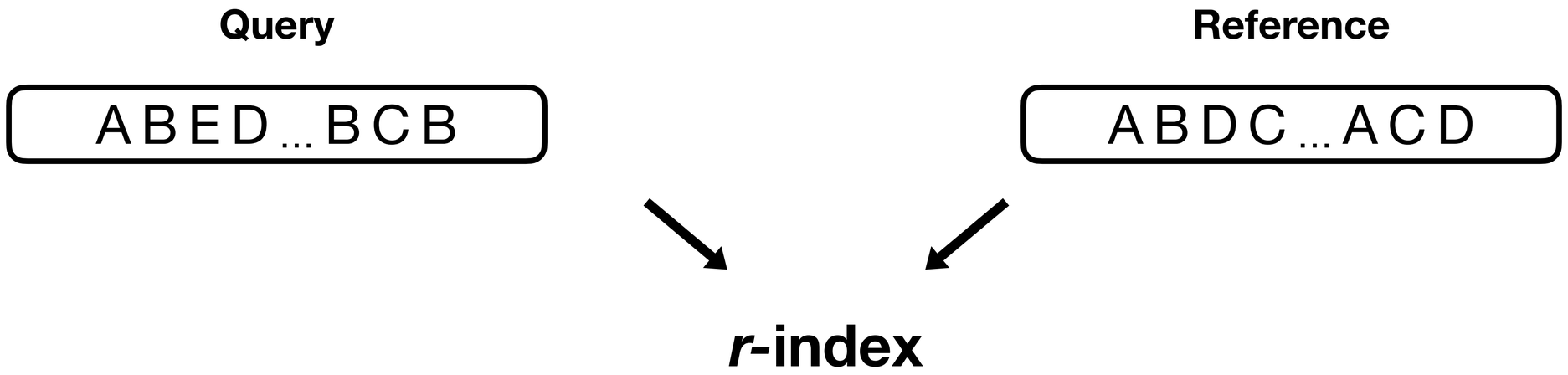
***r*-index**

Methods


Query
Reference
$B_0\ B_1\ B_4\ B_3\ \ldots\ B_1\ B_2\ B_1$
$B_0\ B_1\ B_3\ B_2\ \ldots\ B_0\ B_2\ B_3$
r-index

Methods

$$\Sigma = \{A, B, C, D, E\}$$



Methods

Query

A B E D ... B C B

Reference

A B D C ... A C D

Bin sequence (w/ HPC)

i	BWM	BWT	SA	doc id
12	ABAEEEAEBAAE#AAAACBDEABAE...	E	29	2
13	ABAEEEAEBAAE#AAACCBDEABAE...	E	8	1
14	ABAEEEAEBEAAE$AAEACBDEABA...	E	50	3
15	ACBDEABAEEEAEBAAE#AAACCBD...	E	3	1
16	ACBDEABAEEEAEBEAAE$AAEACB...	A	45	3
17	ACCBDEABAEEEAEBAAE#AAAACB...	A	23	2
18	AE#AAAACBDEABAEEEAEBEAAE$...	A	39	2
19	AE#AAACCBDEABAEEEAEBAAE#A...	A	18	1
20	AE$AAEACBDEABAEEEAEBAAE#A...	A	61	3
21	AEACBDEABAEEEAEBAAE#AAACC...	A	1	1
22	AEBAAE#AAAACBDEABAEEEAEBE...	E	35	2
23	AEBAAE#AAACCBDEABAEEEAEBA...	E	14	1
24	AEBEAAE$AAEACBDEABAEEEAEB...	E	56	3
25	AEEEAEBAAE#AAAACBDEABAEEE...	B	31	2
26	AEEEAEBAAE#AAACCBDEABAEEE...	B	10	1
27	AEEEAEBEAAE$AAEACBDEABAEE...	B	52	3
28	BAAE#AAAACBDEABAEEEAEBEAA...	E	37	2
29	BAAE#AAACCBDEABAEEEAEBAAE...	E	16	1
30	BAEEEAEBAAE#AAAACBDEABAEE...	A	30	2
31	BAEEEAEBAAE#AAACCBDEABAEE...	A	9	1
32	BAEEEAEBEAAE$AAEACBDEABAE...	A	51	3

***r*-index**

doc id	genome
1	*Pseudomonas aeruginosa*
2	*Escherichia coli*
3	*Saccharomyces cerevisiae*

Ahmed, Omar, et al. "**SPUMONI 2**: Improved pangenome classification using a compressed index of minimizer digests." *bioRxiv* (2022).

Methods

Query

A B E D ... B C B

Reference

A B D C ... A C D

Bin sequence (w/ HPC)

i	BWM	BWT	SA	doc id
12	ABAEEEAEBAAE#AAAACBDEABAE...	E	29	2
13	ABAEEEAEBAAE#AAACCBDEABAE...			
14	ABAEEEAEBEAAE$AAEACBDEABA...			
15	ACBDEABAEEEAEBAAE#AAACCBD...	E	3	1
16	ACBDEABAEEEAEBEAAE$AAEACB...	A	45	3
17	ACCBDEABAEEEAEBAAE#AAAACB...			
18	AE#AAAACBDEABAEEEAEBEAAE$...			
19	AE#AAACCBDEABAEEEAEBAAE#A...			
20	AE$AAEACBDEABAEEEAEBAAE#A...			
21	AEACBDEABAEEEAEBAAE#AAACC...	A	1	1
22	AEBAAE#AAAACBDEABAEEEAEBE...	E	35	2
23	AEBAAE#AAACCBDEABAEEEAEBA...			
24	AEBEAAE$AAEACBDEABAEEEAEB...	E	56	3
25	AEEEAEBAAE#AAAACBDEABAEEE...	B	31	2
26	AEEEAEBAAE#AAACCBDEABAEEE...			
27	AEEEAEBEAAE$AAEACBDEABAEE...	B	52	3
28	BAAE#AAAACBDEABAEEEAEBEAA...	E	37	2
29	BAAE#AAACCBDEABAEEEAEBAAE...	E	16	1
30	BAEEEAEBAAE#AAAACBDEABAEE...	A	30	2
31	BAEEEAEBAAE#AAACCBDEABAEE...			
32	BAEEEAEBEAAE$AAEACBDEABAE...	A	51	3

***r*-index**

doc id	genome
1	*Pseudomonas aeruginosa*
2	*Escherichia coli*
3	*Saccharomyces cerevisiae*

Ahmed, Omar, et al. "**SPUMONI 2**: Improved pangenome classification using a compressed index of minimizer digests." *bioRxiv* (2022).

Methods

Query

A B E D ... B C B

Reference

A B D C ... A C D

Bin sequence (w/ HPC)

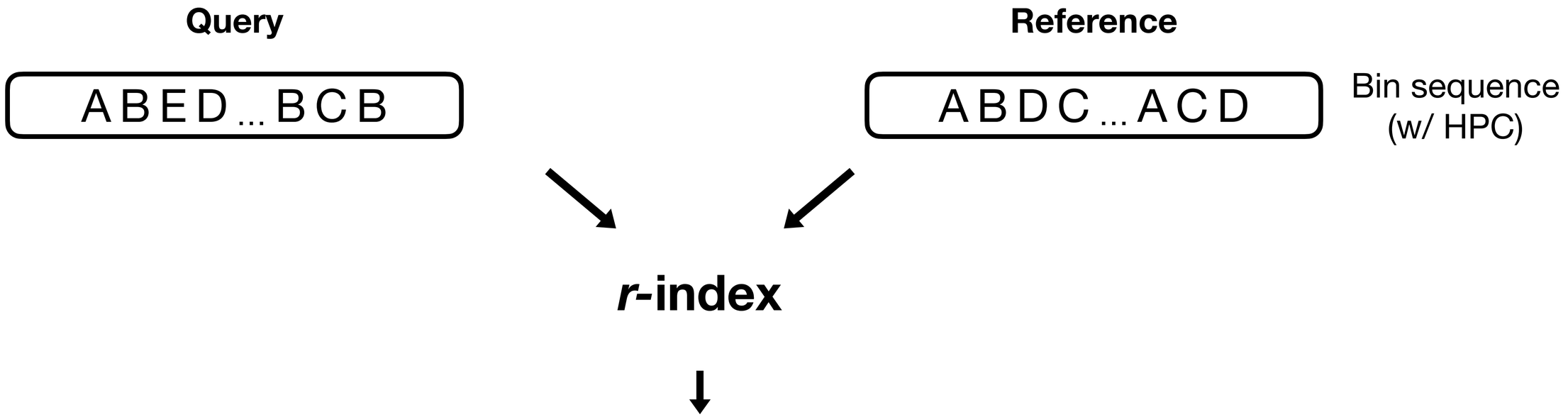
i	BWM	BWT	SA	doc id
12	ABAEEEAEBAAE#AAAACBDEABAE...	E	5034	5
13	ABAEEEAEBAAE#AAACCBDEABAE...			
14	ABAEEEAEBEAAE$AAEACBDEABA...			
15	ACBDEABAEEEAEBAAE#AAACCBD...	E	44920	52
16	ACBDEABAEEEAEBEAAE$AAEACB...	A	4599	23
17	ACCBDEABAEEEAEBAAE#AAAACB...			
18	AE#AAAACBDEABAEEEAEBEAAE$...			
19	AE#AAACCBDEABAEEEAEBAAE#A...			
20	AE$AAEACBDEABAEEEAEBAAE#A...			
21	AEACBDEABAEEEAEBAAE#AAACC...	A	12992	22
22	AEBAAE#AAAACBDEABAEEEAEBE...	E	35227	23
23	AEBAAE#AAACCBDEABAEEEAEBA...			
24	AEBEAAE$AAEACBDEABAEEEAEB...	E	5619	73
25	AEEEAEBAAE#AAAACBDEABAEEE...	B	1021	2
26	AEEEAEBAAE#AAACCBDEABAEEE...			
27	AEEEAEBEAAE$AAEACBDEABAEE...	B	29945	39
28	BAAE#AAAACBDEABAEEEAEBEAA...	E	37395	44
29	BAAE#AAACCBDEABAEEEAEBAAE...	E	29251	61
30	BAEEEAEBAAE#AAAACBDEABAEE...	A	3009	7
31	BAEEEAEBAAE#AAACCBDEABAEE...			
32	BAEEEAEBEAAE$AAEACBDEABAE...	A	1154	3

***r*-index**

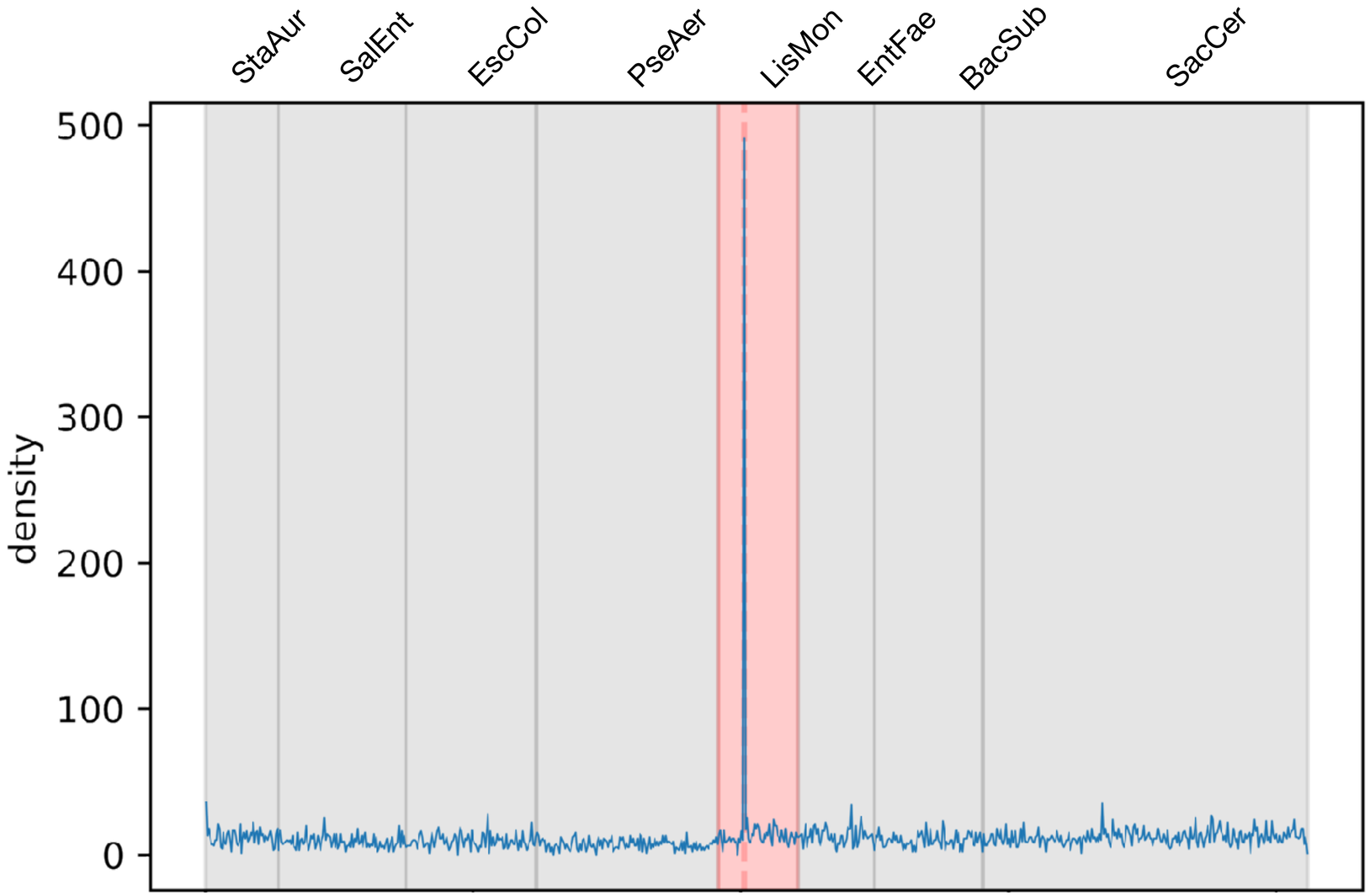
doc id	"shred"
1	P. aeruginosa 1-2Kbp
2	P. aeruginosa 2-3Kbp
	...
75	S. cerevisiae 12.2-12.3Mb
76	S. cerevisiae 12.3-12.4Mb

shredded documents

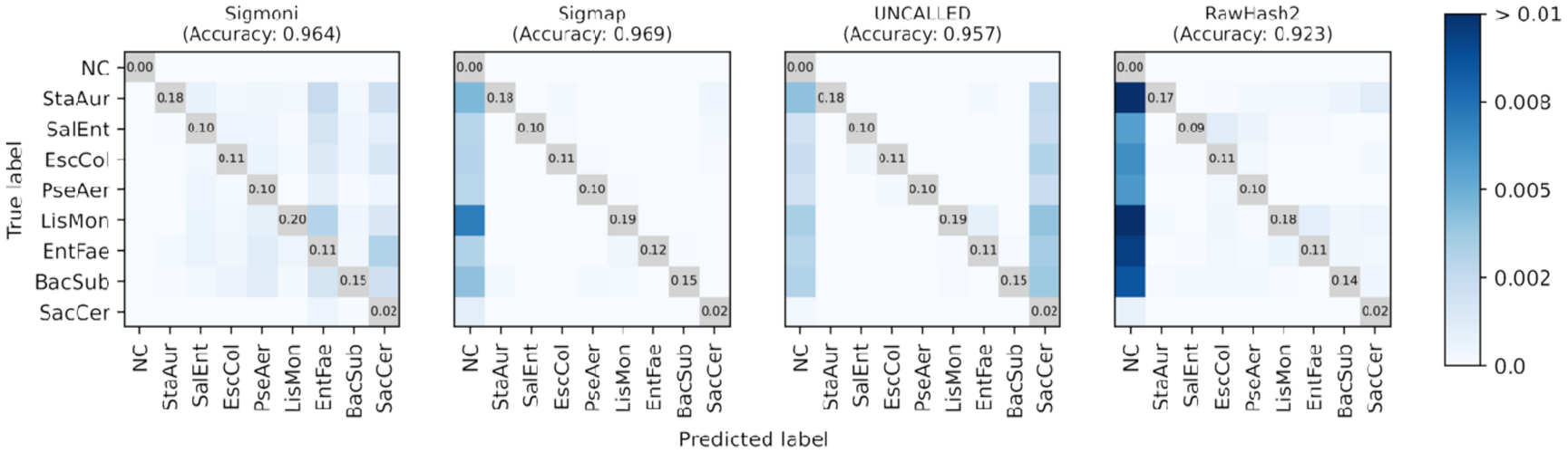
Methods


Query
A B E D ... B C B
Reference
A B D C ... A C D
Bin sequence
(w/ HPC)
r-index
Half MEMs across the query

Methods


StaAur
SalEnt
EscCol
PseAer
LisMon
EntFae
BacSub
SacCer
density
500
400
300
200
100
0

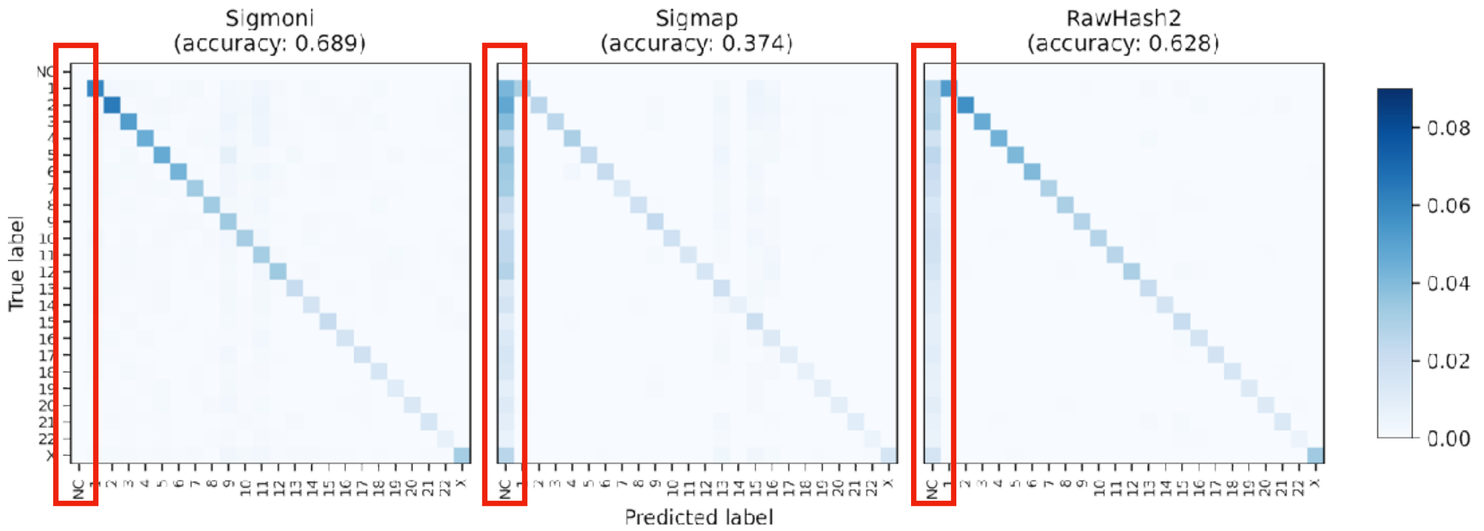
Microbial community read classification



	Precision	Recall	F1 weighted	Unclassified rate	Index size (MB)	Time (s)
Sigmoni	**1.0**	0.928	**0.963**	**0.0**	177.6	**423.3**
Sigmap	0.956	0.949	0.952	0.046	2591.3	891.4
UNCALLED	0.460	**0.986**	0.627	0.044	**72.1**	1303.6
RawHash2	0.876	0.958	0.915	0.095	567.4	653.9

Kovaka., et al (2021). *Nature biotechnology*, *14(1)*.
Firtina., et al (2023). *arXiv*:2309.05771
Zhang., et al (2021). *Bioinformatics*, *37*.

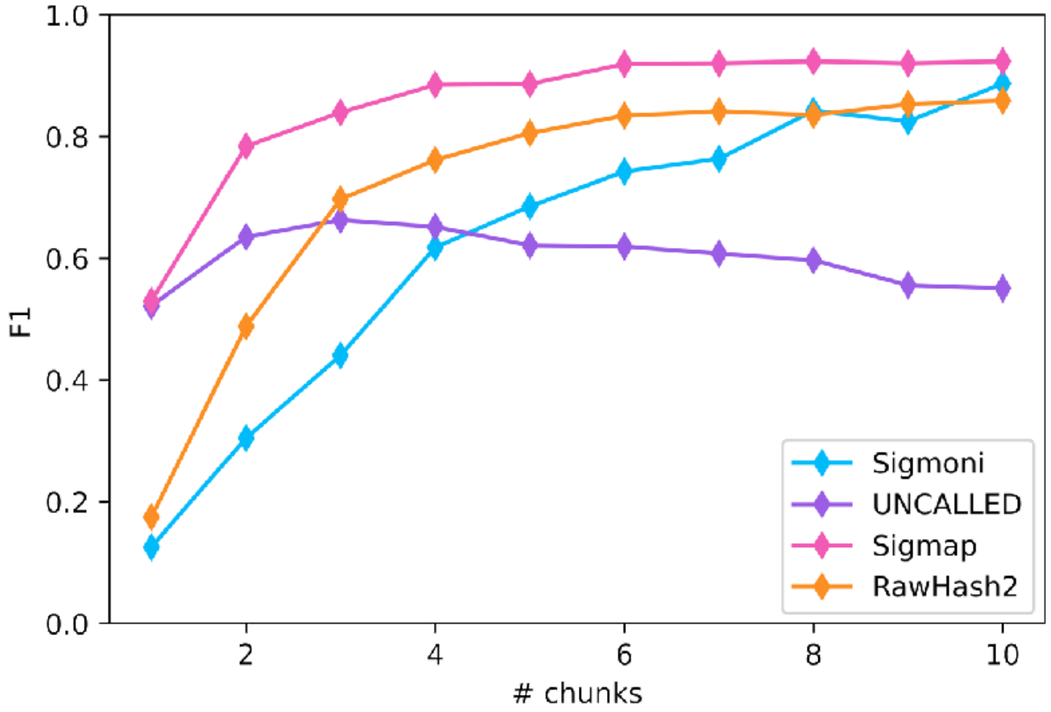
Host depletion

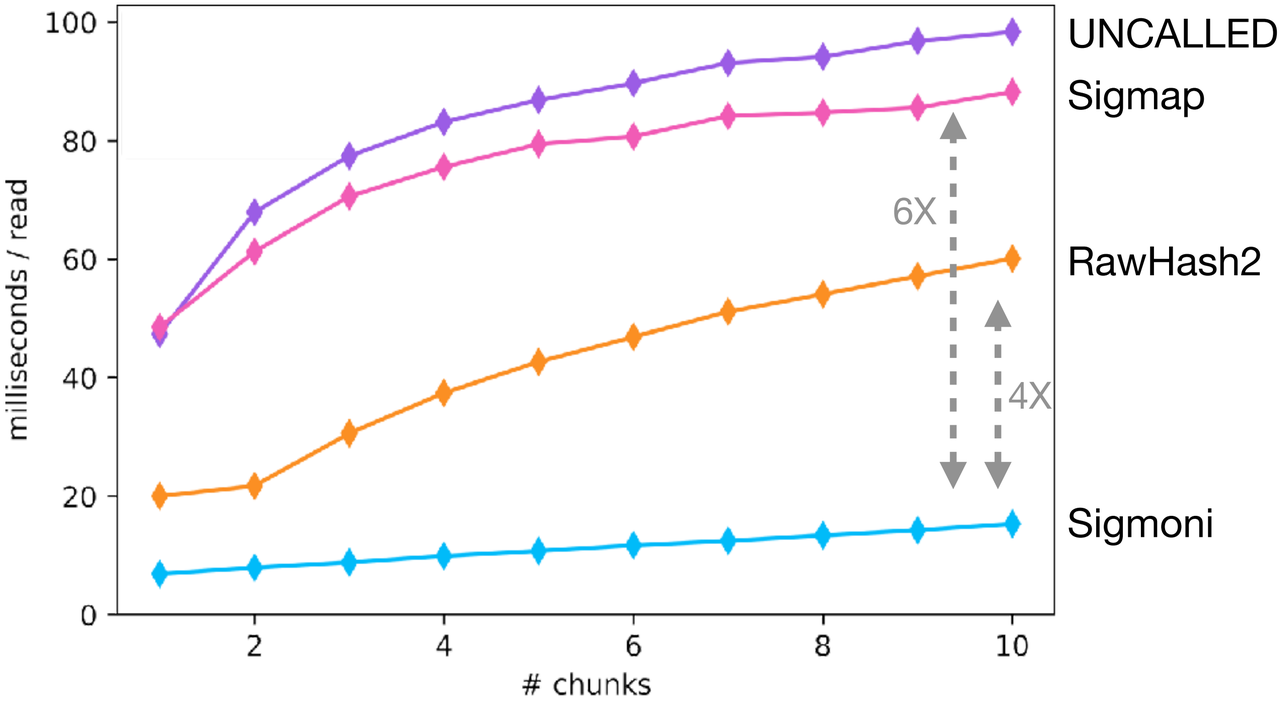


	Precision	Recall	F1 weighted	Unclassified rate	Index size (GB)	Time (s)
Sigmoni	**0.882**	0.979	**0.929**	**0.0**	**14.7**	**537.4**
Sigmap	0.831	0.954	0.888	0.382	175.9	77925.4
UNCALLED	*	*	*	*	*	*
RawHash2	0.839	**0.998**	0.912	0.240	41.1	35159.8

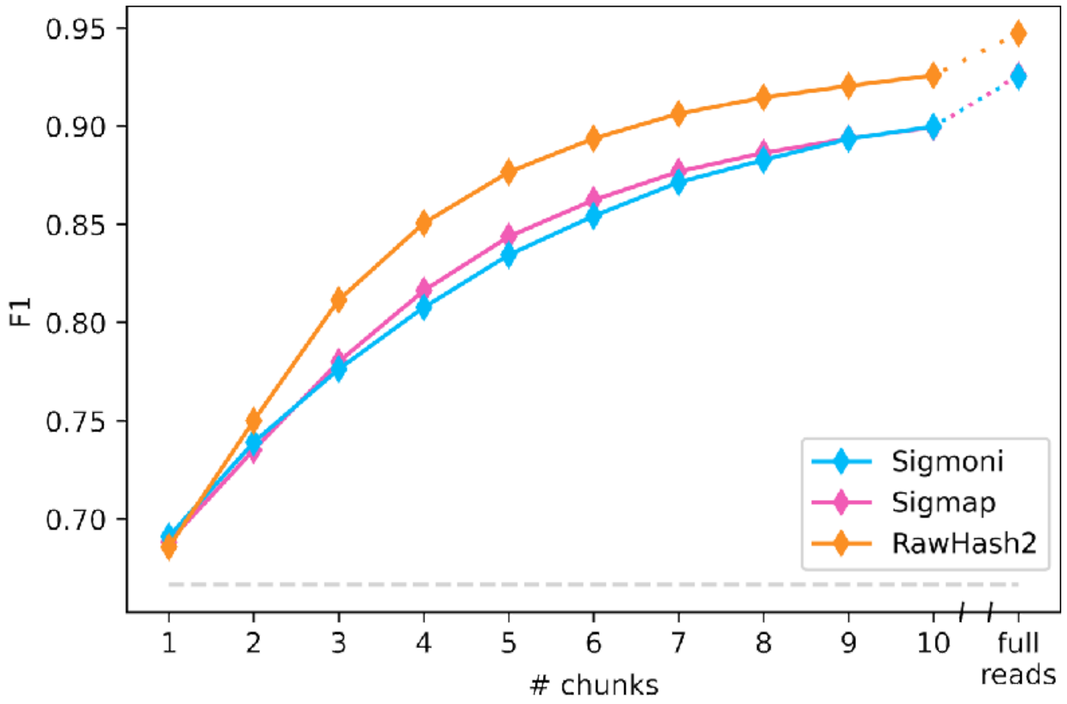
Kovaka., et al (2021). *Nature biotechnology*, *14(1)*.
Firtina., et al (2023). *arXiv*:2309.05771
Zhang., et al (2021). *Bioinformatics*, *37*.

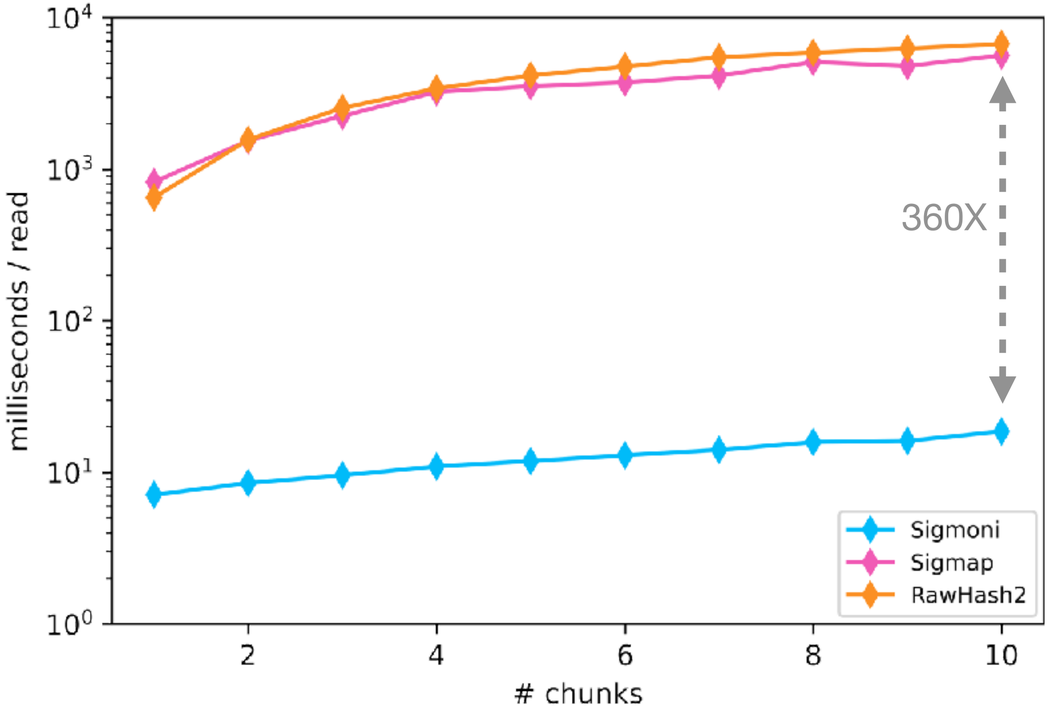
Adaptive sampling for microbial read detection





Adaptive sampling for host depletion

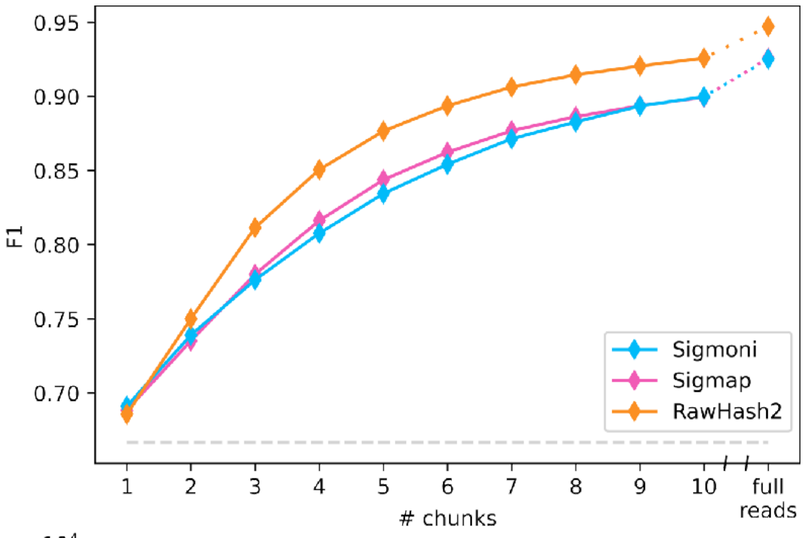


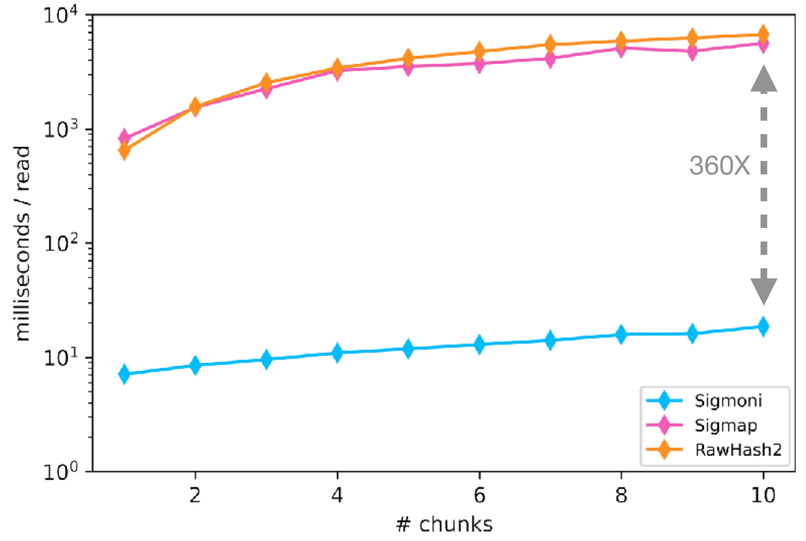


RawHash2
Sigmap

Sigmoni

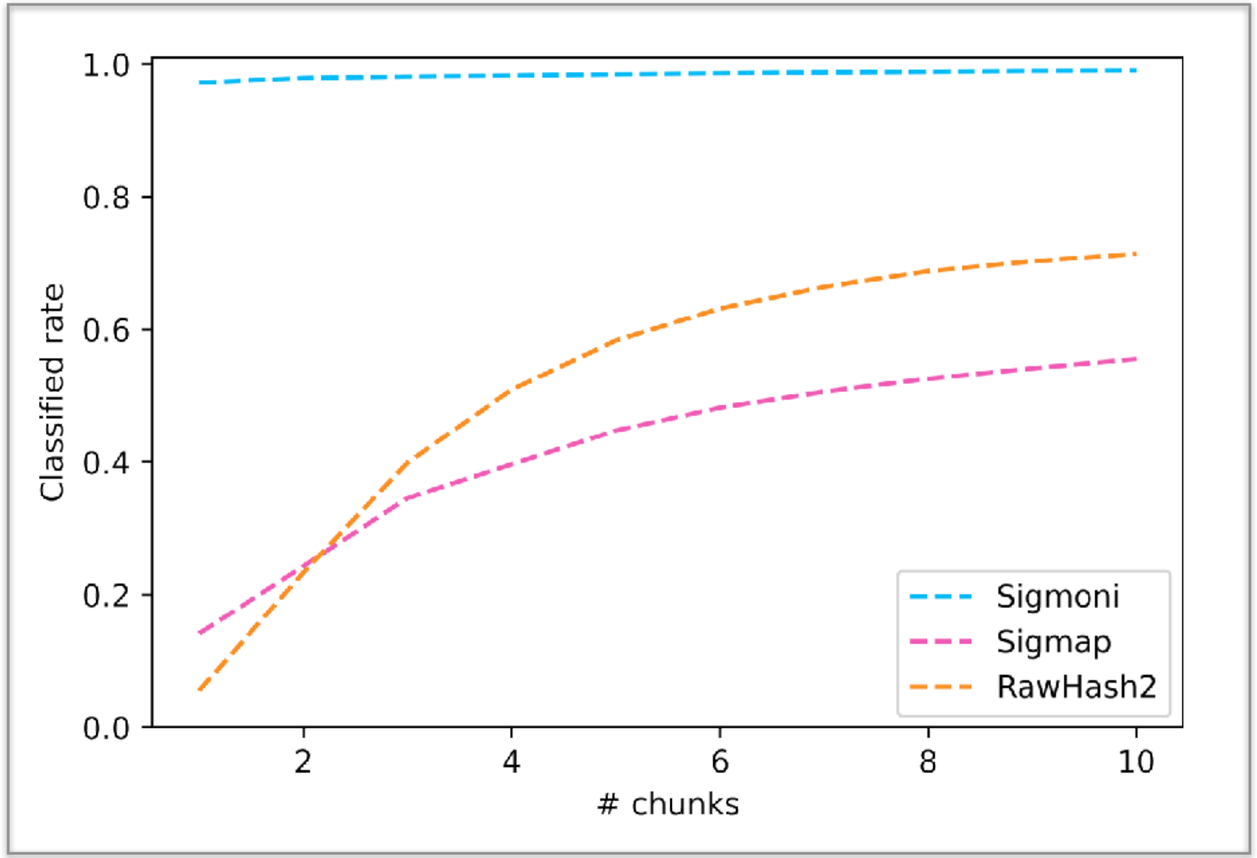
Adaptive sampling for host depletion





RawHash2
Sigmap

Sigmoni



Sigmoni classifies significantly more reads against complex references

r-index enables space-efficient index construction

- Can index over 7000 bacterial genomes in 2.2GB, without affecting query time
- Only signal-based method to index the HPRC assembly collection

	Mock community	Bacterial Pangenome	GRCh38	Human Ref Assemblies	HPRC Pangenome
Reference size	41.9 Mbp	42.2 Gbp	3.09 Gbp	18.29 Gbp	253 Gbp
Sigmoni	178 MB	**2.2 GB**	7.6 GB	**9.6 GB**	**25 GB**
Sigmap	2,591 MB	*	172 GB	786 GB	*
UNCALLED	**72 MB**	*	*	*	*
RawHash2	567 MB	190 GB	40 GB	79 GB	3,563 GB
Minimap2	136 MB	80 GB	**7.2 GB**	41 GB	*

* could not index

Conclusions

- Signal-based classification is difficult due to noise → Sigmoni applies novel signal binning and exact matching to overcome this
- Exact matching can be integrated with approximate location information to map nanopore read signal
- Compressed indexing and rapid read classification enable accurate adaptive sampling against pangenomes, potentially improving reference bias in filtering tasks

Acknowledgements

<u>Langmead Lab</u>
PI: **Dr. Ben Langmead**
Dr. Mohsen Zakeri
Omar Ahmed
Dr. Sina Majidian
Rone Charles
Jessica Bonnie
Nate Brown
Stephen Hwang
Mao-Jan Lin
Naga Sai Kavya Vaddadi

Collaborators:
Dr. Sam Kovaka
Dr. Christina Boucher
Dr. Travis Gagie

Work supported by NIH grants R01HG011392, U01CA25348, NSF BIO grant DBI-2029552, NSF DGE2139757, and Human Frontier Science Program RGP0025

Questions?