May 27-29, 2026

Rectorate of the University of Roma Tre, Rome, Italy

STORAGE AND COMPUTING WITH DNA

BBQ: Improving Consensus Finding in DNA Data Storage with Base-Level Quality Scores

Puru Sharma

National University of Singapore

Sponsored by

DNA DATA STORAGE ALLIANCE
A SNIA Community

European Innovation Council

DigNA

ATLAS

BIOMEMORY
Sustainable Storage Solutions

WD

Organized by

day one

Introduction

Introduction – Problem overview and motivation

The amount of data is growing exponentially, and we are running out of silicon to store it.

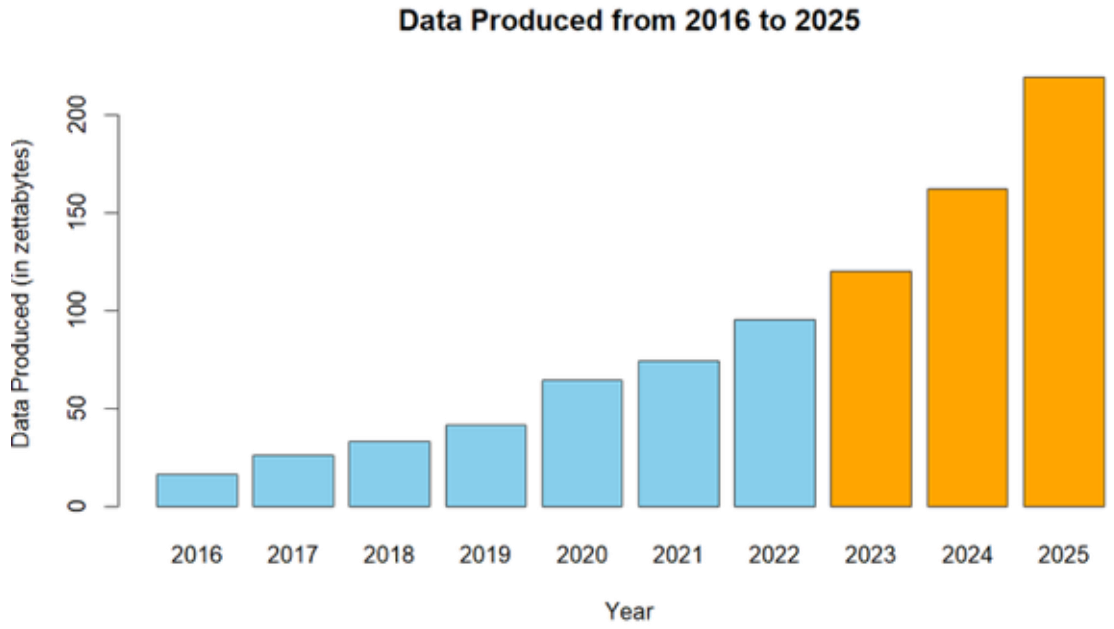


Figure. The amount of data produced each year [Hassini et al., 2023].
Zettabyte = 10^{21} bytes = 1,000,000,000,000,000,000,000 bytes

Hassani, Hossein, and Steve MacFeely. "Driving excellence in official statistics: unleashing the potential of comprehensive digital data governance." *Big Data and Cognitive Computing* 7.3 (2023): 134.

Introduction – Why DNA data storage?

1. Incredible density
   ~ 6 billion bits of info in human cells.
2. Unmatched durability
3. Little Energy Consumption
4. Constantly improving reading/writing interface.

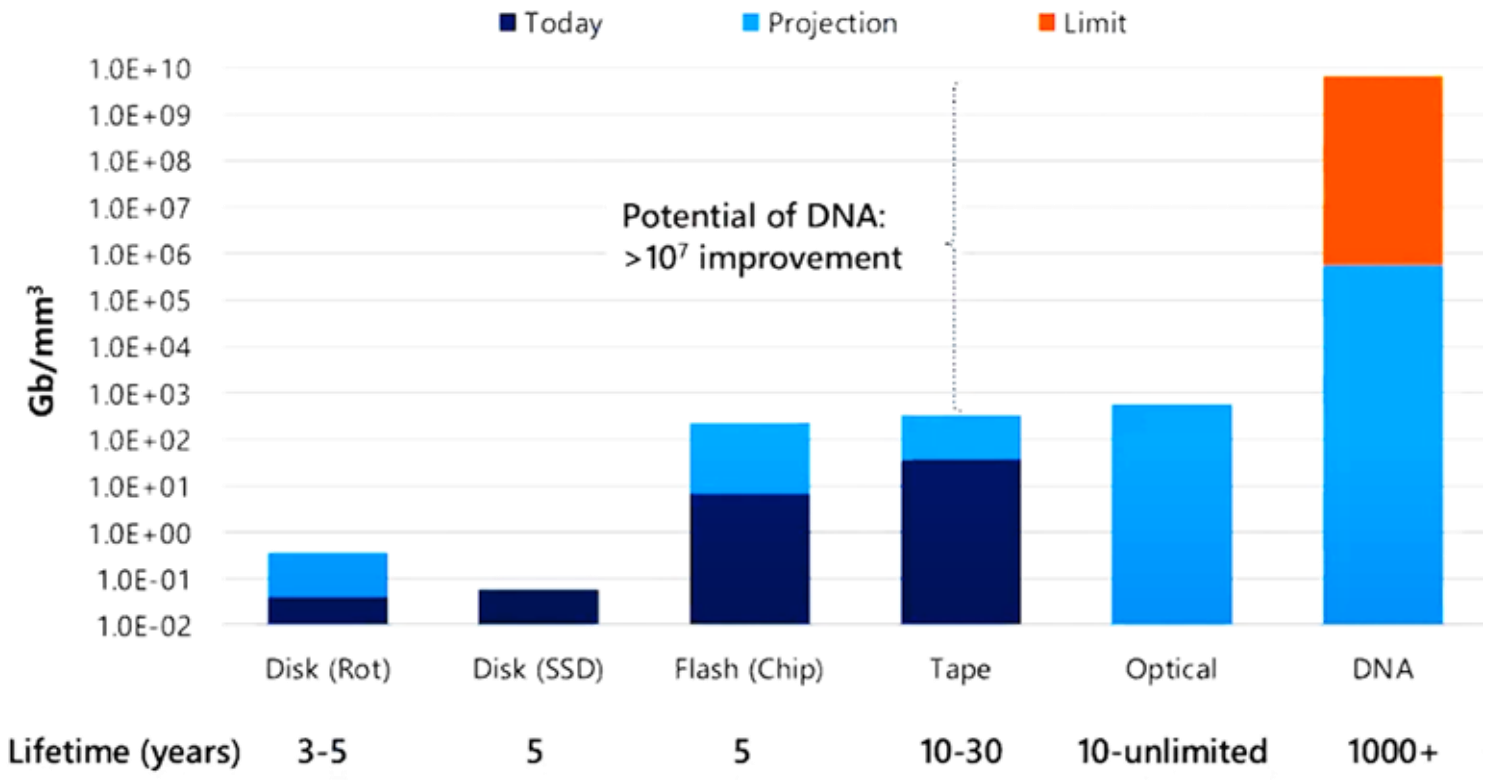


Figure. Storage density and lifetime.

Credit: Keynote by Luis Ceze and Karin Strauss, FAST '21.

Luis Ceze and Karin Strauss. "DNA Data Storage and Near-Molecule Processing for the Yottabyte Era." *Keynote talk on FAST '21.*

Introduction – DNA Storage Pipeline: Writing

Digital data

010010001101
010010101000

Binary

00→A, 01→C,
10→G, 11→T.

CAGATC
CAGGGA

DNA Code

Introduction – DNA Storage Pipeline: Writing

Digital data

010010001101
010010101000

Binary

00→A, 01→C,
10→G, 11→T.

CAGATC
CAGGGA

DNA Code

AACAGA
ACTCCA
AGGGGA

Encoded strands

Typically in length 100-200.

Synthesized DNA

Introduction – DNA Storage Pipeline: Reading

PCR and Sequencing

With Illumina or
Oxford Nanopore

AGTCCA AGGGA

AATAGA GGGGA ACTTCCA

ATCTAGC ACTCCCA

AACGA AACGA

Introduction – DNA Storage Pipeline: Reading

AGTCCA
ACTTCCA
ACTCCCA

AGGGA
GGGGA
AGGCGA

AACGA
AATAGA
AACGA

PCR and Sequencing

With Illumina or Oxford Nanopore

Clustered Reads

Reads that come from the same template are clustered together.

Introduction – DNA Storage Pipeline: Reading

PCR and Sequencing

With Illumina or Oxford Nanopore

AGTCCA
ACTTCCA
ACTCCCA

AGGGA
GGGGA
AGGCGA

AACGA
AATAGA
AACGA

ACTCCA

AGGGGA

AACAGA

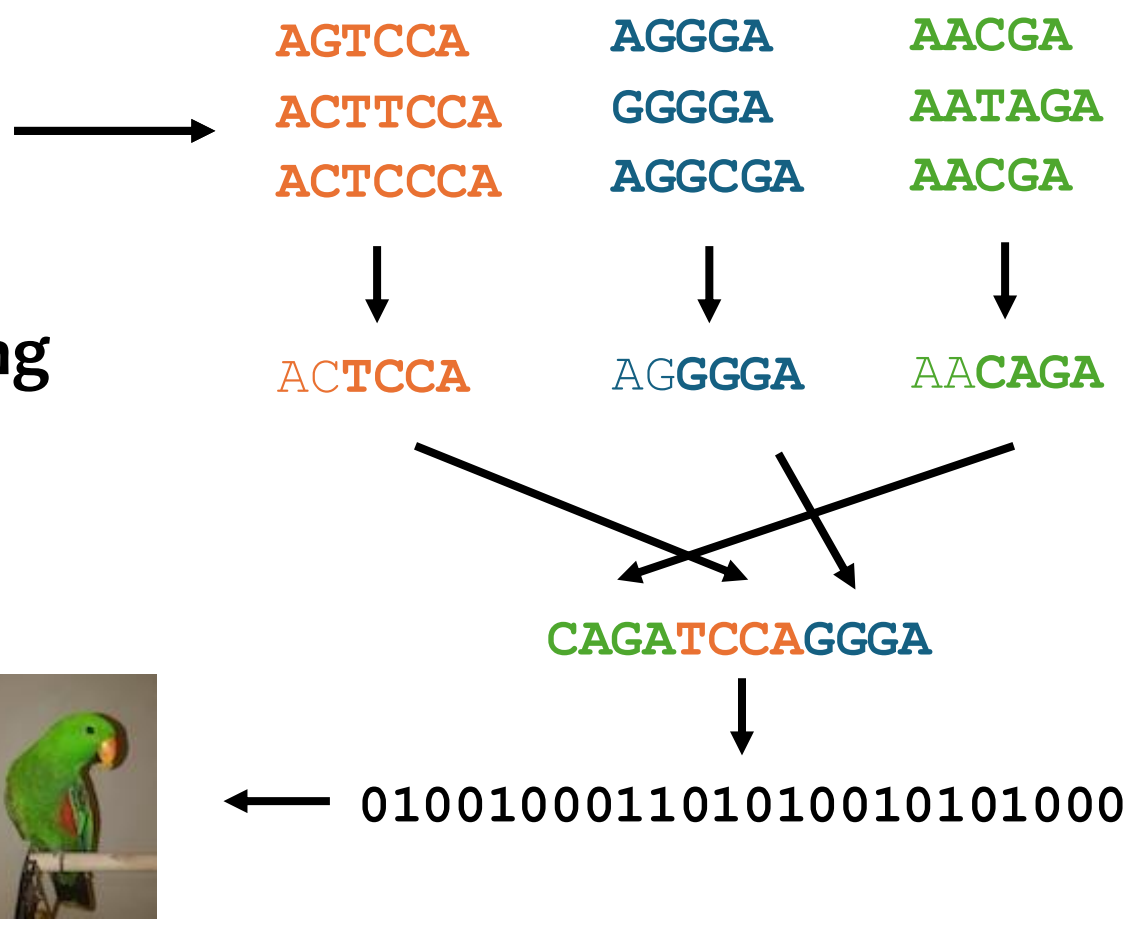
Trace Reconstruction

Consensus construction of each cluster.

Introduction – DNA Storage Pipeline: Reading

PCR and Sequencing

With Illumina or Oxford Nanopore



Decoding

Full decoding of the original data

Introduction – Trace Reconstruction Problem

Problem: Given N *traces*, the noisy copies of original strand x of length L, we aim to reconstruct x with high probability.

$x = \texttt{GTAGTGCCTG}$

$y_1 = \texttt{GTAGGTGCCG}$

$y_2 = \texttt{GTAGTCCTG}$

$y_3 = \texttt{GTAGTGCCTG}$

$y_4 = \texttt{GTAGCGCCAG}$

$y_5 = \texttt{GCATGCTCTG}$

```
GTAGGTGCC-G
GTA-GT-CCTG
GTA-GTGCCTG
GTA-GCGCCAG
GCATGCT-CTG
```

$\hat{x} = \texttt{GTA-GTGCCTG}$

Introduction – Challenges

1. **Cope with high error rate**
    - Errors can be introduced during synthesis, sequencing and clustering.
2. **Deal with low coverage.**
    - Some clusters may be noisy and have very few traces.
3. **Utilize prior knowledge.**
    - The target sequence length L is known.
    - The sequencing technology provides us a quality score for each base called.

Related Work

Method	Representative Tools
Alignment	• MUSCLE [Edgar, 2004], • CPL [Bar-Lev et al., 2025].
Read error correction	• BMA [Batu et al., 2004], • Trellis BMA [Srinivasavaradhan et al., 2021].
Assembly	• DBGPS [Song et al., 2022], • ITR [Sabary et al., 2024].
Deep learning	• DNA-GAN [Zheng et al., 2024], • DNAFormer [Bar-Lev et al., 2025].

Related Work

Potential problems:

- Longer running time.
- Multiple ways of defining the alignment scoring scheme.

Assuming $p_{\text{insertion}} = p_{\text{deletion}} = p_{\text{substitution}}$?

AGTCCACT

AGAACTT

AGTCCCATT

Assuming $p_{\text{substitution}} = 0.001$,
$p_{\text{insertion}} = p_{\text{deletion}} = 0.01$?

Related Work

Potential problems:

- Longer running time.
- Multiple ways of defining the alignment scoring scheme.

AGTCCACT

AGAACTT

$$\begin{cases} \text{match} = 1 \\ \text{mismatch} = -1 \\ \text{indel} = -1 \end{cases}$$

AGTCCCATT

$$\begin{cases} \text{match} = 1 \\ \text{mismatch} = -1 \\ \text{indel} = -0.5 \end{cases}$$

Related Work

Potential problems:

- Hard to capture complications in real data.

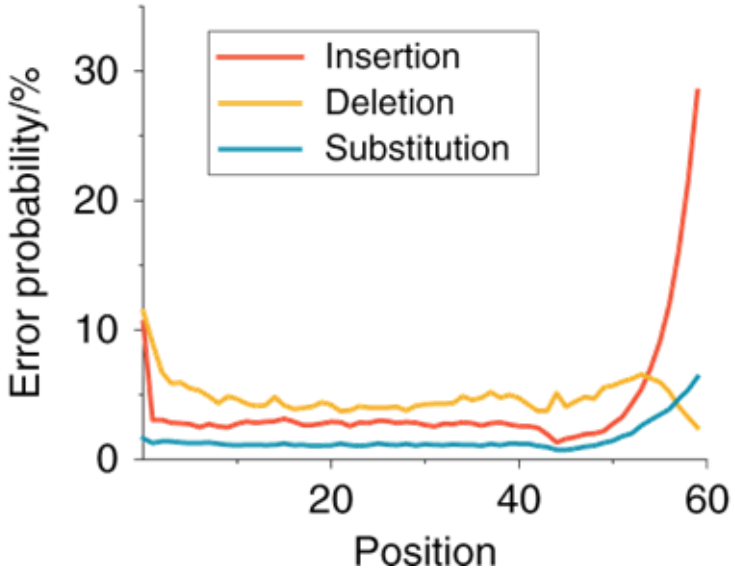


Figure. Error distribution within a read [Antkowiak et al., 2020].

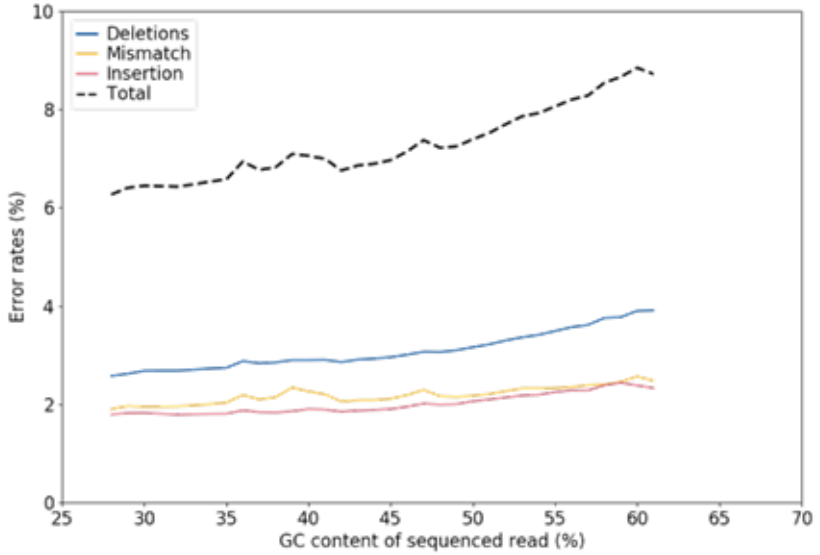


Figure. Higher GC-content may lead to higher error rates [Delahaye et al., 2021].

Antkowiak, Philipp L., et al. "Low cost DNA data storage using photolithographic synthesis and advanced information reconstruction and error correction." *Nature communications* 11.1 (2020): 5345.

Delahaye, Clara, and Jacques Nicolas. "Sequencing DNA with nanopores: Troubles and biases." *PloS one* 16.10 (2021): e0257521.

Related Work

Potential problems:

- Hard to capture complications in real data.

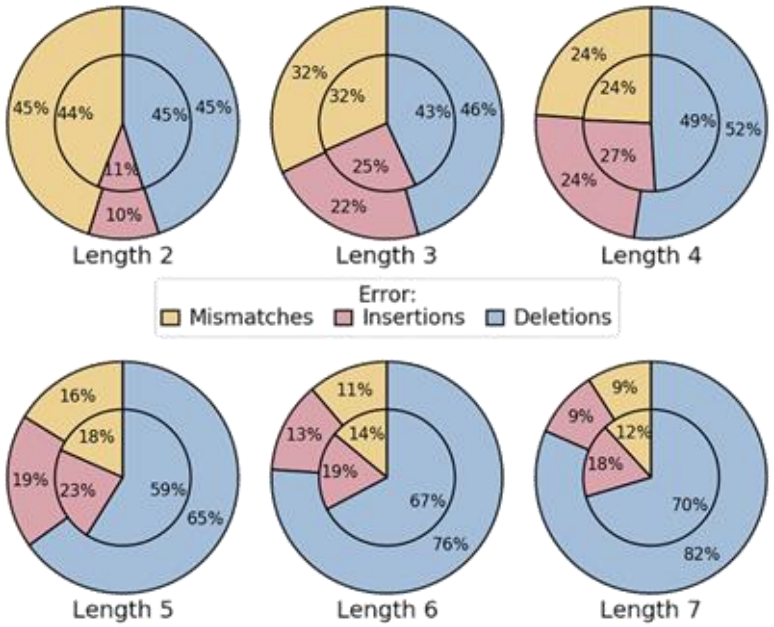


Figure. Error distribution in homopolymer regions [Delahaye et al., 2021].

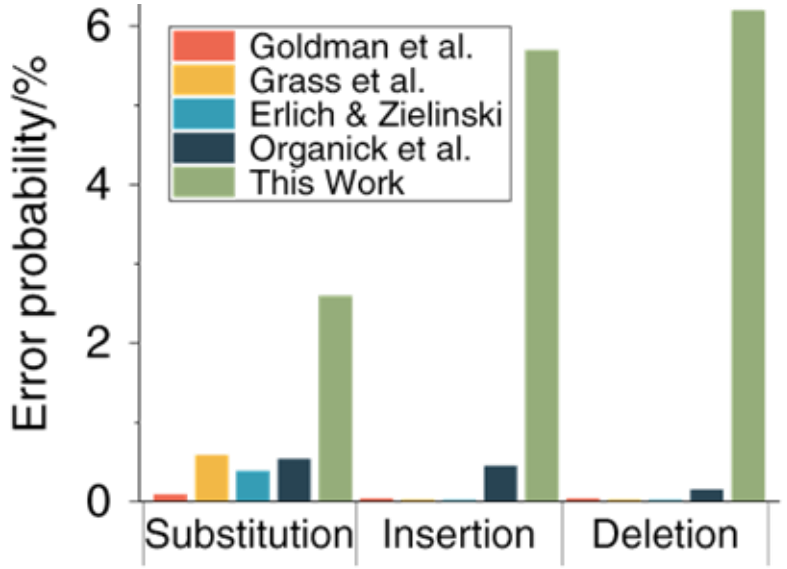


Figure. Error composition in various datasets [Antkowiak et al., 2020].

Related Work

Potential problems:

- Majority aren't always correct, especially given the prior knowledge of the length of x.

AGTCCACT
AGAACTT
AGTCCCATT

$$\begin{cases} \text{match} = 1 \\ \text{mismatch} = -1 \\ \text{indel} = -1 \end{cases}$$

```
AGTCC-ACT
AGAAC--TT
AGTCCCATT
--------
AGTCC-ATT
```

$$\begin{cases} \text{match} = 1 \\ \text{mismatch} = -1 \\ \text{indel} = -0.5 \end{cases}$$

```
AGTCC-ACT-
AGA---ACTT
AGTCCCA-TT
----------
AGTCC-ACTT
```

Related Work

Potential problems:

- Majority aren't always correct, especially given the prior knowledge of the length of x.

AGTCCACT
AGAACTT
AGTCCCATT

$$\begin{cases} \text{match} = 1 \\ \text{mismatch} = -1 \\ \text{indel} = -1 \end{cases}$$

AGTCC-ACT
AGAAC--TT
AGTCCCATT

AGTCC-ATT

$$\begin{cases} \text{match} = 1 \\ \text{mismatch} = -1 \\ \text{indel} = -0.5 \end{cases}$$

AGTCC-ACT-
AGA---ACTT
AGTCCCA-TT

AGTCC-ACTT

Given $L = 10$,

AGTCC-ACT-
AGA---ACTT
AGTCCCA-TT

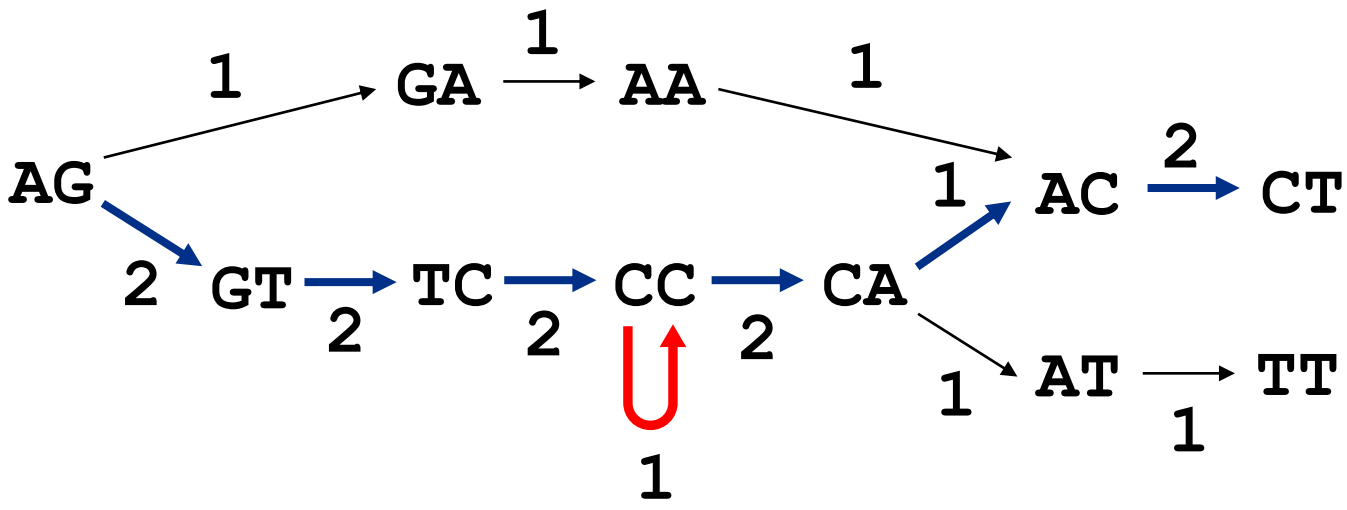
AGTCCCACTT

Related Work

Potential problems:

- Majority aren't always correct, especially given the prior knowledge of the length of x.

AGTCCACT
AGAACTT
AGTCCCATT



Related Work

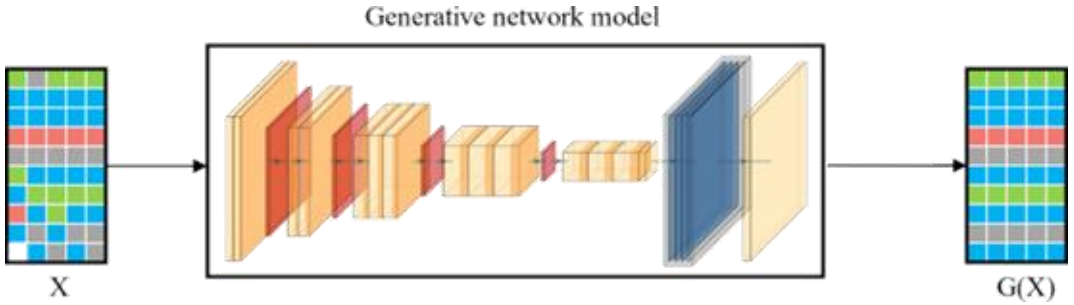


Figure. DNA-GAN [Zheng et al., 2024].

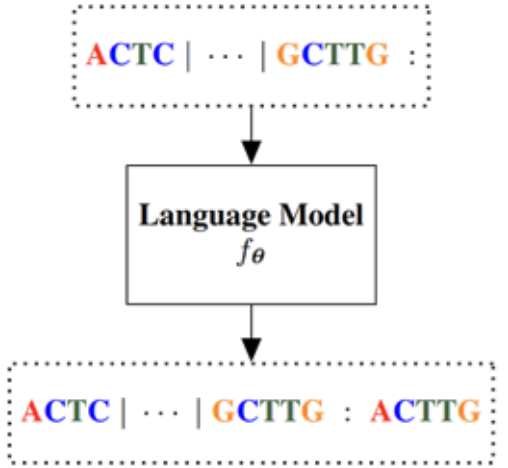


Figure. TReconLM. [Girsch et al., 2025].

- **Problems:** Need for training;
  may not work equally well on unseen error distributions.

Zheng, Xiaodong, et al. "A generative adversarial network for multiple reads reconstruction in DNA storage." *Scientific Reports* 14.1 (2024): 32071.
Girsch, Michael, and Reinhard Heckel. "Trace Reconstruction for DNA Data Storage using Language Models."

Method

Method

GTAGGTGCC-G

GTA-GT-CCTG

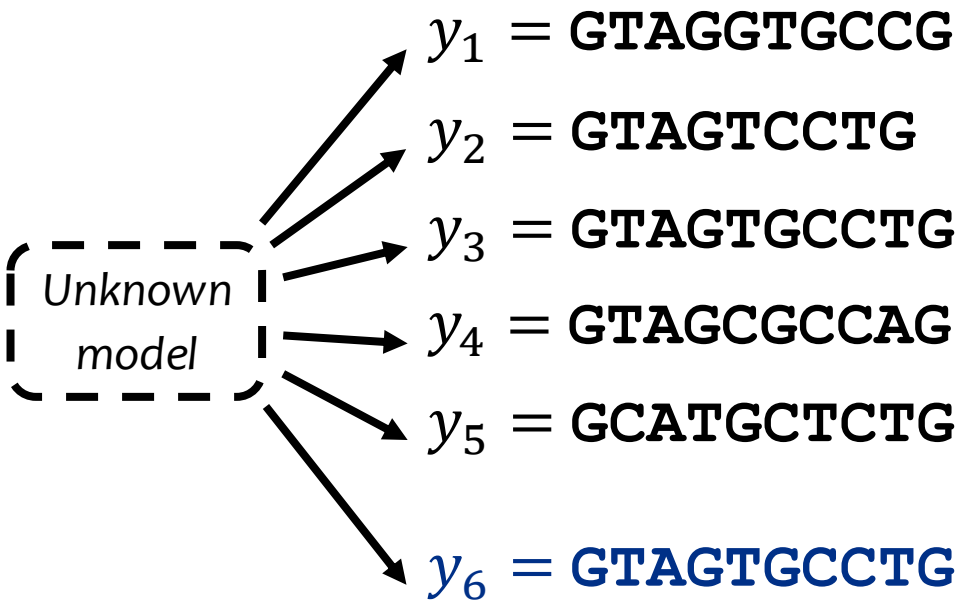
GTA-GTGCCTG

GTA-GCGCCAG

GCATGCT-CTG

$\hat{x}$ = GTA-GTGCCTG

What is the best alignment?



What is the most probable next trace?

- **Idea:** Model the traces as observations of a *k*-th order Markov chain.

Method

Original Traces C

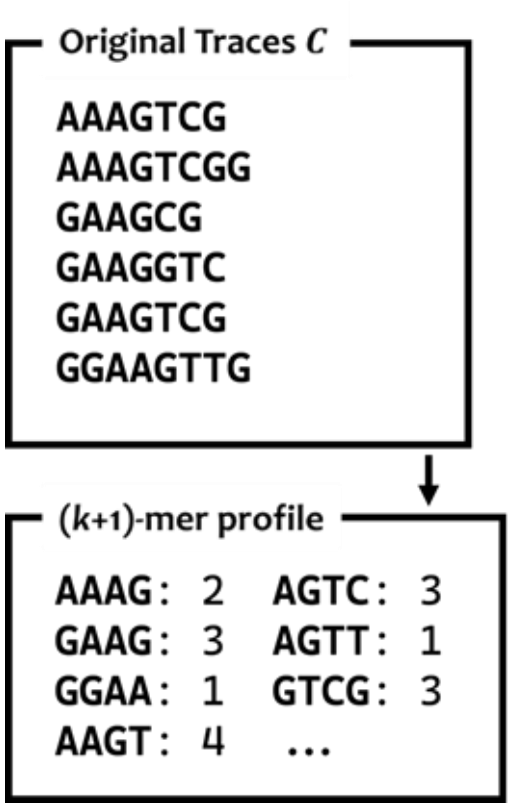
AAAGTCG
AAAGTCGG
GAAGCG
GAAGGTC
GAAGTCG
GGAAGTTG

- k is chosen such that all the k-mers in the reads are unique.
- Estimate the parameters of the Markov chain using maximum likelihood.

$$\Pr[s_{1:4} = \text{AAAG}] = \frac{2}{6}$$

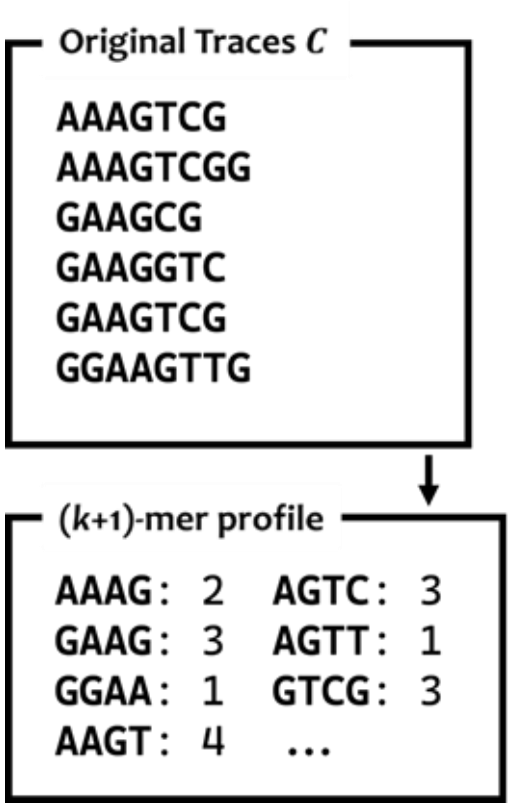
$$\Pr[s_{1:4} = \text{GAAG}] = \frac{3}{6}$$

Method



$$\Pr[\, s_i = \mathrm{C} \mid s_{i-k:i-1} = \mathrm{AGT} \,] = \frac{n(\mathrm{AGTC}) + \mathbf{1}}{n(\mathrm{AGTA}) + n(\mathrm{AGTC}) + n(\mathrm{AGTG}) + n(\mathrm{AGTT}) + \mathbf{4}} = \frac{4}{8}$$

Method



$$\Pr[\, s_i = \mathrm{C} \mid s_{i-k:i-1} = \mathrm{AGT} \,] = \frac{4}{8}$$

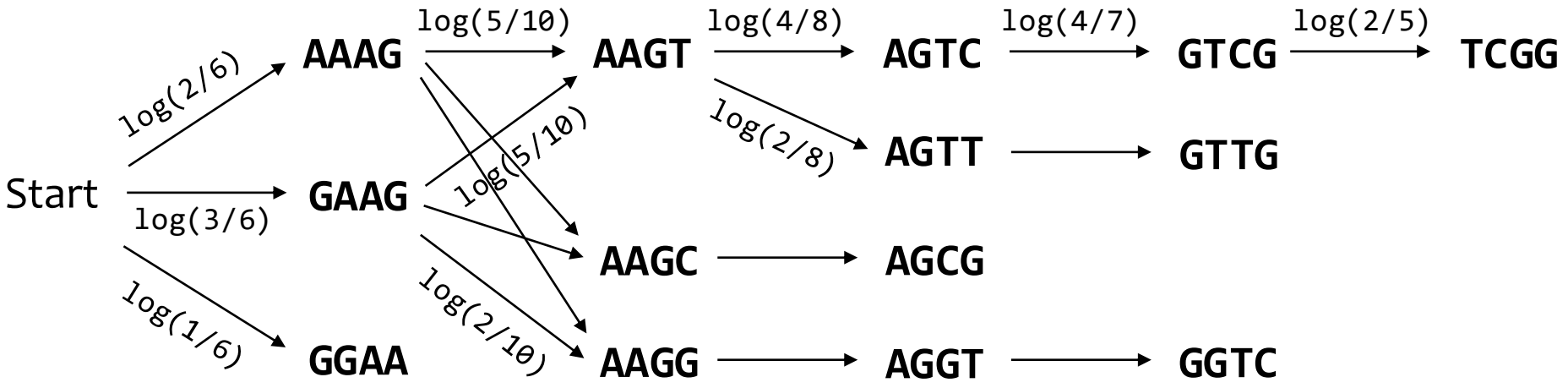
$$\Pr[\, s_i = \mathrm{T} \mid s_{i-k:i-1} = \mathrm{AGT} \,] = \frac{2}{8}$$

$\frac{2}{8}$ chance of substitution to T

- **Intuitively, the parameters of the Markov chain captures position-specific error profiles.**

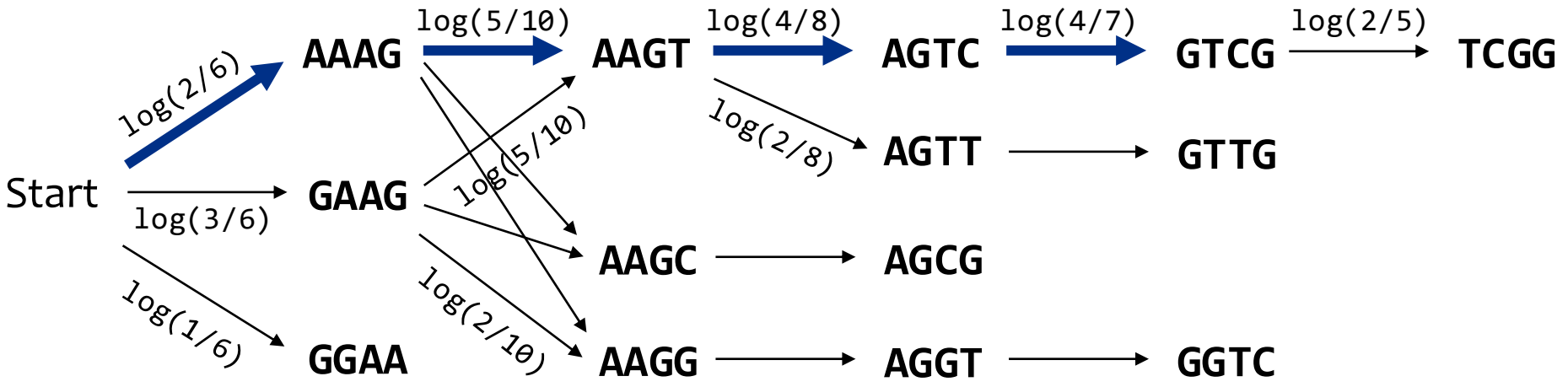
Method

Substitute weights in the de Bruijn graph with the learned parameters of the Markov chain.



Method

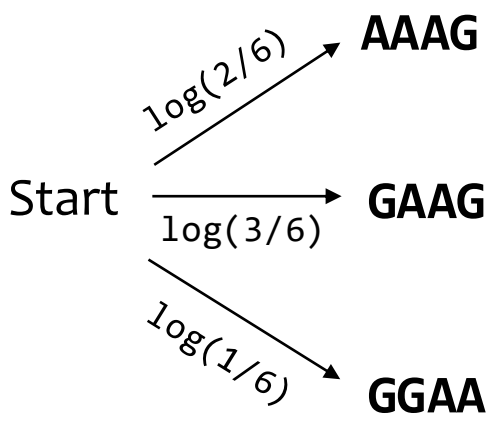
Goal: find the highest-weight path of length $L - k$ in the modified de Bruijn graph.



Sum of edge weight $= \log \Pr[S = \text{AAAGTCG}]$.

Method

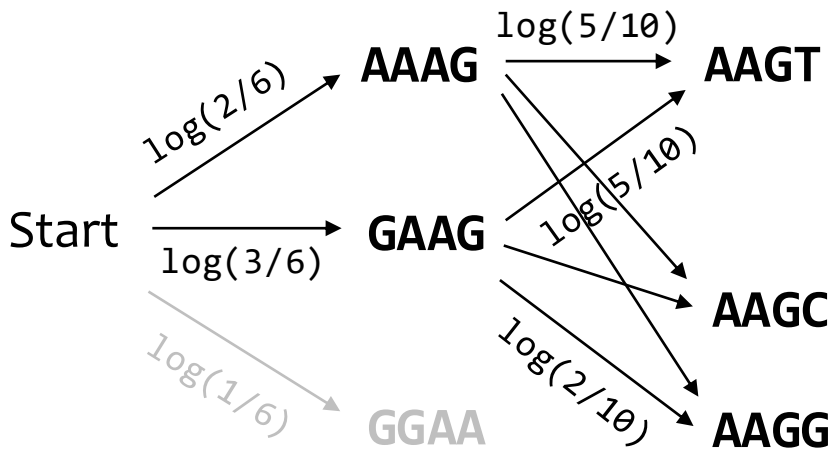
Goal: find the highest-weight path of length $L - k$ in the modified de Bruijn graph.



Beam search: keep only the top B paths (here $B = 2$).

Method

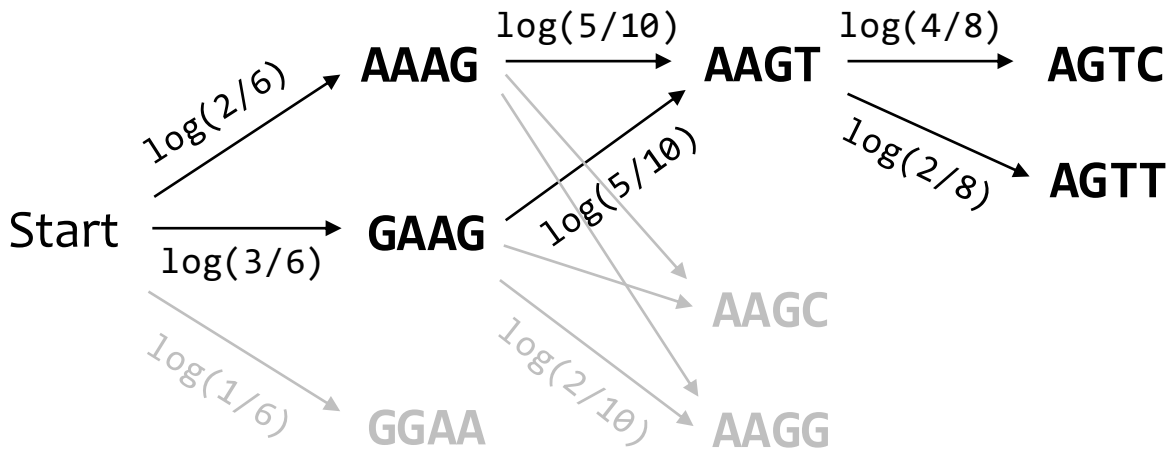
Goal: find the highest-weight path of length $L - k$ in the modified de Bruijn graph.



Beam search: keep only the top B paths (here $B = 2$).

Method

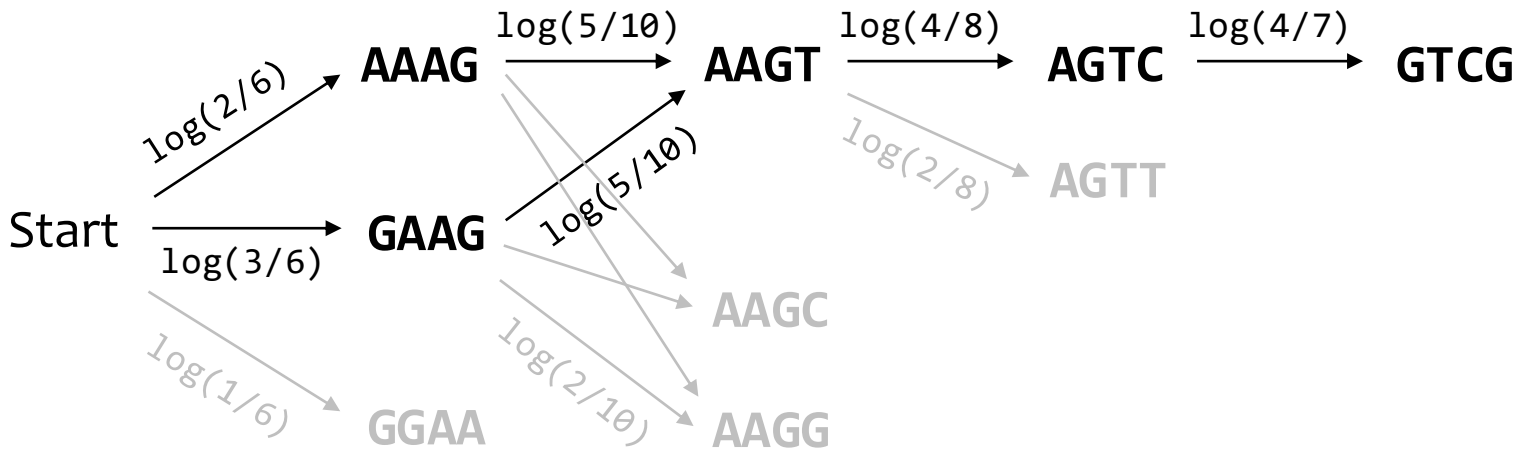
Goal: find the highest-weight path of length $L - k$ in the modified de Bruijn graph.



Beam search: keep only the top B paths (here $B = 2$).

Method

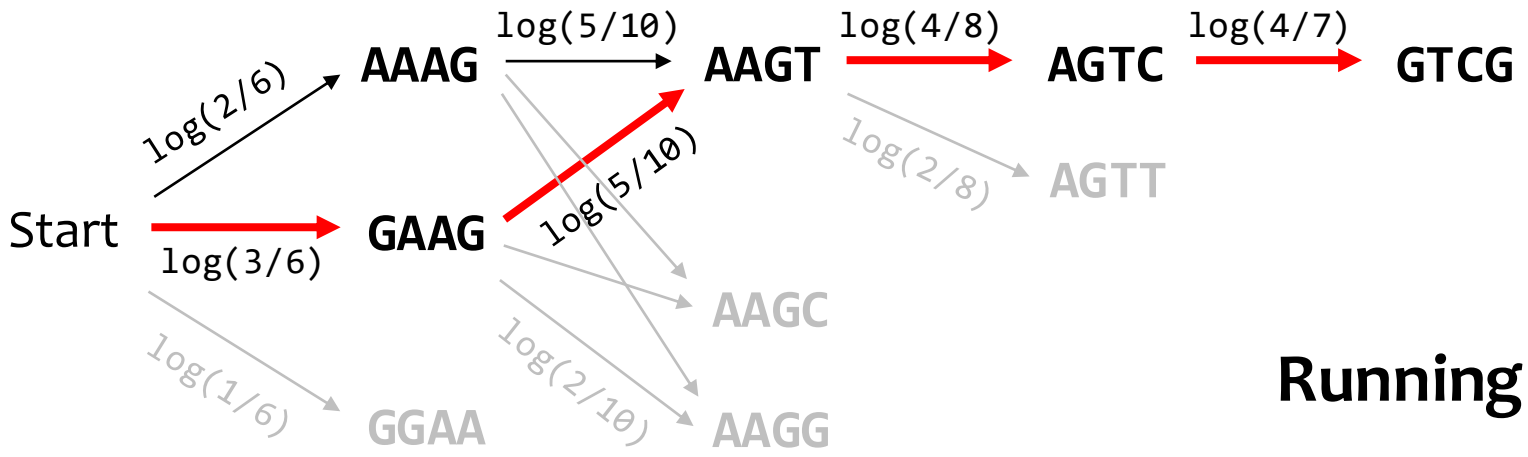
Goal: find the highest-weight path of length $L - k$ in the modified de Bruijn graph.



Beam search: keep only the top B paths (here $B = 2$).

Method

Goal: find the highest-weight path of length $L - k$ in the modified de Bruijn graph.

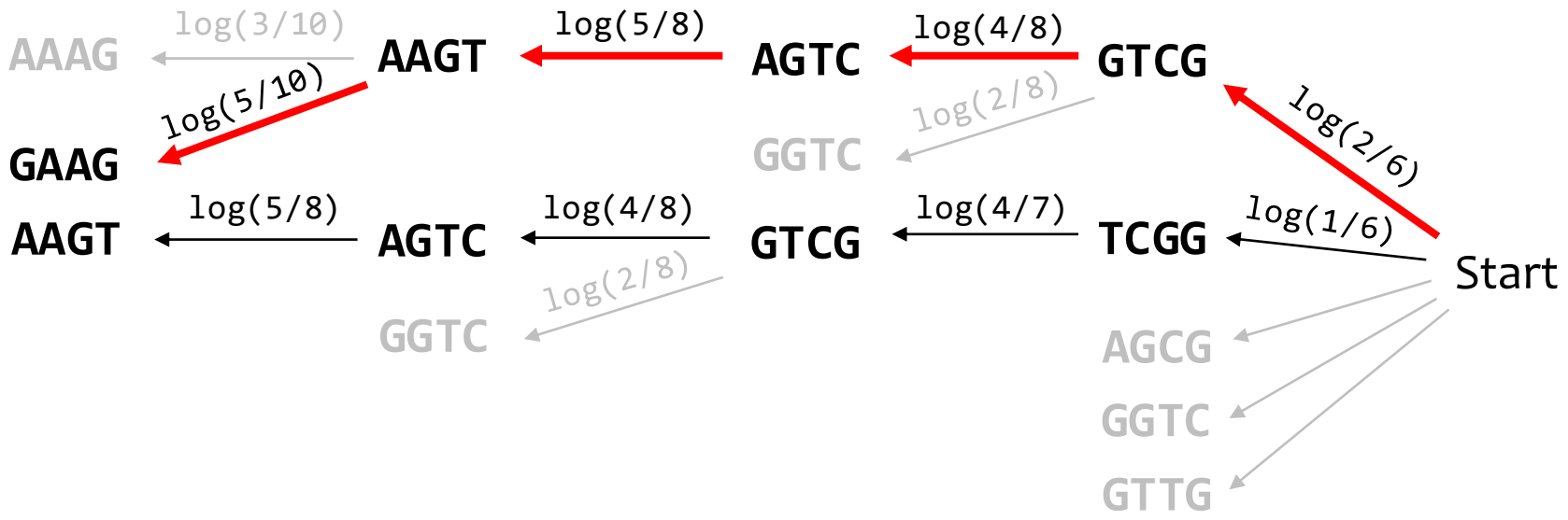


Running time: $O(BNL)$

Beam search: keep only the top B paths, $\boldsymbol{L - k}$ **iterations.**

Method

We can also do the same process from the end of the sequences.



Compare the two returned optimal paths by their weights.

Method

NUS National University of Singapore | Computing

```
@002cd902-0726-4fb8-bf1d-0a056951d76f
GTGAGACTGACGACACAGCAGTGTAAAGCAGTCCAAATGAAGCTTGGATGATTATTGACGTCTCCGTCCAATTCGTATCAGGGACACCCGTTTGATGTGATGGACGAGAATGGATTCTCCTAGACCACAGACCCTGTGCGTGTGTGCGTGACTAGCTAGTACAT
+
+-4=>)JEIIBIOMOQS<RID@C)+((&1?NGNNUUU+*&&'253GMP)PKPK.=+9&,0320,,M8,(UUUUNI3>A*LRQLR)SOP)KU,-QP*K5,/.LO/JGCMTPP@MQ,+08*%+0GPOKS)UTTRP4PPRTOL2E1L)PKL+QG)+KP-*.5+;5.'
@0039c1dc-ab1b-42fa-a601-59ac8ad4b60f
CTAGACGTGCGAGTATACTACTATGCCGAGTTATATTTAAAACTGCATTTAGATTCCGTCGCTAGGCCTACTGATGCCCCAGCTGCCGTTTGGAATTAAGGAAGTGATACCTAGATGTGTGTTCACAGCGACCTTAAATGACACGCAGTCAGACGCGTGT
+
**0,@)?EC=%+ALD4.-*HEJHB,,*).>M6POJRK+,,JOGMGGCJ6,:+-K0B0F+.@-OKMME('@+CF=)2'0/&FC+)QMKORT/PN3N+E7..'564)RUQOJMMNIM.0/*)IRU*OMRNJJNM+UQRQTA*88F-OULUT)MJO-56>5)%
@0041ac21-7266-4a7e-b9c8-1a1a92b729e2
TTAATGTAGCTATCTCTGCTAGCTAAATCGATCAGAAGGTTTTCAAATCTTATAGTGATCTATAGCGCTTGATGGCGGAACGTGTCTTCGCTAAGTGCGGGCGTATGTGCAGCTTTGTGTCCCTGTATCGTTAACTAGGAAGCTGATGTACAGAGATAGACTACAG
+
&$&**6)(%%+0,-/4+%(64;1+>OQ++JORK''&)(3;-%-0&*,*#%'#$$%/'/1&&$*%--(*CB<MPENJKO+U50*%'&*49*(0')-+,/2HG6*1*;8+C93G)/.+)++))+(+%%,(&(%+('('#,213D10QO/+*:;//(*@OKNFB5'&$$
```

Method

READ IDENTIFIER

```
@002cd902-0726-4fb8-bf1d-0a056951d76f
GTGAGACTGACGACACAGCAGTGTAAAGCAGTCCAAATGAAGCTTGGATGATTATTGACGTCTCCGTCCAATTCGTATCAGGGACACCCGTTTGATGTGATGGACGAGAATGGATTCTCCTAGACCACAGACCCTGTGCGTGTGTGCGTGACTAGCTAGTACAT
+
+-4=>)JEIIBIOMOQS<RID@C)+((&1?NGNNUUU+*&&'253GMP)PKPK.=+9&,0320,,M8,(UUUUNI3>A*LRQLR)SOP)KU,-QP*K5,/.LO/JGCMTPP@MQ,+08*%+0GPOKS)UTTRP4PPRTOL2E1L)PKL+QG)+KP-*.5+;5.'
@0039c1dc-ab1b-42fa-a601-59ac8ad4b60f
CTAGACGTGCGAGTATACTACTATGCCGAGTTATATTTAAAACTGCATTTAGATTCCGTCGCTAGGCCTACTGATGCCCCAGCTGCCGTTTGGAATTAAGGAAGTGATACCTAGATGTGTGTTCACAGCGACCTTAAATGACACGCAGTCAGACGCGTGT
+
**0,@)?EC=%+ALD4.-*HEJHB,,*).>M6POJRK+,,JOGMGGCJ6,:+-K0B0F+.@-OKMME('@+CF=)2'0/&FC+)QMKORT/PN3N+E7..'564)RUQOJMMNIM.0/*)IRU*OMRNJJNM+UQRQTA*88F-OULUT)MJO-56>5)%
@0041ac21-7266-4a7e-b9c8-1a1a92b729e2
TTAATGTAGCTATCTCTGCTAGCTAAATCGATCAGAAGGTTTTCAAATCTTATAGTGATCTATAGCGCTTGATGGCGGAACGTGTCTTCGCTAAGTGCGGGCGTATGTGCAGCTTTGTGTCCCTGTATCGTTAACTAGGAAGCTGATGTACAGAGATAGACTACAG
+
&$&**6)(%%+0,-/4+%(64;1+>OQ++JORK''&)(3;-%-0&*,*#%'#$$%/'/1&&$*%--(*CB<MPENJKO+U50*%'&*49*(0')-+,/2HG6*1*;8+C93G)/.+)++))+(+%%,(&(%+('('#,213D10QO/+*:;//(*@OKNFB5'&$$
```

Method

BASECALLED READ

```
@002cd902-0726-4fb8-bf1d-0a056951d76f
GTGAGACTGACGACACAGCAGTGTAAAGCAGTCCAAATGAAGCTTGGATGATTATTGACGTCTCCGTCCAATTCGTATCAGGGACACCCGTTTGATGTGATGGACGAGAATGGATTCTCCTAGACCACAGACCCTGTGCGTGTGTGCGTGACTAGCTAGTACAT
+
+-4=>)JEIIBIOMOQS<RID@C)+((&1?NGNNUUU+*&&'253GMP)PKPK.=+9&,0320,,M8,(UUUUNI3>A*LRQLR)SOP)KU,-QP*K5,/.LO/JGCMTPP@MQ,+08*%+0GPOKS)UTTRP4PPRTOL2E1L)PKL+QG)+KP-*.5+;5.'
@0039c1dc-ab1b-42fa-a601-59ac8ad4b60f
CTAGACGTGCGAGTATACTACTATGCCGAGTTATATTTAAAACTGCATTTAGATTCCGTCGCTAGGCCTACTGATGCCCCAGCTGCCGTTTGGAATTAAGGAAGTGATACCTAGATGTGTGTTCACAGCGACCTTAAATGACACGCAGTCAGACGCGTGT
+
**0,@)?EC=%+ALD4.-*HEJHB,,*).>M6POJRK+,,JOGMGGCJ6,:+-K0B0F+.@-OKMME('@+CF=)2'0/&FC+)QMKORT/PN3N+E7..'564)RUQOJMMNIM.0/*)IRU*OMRNJJNM+UQRQTA*88F-OULUT)MJO-56>5)%
@0041ac21-7266-4a7e-b9c8-1a1a92b729e2
TTAATGTAGCTATCTCTGCTAGCTAAATCGATCAGAAGGTTTTCAAATCTTATAGTGATCTATAGCGCTTGATGGCGGAACGTGTCTTCGCTAAGTGCGGGCGTATGTGCAGCTTTGTGTCCCTGTATCGTTAACTAGGAAGCTGATGTACAGAGATAGACTACAG
+
&$&**6)(%%+0,-/4+%(64;1+>OQ++JORK''&)(3;-%-0&*,*#%'#$$%/'/1&&$*%--(*CB<MPENJKO+U50*%'&*49*(0')-+,/2HG6*1*;8+C93G)/.+)++))+(+%%,(&(%+('('#,213D10QO/+*:;//(*@OKNFB5'&$$
```

Method

NUS National University of Singapore | Computing

```
@002cd902-0726-4fb8-bf1d-0a056951d76f
GTGAGACTGACGACACAGCAGTGTAAAGCAGTCCAAATGAAGCTTGGATGATTATTGACGTCTCCGTCCAATTCGTATCAGGGACACCCGTTTGATGTGATGGACGAGAATGGATTCTCCTAGACCACAGACCCTGTGCGTGTGTGCGTGACTAGCTAGTACAT
+
+-4=>)JEIIBIOMOQS<RID@C)+((&1?NGNNUUU+*&&'253GMP)PKPK.=+9&,0320,,M8,(UUUUNI3>A*LRQLR)SOP)KU,-QP*K5,/.LO/JGCMTPP@MQ,+08*%+0GPOKS)UTTRP4PPRTOL2E1L)PKL+QG)+KP-*.5+;5.'
@0039c1dc-ab1b-42fa-a601-59ac8ad4b60f
CTAGACGTGCGAGTATACTACTATGCCGAGTTATATTTAAAACTGCATTTAGATTCCGTCGCTAGGCCTACTGATGCCCCAGCTGCCGTTTGGAATTAAGGAAGTGATACCTAGATGTGTGTTCACAGCGACCTTAAATGACACGCAGTCAGACGCGTGT
+
**0,@)?EC=%+ALD4.-*HEJHB,,*).>M6POJRK+,,JOGMGGCJ6,:+-K0B0F+.@-OKMME('@+CF=)2'0/&FC+)QMKORT/PN3N+E7..'564)RUQOJMMNIM.0/*)IRU*OMRNJJNM+UQRQTA*88F-OULUT)MJO-56>5)%
@0041ac21-7266-4a7e-b9c8-1a1a92b729e2
TTAATGTAGCTATCTCTGCTAGCTAAATCGATCAGAAGGTTTTCAAATCTTATAGTGATCTATAGCGCTTGATGGCGGAACGTGTCTTCGCTAAGTGCGGGCGTATGTGCAGCTTTGTGTCCCTGTATCGTTAACTAGGAAGCTGATGTACAGAGATAGACTACAG
+
&$&**6)(%%+0,-/4+%(64;1+>OQ++JORK''&)(3;-%-0&*,*#%'#$$%/'/1&&$*%--(*CB<MPENJKO+U50*%'&*49*(0')-+,/2HG6*1*;8+C93G)/.+)++))+(+%%,(&(%+('('#,213D10QO/+*:;//(*@OKNFB5'&$$
```

SEPARATOR

Method

QUALITY STRING

```
@002cd902-0726-4fb8-bf1d-0a056951d76f
GTGAGACTGACGACACAGCAGTGTAAAGCAGTCCAAATGAAGCTTGGATGATTATTGACGTCTCCGTCCAATTCGTATCAGGGACACCCGTTTGATGTGATGGACGAGAATGGATTCTCCTAGACCACAGACCCTGTGCGTGTGTGCGTGACTAGCTAGTACAT
+
+-4=>)JEIIBIOMOQS<RID@C)+((&1?NGNNUUU+*&&'253GMP)PKPK.=+9&,0320,,M8,(UUUUNI3>A*LRQLR)SOP)KU,-QP*K5,/.LO/JGCMTPP@MQ,+08*%+0GPOKS)UTTRP4PPRTOL2E1L)PKL+QG)+KP-*.5+;5.'
@0059c1dc-ab1b-42fa-a661-59ac8a04b66f
CTAGACGTGCGAGTATACTACTATGCCGAGTTATATTTAAAACTGCATTTAGATTCCGTCGCTAGGCCTACTGATGCCCCAGCTGCCGTTTGGAATTAAGGAAGTGATACCTAGATGTGTGTTCACAGCGACCTTAAATGACACGCAGTCAGACGCGTGT
+
**0,@)?EC=%+ALD4.-*HEJHB,,*).>M6POJRK+,,JOGMGGCJ6,:+-K0B0F+.@-OKMME('@+CF=)2'0/&FC+)QMKORT/PN3N+E7..'564)RUQOJMMNIM.0/*)IRU*OMRNJJNM+UQRQTA*88F-OULUT)MJO-56>5)%
@0041ac21-7266-4a7e-b9c8-1a1a92b729e2
TTAATGTAGCTATCTCTGCTAGCTAAATCGATCAGAAGGTTTTCAAATCTTATAGTGATCTATAGCGCTTGATGGCGGAACGTGTCTTCGCTAAGTGCGGGCGTATGTGCAGCTTTGTGTCCCTGTATCGTTAACTAGGAAGCTGATGTACAGAGATAGACTACAG
+
&$&**6)(%%+0,-/4+%(64;1+>OQ++JORK''&)(3;-%-0&*,*#%'#$$%/'/1&&$*%--(*CB<MPENJKO+U50*%'&*49*(0')-+,/2HG6*1*;8+C93G)/.+)++))+(+%%,(&(%+('('#,213D10QO/+*:;//(*@OKNFB5'&$$
```

Method

```
@002cd902-0726-4fb8-bf1d-0a056951d76f
GTGAGACTGACGACACAGCAGTGTAAAGCAGTCCAAATGAAGCTTGGATGATTATTGACGTCTCCGTCCAATTCGTATCAGGGACACCCGTTTGATGTGATGGACGAGAATGGATTCTCCTAGACCACAGACCCTGTGCGTGTGTGCGTGACTAGCTAGTACAT
+
+-4#>)JEIIBIOMOQS<RID@C)+((&1?NGNNUUU+*&&'253GMP)PKPK.=+9&,0320,,M8,(UUUUNI3>A*LRQLR)SOP)KU,-QP*K5,/.LO/JGCMTPP@MQ,+08*%+0GPOKS)UTTRP4PPRTOL2E1L)PKL+QG)+KP-*.5+;5.'
@0059c1dc-ab1b-42fa-a601-59ac8ad4b60f
CTAGACGTGCGAGTATACTACTATGCCGAGTTATATTTAAAACTGCATTTAGATTCCGTCGCTAGGCCTACTGATGCCCCAGCTGCCGTTTGGAATTAAGGAAGTGATACCTAGATGTGTGTTCACAGCGACCTTAAATGACACGCAGTCAGACGCGTGT
+
**0,@)?EC=%+ALD4.-*HEJHB,,*).>M6POJRK+,,JOGMGGCJ6,:+-K0B0F+.@-OKMME('@+CF=)2'0/&FC+)QMKORT/PN3N+E7..'564)RUQOJMMNIM.0/*)IRU*OMRNJJNM+UQRQTA*88F-OULUT)MJO-56>5)%
@0041ac21-7266-4a7e-b9c8-1a1a92b729e2
TTAATGTAGCTATCTCTGCTAGCTAAATCGATCAGAAGGTTTTCAAATCTTATAGTGATCTATAGCGCTTGATGGCGGAACGTGTCTTCGCTAAGTGCGGGCGTATGTGCAGCTTTGTGTCCCTGTATCGTTAACTAGGAAGCTGATGTACAGAGATAGACTACAG
+
&$&**6)(%%+0,-/4+%(64;1+>OQ++JORK''&)(3;-%-0&*,*#%'#$$%/'/1&&$**%--(*CB<MPENJKO+U50**%'&*49*(0')-+,/2HG6*1*;8+C93G)/.+)++))+(+%%,(&(%+('('#,213D10QO/+*:;//(*@OKNFB5'&$$
```

Basecalled Base	ASCII Quality Score	Q-Score	P(error)	P(correct)
G	+			
T	-			
G	4			

Method

```
@002cd902-0726-4fb8-bf1d-0a056951d76f
GTGAGACTGACGACACAGCAGTGTAAAGCAGTCCAAATGAAGCTTGGATGATTATTGACGTCTCCGTCCAATTCGTATCAGGGACACCCGTTTGATGTGATGGACGAGAATGGATTCTCCTAGACCACAGACCCTGTGCGTGTGTGCGTGACTAGCTAGTACAT
+
+-4#>)JEIIBIOMOQS<RID@C)+((&1?NGNNUUU+*&&'253GMP)PKPK.=+9&,0320,,M8,(UUUUNI3>A*LRQLR)SOP)KU,-QP*K5,/.LO/JGCMTPP@MQ,+08*%+0GPOKS)UTTRP4PPRTOL2E1L)PKL+QG)+KP-*.5+;5.'
@0059c1dc-ab1b-42fa-a601-59ac8ad4b60f
CTAGACGTGCGAGTATACTACTATGCCGAGTTATATTTAAAACTGCATTTAGATTCCGTCGCTAGGCCTACTGATGCCCCAGCTGCCGTTTGGAATTAAGGAAGTGATACCTAGATGTGTGTTCACAGCGACCTTAAATGACACGCAGTCAGACGCGTGT
+
**0,@)?EC=%+ALD4.-*HEJHB,,*).>M6POJRK+,,JOGMGGCJ6,:+-K0B0F+.@-OKMME('@+CF=)2'0/&FC+)QMKORT/PN3N+E7..'564)RUQOJMMNIM.0/*)IRU*OMRNJJNM+UQRQTA*88F-OULUT)MJO-56>5)%
@0041ac21-7266-4a7e-b9c8-1a1a92b729e2
TTAATGTAGCTATCTCTGCTAGCTAAATCGATCAGAAGGTTTTCAAATCTTATAGTGATCTATAGCGCTTGATGGCGGAACGTGTCTTCGCTAAGTGCGGGCGTATGTGCAGCTTTGTGTCCCTGTATCGTTAACTAGGAAGCTGATGTACAGAGATAGACTACAG
+
&$&**6)(%%+0,-/4+%(64;1+>OQ++JORK''&)(3;-%-0&*,*#%'#$$%/'/1&&$*%--(*CB<MPENJKO+U50*%'&*49*(0')-+,/2HG6*1*;8+C93G)/.+)++))+(+%%,(&(%+('('#,213D10QO/+*:;//(*@OKNFB5'&$$
```

$$\text{Q-score} = \text{ASCII value} - 33$$

$$P(error) = 10^{-Q/10}$$

$$P(correct) = 1 - P(error)$$

Basecalled Base	ASCII Quality Score	Q-Score	P(error)	P(correct)
G	+			
T	-			
G	4			

Method

```
@002cd902-0726-4fb8-bf1d-0a056951d76f
GTGAGACTGACGACACAGCAGTGTAAAGCAGTCCAAATGAAGCTTGGATGATTATTGACGTCTCCGTCCAATTCGTATCAGGGACACCCGTTTGATGTGATGGACGAGAATGGATTCTCCTAGACCACAGACCCTGTGCGTGTGTGCGTGACTAGCTAGTACAT
+
+-4#>)JEIIBIOMOQS<RID@C)+((&1?NGNNUUU+*&&'253GMP)PKPK.=+9&,0320,,M8,(UUUUNI3>A*LRQLR)SOP)KU,-QP*K5,/.LO/JGCMTPP@MQ,+08*%+0GPOKS)UTTRP4PPRTOL2E1L)PKL+QG)+KP-*.5+;5.'
@0059c1dc-ab1b-42fa-a601-59ac8ad4b60f
CTAGACGTGCGAGTATACTACTATGCCGAGTTATATTTAAAACTGCATTTAGATTCCGTCGCTAGGCCTACTGATGCCCCAGCTGCCGTTTGGAATTAAGGAAGTGATACCTAGATGTGTGTTCACAGCGACCTTAAATGACACGCAGTCAGACGCGTGT
+
**0,@)?EC=%+ALD4.-*HEJHB,,*).>M6POJRK+,,JOGMGGCJ6,:+-K0B0F+.@-OKMME('@+CF=)2'0/&FC+)QMKORT/PN3N+E7..'564)RUQOJMMNIM.0/*)IRU*OMRNJJNM+UQRQTA*88F-OULUT)MJO-56>5)%
@0041ac21-7266-4a7e-b9c8-1a1a92b729e2
TTAATGTAGCTATCTCTGCTAGCTAAATCGATCAGAAGGTTTTCAAATCTTATAGTGATCTATAGCGCTTGATGGCGGAACGTGTCTTCGCTAAGTGCGGGCGTATGTGCAGCTTTGTGTCCCTGTATCGTTAACTAGGAAGCTGATGTACAGAGATAGACTACAG
+
&$&**6)(%%+0,-/4+%(64;1+>OQ++JORK''&)(3;-%-0&*,*#%'#$$%/'/1&&$*%--(*CB<MPENJKO+U50*%'&*49*(0')-+,/2HG6*1*;8+C93G)/.+)++))+(+%%,(&(%+('('#,213D10QO/+*:;//(*@OKNFB5'&$$
```

$$\text{Q-score} = \text{ASCII value} - 33$$

$$P(error) = 10^{-Q/10}$$

$$P(correct) = 1 - P(error)$$

Basecalled Base	ASCII Quality Score	Q-Score	P(error)	P(correct)
G	+	**10**	**0.1000**	**0.9000**
T	-	**12**	**0.0631**	**0.9369**
G	4	**19**	**0.0126**	**0.9874**

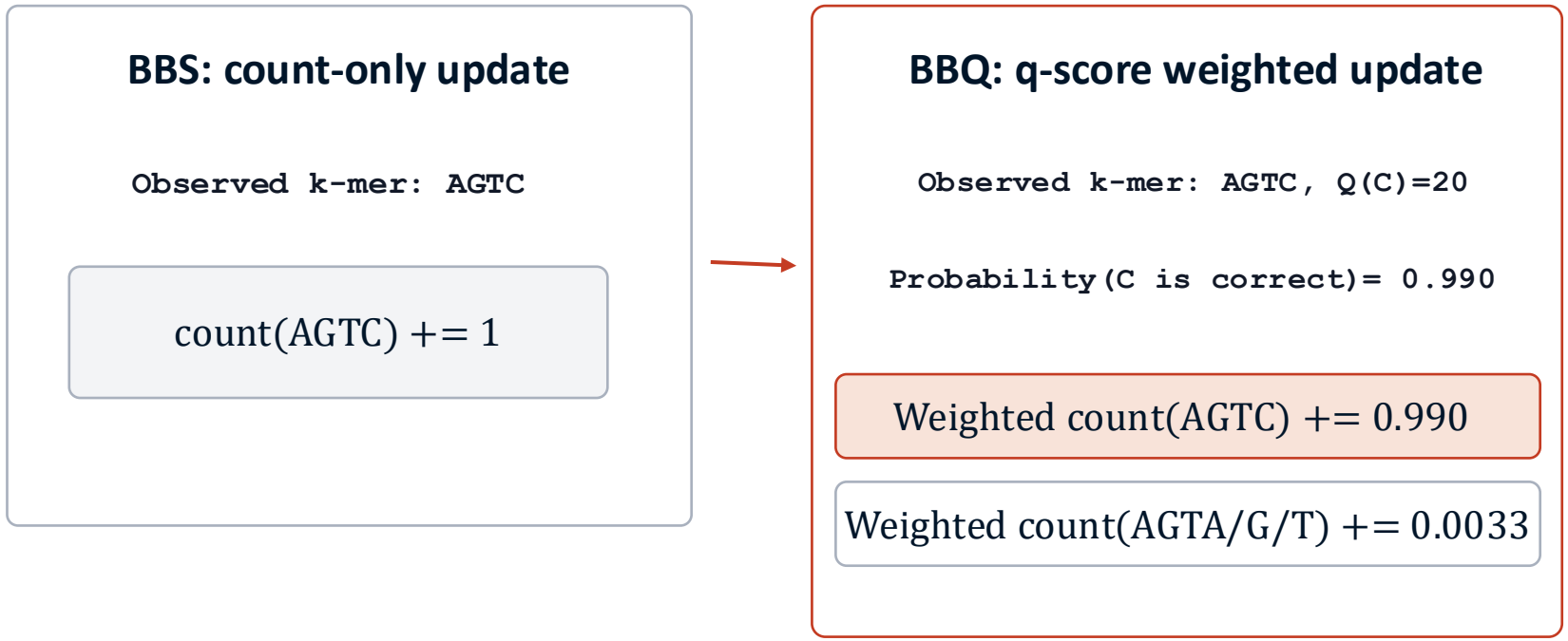
Method

BBS: count-only update

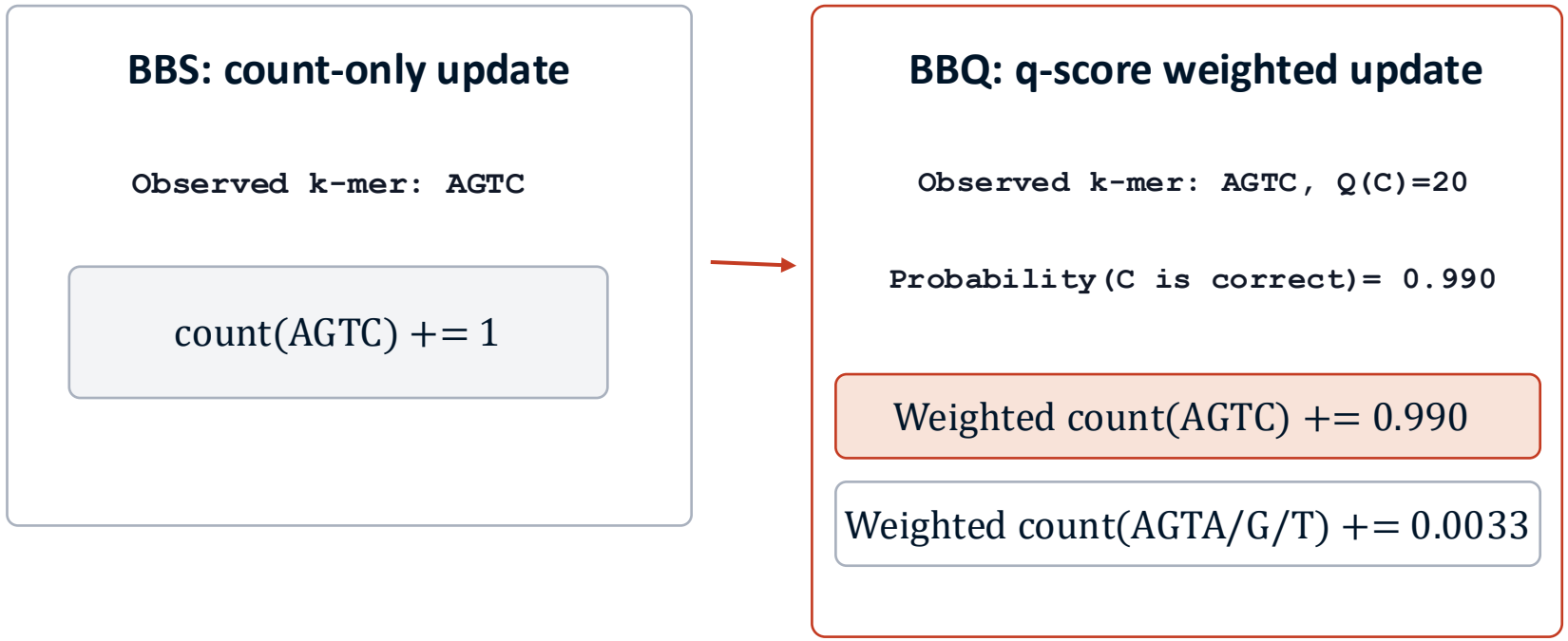
`Observed k-mer: AGTC`

$$\text{count}(\text{AGTC}) \mathrel{+}= 1$$

Method


BBS: count-only update
Observed k-mer: AGTC
count(AGTC) += 1
BBQ: q-score weighted update
Observed k-mer: AGTC, Q(C)=20
Probability(C is correct)= 0.990
Weighted count(AGTC) += 0.990
Weighted count(AGTA/G/T) += 0.0033

Method



- k is chosen exactly as in BBS.
- The same pseudo-count smoothing can be used.
- Each transition still means: keep the prefix and append one base.

Results

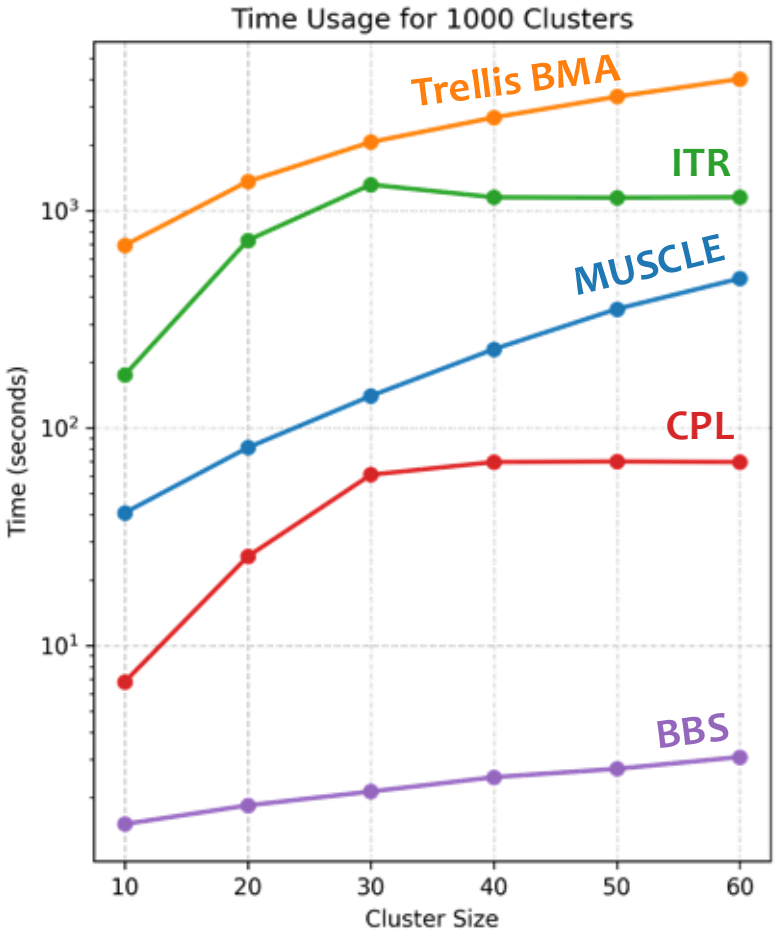
Results

Real datasets from different sources.

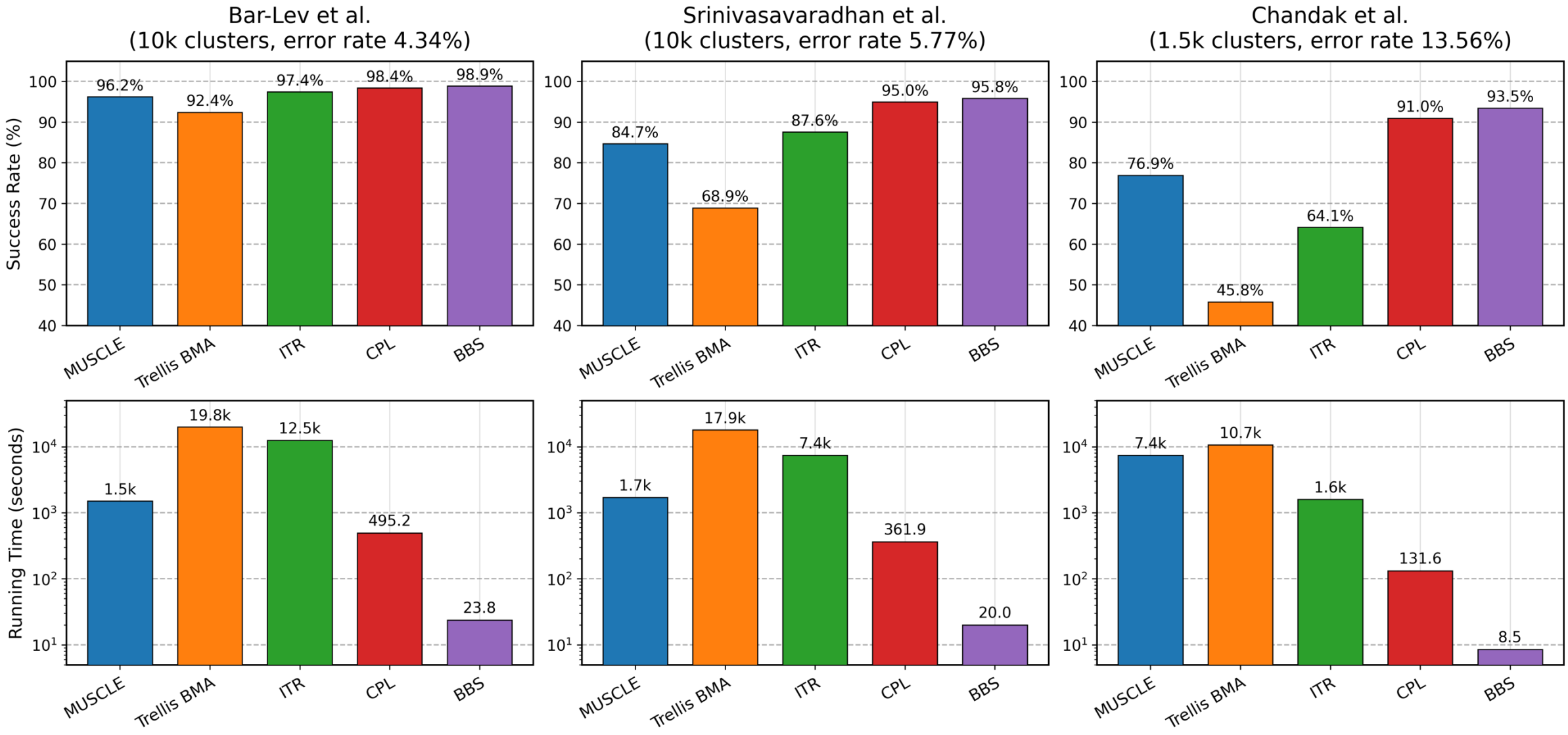
	Bar-Lev et al., 2025	Srinivasavaradhan et al., 2021	Chandak et al., 2020
# Clusters	10000	9984	1466
Synthesis Technology	Twist Bioscience	Twist Bioscience	CustomArray
Sequencing Technology	MinION	MinION	MinION
Clustering algorithm	[Bar-Lev et al., 2025]	[Rashtchian et al., 2018]	Perfect
L	140	110	108
Coverage	21.37	27.01	114.29
Read error rate	4.34%	5.77%	13.56%

Results

Algorithm	Method
MUSCLE [Edgar, 2004] [Antkowiak et al., 2020]	MSA
Trellis BMA [Srinivasavaradhan et al., 2021]	Read correction
ITR [Sabary et al., 2024]	Assembly
CPL [Bar-Lev et al., 2025]	Pairwise alignment
BBS & BBQ (this work, beam width $B = 20$)	k-th order Markov chain

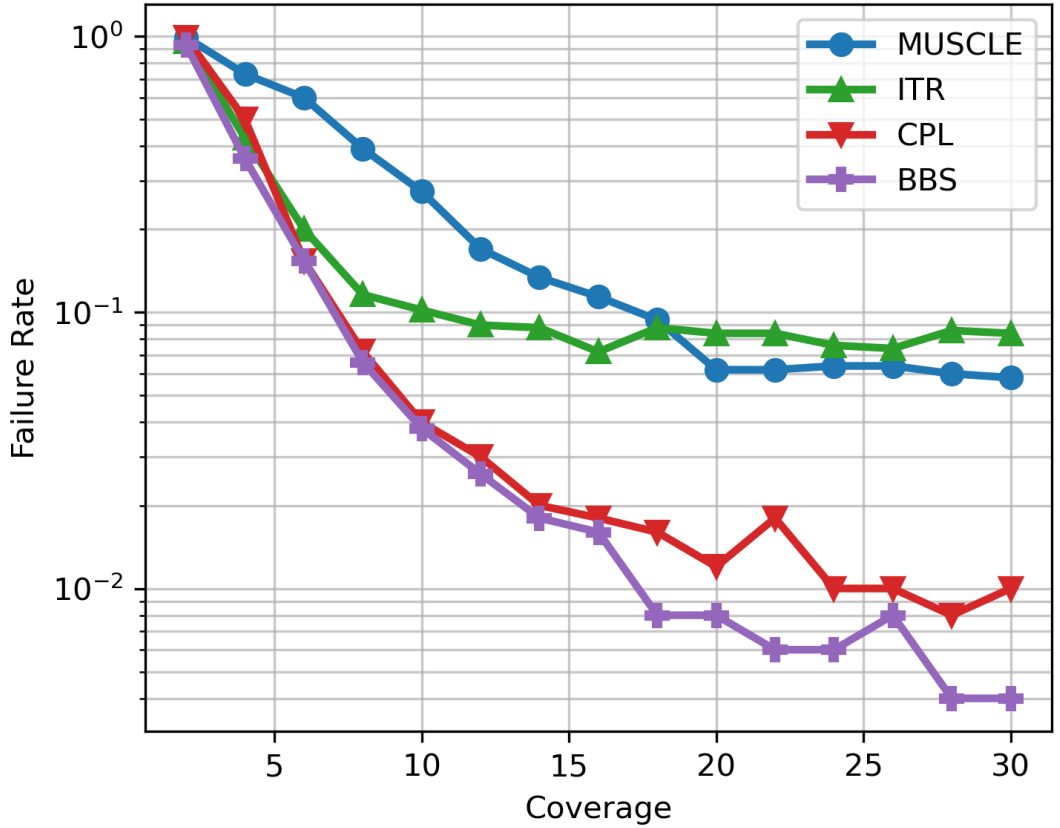


Results


Bar-Lev et al.
(10k clusters, error rate 4.34%)
Srinivasavaradhan et al.
(10k clusters, error rate 5.77%)
Chandak et al.
(1.5k clusters, error rate 13.56%)
Success Rate (%)
96.2%
92.4%
97.4%
98.4%
98.9%
84.7%
68.9%
87.6%
95.0%
95.8%
76.9%
45.8%
64.1%
91.0%
93.5%
Running Time (seconds)
1.5k
19.8k
12.5k
495.2
23.8
1.7k
17.9k
7.4k
361.9
20.0
7.4k
10.7k
1.6k
131.6
8.5
MUSCLE
Trellis BMA
ITR
CPL
BBS

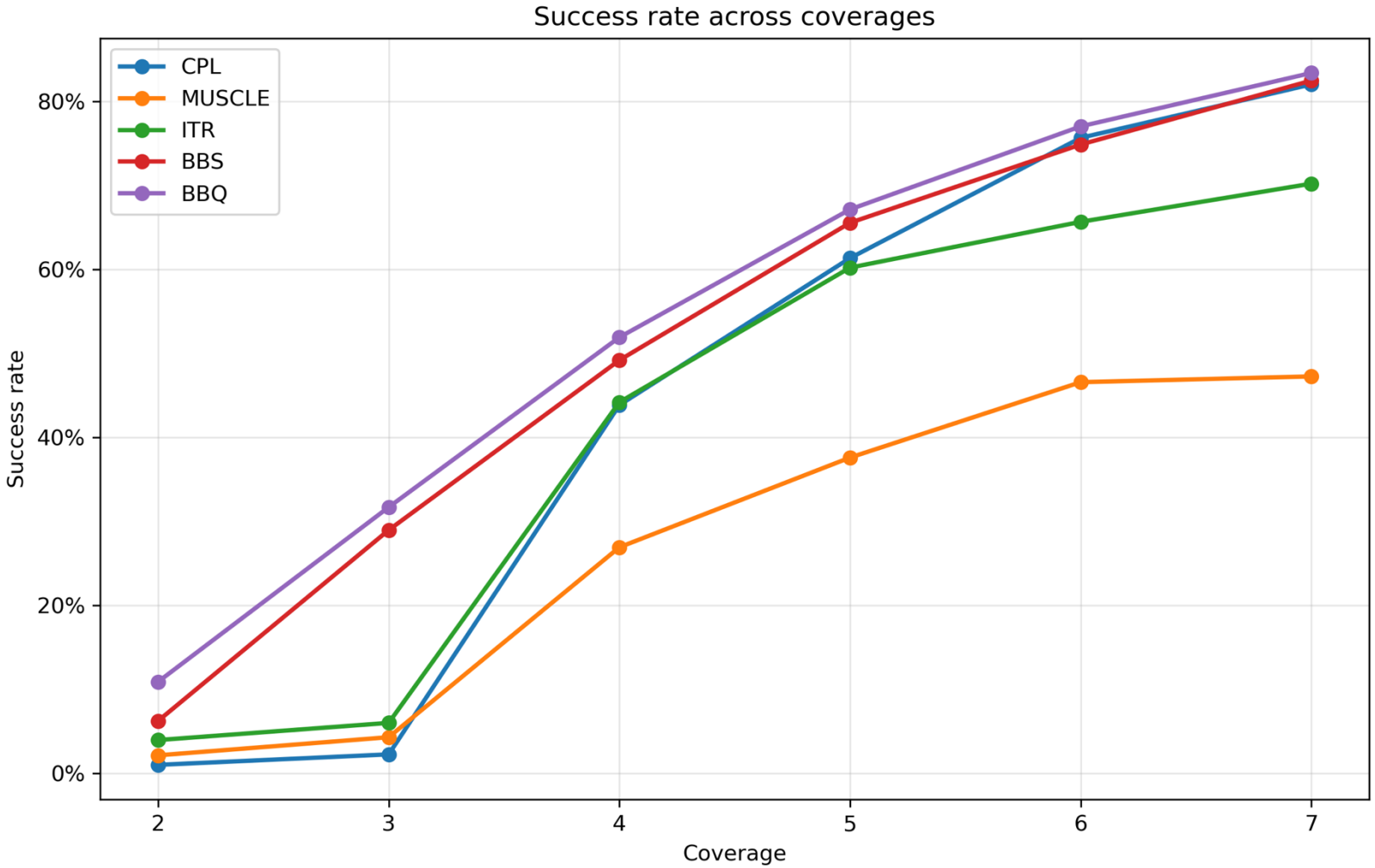
Results

(Lower is better)

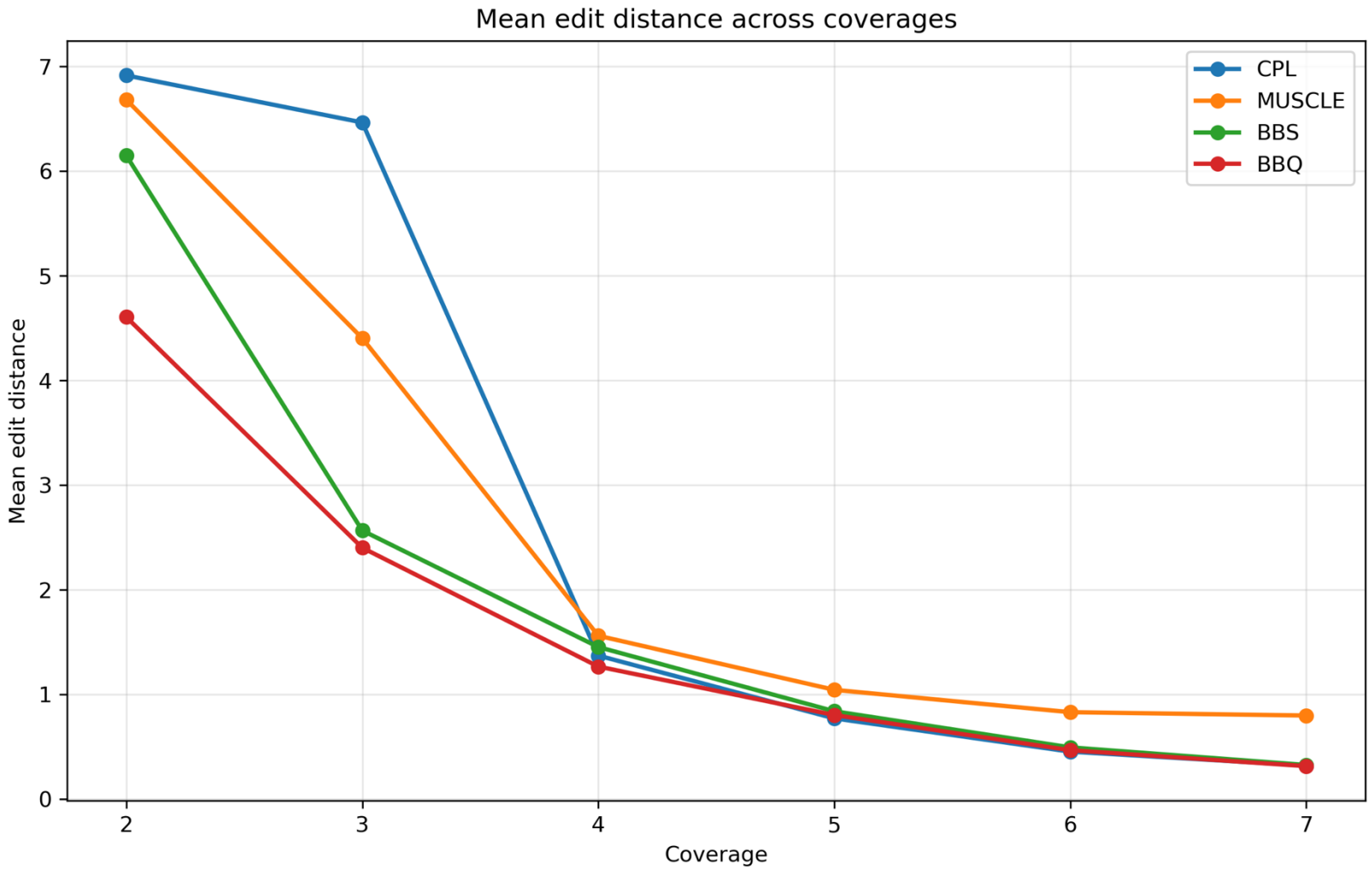


BBS performs well even with small coverage.

Results


Success rate across coverages
CPL
MUSCLE
ITR
BBS
BBQ
80%
60%
40%
20%
0%
Success rate
2
3
4
5
6
7
Coverage

Results


Mean edit distance across coverages
CPL
MUSCLE
BBS
BBQ
Mean edit distance
Coverage
0
1
2
3
4
5
6
7
2
3
4
5
6
7

Conclusion

Takeaways:

- Modeling the reads as outputs of a *k*-th order Markov chain is an effective way to capture position-specific error rates.
- Beam search is very fast compared to traditional methods based on alignment or assembly, and has almost no loss in accuracy.
- The path weights of returned candidates are good indicators of reconstruction quality.
- The q-score in FASTQ files is an important tool to achieve higher reconstruction accuracy at low coverages.

Acknowledgement

Thank you very much for your attention!

My collaborators:

Zhenhao Gu

Gary Goh Yipeng

Weng-Fai Wong

Special thanks: Djordje Jevdic , Vasily Shenshin, Yannick Rondelez

NUS | Computing

STORAGE AND COMPUTING WITH DNA

Thank you

Sponsored by

DNA DATA STORAGE ALLIANCE
A SNIA Community

European Innovation Council

DigNA

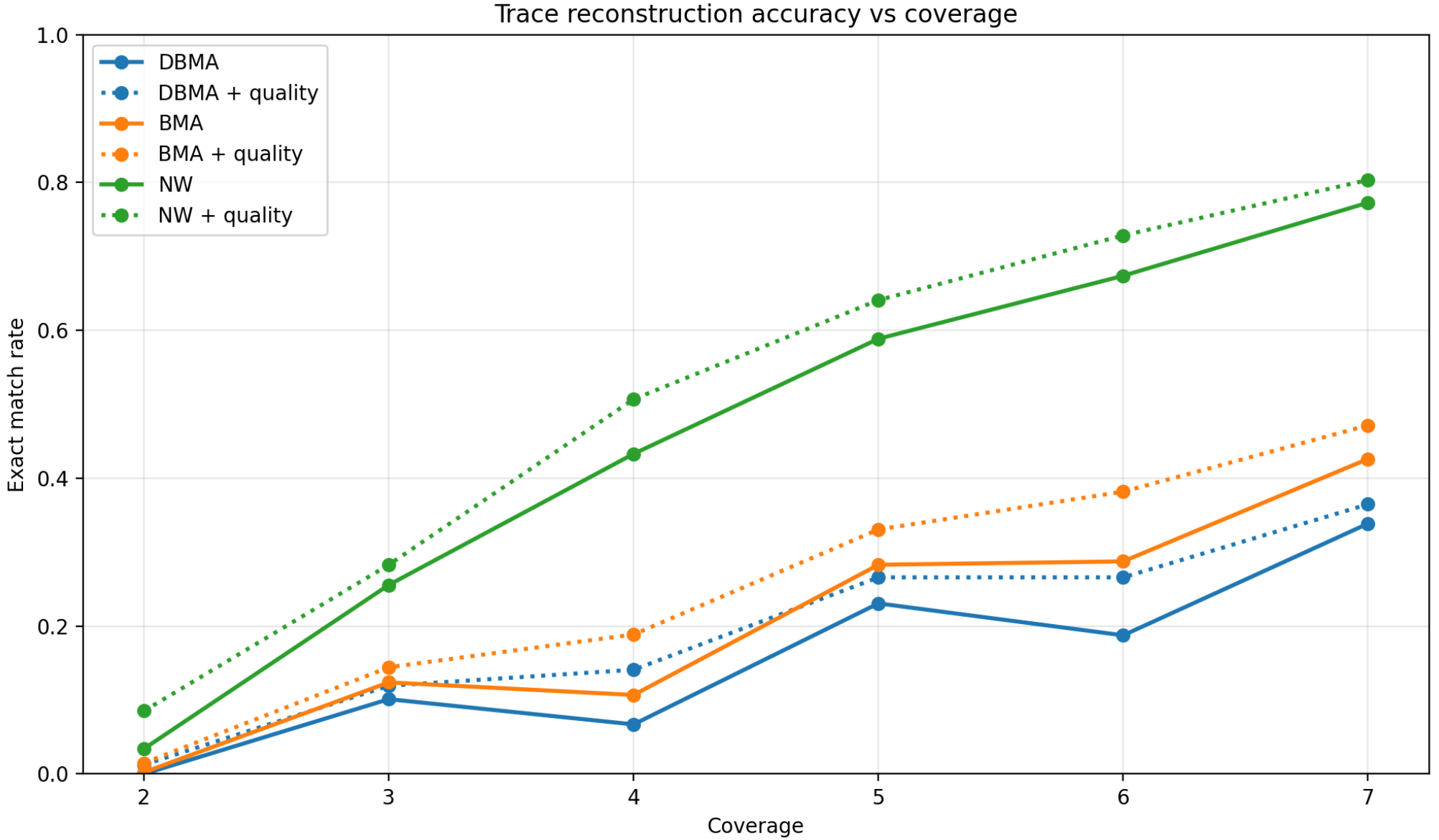
ATLAS

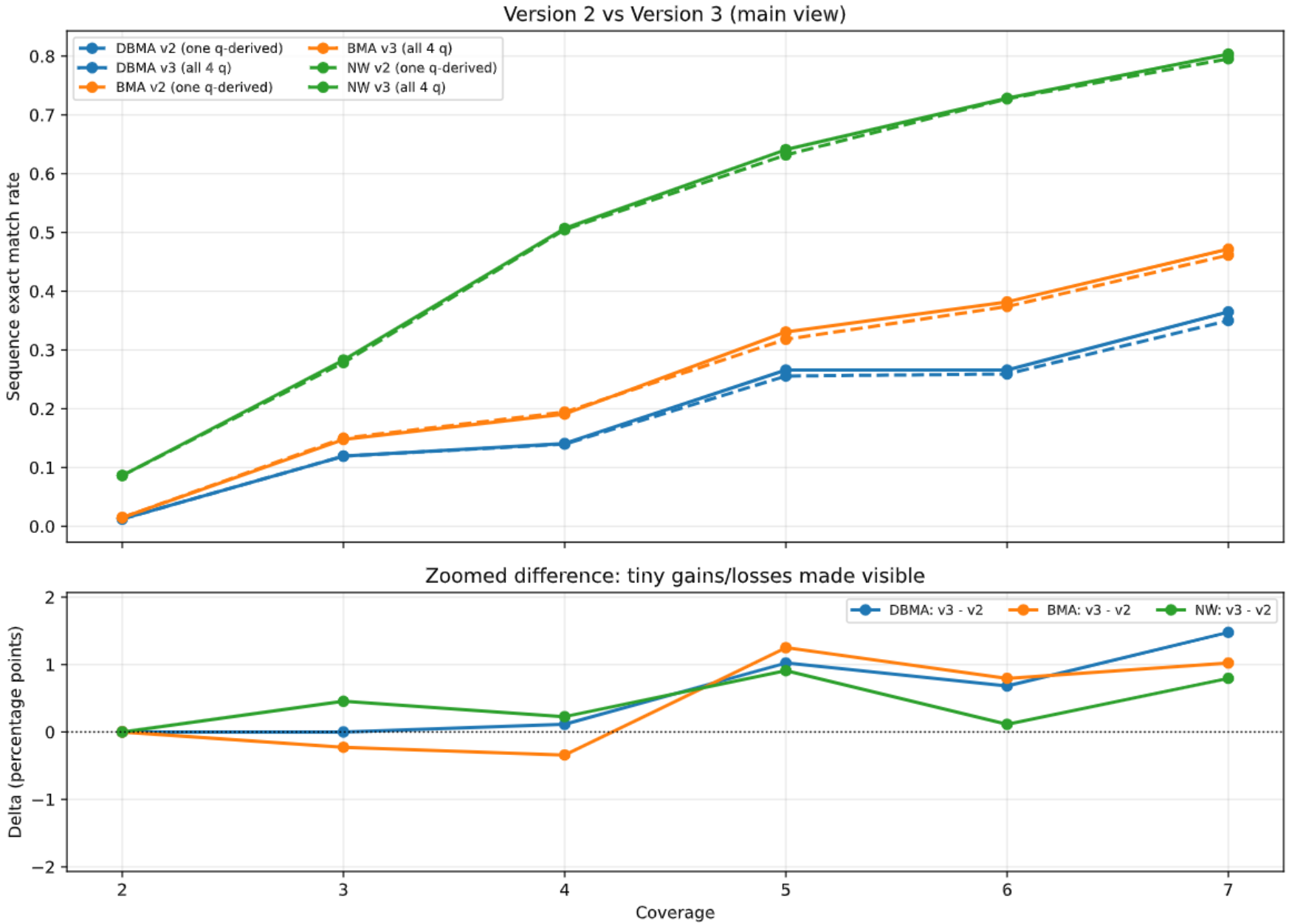
BIOMEMORY
Sustainable Storage Solutions

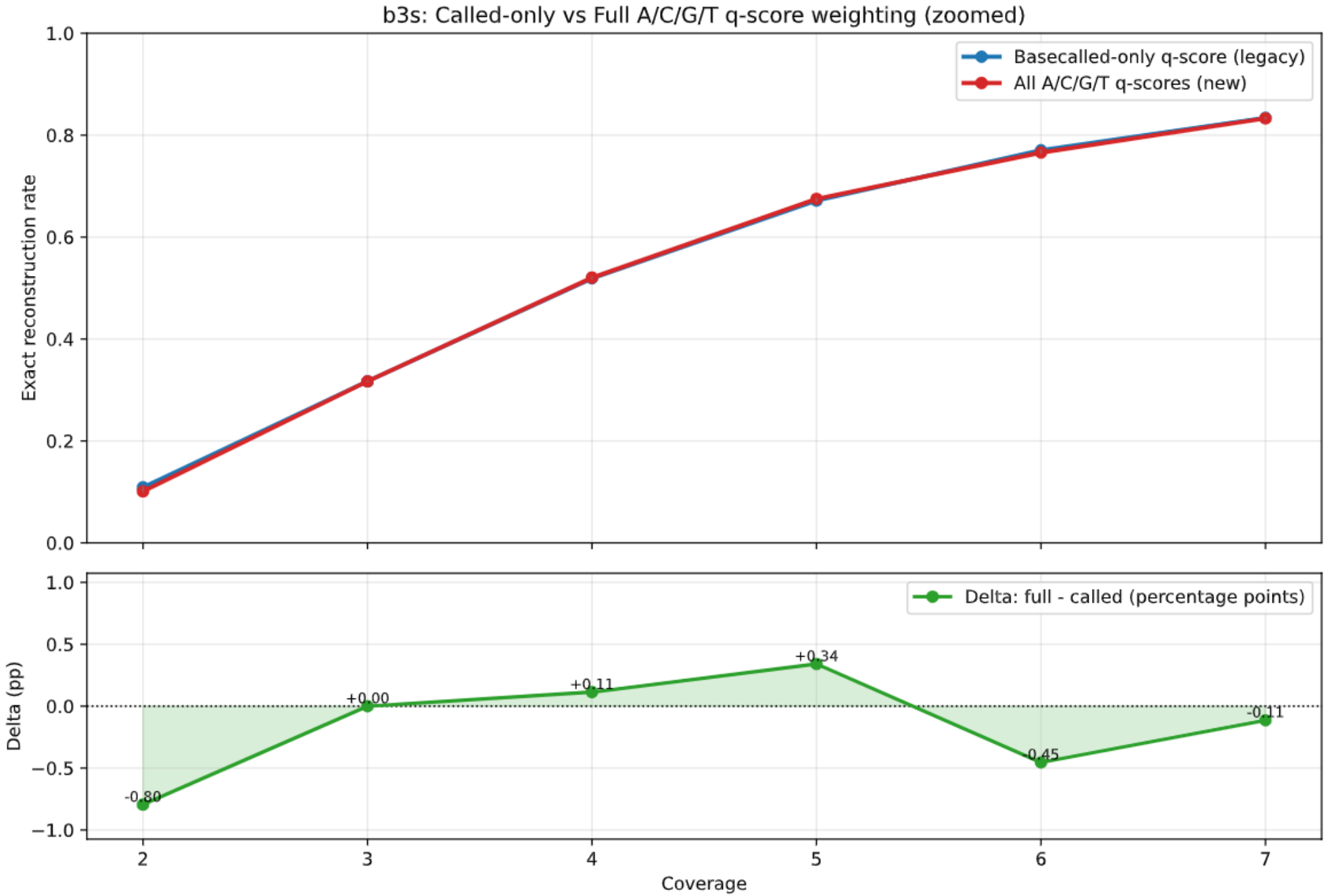
WD

Organized by

day one


Trace reconstruction accuracy vs coverage
DBMA
DBMA + quality
BMA
BMA + quality
NW
NW + quality
Exact match rate
Coverage
0.0
0.2
0.4
0.6
0.8
1.0
2
3
4
5
6
7


Version 2 vs Version 3 (main view)
DBMA v2 (one q-derived)
DBMA v3 (all 4 q)
BMA v2 (one q-derived)
BMA v3 (all 4 q)
NW v2 (one q-derived)
NW v3 (all 4 q)
Sequence exact match rate
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Zoomed difference: tiny gains/losses made visible
DBMA: v3 - v2
BMA: v3 - v2
NW: v3 - v2
Delta (percentage points)
2
1
0
−1
−2
2
3
4
5
6
7
Coverage


b3s: Called-only vs Full A/C/G/T q-score weighting (zoomed)
Basecalled-only q-score (legacy)
All A/C/G/T q-scores (new)
Exact reconstruction rate
1.0
0.8
0.6
0.4
0.2
0.0
Delta: full - called (percentage points)
Delta (pp)
1.0
0.5
0.0
−0.5
−1.0
-0.80
+0.00
+0.11
+0.34
-0.45
-0.11
2
3
4
5
6
7
Coverage