Stanford

Improved read/write cost tradeoff in DNA-based data storage using LDPC codes

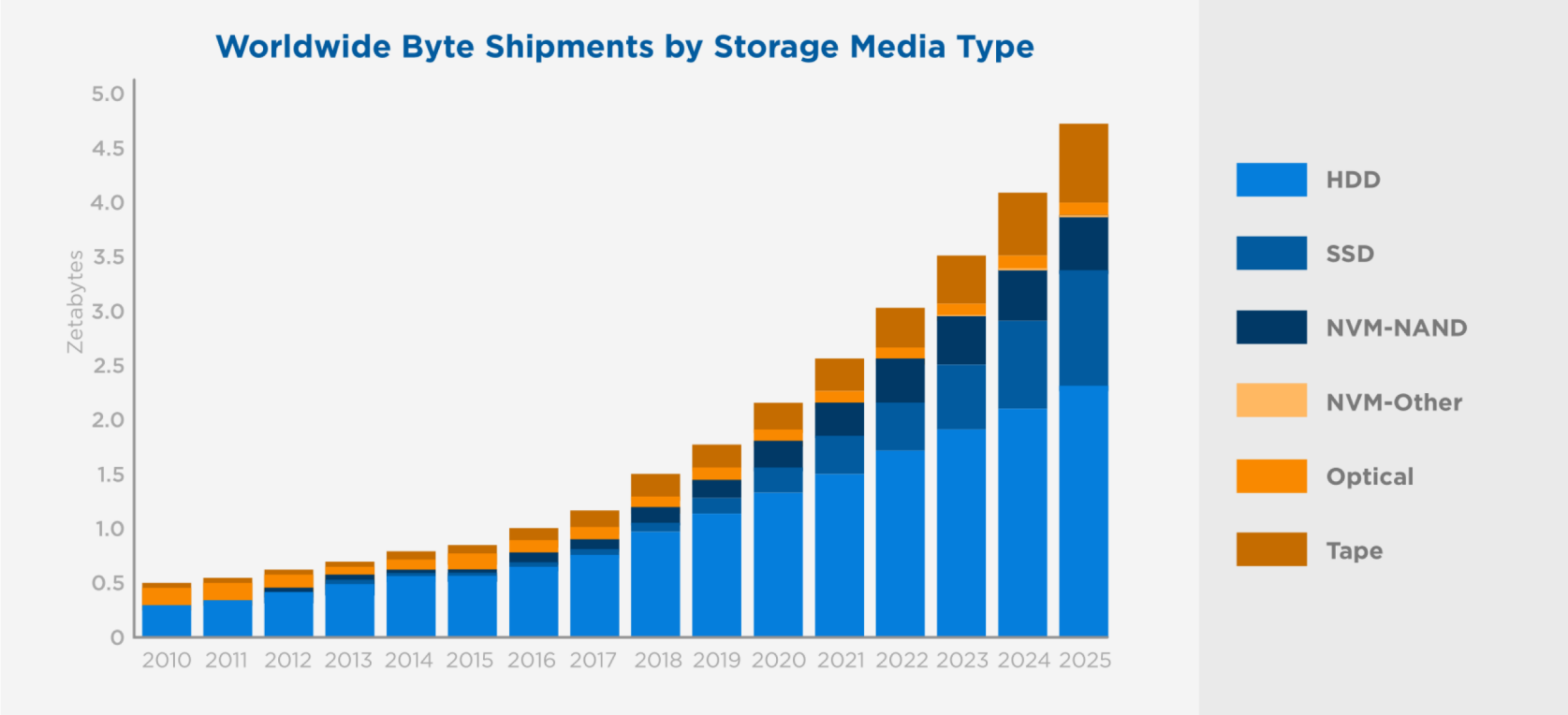
Shubham Chandak
Stanford University
Allerton 2019

Outline

- Motivation
- DNA storage setup
- Theoretical analysis
- Proposed framework
- Results
- Conclusions

Motivation

The amount of stored data is growing exponentially:



Source: Data Age 2025, sponsored by Seagate with data from IDC Global DataSphere, Nov 2018

Source: https://www.seagate.com/our-story/data-age-2025/

200 Petabyte

200 Petabyte

40,000 x 5 TByte HDDs
40 tons

10s of years

200 Petabyte

40,000 x 5 TByte HDDs
40 tons

10s of years

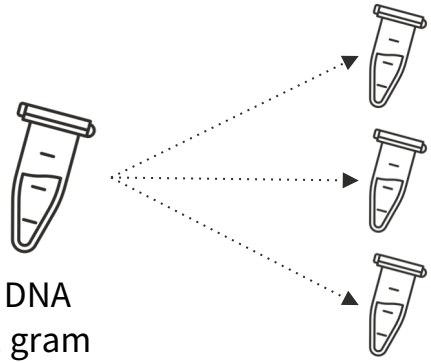
DNA
1 gram

1,000s of years

200 Petabyte

40,000 x 5 TByte HDDs
40 tons

10s of years



DNA
1 gram

1,000s of years

Easy duplication

July 2, 2019

Hot News for the Summer from CATALOG

POSTED BY : SEAN MIHM / 0 COMMENTS / UNDER : UNCATEGORIZED

CATALOG Encodes Wikipedia Into DNA!

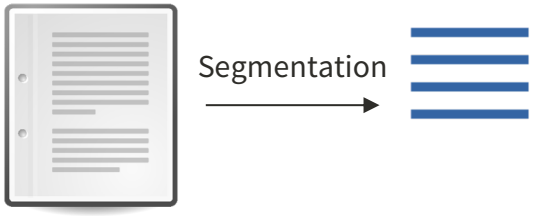
https://catalogdna.com/uncategorized/hot-news-for-the-summer-from-catalog/

DNA storage setup

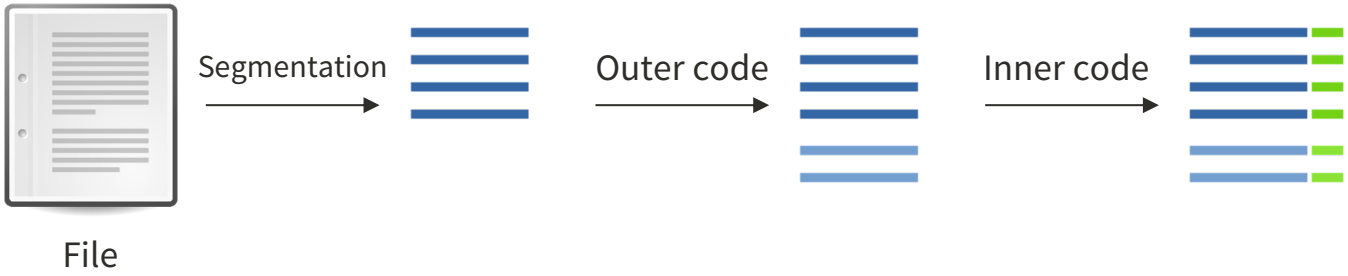
How to store data in DNA sequences?

File

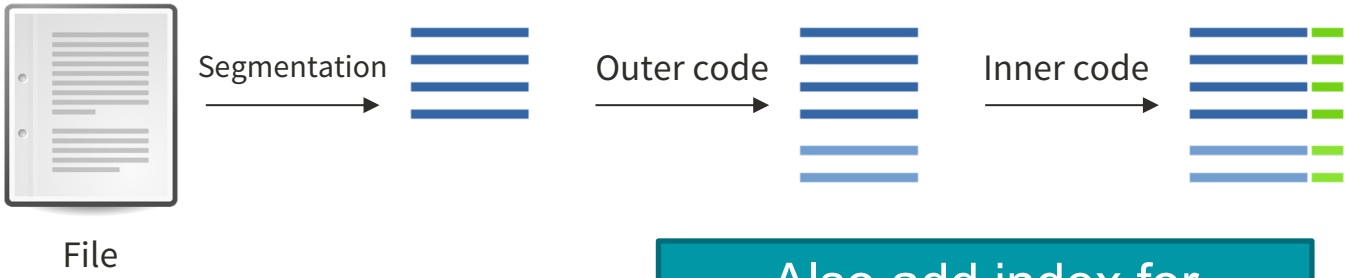
How to store data in DNA sequences?



How to store data in DNA sequences?

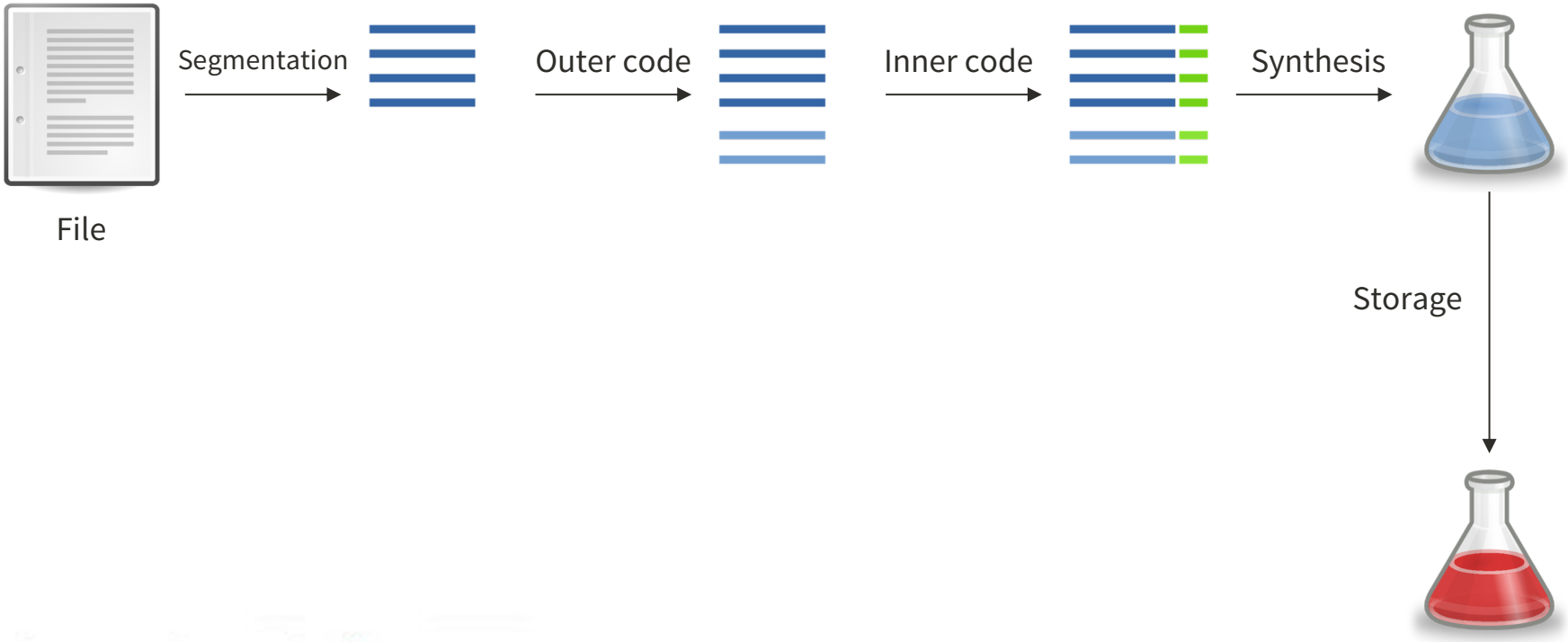


How to store data in DNA sequences?



Also add index for recovering order of segments

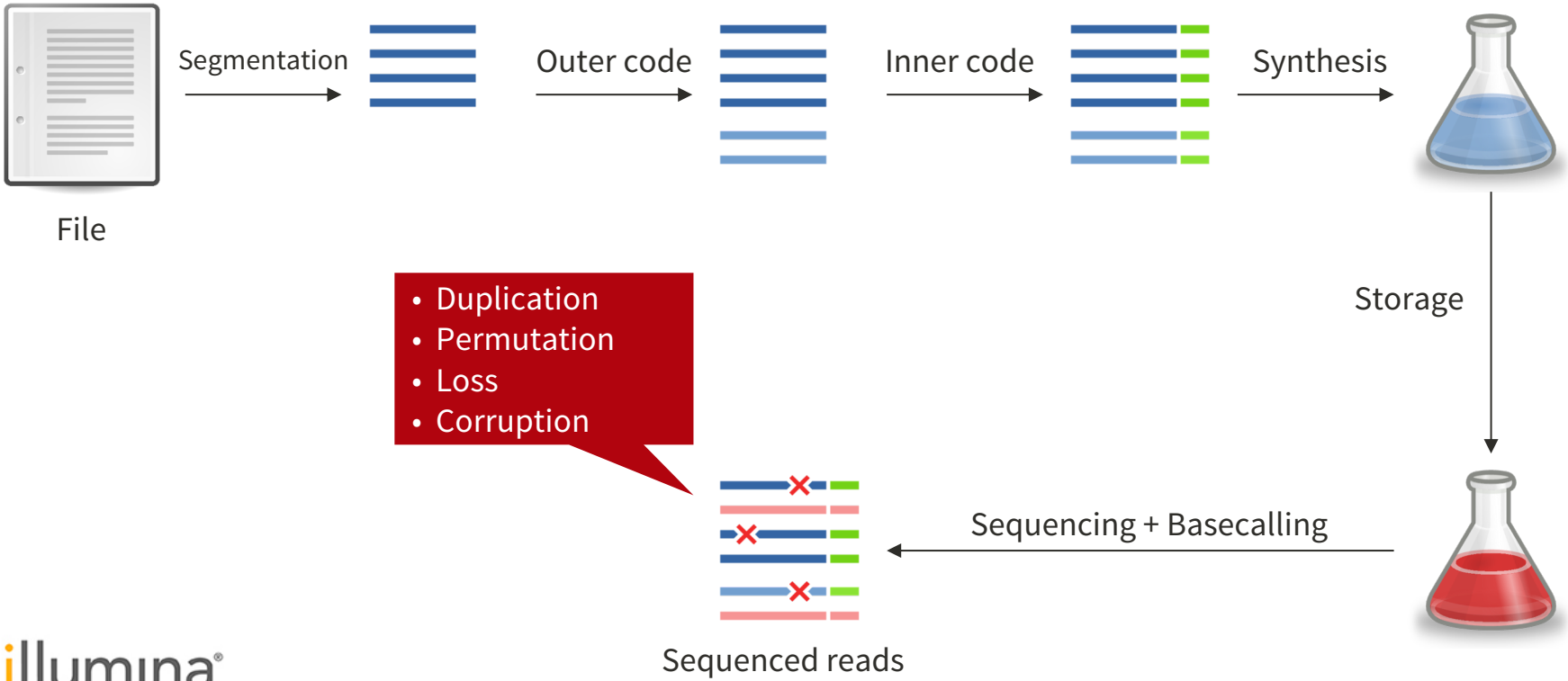
How to store data in DNA sequences?



CustomArray | GenScript Make Research Easy

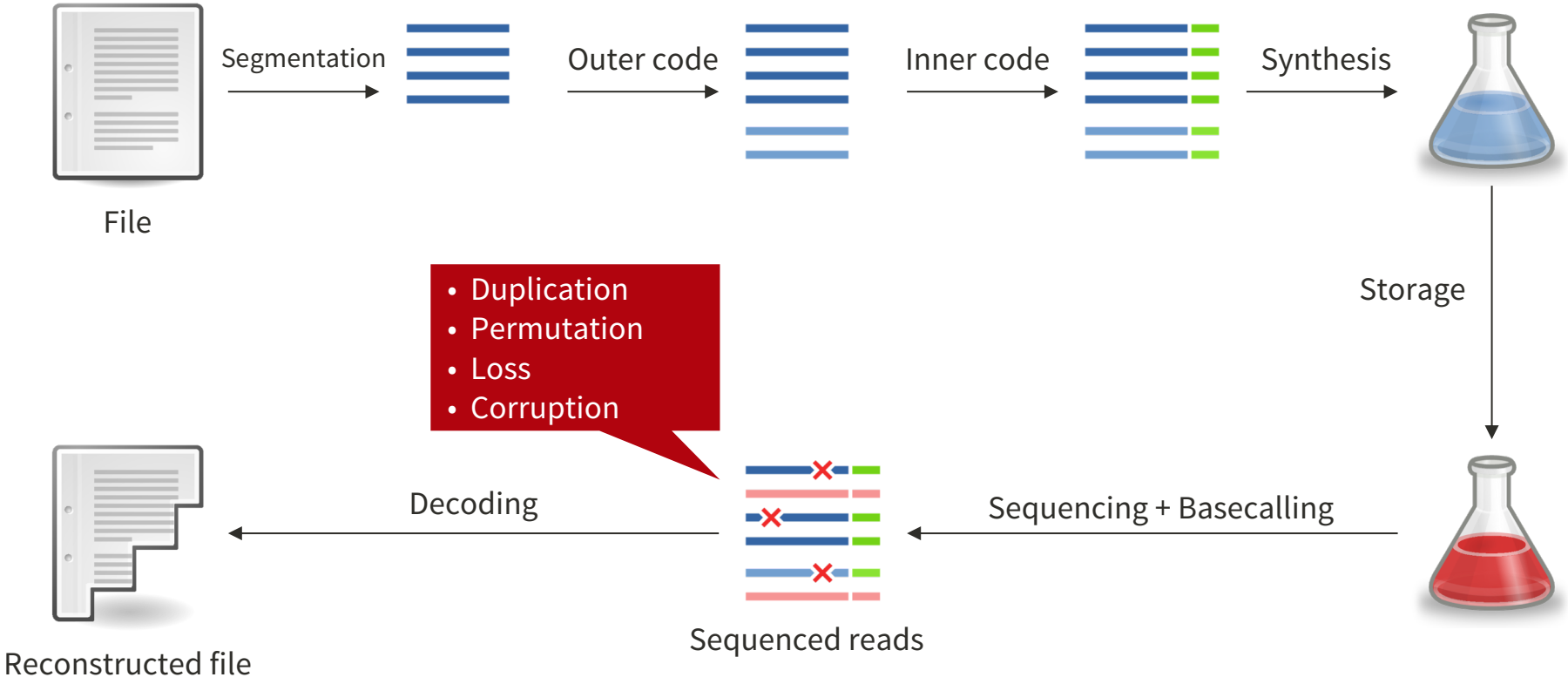
http://www.customarrayinc.com/

How to store data in DNA sequences?

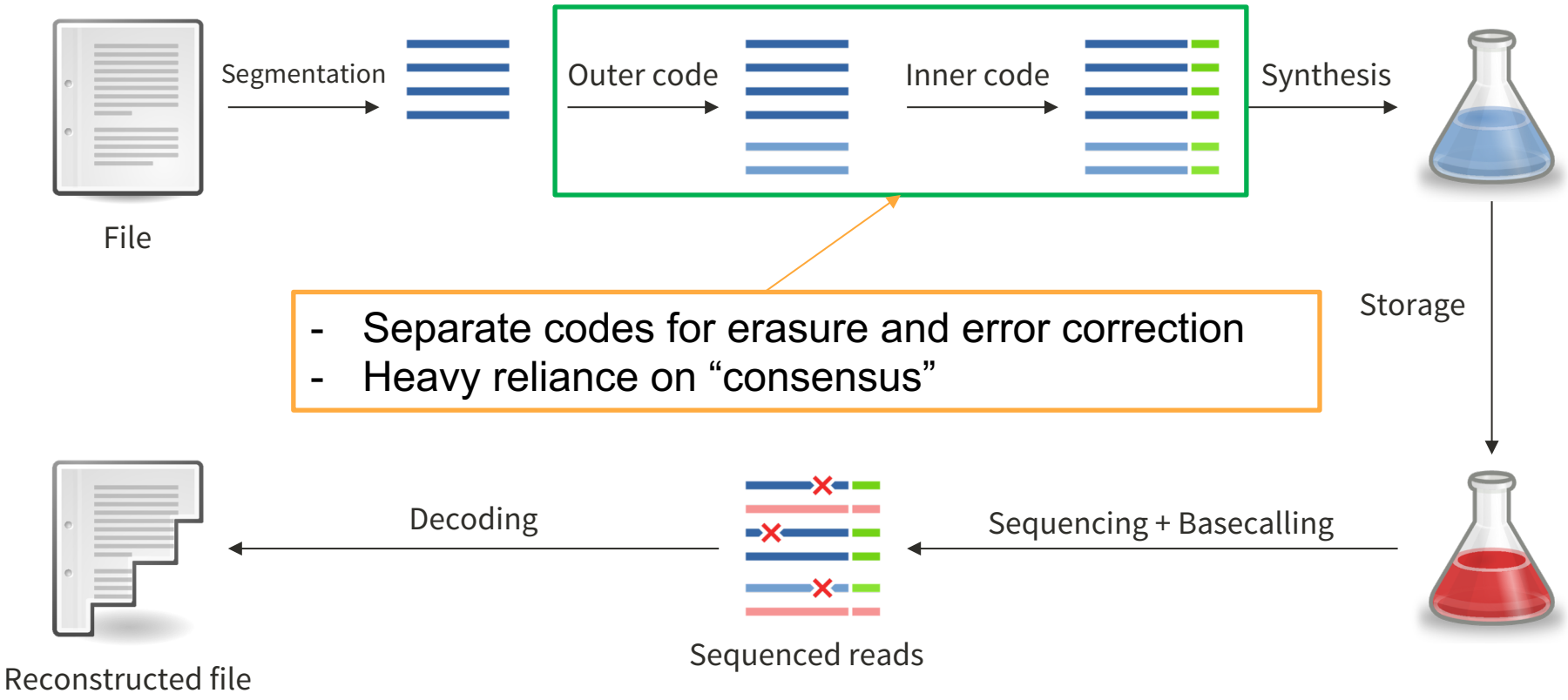


illumina®

How to store data in DNA sequences?



How to store data in DNA sequences?



Previous works

- Multiple previous works focusing on:
  - Error correction coding
  - Random access to subsets of synthesized sequences using PCR primers
  - Scalable and cost effective synthesis techniques
  - Different sequencing platforms
  - Theoretical analysis

1. Yazdi, SM Hossein Tabatabaei, et al. "A rewritable, random-access DNA-based storage system." *Scientific reports* 5 (2015): 14138.
2. Erlich, Yaniv, and Dina Zielinski. "DNA Fountain enables a robust and efficient storage architecture." Science 355.6328 (2017): 950-954.
3. Organick, Lee, et al. "Random access in large-scale DNA data storage." Nature biotechnology 36.3 (2018): 242.
4. Blawat, Meinolf, et al. "Forward error correction for DNA data storage." Procedia Computer Science 80 (2016): 1011-1022.
5. Church, George M., Yuan Gao, and Sriram Kosuri. "Next-generation digital information storage in DNA." *Science* 337.6102 (2012): 1628-1628.
6. Heckel, Reinhard, et al. "Fundamental limits of DNA storage systems." *2017 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2017.
7. Tomek, Kyle J., et al. "Driving the scalability of DNA-based information storage systems." *ACS synthetic biology* (2019).
8. Lenz, Andreas, et al. "Coding over sets for DNA storage." *2018 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2018.
9. Lee, Henry H., et al. "Terminator-free template-independent enzymatic DNA synthesis for digital information storage." *Nature communications* 10.1 (2019): 2383.

Theoretical analysis

Read-write cost tradeoff

- Fundamental quantities from a coding theory perspective:
  - Writing cost (bases synthesized/message bit)
  - Reading cost (bases sequenced/message bit)
  - *Note*: “Coverage” (= bases sequenced/bases synthesized) doesn’t capture the actual reading cost.

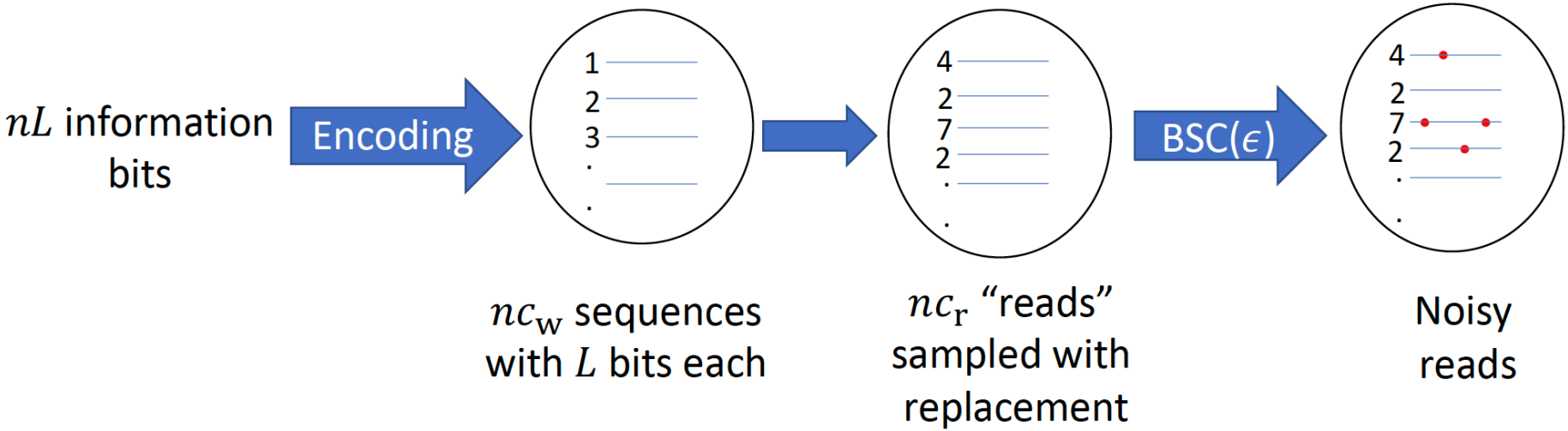
Read-write cost tradeoff

- Fundamental quantities from a coding theory perspective:
  - Writing cost (bases synthesized/message bit)
  - Reading cost (bases sequenced/message bit)
  - *Note*: “Coverage” (= bases sequenced/bases synthesized) doesn’t capture the actual reading cost.
- Fixed sequence length means asymptotic information capacity = 0!

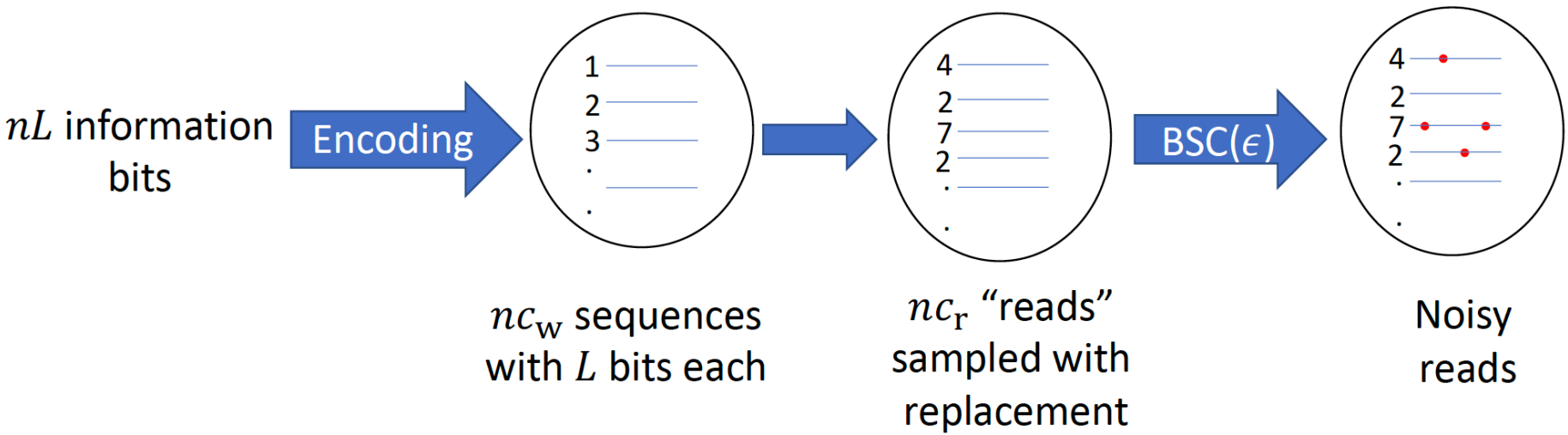
Read-write cost tradeoff

- Fundamental quantities from a coding theory perspective:
  - Writing cost (bases synthesized/message bit)
  - Reading cost (bases sequenced/message bit)
  - *Note*: “Coverage” (= bases sequenced/bases synthesized) doesn’t capture the actual reading cost.
- Fixed sequence length means asymptotic information capacity = 0!
  - Previous works assumed sequence length growing logarithmically in number of sequences
  - Does not capture the limitations posed by short sequence length

Simplified model for analysis



Simplified model for analysis



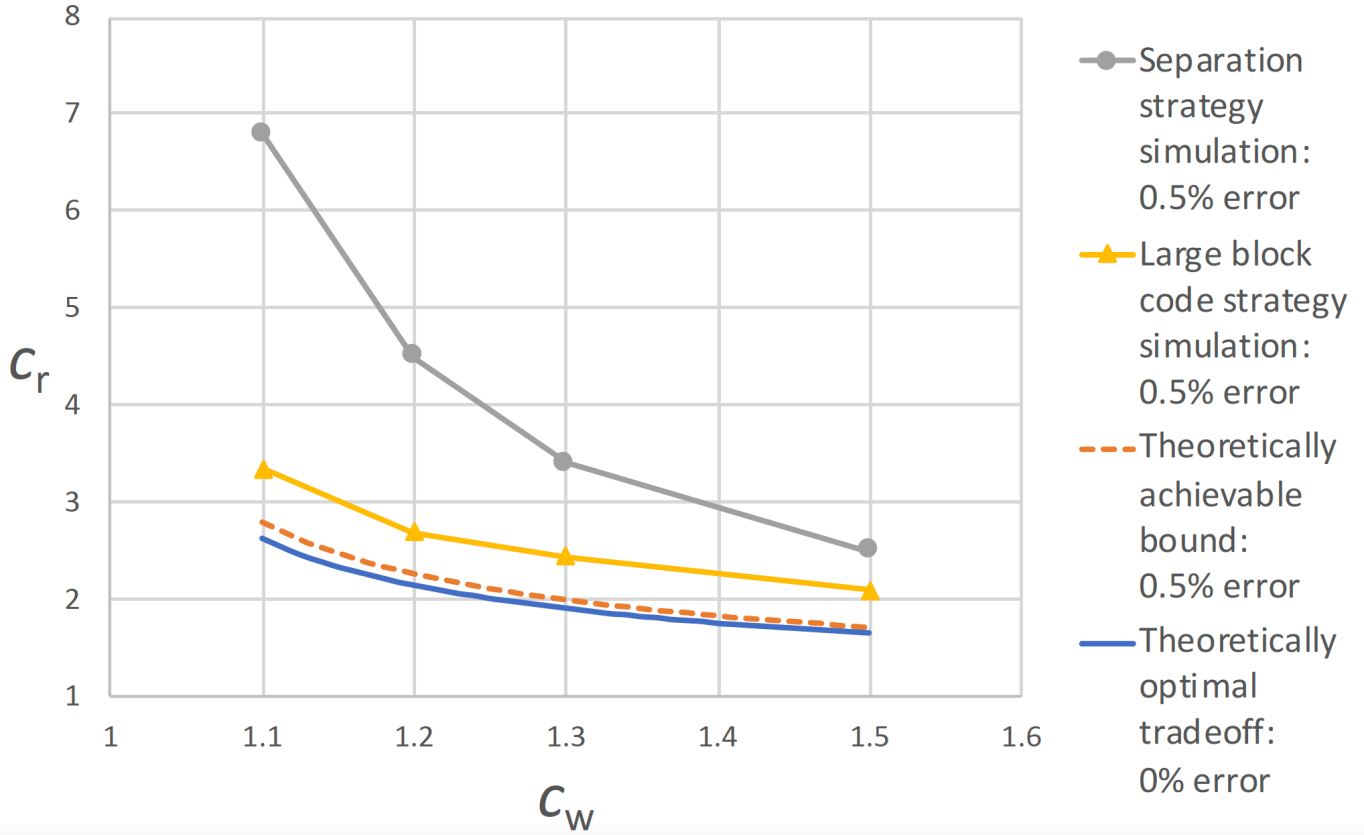
Use a memoryless approximation and obtain asymptotically achievable tradeoff between c_w and c_r

Two strategies

Strategy 1: Inner/outer code separation

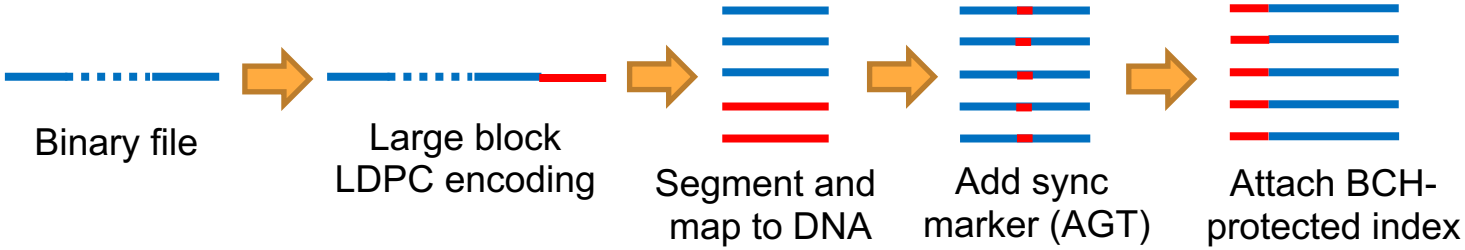
Strategy 2: Single large block code

Simulation results



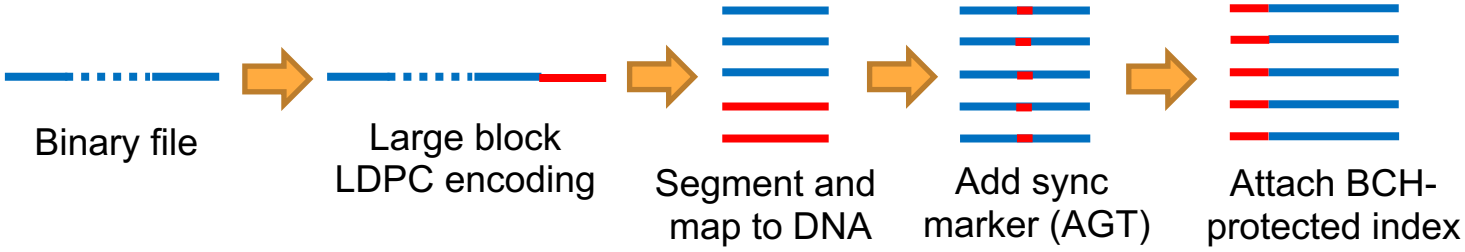
Proposed framework

Proposed approach

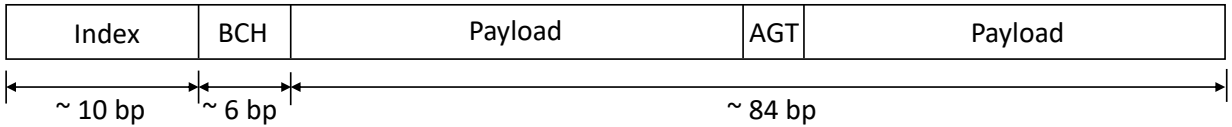


Encoding

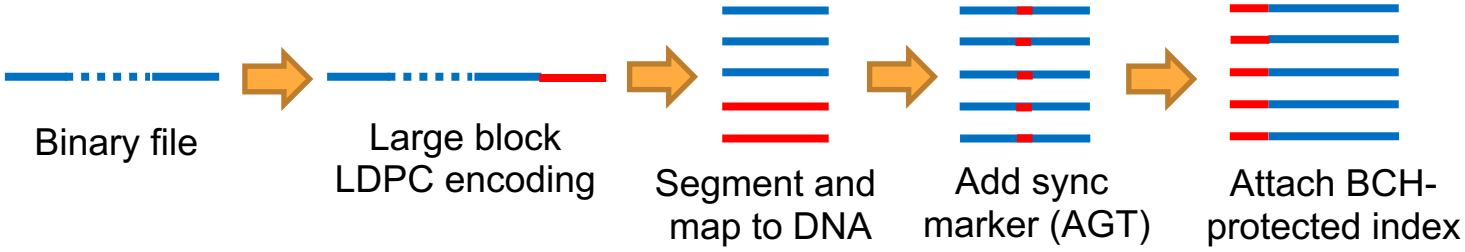
Proposed approach



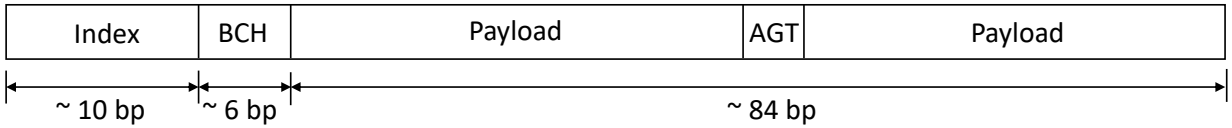
Encoding

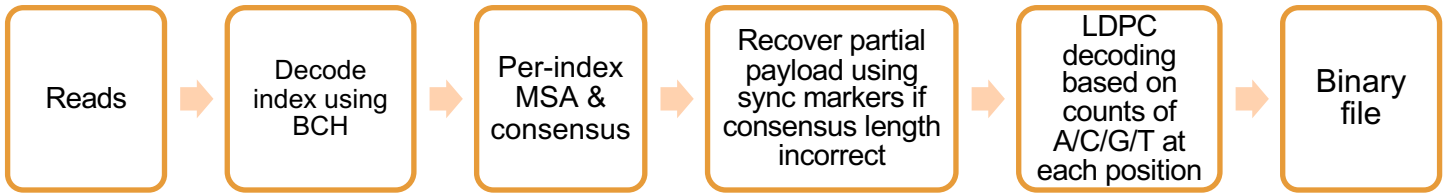


Proposed approach



Encoding





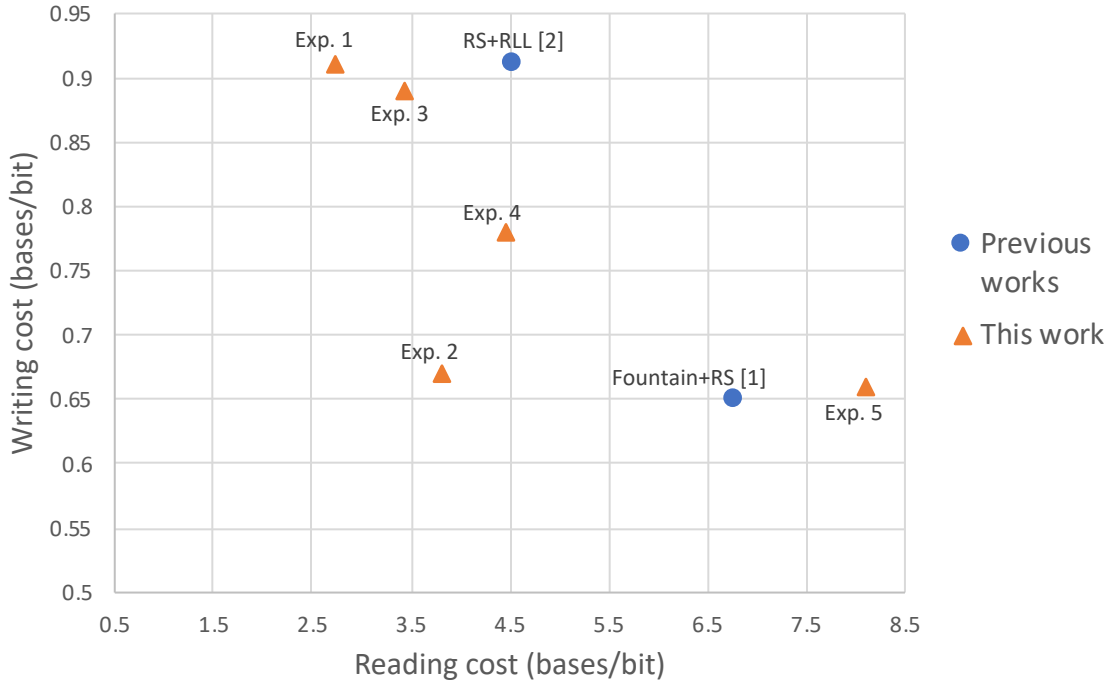
Decoding

Results

Experimental Parameters

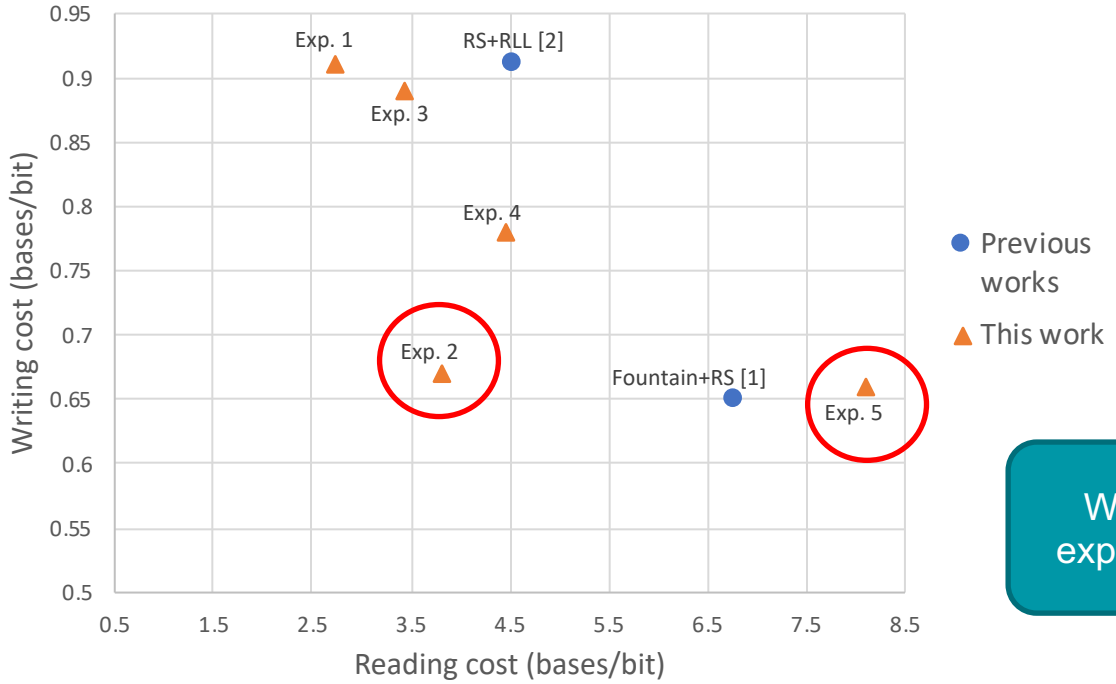
- Multiple parameter experiments, storing around 200 KB data each.
- CustomArray synthesis, length 150 including primers.
- Sequenced with Illumina iSeq.
- Total error rate around 1.3% (substitution: 0.4%, deletion: 0.85%, insertion: 0.05%) – *cheaper* and *noisier* synthesis as compared to previous works.

Experimental Results



1. Y. Erlich and D. Zielinski, “DNA Fountain enables a robust and efficient storage architecture," *Science*, vol. 355, no. 6328, pp. 950-954, 2017.
2. L. Organick *et al.*, “Random access in large-scale DNA data storage," *Nature biotechnology*, vol. 36, no. 3, p. 242, 2018.

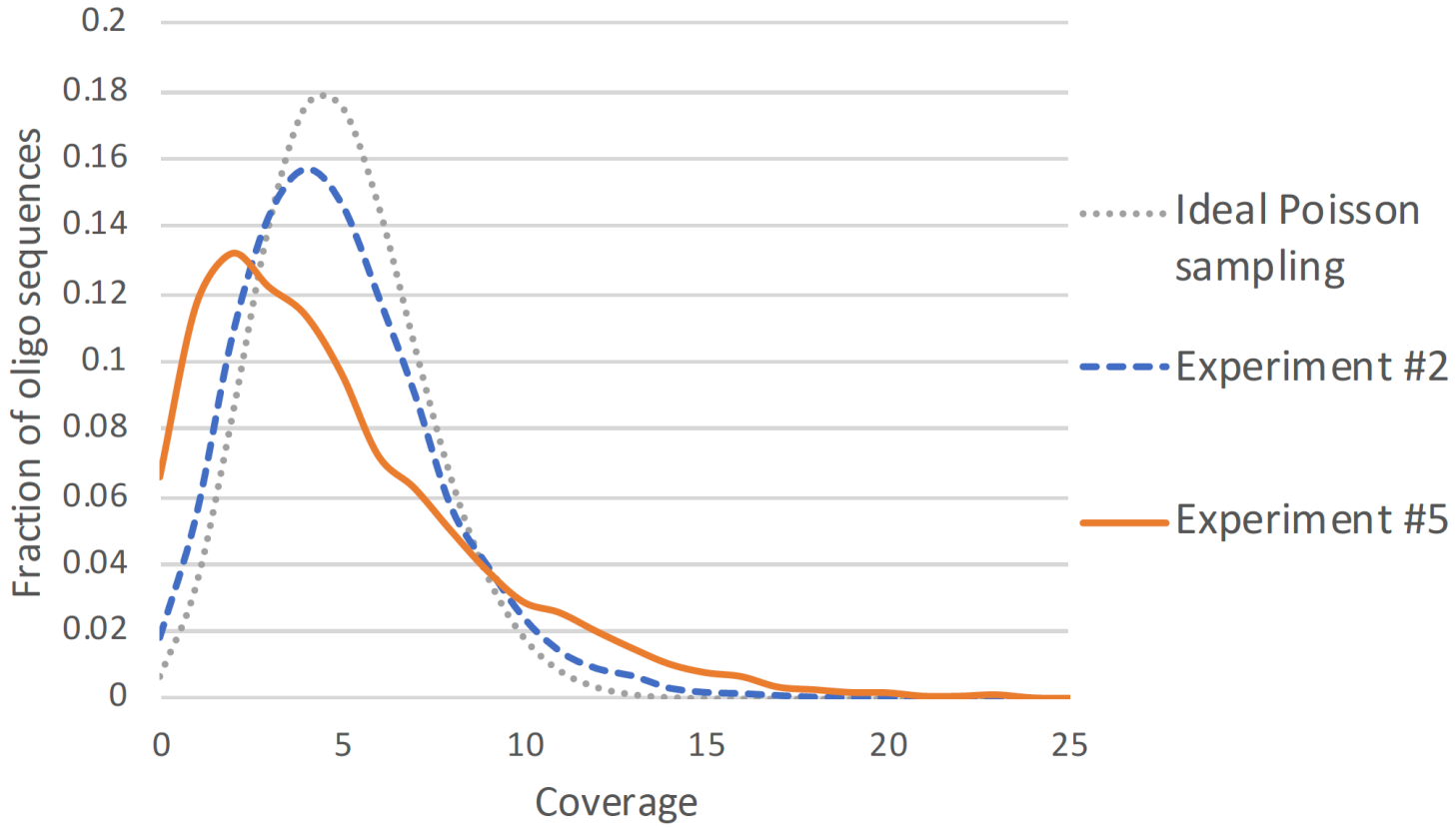
Experimental Results



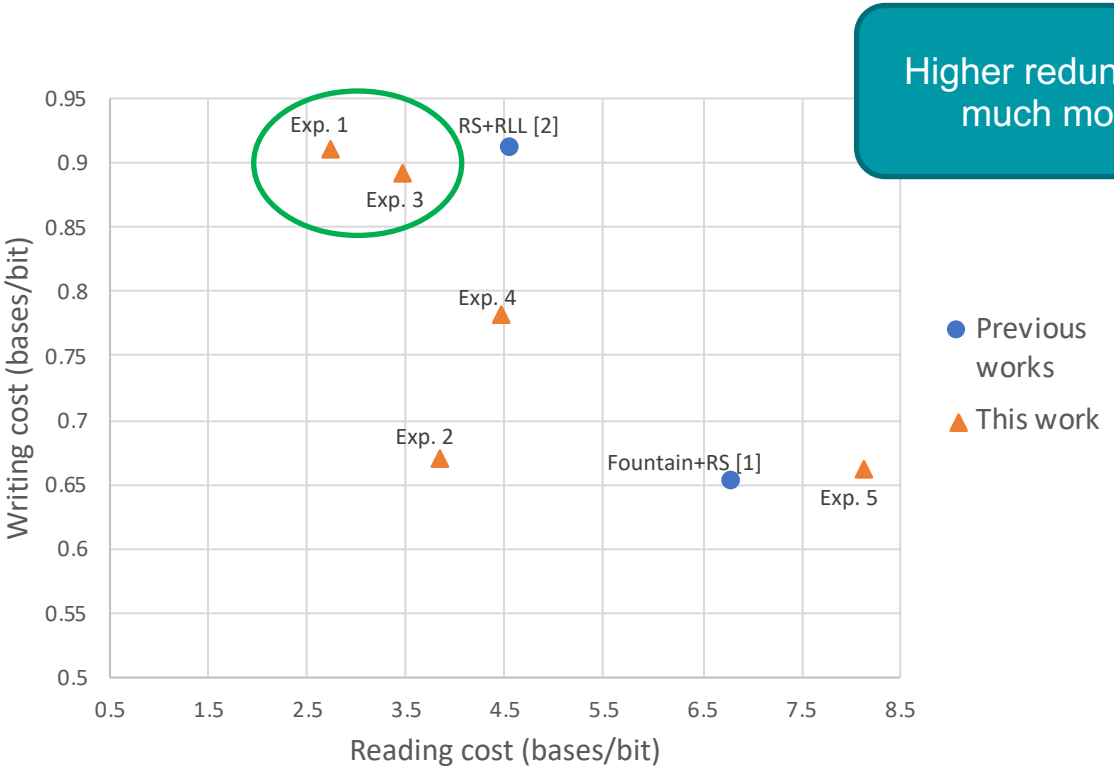
What happened in experiments 2 and 5?

1. Y. Erlich and D. Zielinski, “DNA Fountain enables a robust and efficient storage architecture," *Science*, vol. 355, no. 6328, pp. 950-954, 2017.
2. L. Organick *et al.*, “Random access in large-scale DNA data storage," *Nature biotechnology*, vol. 36, no. 3, p. 242, 2018.

Coverage variation



Experimental Results



Experimental Results

Higher redundancy codes much more robust!

More analysis in paper

Conclusions

- Introduced novel coding schemes for Illumina sequencing based DNA storage
    - Improved read/write cost tradeoff despite noisier synthesis
- Code and data: https://github.com/shubhamchandak94/LDPC_DNA_storage
- Biorxiv: https://www.biorxiv.org/content/10.1101/770032v1

Future work

- Possibilities for improvement:
  - Optimized LDPC codes, e.g., using protographs
  - Better codes for insertion/deletion: LDPC with markers, VT codes
  - Check out q-ary VT codes implementation: https://github.com/shubhamchandak94/VT_codes/

Future work

- Possibilities for improvement:
  - Optimized LDPC codes, e.g., using protographs
  - Better codes for insertion/deletion: LDPC with markers, VT codes
  - Check out q-ary VT codes implementation: https://github.com/shubhamchandak94/VT_codes/
- Plan to integrate these with random access and repeated reading.

Future work

- Possibilities for improvement:
  - Optimized LDPC codes, e.g., using protographs
  - Better codes for insertion/deletion: LDPC with markers, VT codes
  - Check out q-ary VT codes implementation: https://github.com/shubhamchandak94/VT_codes/
- Plan to integrate these with random access and repeated reading.
- Long term vision: Nanopore sequencing + cheaper and noisier synthesis techniques

Team and funding

Shubham Chandak

Kedar Tatwawadi

Joachim Neu

Jay Mardia

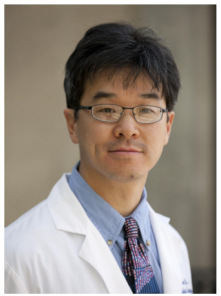
Billy Lau

Matt Kubit

Peter Griffin

Tsachy Weissman

Mary Wootters



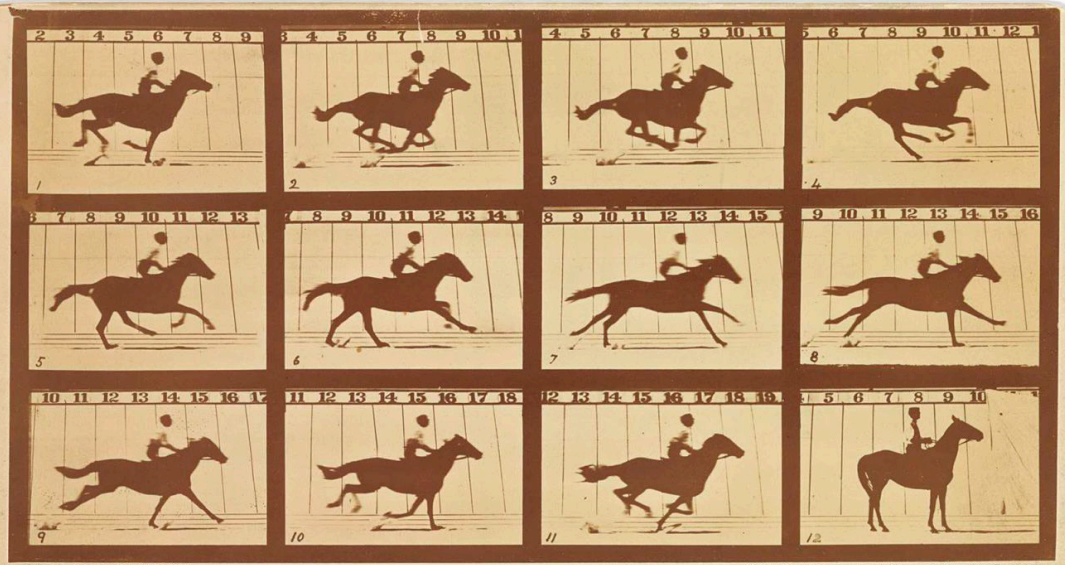
Hanlee Ji

SemiSynBio: Highly scalable random access DNA data storage with nanopore-based reading

Beckman Center Innovative Technology Seed Grant

Scalable Long-Term DNA Storage with Error Correction and Random-Access Retrieval

National Institutes of Health



MORSE'S Gallery, 417 Montgomery St., San Francisco

THE HORSE IN MOTION.

Illustrated by

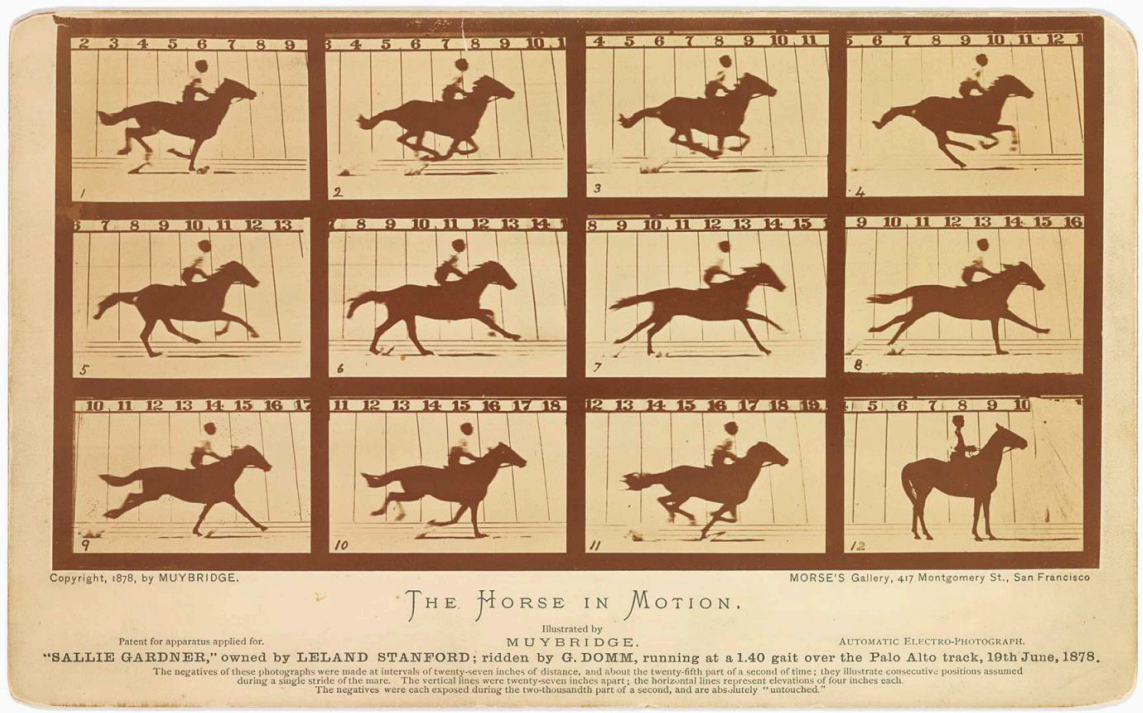
MUYBRIDGE.

Patent for apparatus applied for. AUTOMATIC ELECTRO-PHOTOGRAPH.

"SALLIE GARDNER," owned by LELAND STANFORD; ridden by G. DOMM, running at a 1.40 gait over the Palo Alto track, 19th June, 1878.

The negatives of these photographs were made at intervals of twenty-seven inches of distance, and about the twenty-fifth part of a second of time; they illustrate consecutive positions assumed during a single stride of the mare. The vertical lines were twenty-seven inches apart; the horizontal lines represent elevations of four inches each.
The negatives were each exposed during the two-thousandth part of a second, and are absolutely "untouched."

Thank You!



Biorxiv: https://www.biorxiv.org/content/10.1101/770032v1

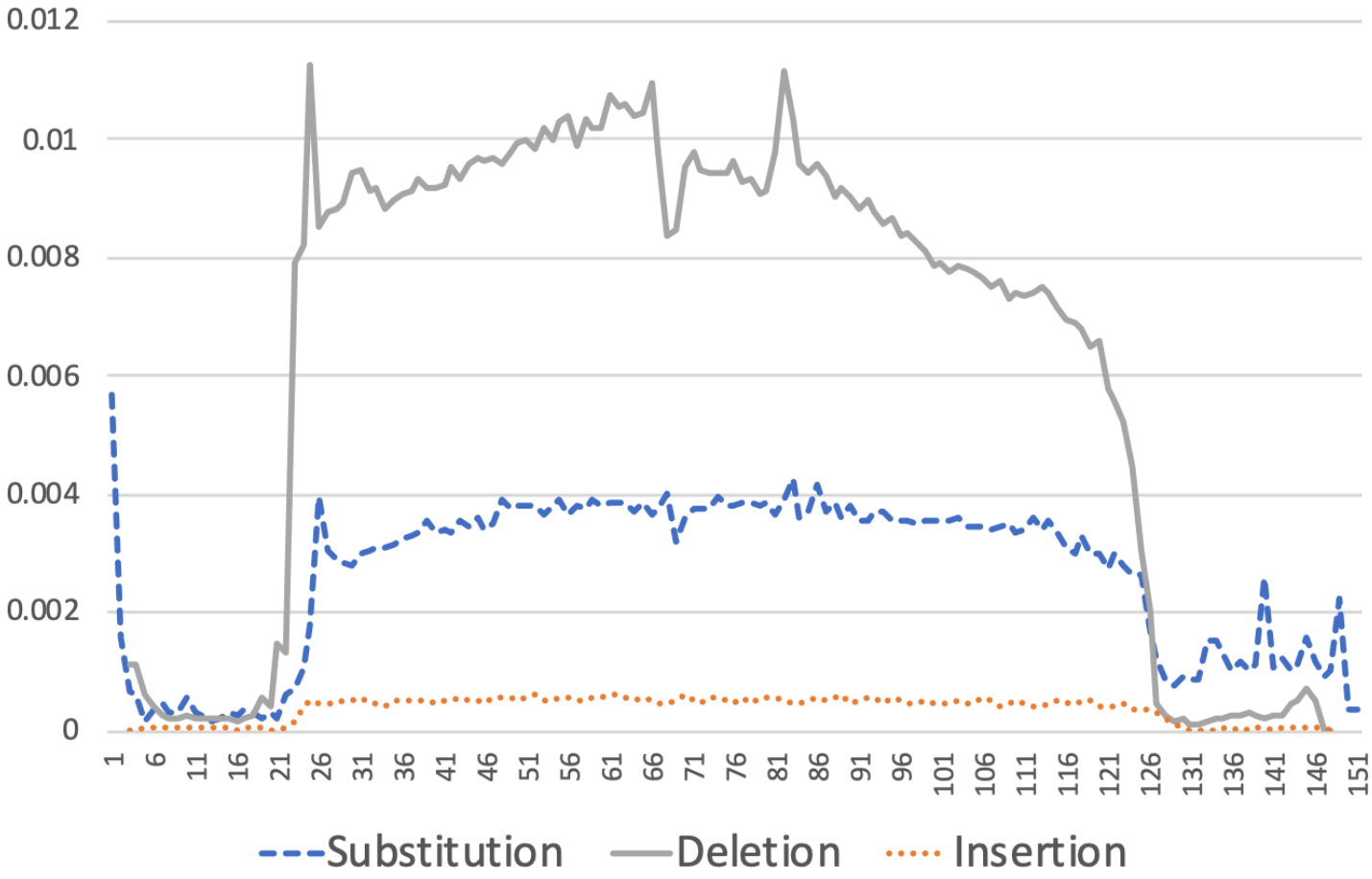
Backup

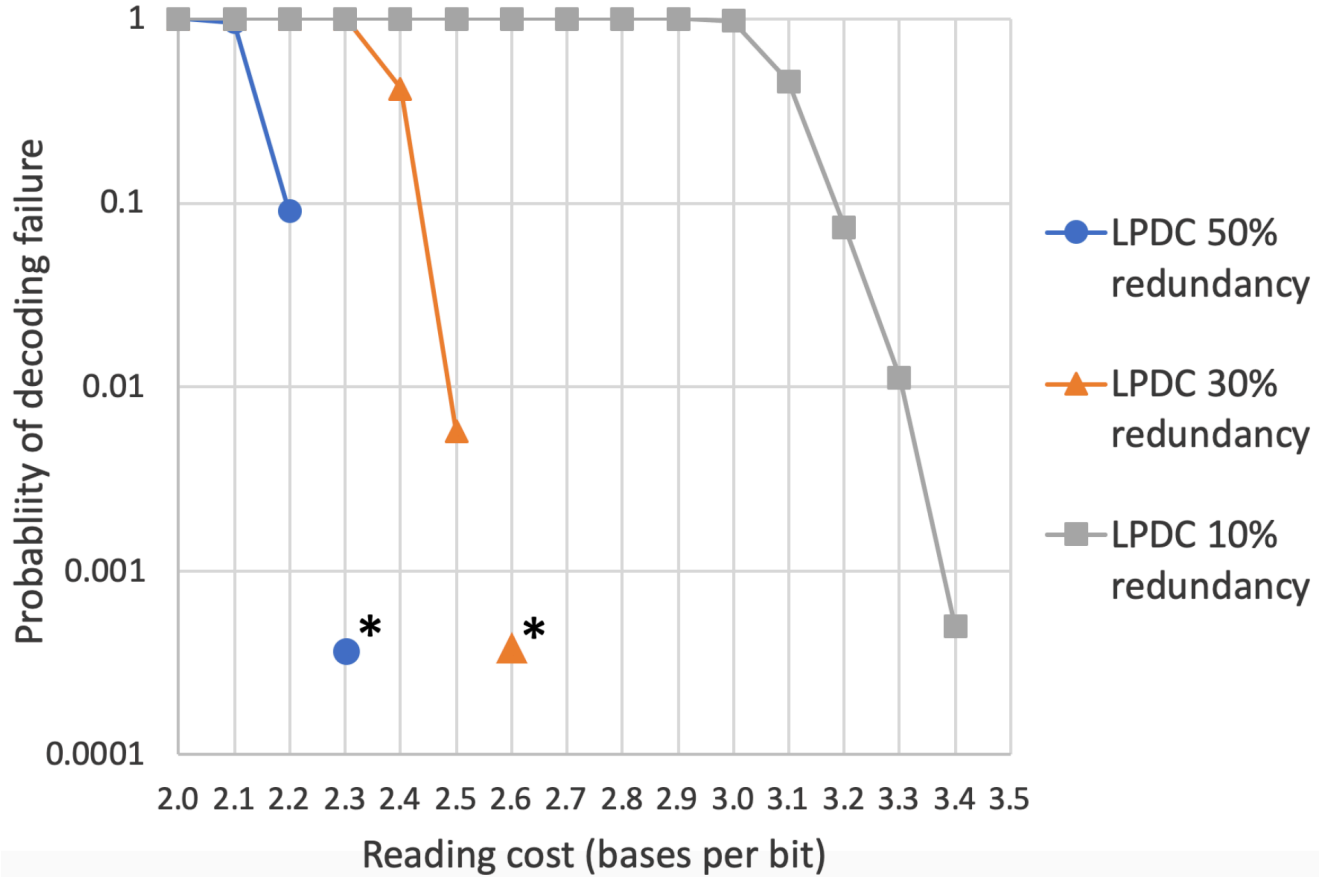
We first compute the optimal tradeoff between c_w and c_r when $\epsilon = 0$, i.e., the reads are error-free. In this case, for large enough n, we can use the Poisson(λ) approximation for the number of times each sequence is observed with $\lambda = c_r/c_w$. Since the probability of seeing zero copies of a sequence is $e^{-\lambda}$, this gives us an erasure channel with capacity $1 - e^{-\lambda}$ [20]. For reliable recovery, we need that the rate $1/c_w$ be less than the capacity. This gives us

$$c_r \geq c_w \log_e \frac{c_w}{c_w - 1}$$

$$P((k_0, k_1) \mid 0) = \frac{e^{-\lambda}\lambda^{k_0+k_1}}{(k_0+k_1)!}\binom{k_0+k_1}{k_0}(1-\epsilon)^{k_0}\epsilon^{k_1}$$

$$LLR(k_0, k_1) = \ln \frac{P((k_0, k_1) \mid 0)}{P((k_0, k_1) \mid 1)} = (k_0 - k_1) \ln \frac{1-\epsilon}{\epsilon}$$


0.012
0.01
0.008
0.006
0.004
0.002
0
1
6
11
16
21
26
31
36
41
46
51
56
61
66
71
76
81
86
91
96
101
106
111
116
121
126
131
136
141
146
151
Substitution
Deletion
Insertion


1
0.1
0.01
0.001
0.0001
Probablility of decoding failure
2.0 2.1 2.2 2.3 2.4 2.5 2.6 2.7 2.8 2.9 3.0 3.1 3.2 3.3 3.4 3.5
Reading cost (bases per bit)
*
*
LPDC 50% redundancy
LPDC 30% redundancy
LPDC 10% redundancy