Stanford

Error Correcting Codes for DNA based Data Storage

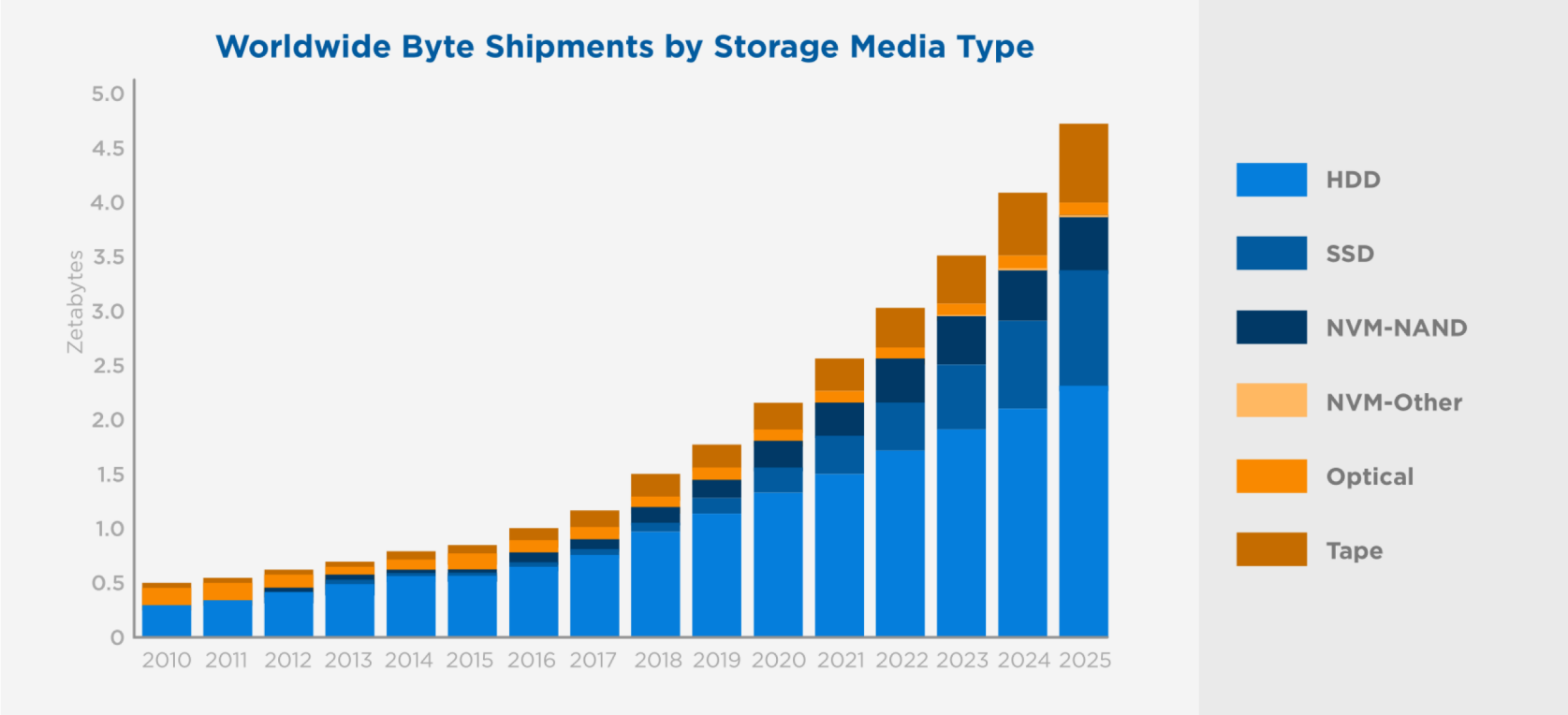
Shubham Chandak
Stanford University
ISMB/ECCB 2019

Outline

- Motivation
- DNA storage setup
- Illumina sequencing-based DNA storage
- Nanopore sequencing-based DNA storage
- 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

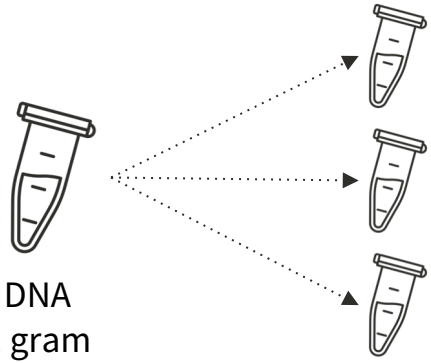
DNA
1 gram

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/

How to store data in DNA sequences?

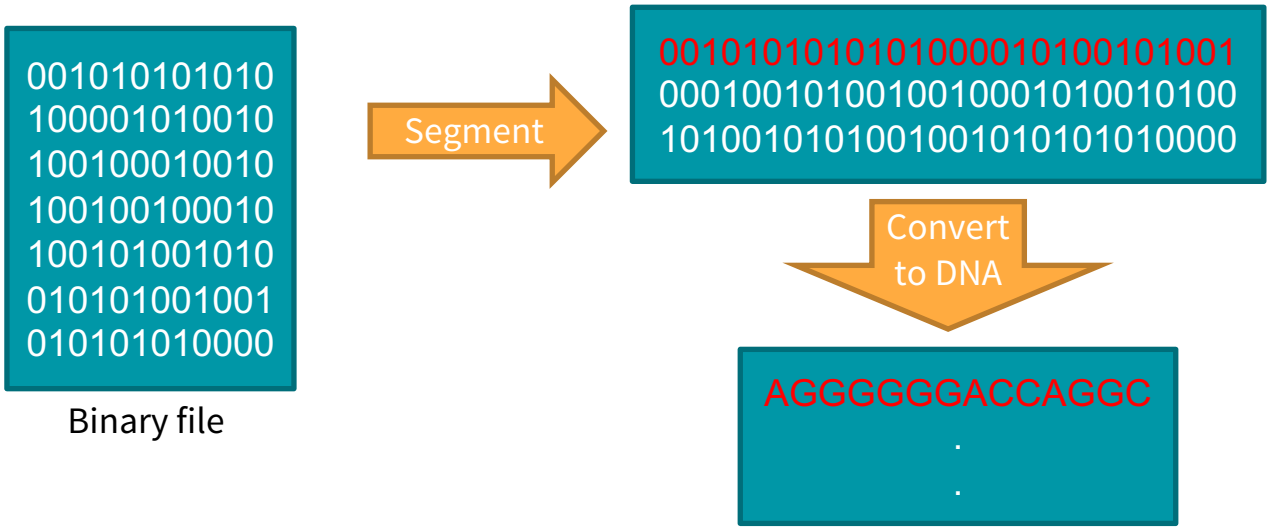
How to store data in DNA sequences?

- Ability to synthesize short ssDNA oligonucleotides (~150 nt) at scale.

CustomArray | GenScript Make Research Easy http://www.customarrayinc.com/

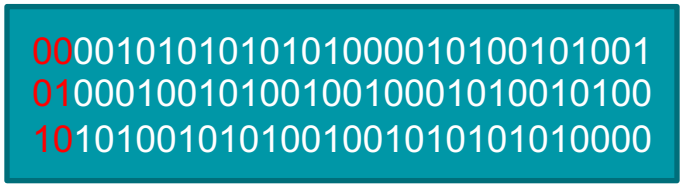
How to store data in DNA sequences?

- Ability to synthesize short ssDNA oligonucleotides (~150 nt) at scale.
- Convert binary data to A/C/G/T alphabet: e.g., 00 – A, 01 – C, etc.



How to store data in DNA sequences?

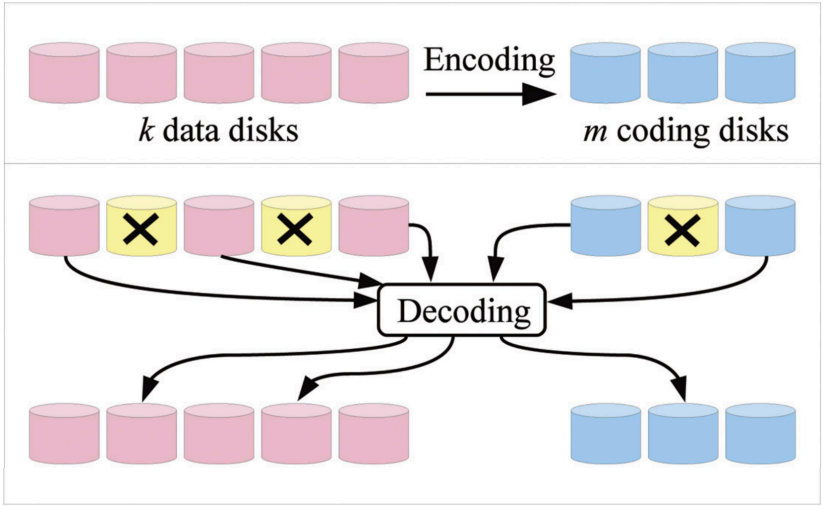
- Ability to synthesize short ssDNA oligonucleotides (~150 nt) at scale.
- Convert binary data to A/C/G/T alphabet: e.g., 00 – A, 01 – C, etc.
- But order of sequences lost in the solution – need to add index to each segment.



Length of index in binary segment at least $\log_2$(number of segments)

How to store data in DNA sequences?

- Ability to synthesize short ssDNA oligonucleotides (~150 nt) at scale.
- Convert binary data to A/C/G/T alphabet: e.g., 00 – A, 01 – C, etc.
- But order of sequences lost in the solution – need to add index to each segment.
- Some sequences have zero coverage while sequencing – erasure coding+coverage.



Also used in traditional storage systems (e.g., RAID)

Figure source: https://www.usenix.org/system/files/login/articles/10_plank-online.pdf

How to store data in DNA sequences?

- Ability to synthesize short ssDNA oligonucleotides (~150 nt) at scale.
- Convert binary data to A/C/G/T alphabet: e.g., 00 – A, 01 – C, etc.
- But order of sequences lost in the solution – need to add index to each segment.
- Some sequences have zero coverage while sequencing – erasure coding+coverage.
- Sequencing and synthesis cause errors – substitutions, insertions and deletions – error correction coding+coverage.

How to store data in DNA sequences?

- Ability to synthesize short ssDNA oligonucleotides (~150 nt) at scale.
- Convert binary data to A/C/G/T alphabet: e.g., 00 – A, 01 – C, etc.
- But order of sequences lost in the solution – need to add index to each segment.
- Some sequences have zero coverage while sequencing – erasure coding+coverage.
- Sequencing and synthesis cause errors – substitutions, insertions and deletions – error correction coding+coverage.
- Error correction studied extensively for communication and traditional data storage systems – information theory and coding theory.

How to store data in DNA sequences?

- Ability to synthesize short ssDNA oligonucleotides (~150 nt) at scale.
- Convert binary data to A/C/G/T alphabet: e.g., 00 – A, 01 – C, etc.
- But order of sequences lost in the solution – need to add index to each segment.
- Some sequences have zero coverage while sequencing – erasure coding+coverage.
- Sequencing and synthesis cause errors – substitutions, insertions and deletions – error correction coding+coverage.
- Error correction studied extensively for communication and traditional data storage systems – information theory and coding theory.

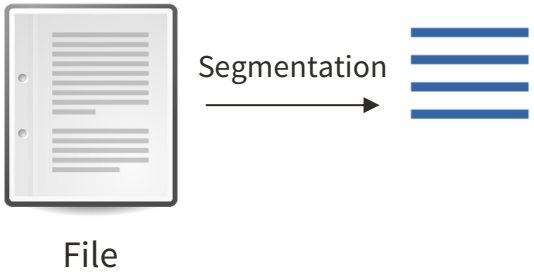
Error/Erasure Correcting Codes enable reliable data recovery even for noisy, low cost synthesis and sequencing – likely to be the future of DNA storage.

DNA storage setup

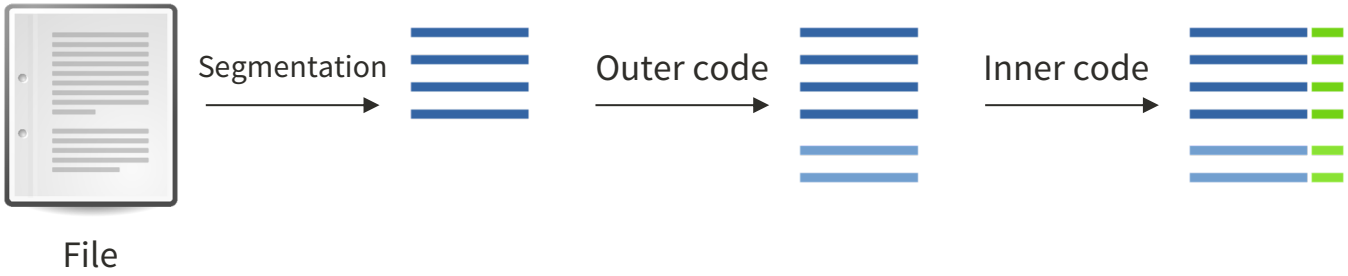
Typical DNA Storage System

File

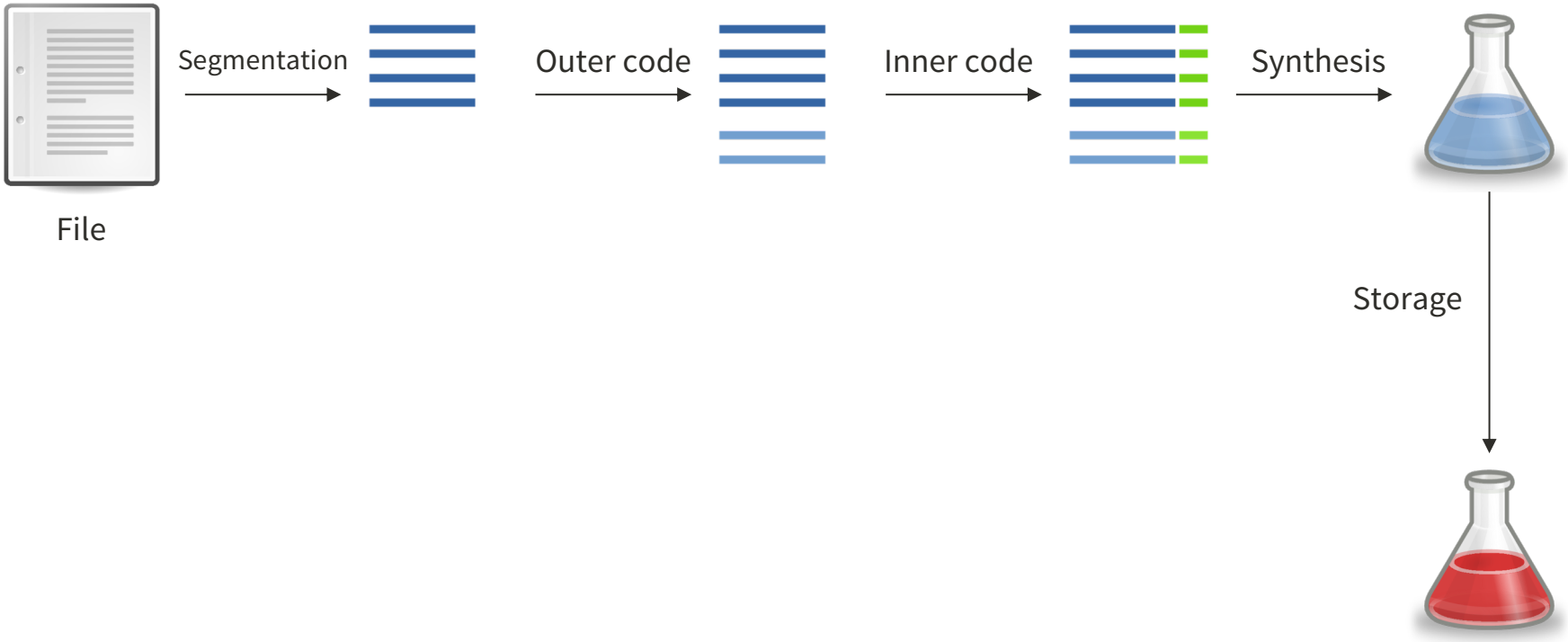
Typical DNA Storage System



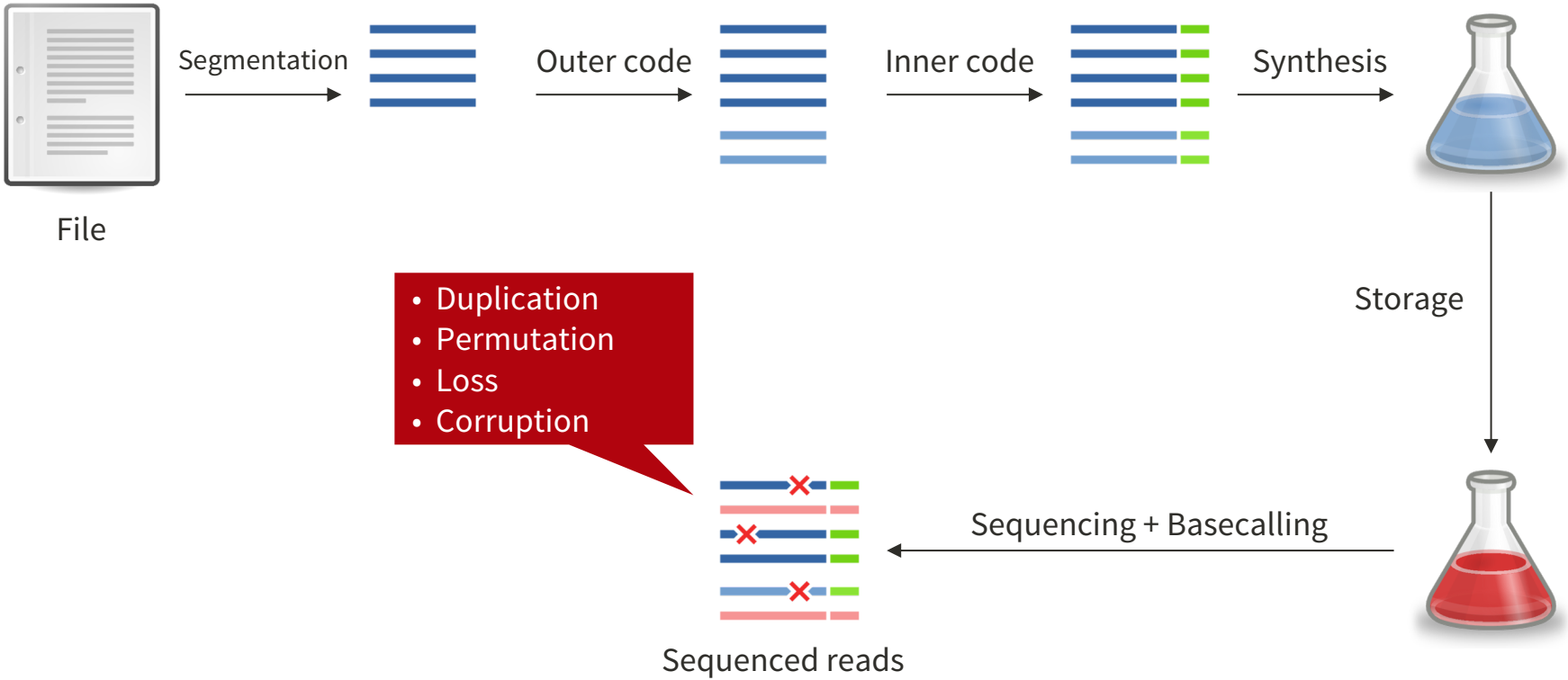
Typical DNA Storage System



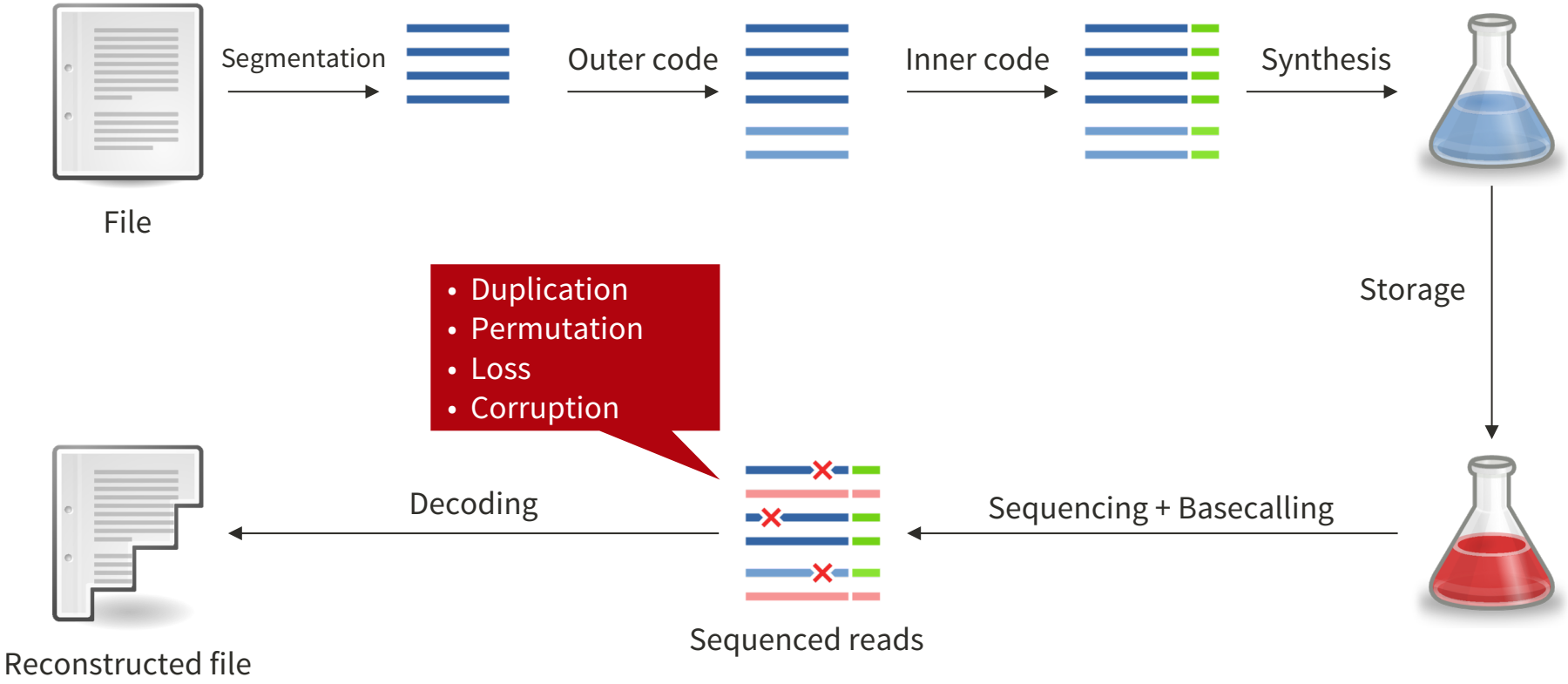
Typical DNA Storage System



Typical DNA Storage System



Typical DNA Storage System



2nd gen sequencing

Illumina sequencing

Portability	❌
Real-time	❌
Long reads	❌
Throughput	✅
Error rates	**< 1%**

{ mostly substitutions

3rd gen sequencing

Nanopore sequencing

Portability	✅
Real-time	✅
Long reads	✅
Throughput	❌
Error rates	**10 - 15%**

{ insertions, deletions, substitutions

Previous works

- Multiple previous works focusing on:
  - Error correction coding
  - Random access of subsets of 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.

Our contribution

- Fundamental quantities to evaluate a DNA storage system:
  - Writing cost (bases synthesized/message bit)
  - Reading cost (bases sequenced/message bit) (*not* coverage)

Our contribution

- Fundamental quantities to evaluate a DNA storage system:
  - Writing cost (bases synthesized/message bit)
  - Reading cost (bases sequenced/message bit) (*not* coverage)

- Study theoretical tradeoff between writing cost and reading cost.

Our contribution

- Fundamental quantities to evaluate a DNA storage system:
  - Writing cost (bases synthesized/message bit)
  - Reading cost (bases sequenced/message bit) (*not* coverage)

- Study theoretical tradeoff between writing cost and reading cost.
- Achieve better tradeoff by reducing reliance on high coverage.

Our contribution

- Fundamental quantities to evaluate a DNA storage system:
  - Writing cost (bases synthesized/message bit)
  - Reading cost (bases sequenced/message bit) (*not* coverage)

- Study theoretical tradeoff between writing cost and reading cost.
- Achieve better tradeoff by reducing reliance on high coverage.
- Break inner-outer code separation which is theoretically suboptimal for short sequences.

Our contribution

- Fundamental quantities to evaluate a DNA storage system:
  - Writing cost (bases synthesized/message bit)
  - Reading cost (bases sequenced/message bit) (*not* coverage)

- Study theoretical tradeoff between writing cost and reading cost.
- Achieve better tradeoff by reducing reliance on high coverage.
- Break inner-outer code separation which is theoretically suboptimal for short sequences.
- Basecaller-decoder integration for nanopore to exploit additional information in raw current signal.

Illumina sequencing-based DNA storage

Key idea

Key idea

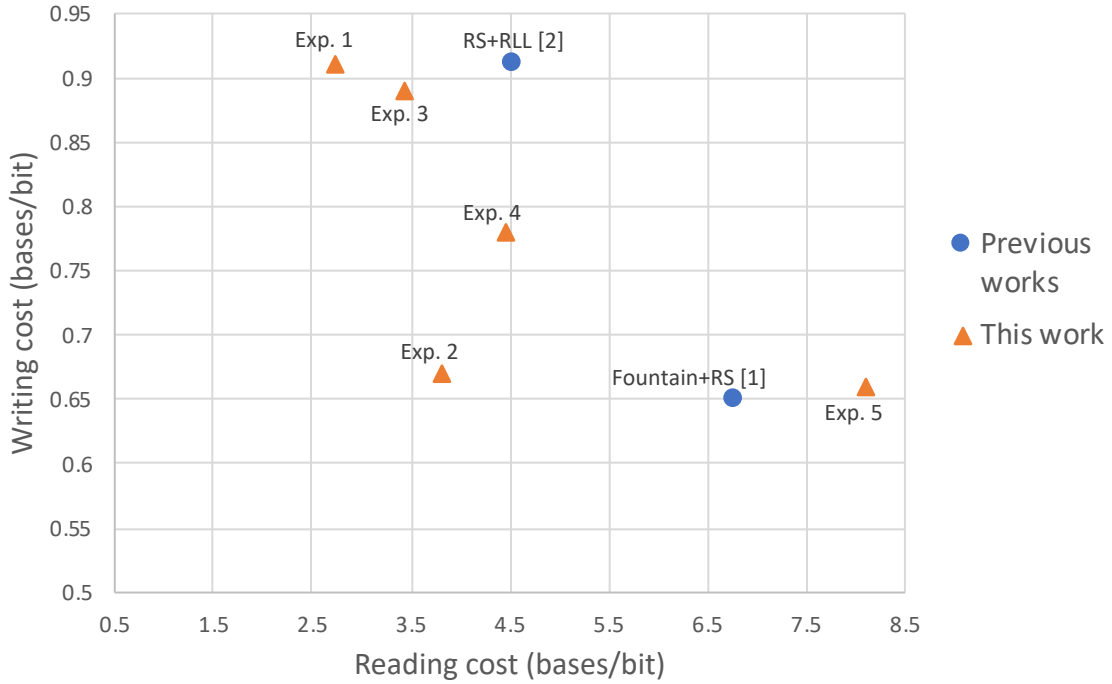
✗ Strategy 1: Inner/outer code separation

✓ Strategy 2: Single large block code (LDPC)

Experimental Results

- 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.
- Approach combines LDPC codes with heuristics for handling deletion errors.

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.

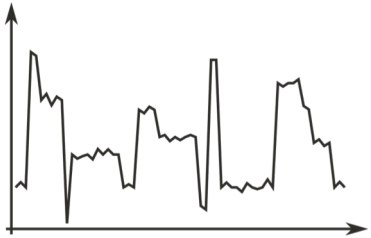
Nanopore sequencing-based DNA storage

Nanopore Sequencing Model

… ACGTACGTACGT …

Nanopore sequencing channel

- Memory (inter-symbol interference)
- Base skips
- Fading
- Random symbol duration
- Noise



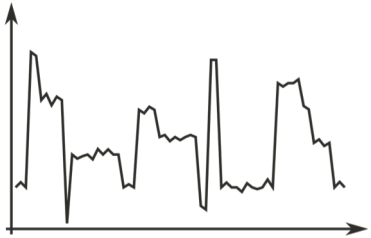
Source: "Models and Information-Theoretic Bounds for Nanopore Sequencing", Wei Mao et al., IEEE Trans. Inf. Theory 2017

Nanopore Sequencing Model

… ACGTACGTACGT …

Nanopore sequencing channel

- Memory (inter-symbol interference)
- Base skips
- Fading
- Random symbol duration
- Noise



VERY HARD TO MODEL AND ANALYZE FAITHFULLY

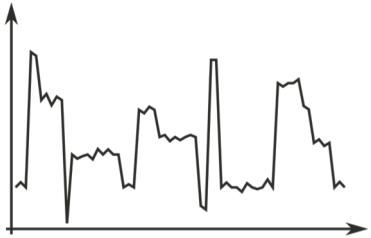
Source: "Models and Information-Theoretic Bounds for Nanopore Sequencing", Wei Mao et al., IEEE Trans. Inf. Theory 2017

Nanopore Sequencing Model

… ACGTACGTACGT …

Nanopore sequencing channel

- Memory (inter-symbol interference)
- Base skips
- Fading
- Random symbol duration
- Noise



VERY HARD TO MODEL AND ANALYZE FAITHFULLY

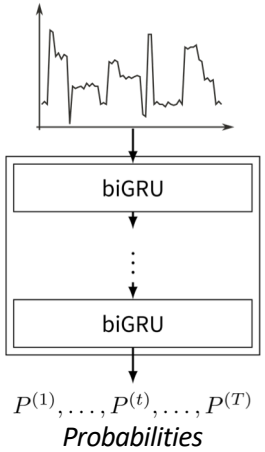
COMBINE STRENGTHS OF MACHINE LEARNING & CODING THEORY!

Source: "Models and Information-Theoretic Bounds for Nanopore Sequencing", Wei Mao et al., IEEE Trans. Inf. Theory 2017

Our approach

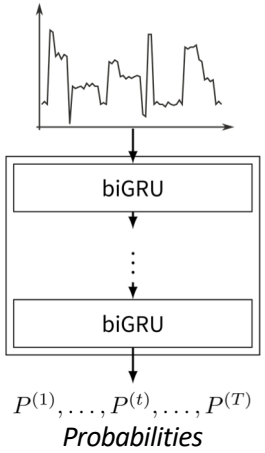
Our approach

Using Flappie basecaller (Oxford Nanopore)



Our approach

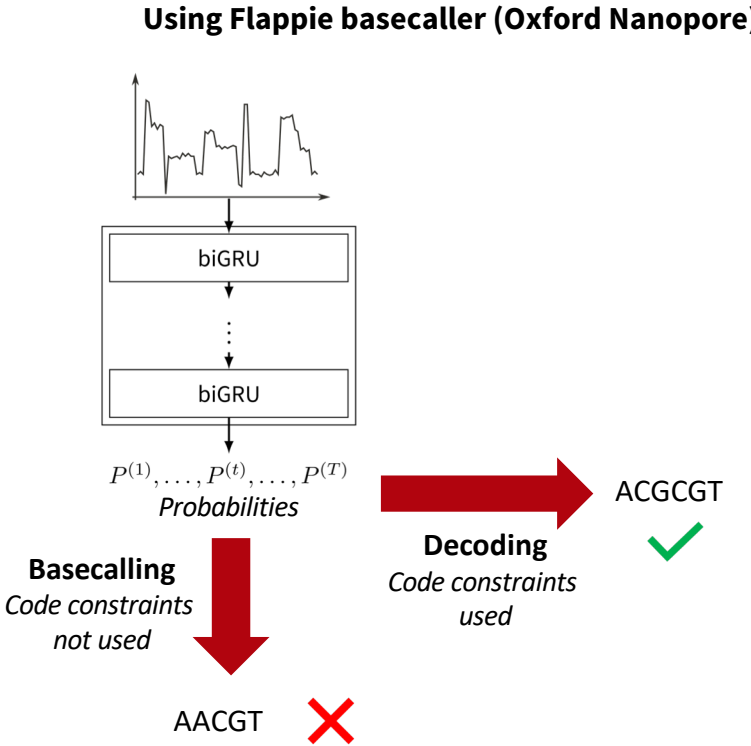
Using Flappie basecaller (Oxford Nanopore)



Basecalling
Code constraints not used

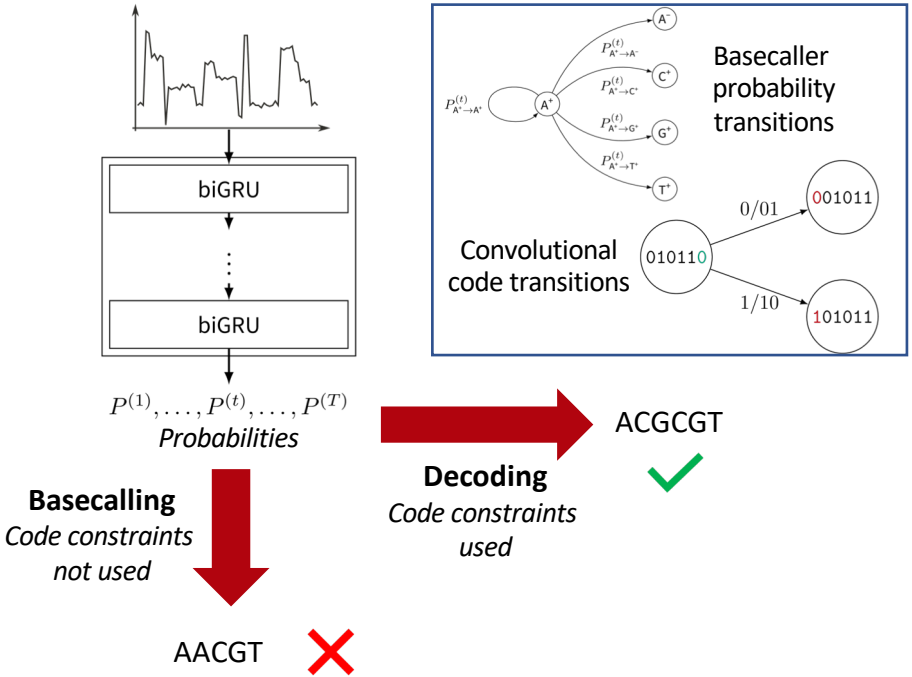
AACGT ✗

Our approach



Our approach

Using Flappie basecaller (Oxford Nanopore)



Preliminary Results

- Around **3x-6x** lower reading costs than state-of-the-art [1].

1. L. Organick *et al.*, “Random access in large-scale DNA data storage," *Nature biotechnology*, vol. 36, no. 3, p. 242, 2018.

Preliminary Results

- Around **3x-6x** lower reading costs than state-of-the-art [1].
- Significant fraction of sequences decoded from single read - theoretically impossible using basecalled sequence with 10-15% error.

1. L. Organick *et al.*, “Random access in large-scale DNA data storage," *Nature biotechnology*, vol. 36, no. 3, p. 242, 2018.

Preliminary Results

- Around **3x-6x** lower reading costs than state-of-the-art [1].
- Significant fraction of sequences decoded from single read - theoretically impossible using basecalled sequence with 10-15% error.
- Suggests that **raw signal** carries much **more information** than **basecalled sequence** - this can help other bioinformatics applications as well.

1. L. Organick *et al.*, “Random access in large-scale DNA data storage," *Nature biotechnology*, vol. 36, no. 3, p. 242, 2018.

Conclusions and future work

- Introduced novel coding schemes for both Illumina and nanopore based storage.

Conclusions and future work

- Introduced novel coding schemes for both Illumina and nanopore based storage.
- Plan to integrate these with random access and repeated reading.

Conclusions and future work

- Introduced novel coding schemes for both Illumina and nanopore based storage.
- Plan to integrate these with random access and repeated reading.
- Long term vision: Nanopore sequencing + cheaper and noisier synthesis techniques:
    - Basecaller-decoder integration works with various synthesis strategies, e.g., k-mer by k-mer

Conclusions and future work

- Introduced novel coding schemes for both Illumina and nanopore based storage.
- Plan to integrate these with random access and repeated reading.
- Long term vision: Nanopore sequencing + cheaper and noisier synthesis techniques:
    - Basecaller-decoder integration works with various synthesis strategies, e.g., k-mer by k-mer
- Core idea behind basecaller-decoder integration applicable beyond DNA storage:
    - Bioinformatics (soft-information based processing) - e.g., nanopolish
    - Communication (coding for complex and hard-to-model channels)

Team and funding

Shubham Chandak

Kedar Tatwawadi

Joachim Neu

Jay Mardia

Billy Lau

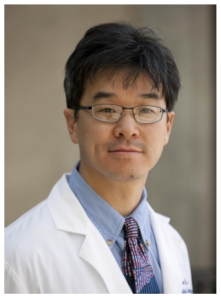
Matt Kubit

Dmitri Pavlichin

Peter Griffin

Tsachy Weissman

Mary Wootters



Hanlee Ji

Team and funding

Shubham Chandak

Kedar Tatwawadi

Joachim Neu

Jay Mardia

Billy Lau

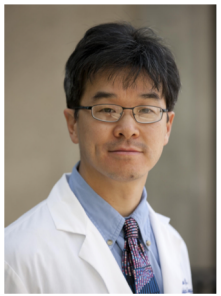
Matt Kubit

Dmitri Pavlichin

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

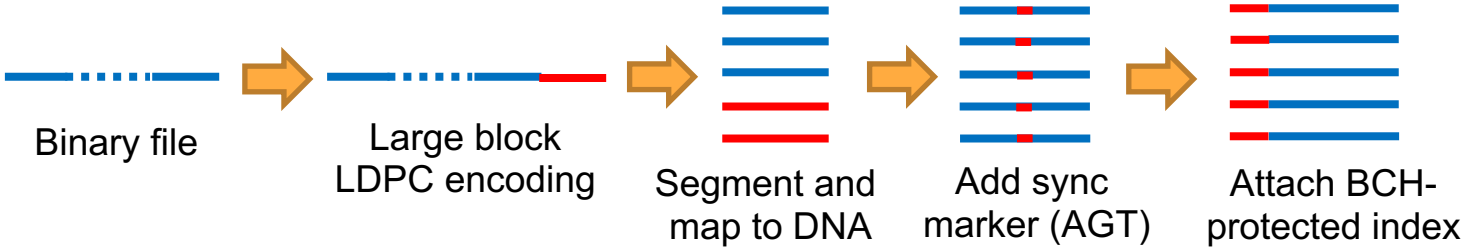
Scalable Long-Term DNA Storage with Error Correction and Random-Access Retrieval

National Institutes of Health

Thank You

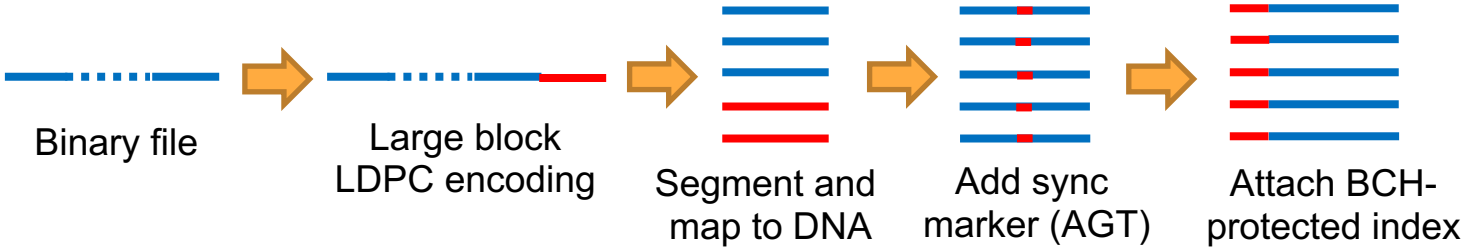
Poster session today 6pm-8pm: V-071

Proposed approach - schematics

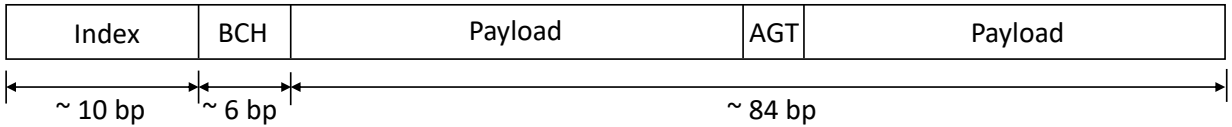


Encoding

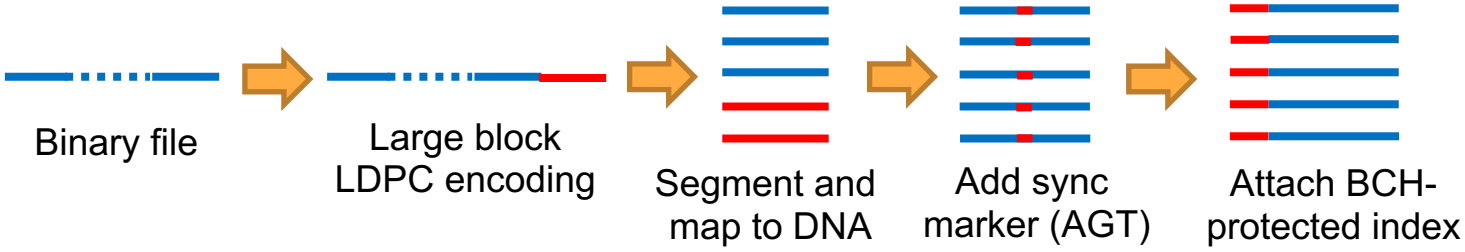
Proposed approach - schematics



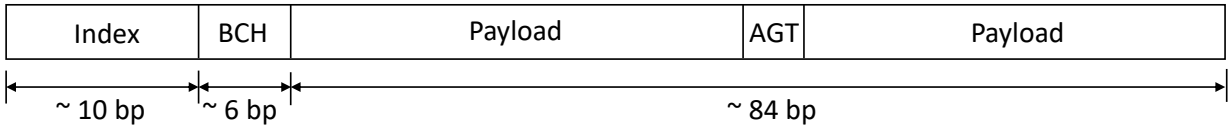
Encoding

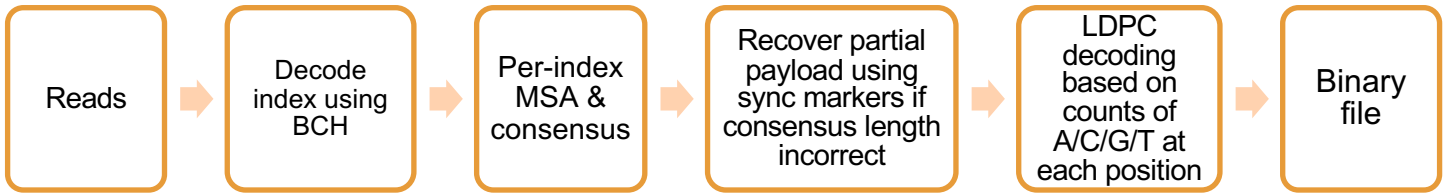


Proposed approach - schematics



Encoding

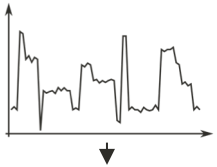




Decoding

Our approach

Stanford



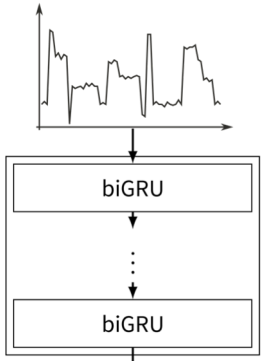
Deep neural network (DNN) basecaller (state-of-the-art)

Soft information

Viterbi convolutional decoder

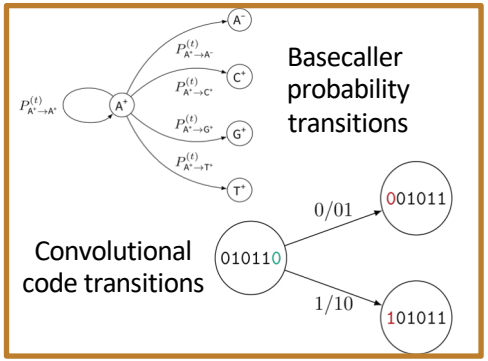
10111 …
10011 …
10101 …

Using Flappie basecaller (Oxford Nanopore)



$P^{(1)}, \ldots, P^{(t)}, \ldots, P^{(T)}$

Probabilities



Basecalling
Code constraints not used

AACGT ✗

Decoding
Code constraints used

ACGCGT ✓