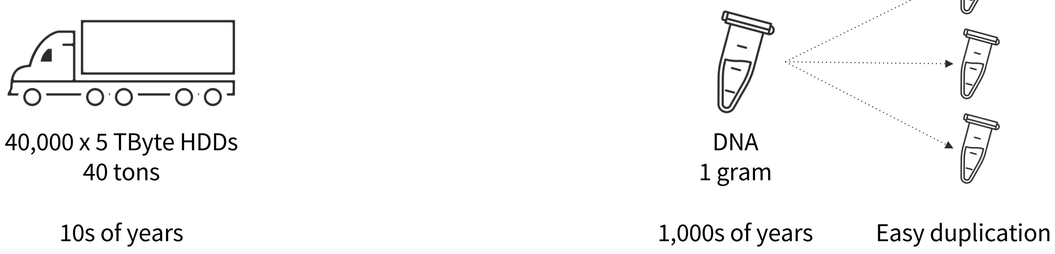
Error Correcting Codes for DNA based Data Storage

Shubham Chandak, Kedar Tatwawadi, Billy Lau, Jay Mardia, Matthew Kubit, Joachim Neu,
Peter Griffin, Mary Wootters, Tsachy Weissman, Hanlee Ji
Stanford University

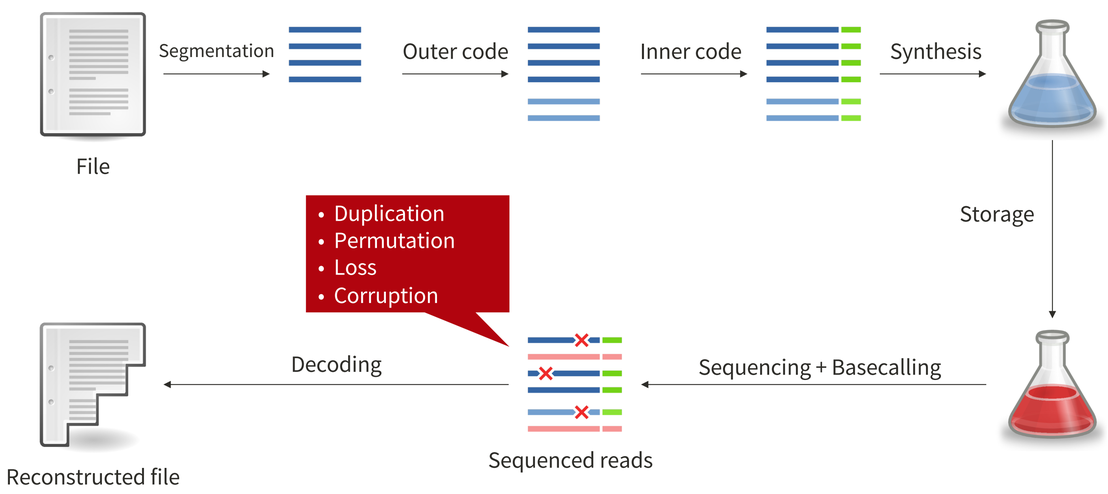
Motivation

- High storage densities (100s of Petabytes per gram).
- Long-term durability (1000s of years).
- Easy duplication.
- Random access capabilities.
- Storage medium of choice for life on Earth.
- Ideal as an archival medium to store the knowledge gained by humanity over the millennia.

200 Petabyte Data



Typical DNA Storage System



Error Correcting Codes enable reliable data recovery even for noisy, low cost synthesis and sequencing.

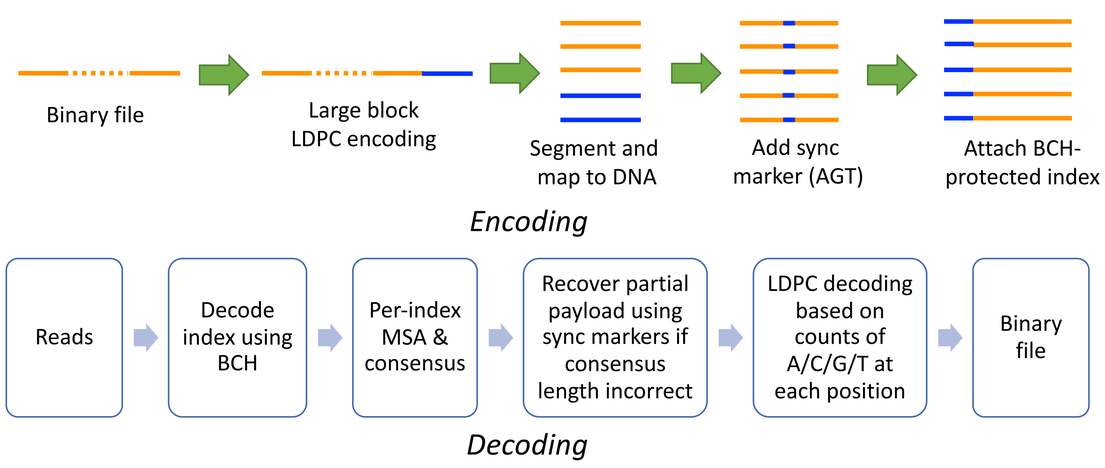
Our Contributions

Fundamental quantities to evaluate a DNA storage system:

- **Writing cost** (bases synthesized/message bit)
- **Reading cost** (bases sequenced/message bit) (*not* coverage)

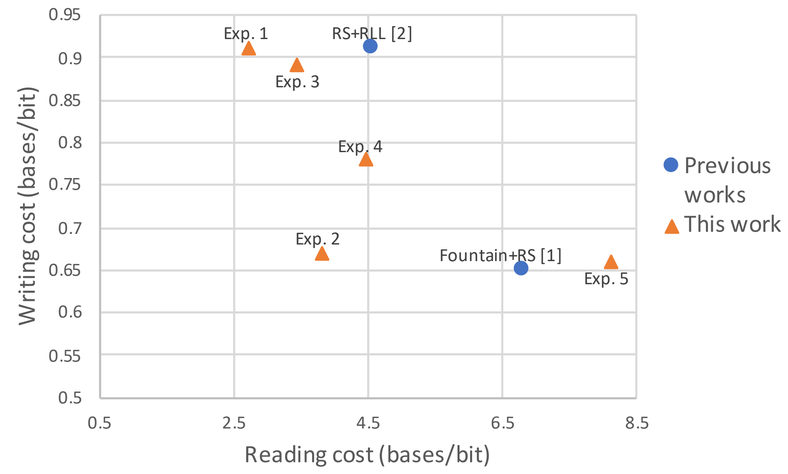
- Study theoretical tradeoff between writing cost and reading cost.
- Develop systems that yield better tradeoffs for both Illumina and Nanopore sequencing.
- Break inner-outer code separation which is theoretically suboptimal for short sequences.
- Basecaller-decoder integration for nanopore to exploit additional information in raw current signal.

Illumina Sequencing - Proposed Schematic



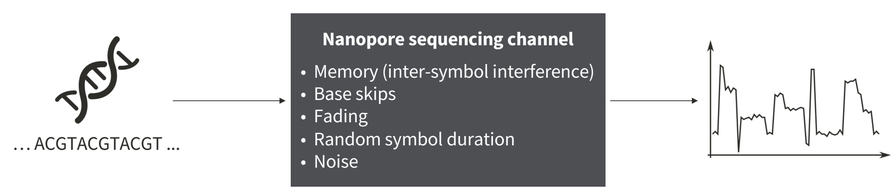
Experimental Results

- Performed experiment with different code parameters, storing around 200 KB data each.
- Oligonucleotide pools synthesized with CustomArray, length 150 including primers.
- Sequenced with Illumina iSeq.
- Total error rate around 1.3% (substitution: 0.4%, deletion: 0.85%, insertion: 0.05%).
- **Improved read/write cost tradeoff** than previous works despite **higher error rates and coverage variance** due to **cheaper synthesis**.



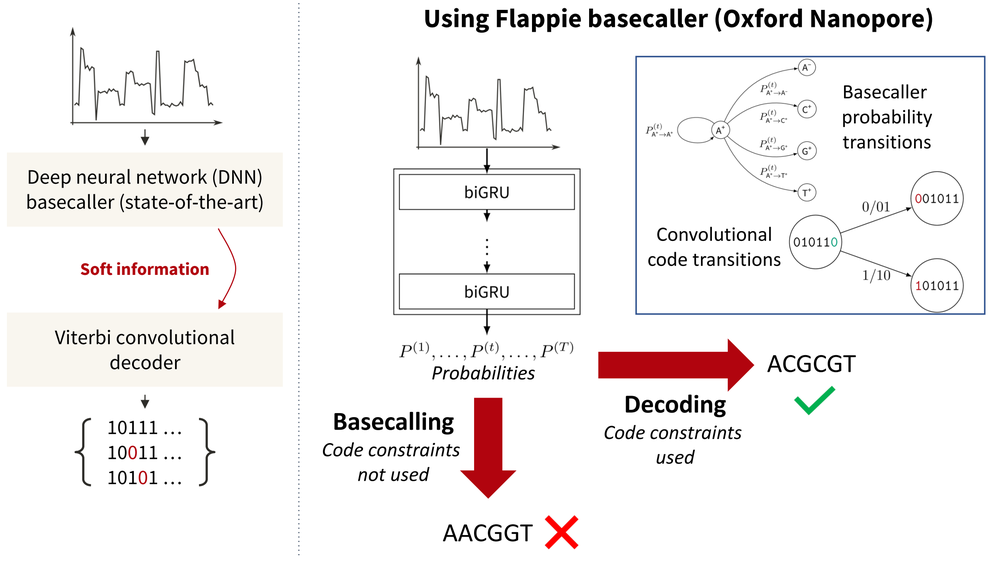
Nanopore Sequencing - Ongoing Work

- Very high error rates after basecalling: 10-15 % (mostly indels).
- Advantages include portability, real-time sequencing, long reads.
- Previous works use very high coverage - *suboptimal*.
- **Proposed approach** - use additional information present in raw current signal.



VERY HARD TO MODEL AND ANALYZE FAITHFULLY

COMBINE STRENGTHS OF MACHINE LEARNING & CODING THEORY!



Preliminary results:

- Around 3x-6x lower reading costs than [2].
- More than 50% 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.

Funding

- NSF/SRC (award number 1807371) under the SemiSynBio program.
- Beckman Technology Development Seed Grant.
- National Institutes of Health (NIH) grant NHGRI P01 HG000205.

References

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.