Random Access and Similarity Search in DNA Data Storage

Yuan-Jyue Chen[1], Callista Bee[2], Christopher N. Takahashi[2], Lee Organick[2], Siena Dumas Ang[1], David Ward[2], Xiaomeng Liu[2], Patrick Weiss[3], Bill Peck[3], Georg Seelig[2,4], Luis Ceze[2], Karin Strauss[1]

[1] Microsoft Research, Redmond, Washington, 98052

[2] Paul G. Allen School of Computer Science and Engineering, University of Washington, Seattle, Washington, 98195

[3] Twist Bioscience, San Francisco, California, 94158

[4] Electrical Engineering Department, University of Washington, Seattle, Washington, 98195

1 Project Goal

In order to **efficiently retrieve information** from DNA data storage, we developed two different molecular methods: **random access** and **similarity search**. Random access can retrieve individual files by their **unique identifiers**. Similarity search can retrieve data by their **contents**.

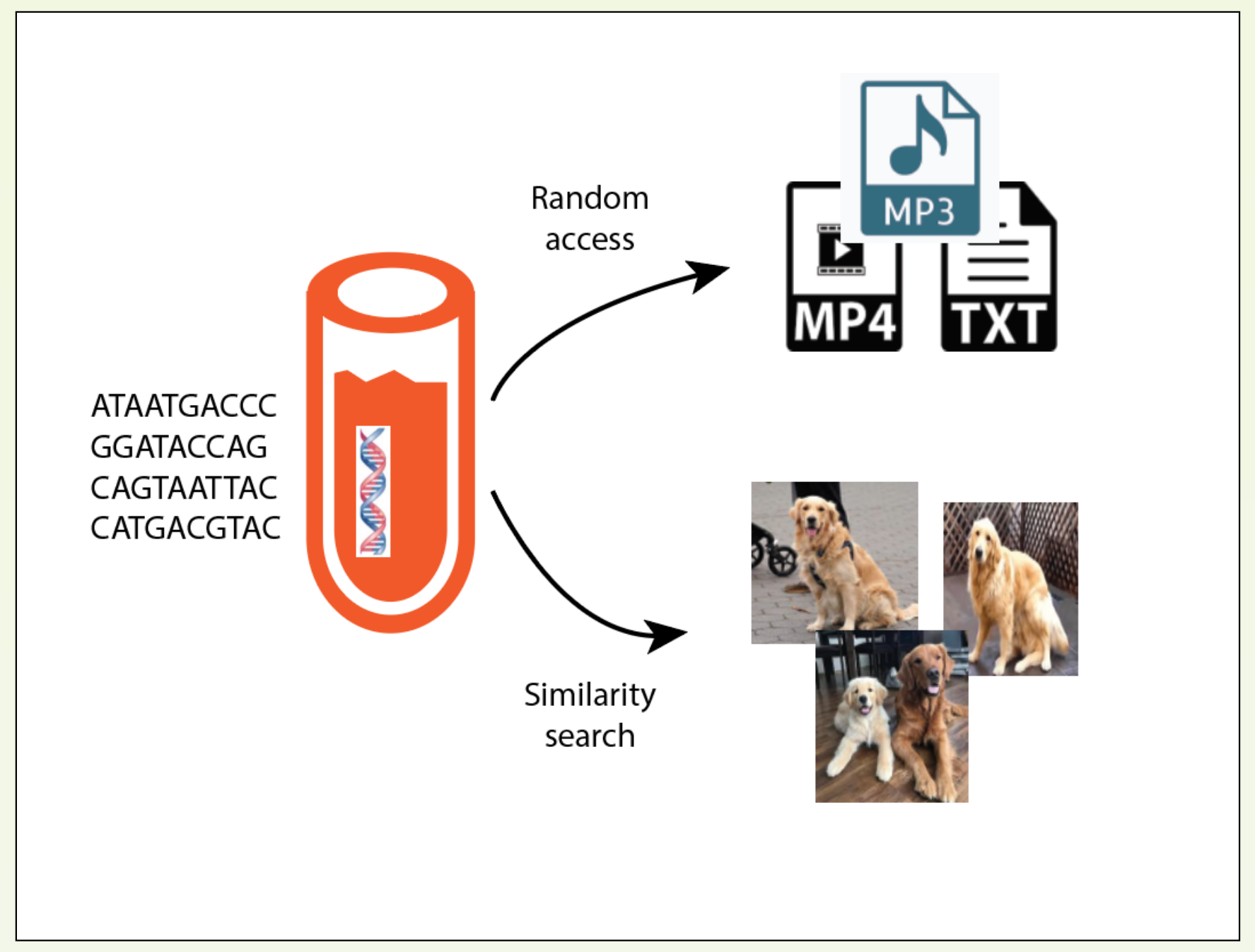

Figure 1. DNA database search. Digital data are stored in DNA. Random access retrieves individual files by their ID. Visually similar images can be retrieved through content-based similarity search.

2 Molecular Bias in Storage System with Random Access

- DNA storage systems showed significant **bias (uneven copy number distribution),** causing excessing number of missing sequences.
- **Identified two significant bias sources** in DNA storage: synthesis bias and PCR stochasticity.
- **Built the first process-wide model for storing data in DNA** that quantitatively shows how oligo copy distribution becomes more biased throughout the storage process.

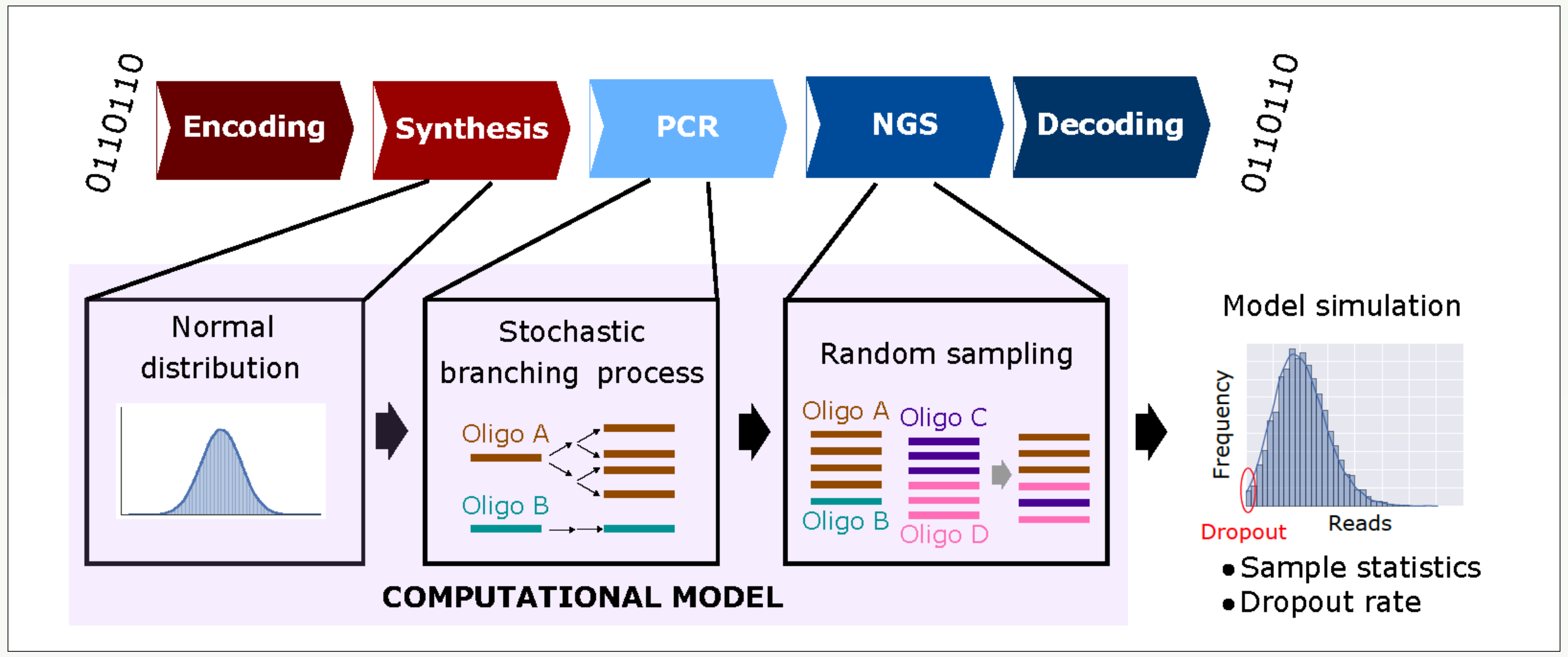

Figure 2. Quantitative Model. Digital data are **encoded** to DNA sequences; converted to DNA molecules using **array-based synthesis** technology; **PCR amplified;** read using **next generation sequencing (NGS)**. A computational model helps researchers understand trade-offs in architecting a DNA data storage system.

Lee et al, Nature Biotechnology (2018).

Chen et al, Nature Communications (2020).

3 Content-based Similarity Search

Computational process that converts similar images to similar DNA sequences (Fig. 3a).

- Images are extracted to high-dimensional features.
- Features are then converted to DNA sequences using a neural network.
- A neural network is optimized to minimize unwanted binding.

A **magnetic beads-based method** filters similar images of a query image in a DNA database (Fig. 3b).

Experimental results show higher read counts are associated with images closer to the query from a 1.6 million image DNA pool (Fig. 3c).

➢ Relevant images can be retrieved while conserving sequencing resources.

Callista Bee et al., bioRxiv (2020).

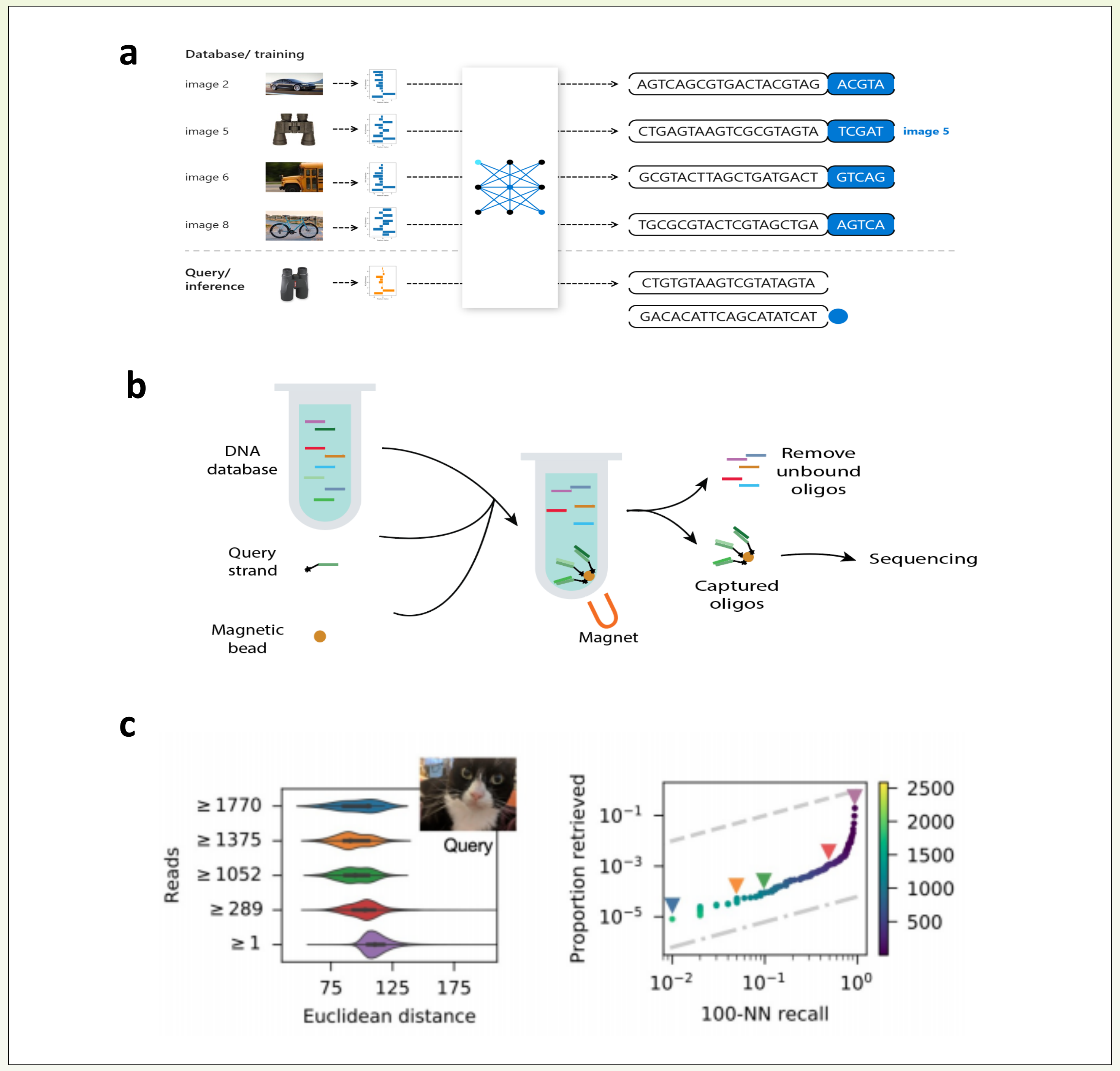

Figure 3. Content based similarity search. (**a**) Computational process to convert images to DNA sequences. (**b**) Experimental method to retrieve similar images. (**c**) Experimental data. Left: Euclidean distance of images against sequencing reads. Right: 100-neighest neighbor recall against proportion of images retrieved.

4 Summary

➢ **Random Access**. Model will help researchers rationally optimize both physical and sequencing redundancy for reliable data decoding, a **significant step towards engineering robust, efficient DNA storage systems.**

➢ **Similarity Search.** Learned encodings and the magnetic beads-based method enable content-based image similarity search from a database of 1.6 million images encoded in synthetic DNA, **an important demonstration of DNA computing in large-scale DNA storage systems.**

6 Acknowledgements

We thank Patrick Finn, Siyuan Chen, Andrew Stewart, Bernadette Arias, and Emily Leproust, from Twist Bioscience for supplying the array-synthesized DNA, helping improve our lab protocols, and offering input to our data analysis. This work was supported in part by a sponsored research agreement by Microsoft, NSF award CCF-1409831 to LC and GS, by NSF award CCF-1317653 to GS, and a DAPRA grant under the Molecular Informatics program.