SNIA DEVELOPER CONFERENCE

SDC 24

BY Developers FOR Developers

September 16-18, 2024
Santa Clara, CA

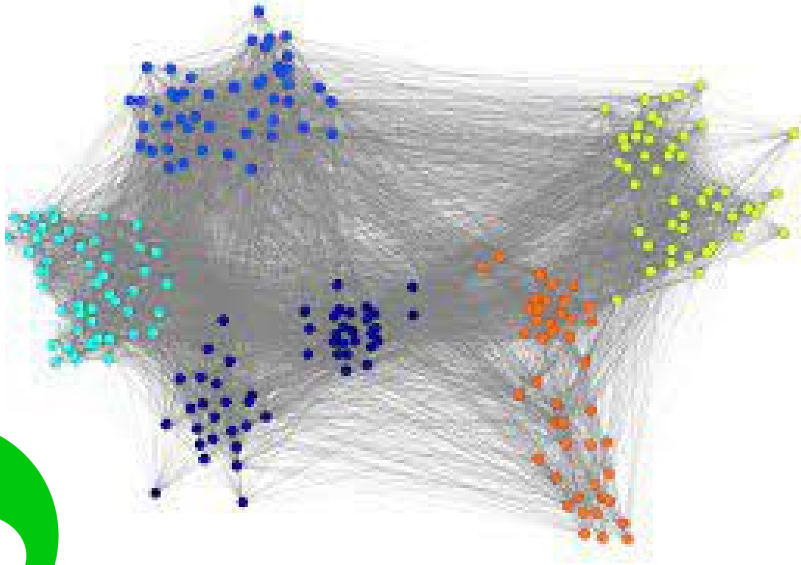
An Encoding Scheme to Enlarge Practical DNA Storage Capacity by Reducing Primer-Payload Collisions

Bingzhe Li[^], Yixun Wei[*], and David Du
[^]University of Texas at Dallas[*]
[*]University of Minnesota, Twin Cities

Outlines

- Introduction and motivation
- Background
- CAC algorithm
- Experimental result
- Conclusion

Big Data

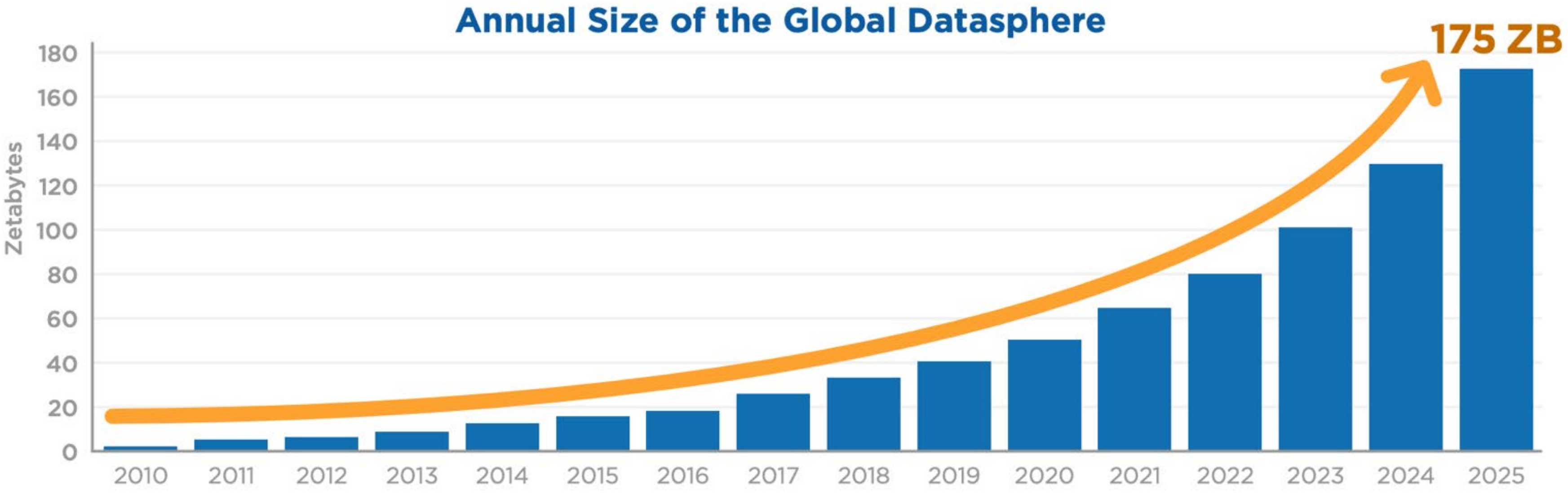


Big Data Era

Data is **doubled** almost every **2 years**

44 Zettabytes in 2020

175 Zettabytes in 2025

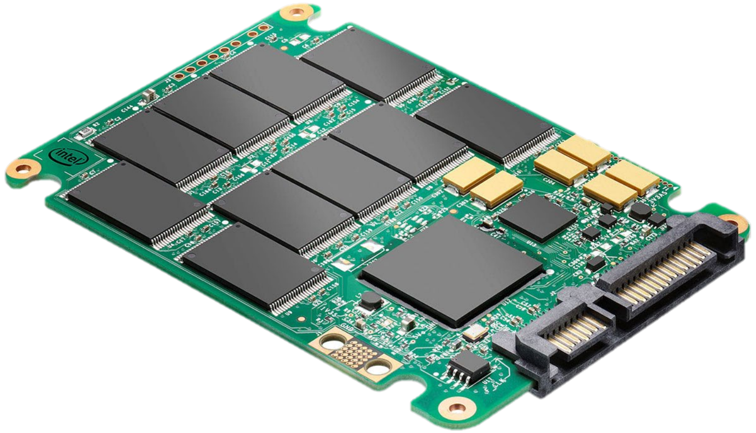


Annual Size of the Global Datasphere

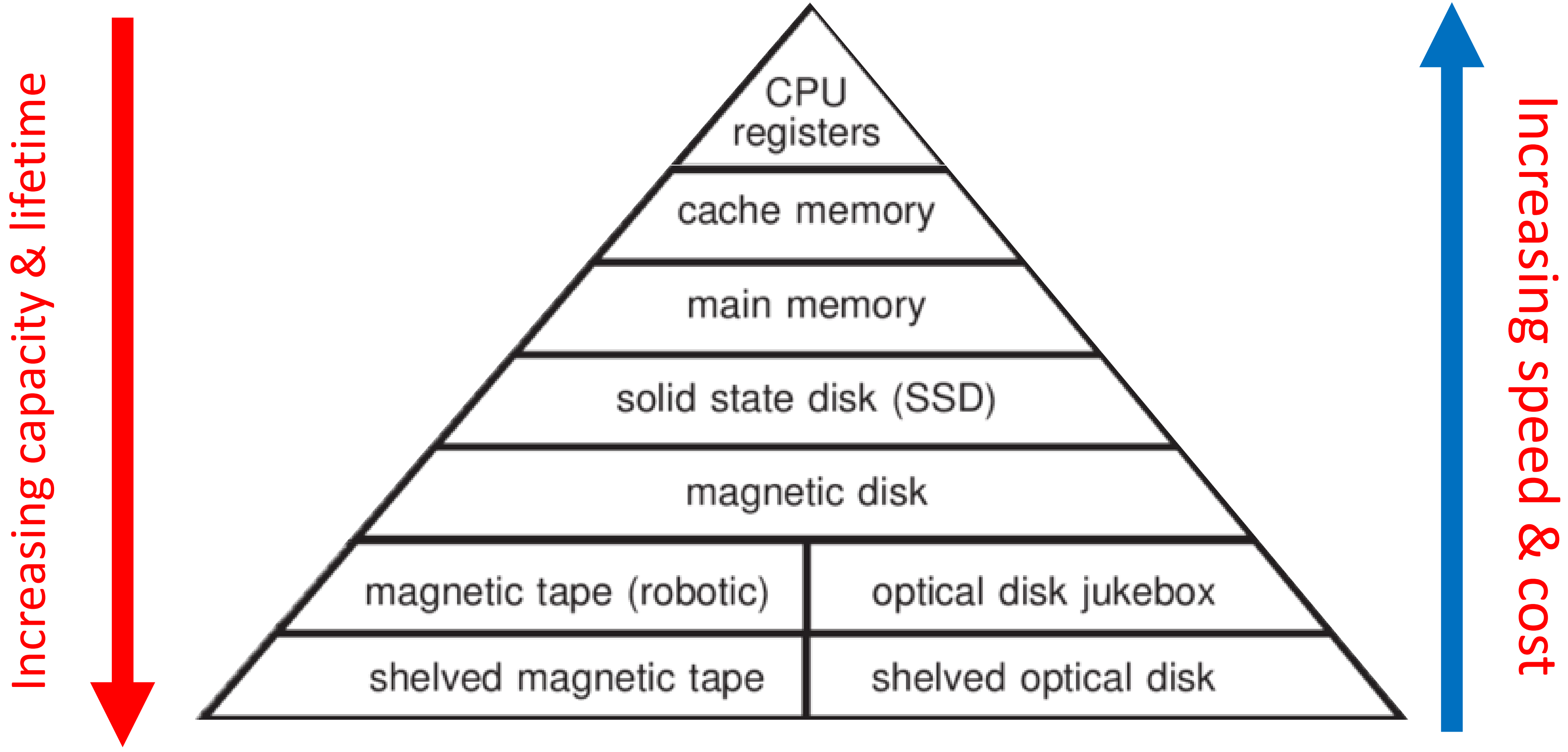
Source: Data Age 2025, sponsored by Seagate with data from IDC Global DataSphere, Nov 2018

Image from: https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf

Storing Digital Data

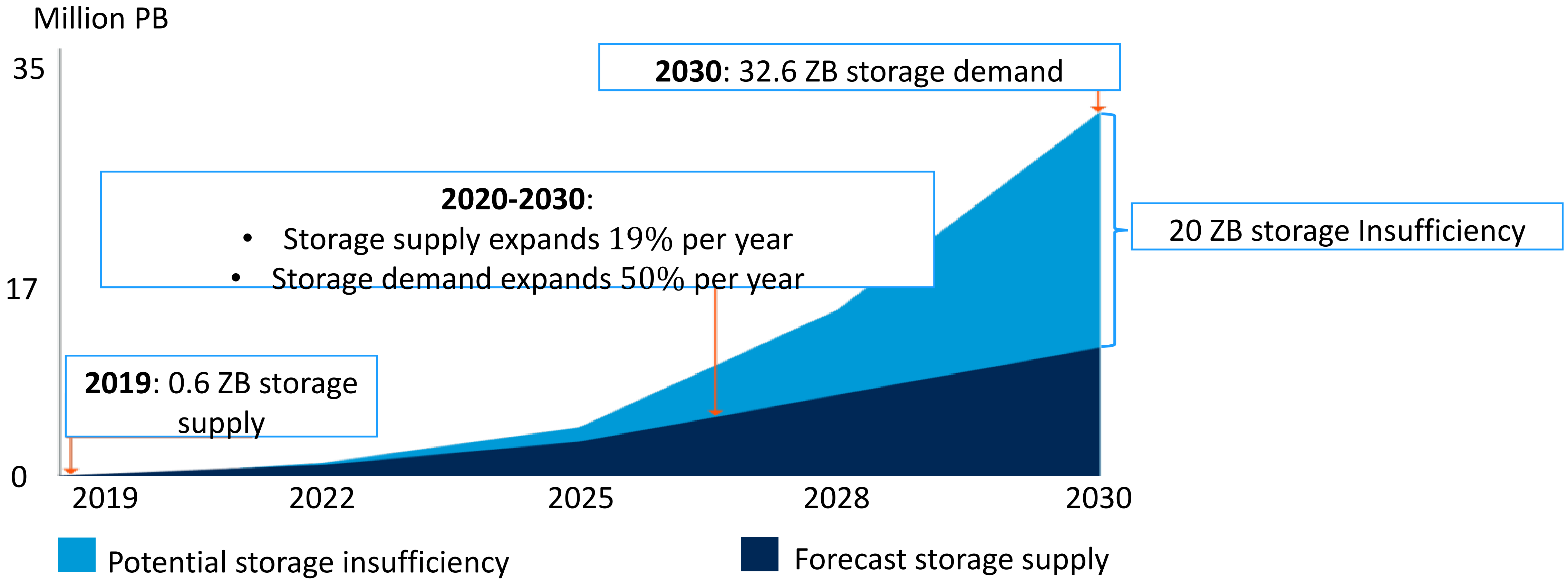


Storage Devices



Challenges of Existing Archival Storage #1

- **The supply of storage is not keeping pace with the increasing demand**[1]



2030: 32.6 ZB storage demand

2020-2030:

- Storage supply expands 19% per year
- Storage demand expands 50% per year

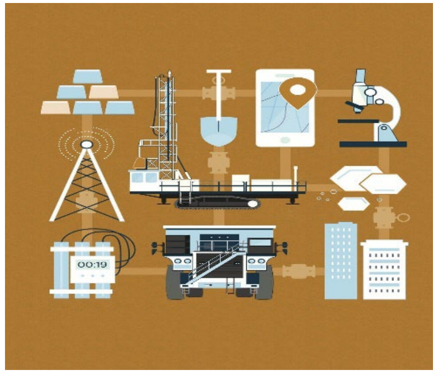
20 ZB storage Insufficiency

2019: 0.6 ZB storage supply

[1] IDC, Worldwide Global StorageSphere Forecast, 2021–2025: To Save or Not to Save Data, That Is the Question, IDC Doc #US47509621, March 2021

Challenges of Existing Archival Storage #2

- **Existing storage media cannot preserve data long enough**



Video streaming, Smart cities, Healthcare, Scientific discovery……

In order for future data mining, preserve data **in a long duration** →

Typical Storage Media & Lifespan

magnetic tape
- up to 15 years

hard disk
- 3-5 years

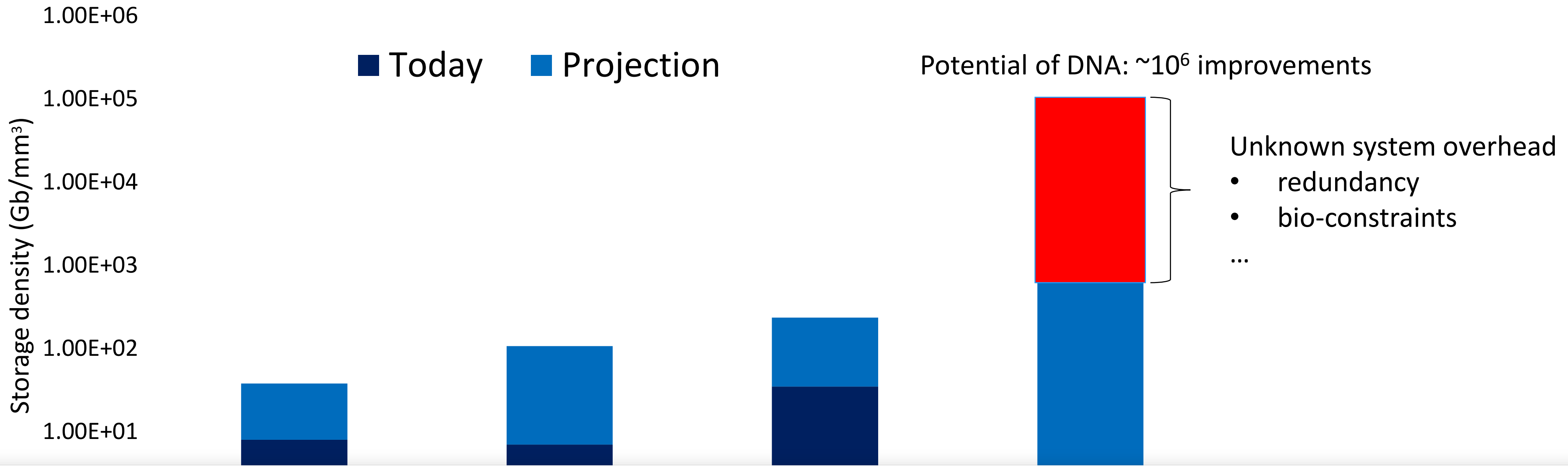
flash SSD
- 5-10 years (depends on write cycles)

Data has to be migrated from obsolete worn-out devices to new devices every a few years --> huge maintenance cost

DNA as Archival Storage Media

- **DNA's long durability saves enormous maintenance cost**
- **DNA's ultra-high storage density can potentially meet people's storage demand**

Storge density comparison – HDD, Flash, Tape, and DNA[2]



Potential of DNA: ~10^6 improvements

Unknown system overhead

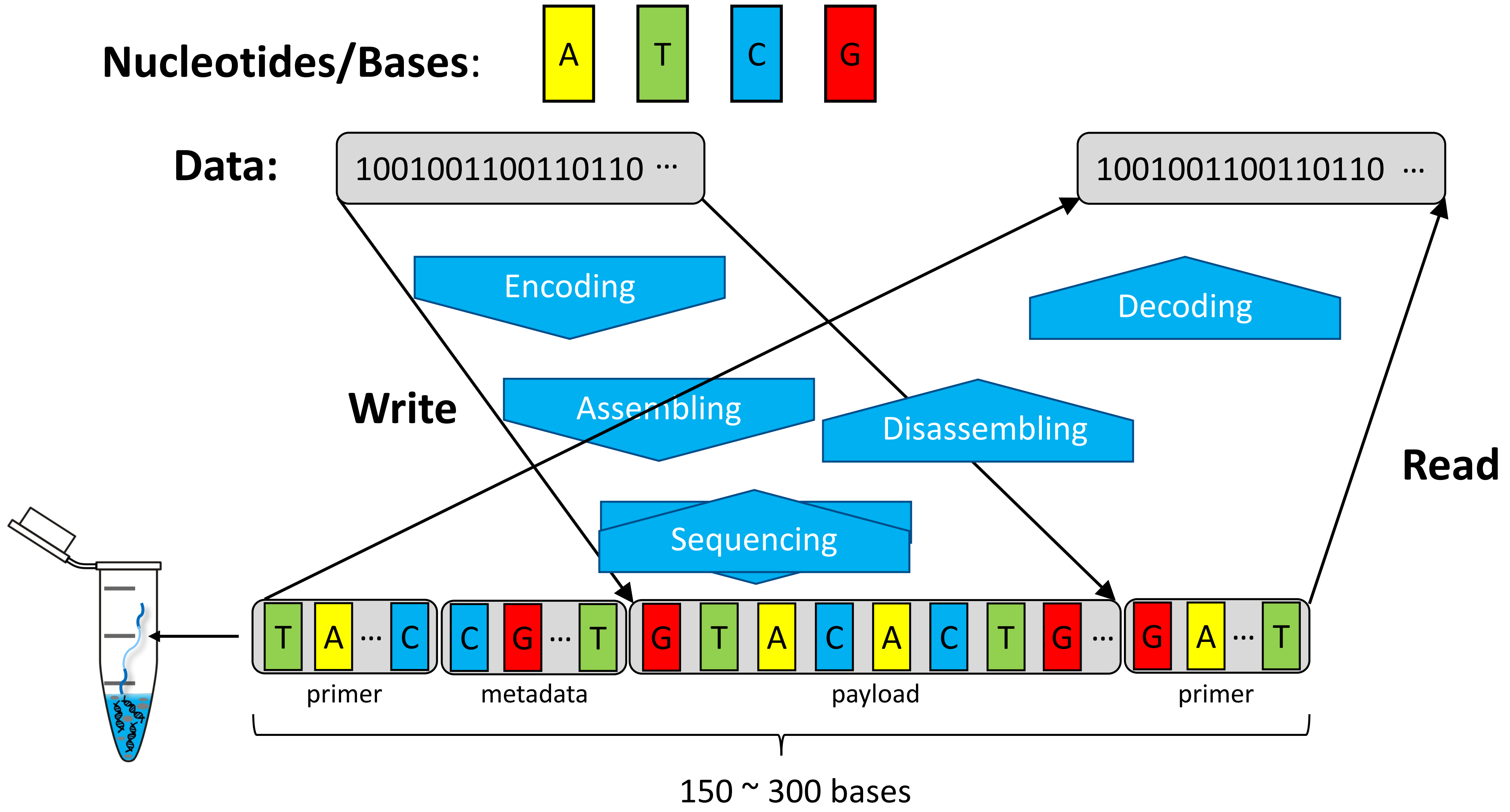
- redundancy
- bio-constraints

…

Explore the practical DNA storage capacity: the challenge and the solution

[2] DNA Data Storage Alliance https://dnastoragealliance.org/

What is DNA Storage?



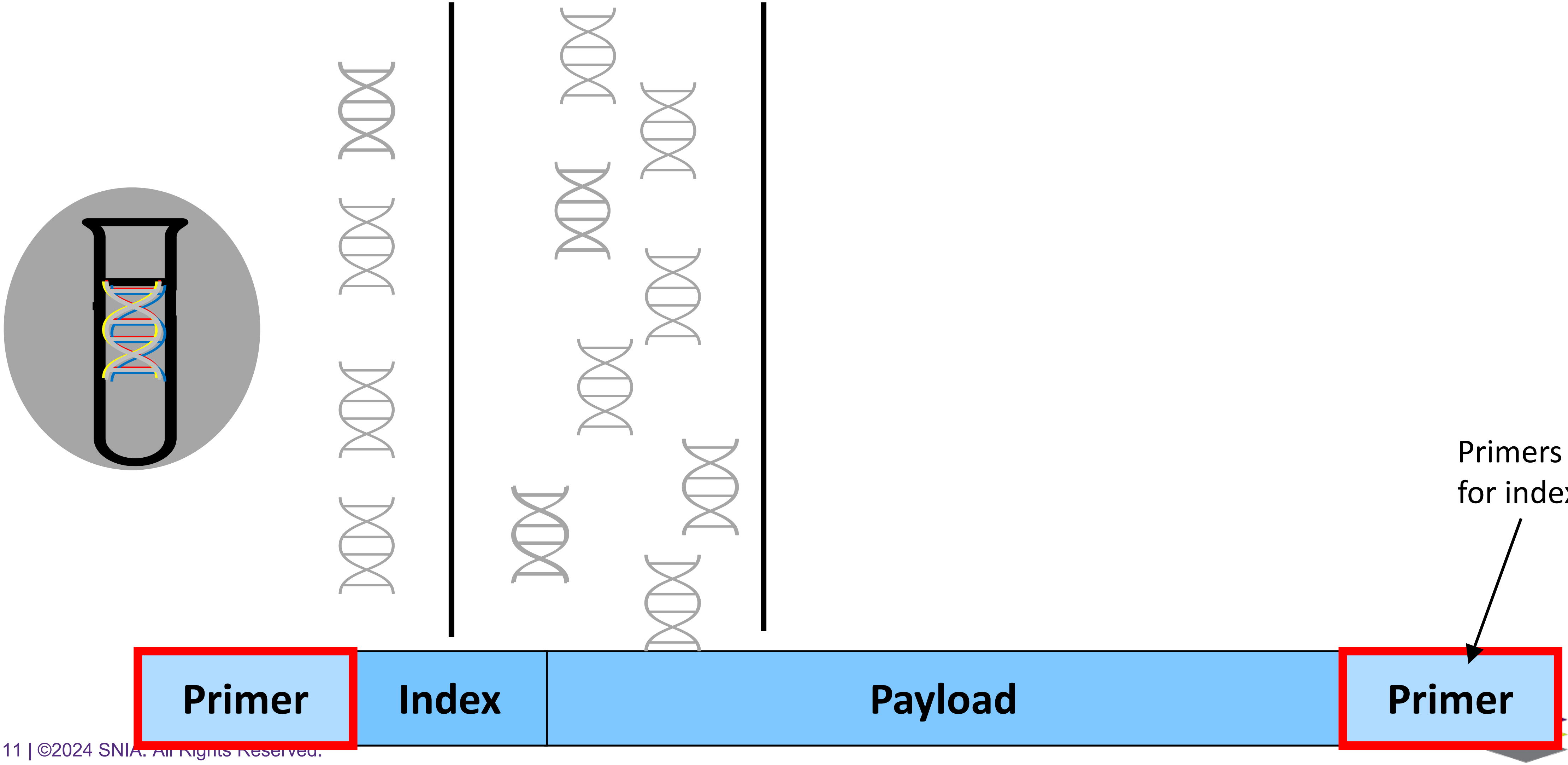
Simple encoding:

Bit	Base
00	A
01	T
10	G
11	C

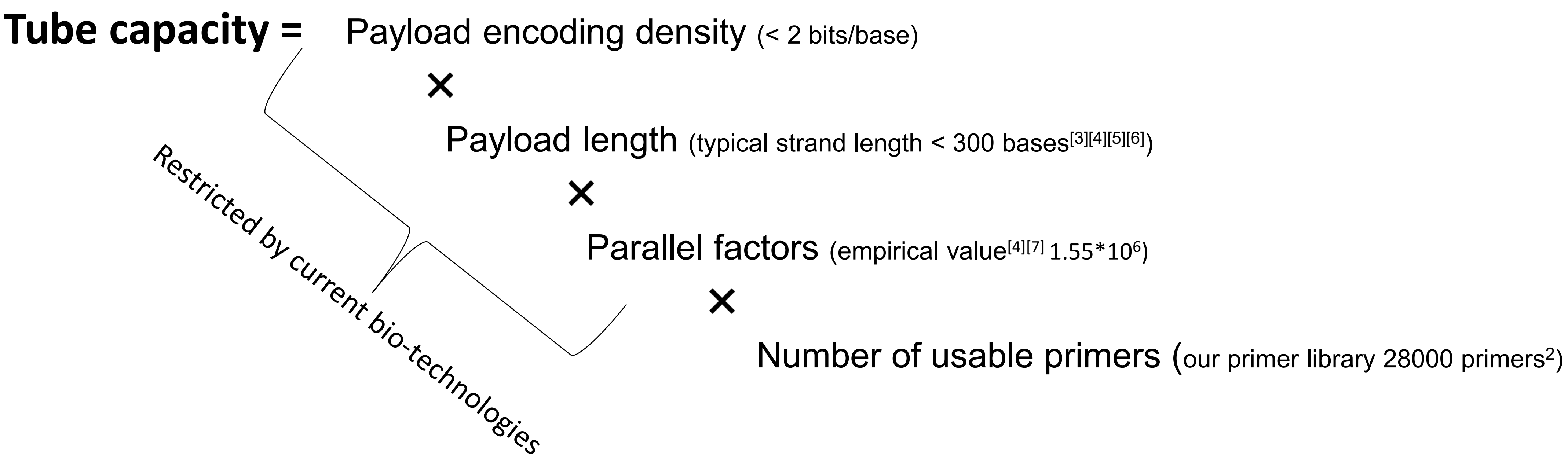
DNA storage capacity can be measured by the **DNA tube storage capacity**

Polymerase Chain Reaction (PCR) and Random-Access

PCR amplifies the target DNA strands for random access



Factors Affecting DNA Tube Storage Capacity



Tube capacity = Payload encoding density (< 2 bits/base)

×

Payload length (typical strand length < 300 bases[3][4][5][6])

×

Parallel factors (empirical value[4][7] $1.55*10^6$)

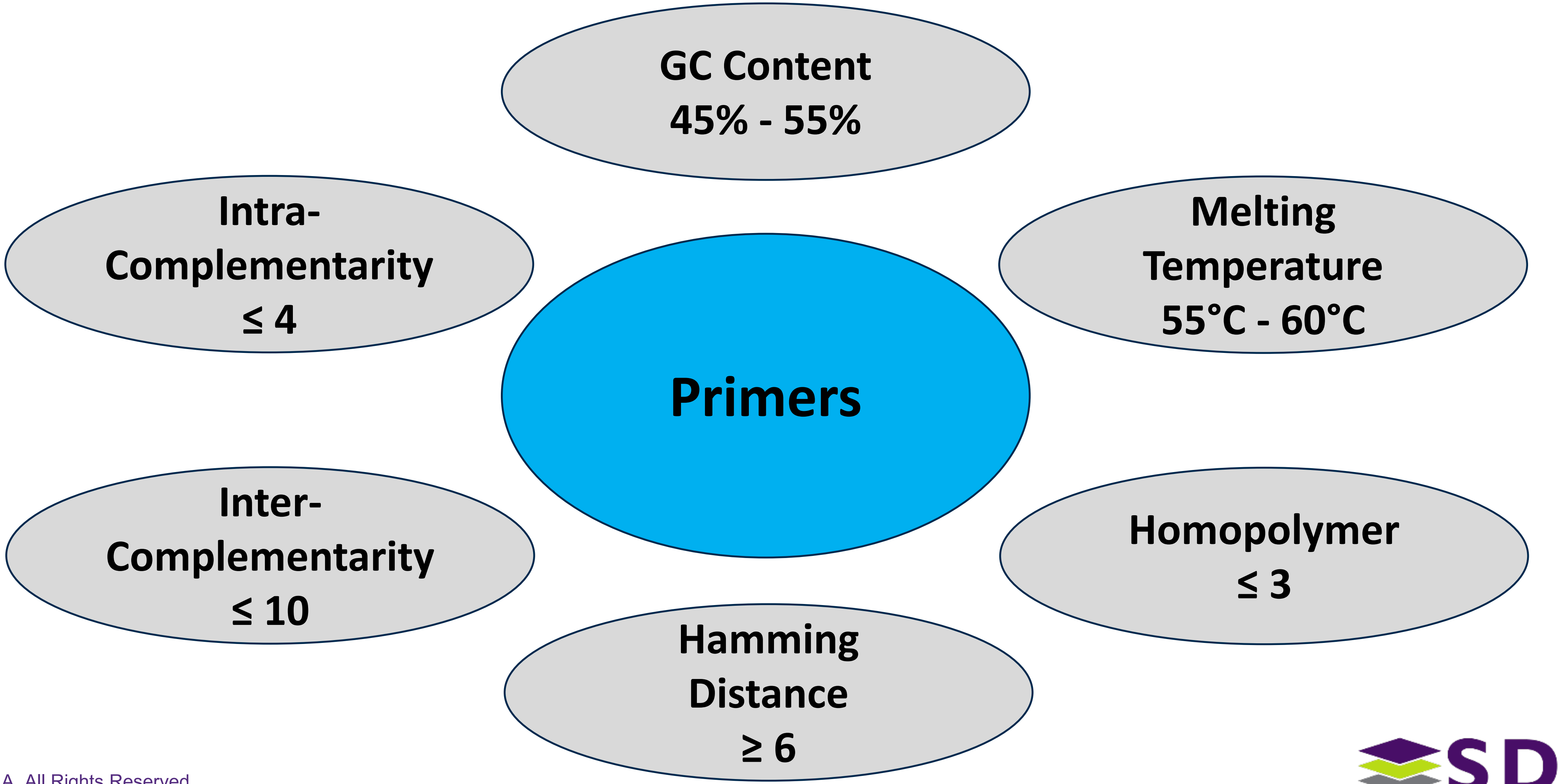
×

Number of usable primers (our primer library 28000 primers[2])

Restricted by current bio-technologies

[3] Church, George M., Yuan Gao, and Sriram Kosuri. ”Next-generation digital information storage in DNA.” Science 337, no. 6102 (2012): 1628-1628.
[4] Blawat, Meinolf, Klaus Gaedke, Xiao-Ming Chen, Brian Turczyk, Samuel Inverso, Benjamin W. Pruitt, and George M. Church. ”Forward error correction for DNA data storage.” Procedia Computer Science 80 (2016): 1011-1022.
[5] Grass, Robert N., Reinhard Heckel, Daniela Paunescu. ”Robust chemical preservation of digital information on DNA in silica with error-correcting codes.” Angewandte Chemie International Edition 54, no. 8 (2015): 2552-2555.
[6] Erlich, Yaniv, and Dina Zielinski. ”DNA Fountain enables a robust and efficient storage architecture.” Science 355.6328 (2017): 950-954
[7] Li, Bingzhe, Nae Young Song, Li Ou, and David HC Du. ”Can We Store the Whole World’s Data in DNA Storage?.” In 12th USENIX Workshop on Hot Topics in Storage and File Systems (HotStorage 20). 2020.
[8] Organick, Lee, Siena Dumas Ang, Yuan-Jyue Chen, Randolph Lopez, Sergey Yekhanin, Konstantin Makarychev, Miklos Z. Racz et al. ”Random access in large-scale DNA data storage.” Nature biotechnology 36, no. 3 (2018): 242-248
[2] Largest primer library in DNA storage

Primer Design Rule



Primers

GC Content
45% - 55%

Melting Temperature
55°C - 60°C

Homopolymer
≤ 3

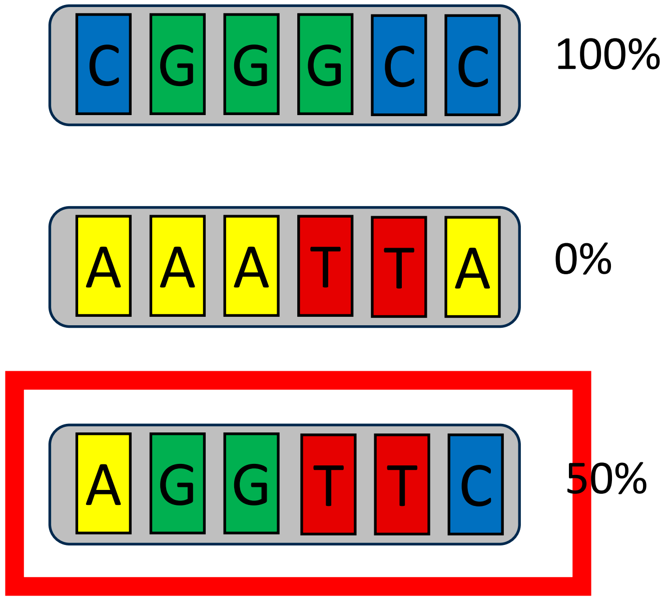
Hamming Distance
≥ 6

Inter-Complementarity
≤ 10

Intra-Complementarity
≤ 4

Bio-Constraints Examples

GC Contents: 45% - 55%

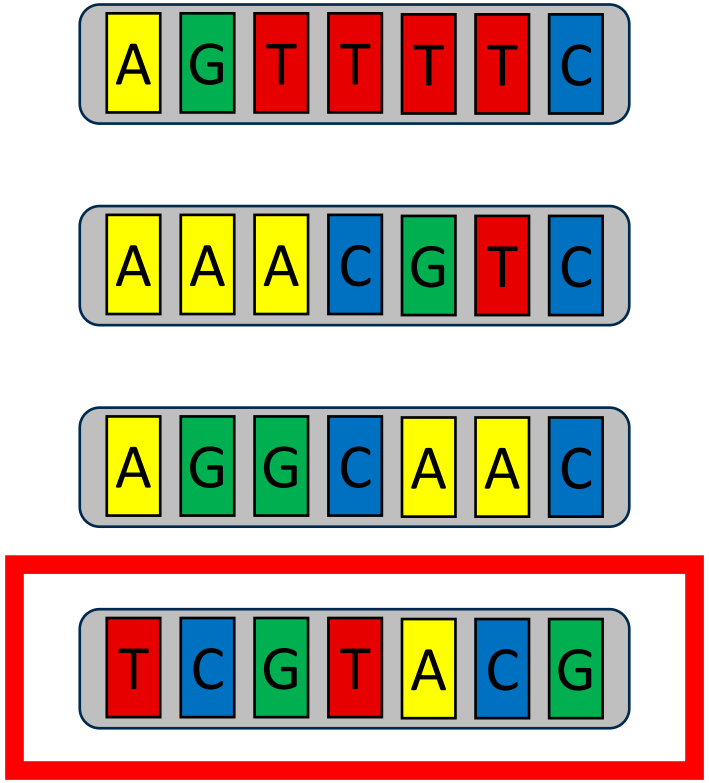


C G G G C C 100%

A A A T T A 0%

A G G T T C 50%

Homopolymers



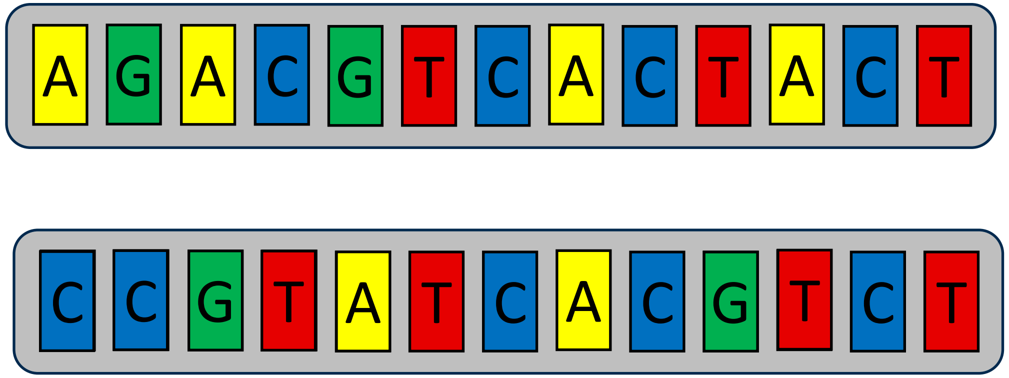
A G T T T T C

A A A C G T C

A G G C A A C

T C G T A C G

Hamming Distance



A G A C G T C A C T A C T

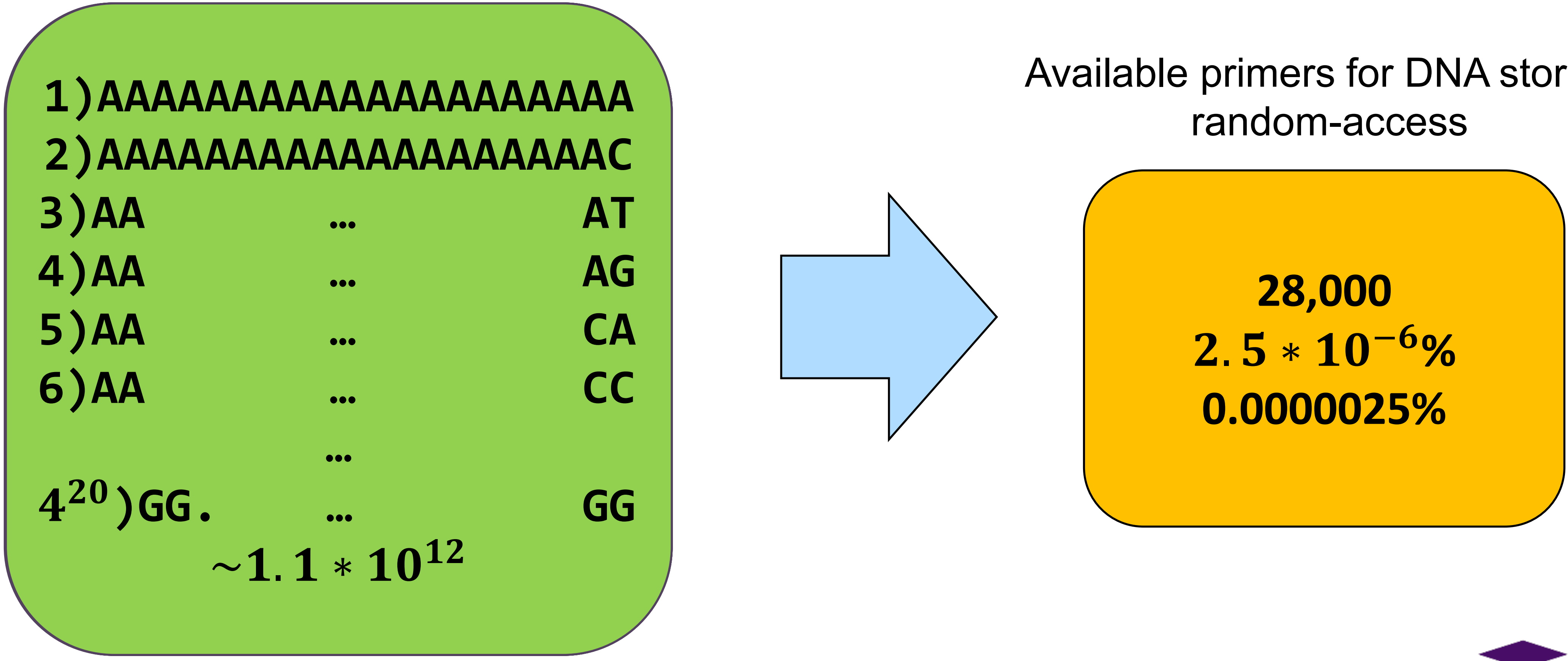
C C G T A T C A C G T C T

[redacted]

Primers with Primer-Primer Collision

Possible DNA Strands

20 Nucleotide Primers



1)AAAAAAAAAAAAAAAAAAAA
2)AAAAAAAAAAAAAAAAAAAC
3)AA … AT
4)AA … AG
5)AA … CA
6)AA … CC
…
4^{20})GG. … GG

$\sim 1.1 * 10^{12}$

Available primers for DNA storage random-access

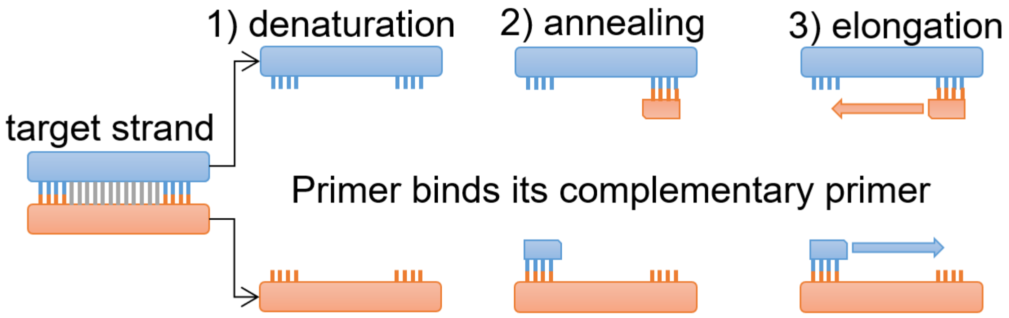
28,000
$2.5 * 10^{-6}\%$
0.0000025%

Primer-payload Collision

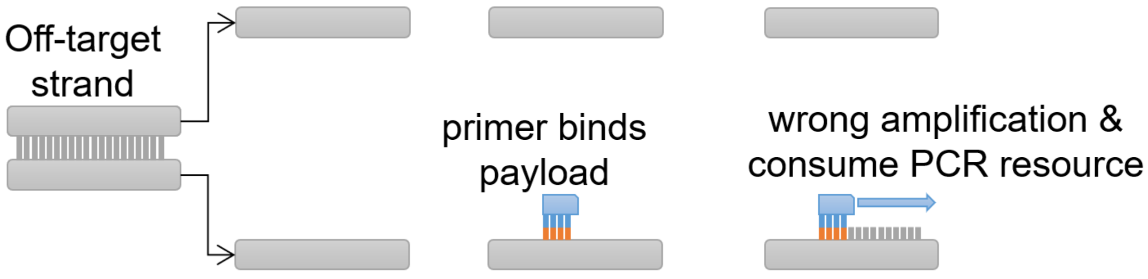
A pair of long almost identical sub-sequences between a primer and any payload stored in the tube [9]

- >=12 bases
- allows at most two mismatches or gaps

Standard PCR



PCR with primer-payload collision

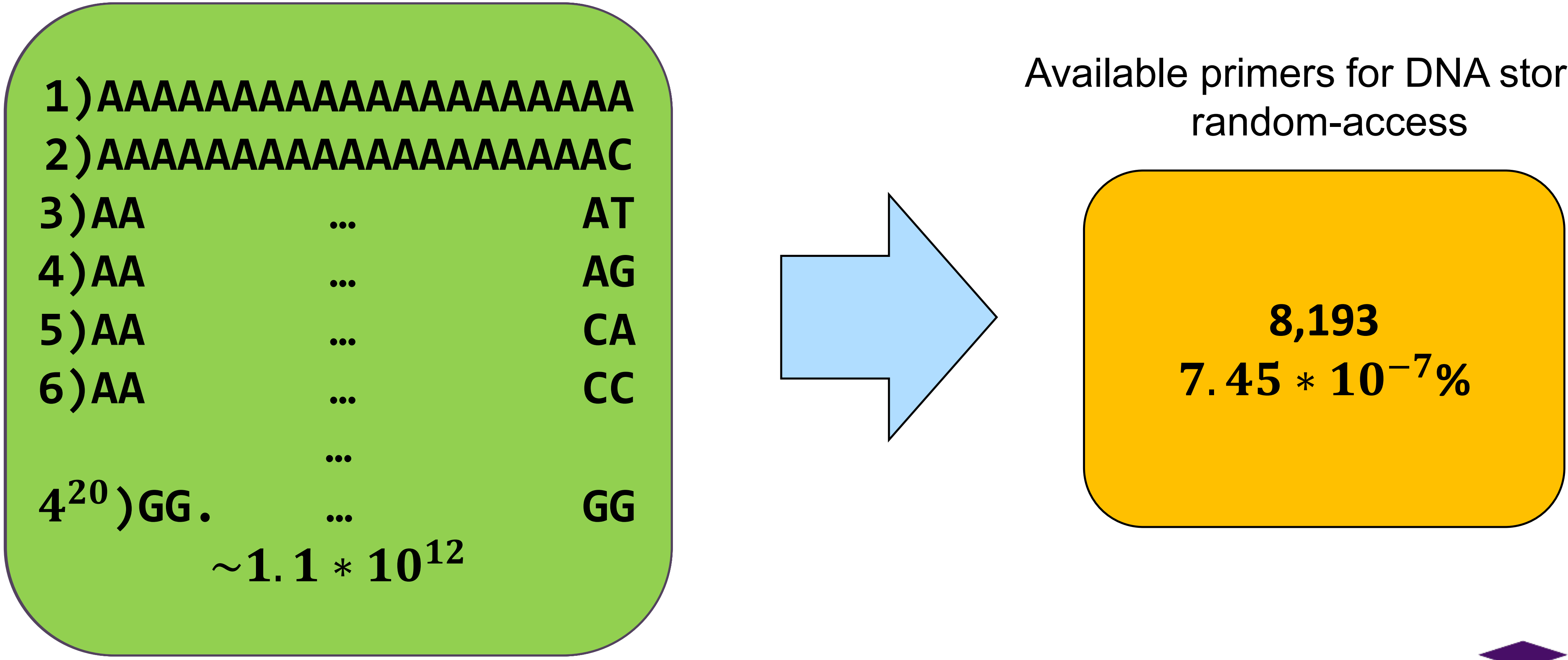


A primer must be disabled if it has collision with any payload in the tube

Primers with Primer-Payload Collision

Possible DNA Strands

20 Nucleotide Primers



1)AAAAAAAAAAAAAAAAAAAA
2)AAAAAAAAAAAAAAAAAAAC
3)AA … AT
4)AA … AG
5)AA … CA
6)AA … CC
…
4^{20})GG. … GG

$\sim 1.1 * 10^{12}$

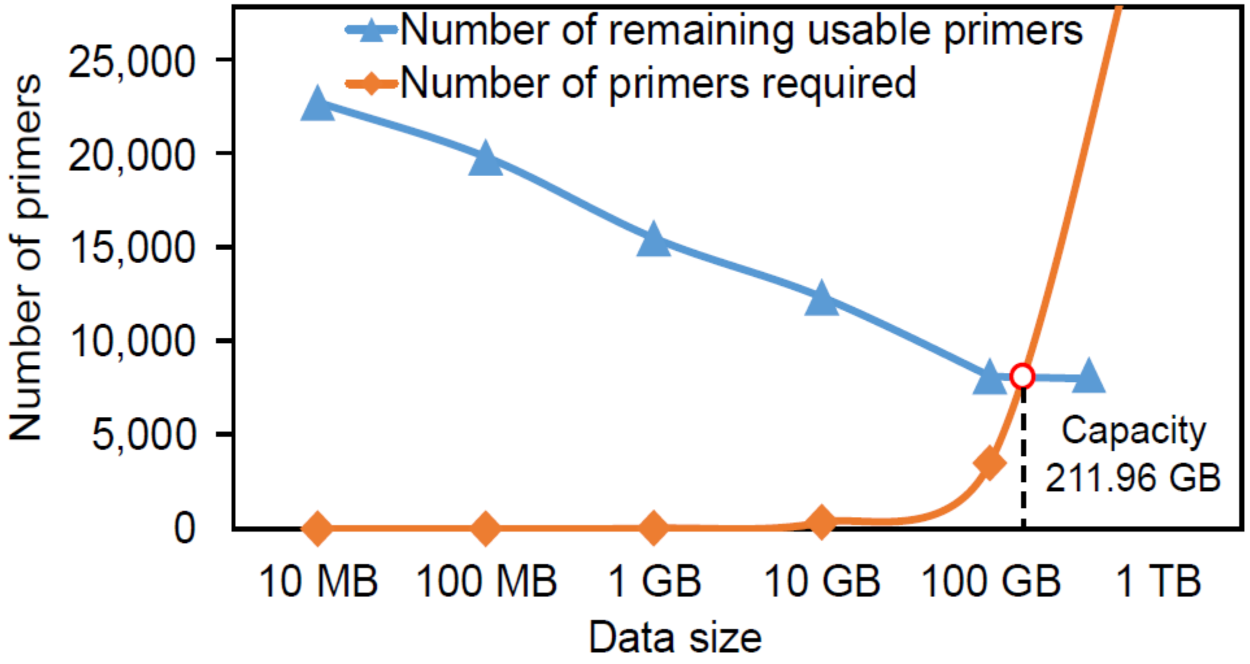
Available primers for DNA storage random-access

8,193

$7.45 * 10^{-7}\%$

Primer Reduction Because of Primer-payload Collision

Figure 1. The number of usable primers decreases as the storage data size increases[1]

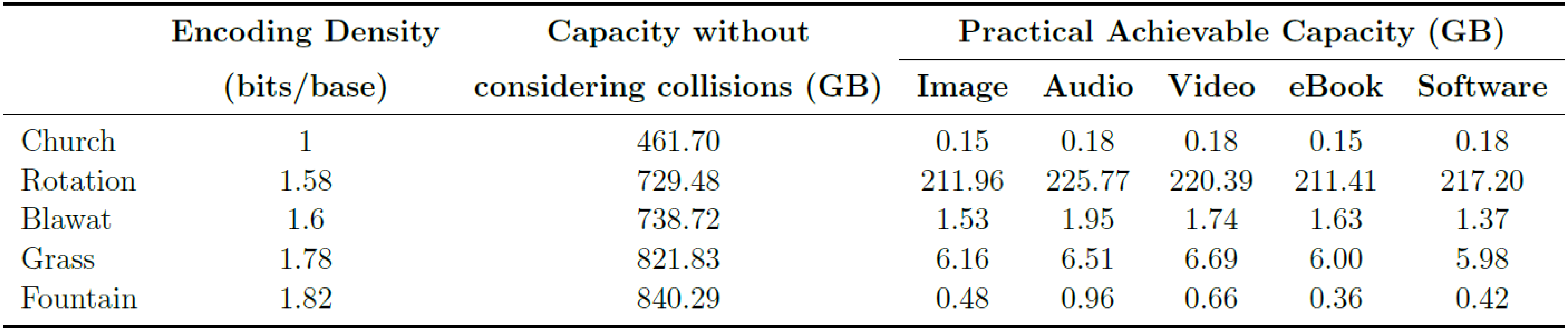


A primer must be disabled if it has collision with any payload in the tube

[3]The experiment is based on ImageNet and Rotation code

Practical DNA Tube Storage Capacities

- Implement five state of art encoding schemes
- Collect five types of data (https://archive.org), each type 500GB



	Encoding Density (bits/base)	Capacity without considering collisions (GB)	Practical Achievable Capacity (GB)				
			Image	Audio	Video	eBook	Software
Church	1	461.70	0.15	0.18	0.18	0.15	0.18
Rotation	1.58	729.48	211.96	225.77	220.39	211.41	217.20
Blawat	1.6	738.72	1.53	1.95	1.74	1.63	1.37
Grass	1.78	821.83	6.16	6.51	6.69	6.00	5.98
Fountain	1.82	840.29	0.48	0.96	0.66	0.36	0.42

Due to fewer usable primers

- **Great capacity reduction (70%~99%) due to primer-payload collision**
  - **All five encoding schemes**
  - **All five data types**

Collision is detected by NCBI BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi)

Collision Aware Coding (CAC) Scheme

It pays to trade some encoding density to design payloads with special patterns to reduce possible collisions with primers

Sequence pattern of primers

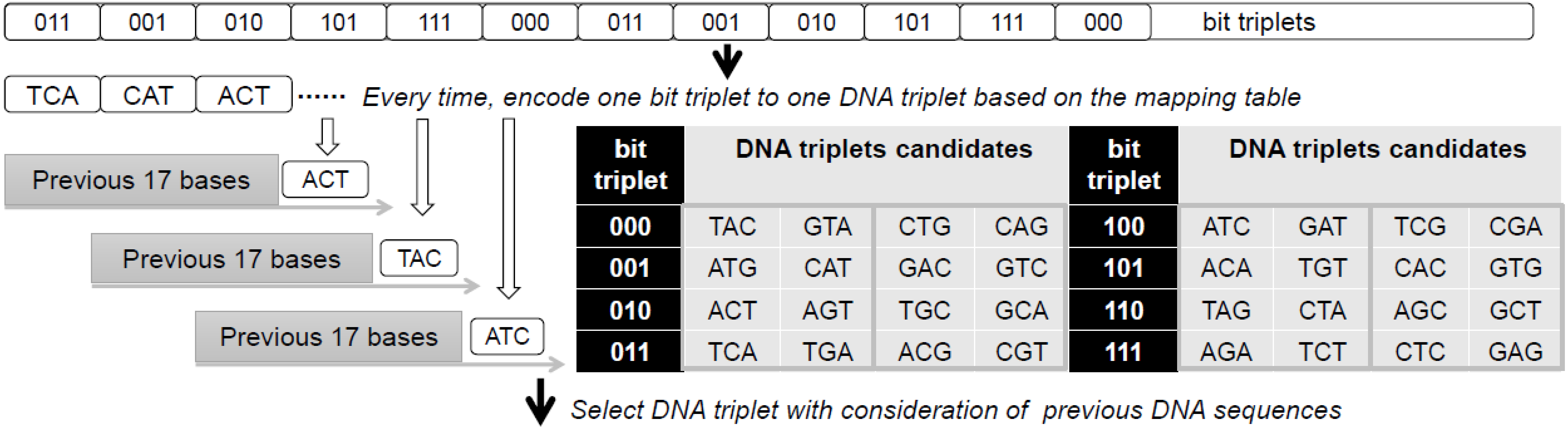
- do not have any homopolymers of A/T/C/G with length > 3
- do not have any consecutive complementary sequences with length > 4 (e.g., … AAACC…GGTTT …)
- GC content ∈ [0.45, 0.55]

~~different from primer pattern~~ or
further tighten the pattern (payload can only be similar to a subset of primers)

Sequence pattern of payloads

- No homopolymers.
- Any 20 base subsequences of payloads should have fewer or no consecutive complementary sequences.
- Any 20 base subsequences of payloads should have more balanced GC content (closer to 0.5).

Collision Aware Coding (CAC) Scheme cont.



Every time, encode one bit triplet to one DNA triplet based on the mapping table

bit triplet	DNA triplets candidates				bit triplet	DNA triplets candidates			
000	TAC	GTA	CTG	CAG	**100**	ATC	GAT	TCG	CGA
001	ATG	CAT	GAC	GTC	**101**	ACA	TGT	CAC	GTG
010	ACT	AGT	TGC	GCA	**110**	TAG	CTA	AGC	GCT
011	TCA	TGA	ACG	CGT	**111**	AGA	TCT	CTC	GAG

Select DNA triplet with consideration of previous DNA sequences

Look back at the last base -- avoid forming homopolymers with the last base
Look back on the previous 17 bases -- less complementary sequences and balanced GC content

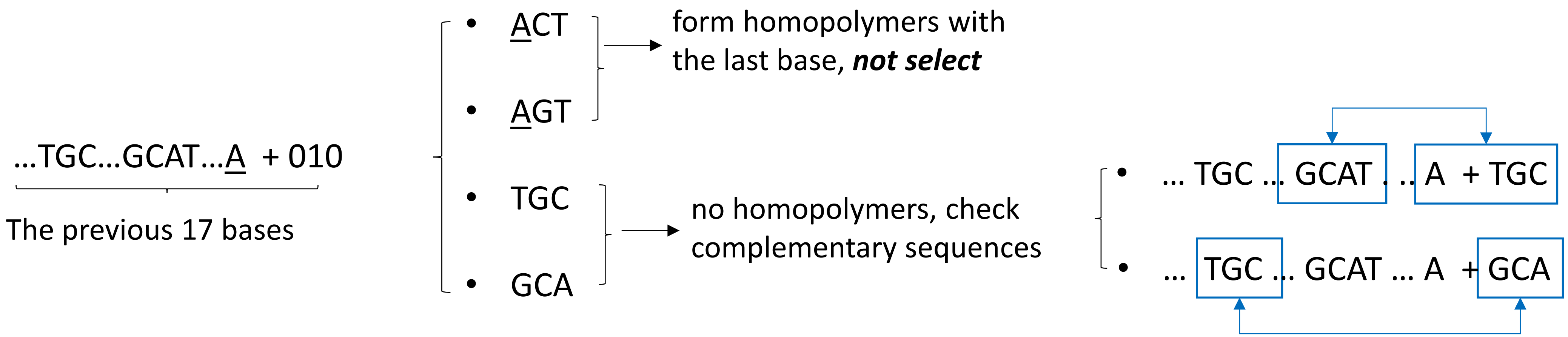
Challenges:

- How to build the encoding table so that each bit triplet can always be encoded as a proper DNA triplet
- With the encoding table how to select a proper DNA triplet

Select A DNA Triplet From Encoding Table

- Initially each bit triplet has 4 DNA triplet candidates

- Avoid homopolymers > Less complementary sequences > Strict balanced GC content



…TGC…GCAT…A + 010

The previous 17 bases

- ACT
- AGT
- TGC
- GCA

form homopolymers with the last base, ***not select***

no homopolymers, check complementary sequences

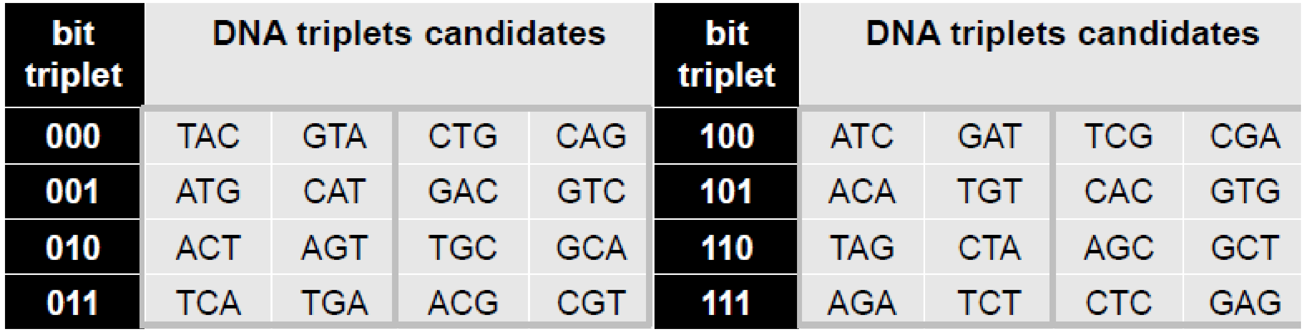
- … TGC … GCAT … A + TGC
- … TGC … GCAT … A + GCA

shorter complementary sequence, ***selected***.

otherwise, further check GC content, select the one with more balanced GC.

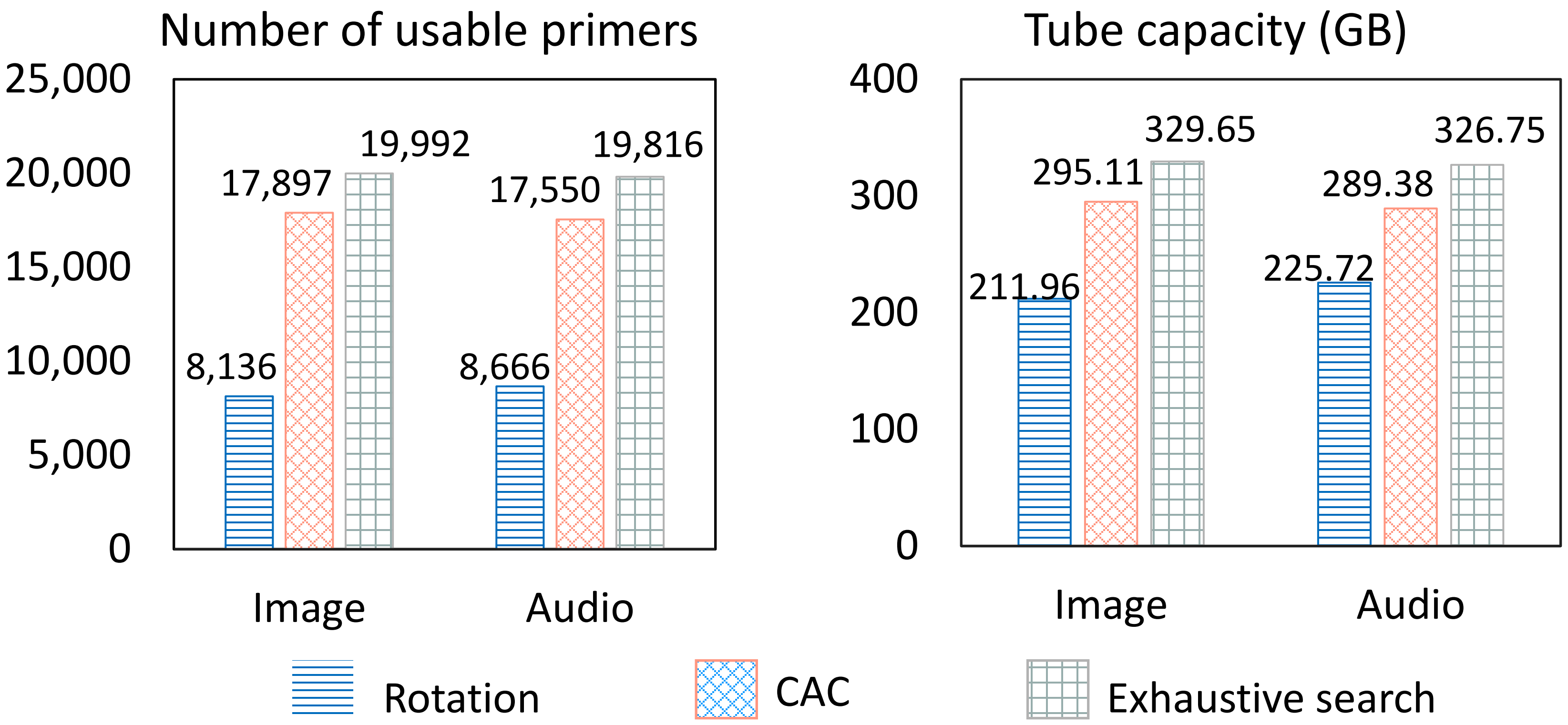
Encoding Table: From Bit Triplet To Base Triplet

- Principle 1 – no internal homopolymers (e.g., TTA is not allowed)
- Principle 2 – different starting base:
  - each bit triplet should have candidates starting with different bases (e.g., 000 can be encoded as T_ _ & G_ _ & C_ _)
- Principle 3 – complementary pairs:
  - each DNA triplet, together with its complementary peer, should be candidates of the same bit triplet. (e.g., candidate 1&2, 3&4)
- Principle 4 – GC balanced candidates:
  - each bit triplet should have candidates with different GC portions
  - e.g., candidate 1&2 have one GC, candidate 3&4 have two GC



bit triplet	DNA triplets candidates				bit triplet	DNA triplets candidates			
000	TAC	GTA	CTG	CAG	100	ATC	GAT	TCG	CGA
001	ATG	CAT	GAC	GTC	101	ACA	TGT	CAC	GTG
010	ACT	AGT	TGC	GCA	110	TAG	CTA	AGC	GCT
011	TCA	TGA	ACG	CGT	111	AGA	TCT	CTC	GAG

Experimental Result - Capacity



Rotation(29% of primer library, 211GB) << CAC(65% of primer library, 295GB) < Exhaustive search(70% of primer library, 329GB)

① high encoding density but low collision avoidance

② balance between encoding density & collision avoidance

③ search a DNA candidate from all DNA triplets; not decodable & low encoding speed

Experimental Result – Execution Time

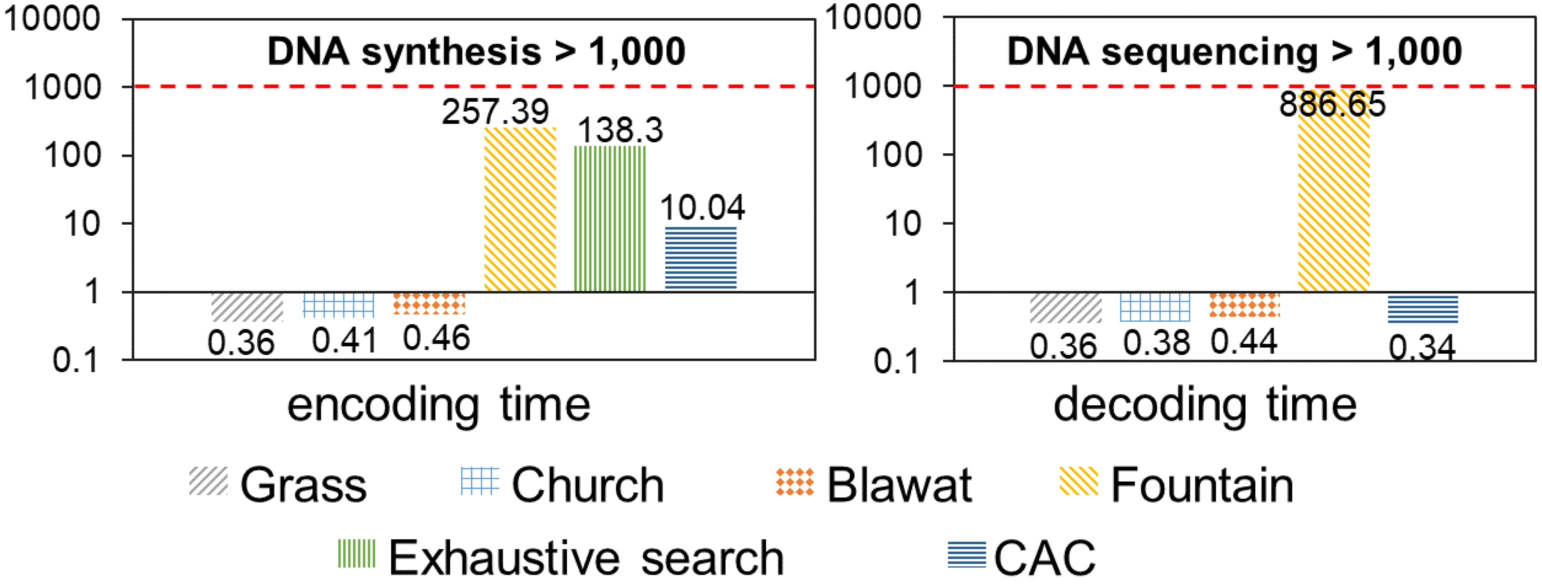


Figure: Encoding and decoding time when processing a 135MB video file (normalized based on Rotation code)

- DNA sequencing: hundreds Kilo base/s ~ Mega base/s[2][3]
- DNA synthesis: Kilo base/s ~ tens Kilo base/s[4]

Potential optimizations

- simple & repeated calculation -> GPU/FPGA
- in-memory buffer / partially buffer the encoding decisions

[11] Doricchi, Andrea, et al. "Emerging approaches to DNA data storage: Challenges and prospects." *ACS nano* 16.11 (2022): 17552-17571.

[12] Yiming Dong, Fajia Sun, Zhi Ping, Qi Ouyang, and Long Qian. Dna storage: research landscape and future prospects. National Science Review, 7(6):1092–1107, 2020

[13] Hao, Y.; Li, Q.; Fan, C.; Wang, F. Data Storage Based on DNA. Small Struct 2021, 2 (2), 2000046.

Experimental Result – Encoding Density

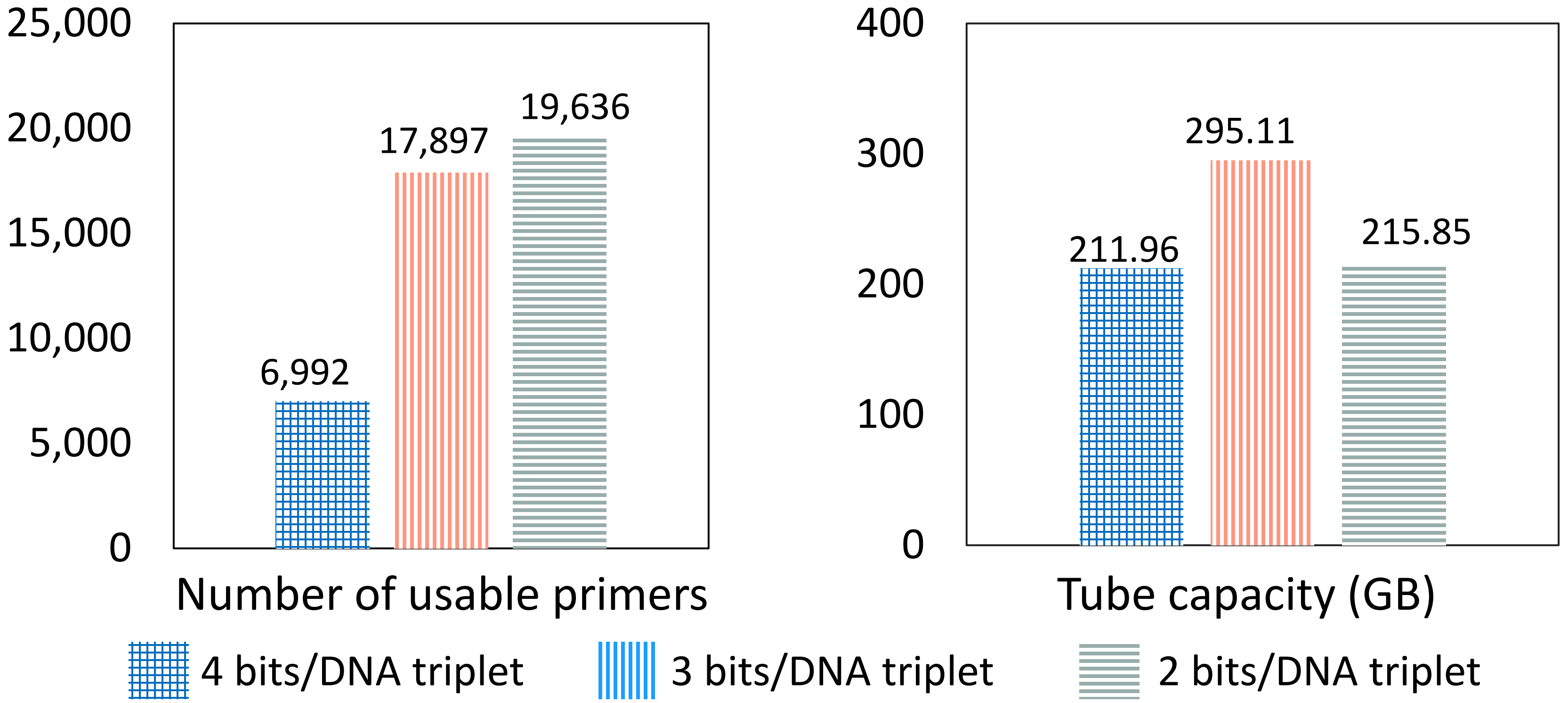


Figure: Number of usable primers and tube capacity for CAC-like encoding schemes with different encoding density

	4 bits/DNA triplet	3 bits/DNA triplet	2 bits/DNA triplet
Number of candidates	2	4	8
Collision avoidance	low	medium	high
Encoding density	1.33 bits/base	1 bit/base	0.66 bit/base

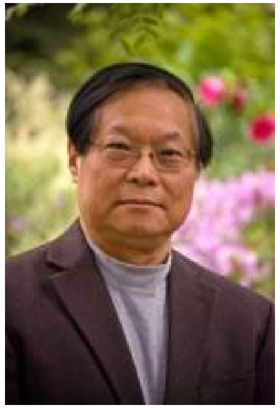
Conclusion

- Practical DNA storage capacity is much lower than expectation.
- Propose a new collision aware encoding (CAC) to improve the capacity.
- A new mapping table is proposed.
- CAC can improve the number of primers by ~2X and the capacity by ~40%.

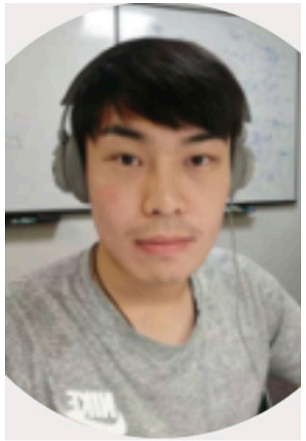
Further Improvement

- New encoding scheme to avoid primer payload collision
- Error correction code
- More primer generation
- Higher encoding density
- …

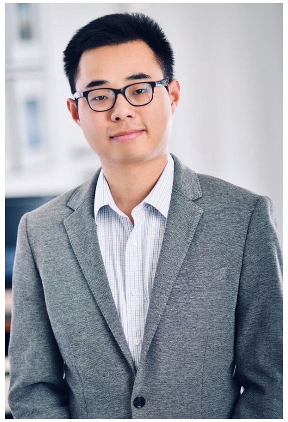
Acknowledgement



Prof. David Du



Yixun Wei



Dr. Li Ou

Yi Li

Alex Sensintaiffar

NSF

DNA DATA STORAGE ALLIANCE

SNIA®

Lab4ISC

Lab for Intelligent Storage and Computing

Thanks!

Q&A