Towards practical, high-capacity, low-maintenance information storage in synthesized DNA

Nick Goldman[1], Paul Bertone[1], Siyuan Chen[2], Christophe Dessimoz[1], Emily M. LeProust[2], Botond Sipos[1] & Ewan Birney[1]

Presented by Divya and Natalie

March 13th, 2013

20.385

Nature (2013) 494: 77–80 doi:10.1038/nature11875

Who needs digital storage?

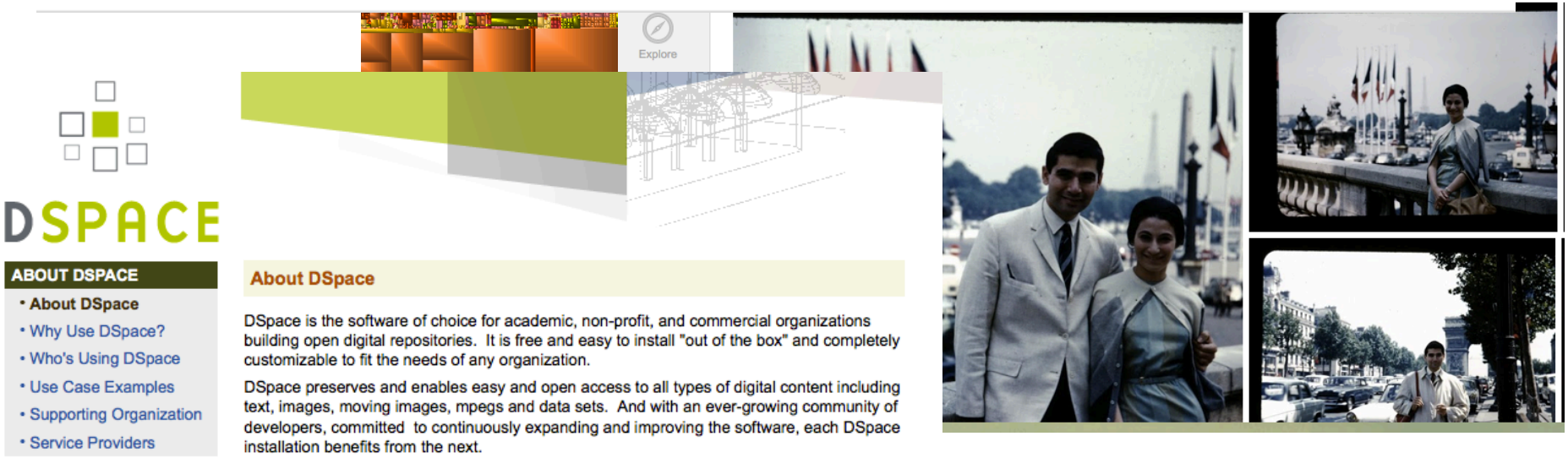
iCloud: Purchasing iCloud storage and billing

Summary

iCloud customers are provided with 5 GB of free cloud storage. Purchased music, movies, TV shows, apps, and books, as well as photos in your Photo Stream don't count against your 5 GB of free storage.

Products Affected

iCloud

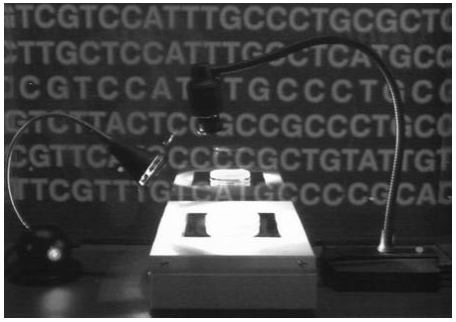


Other examples of DNA storage

Eduardo Kac, 1999

GENESIS

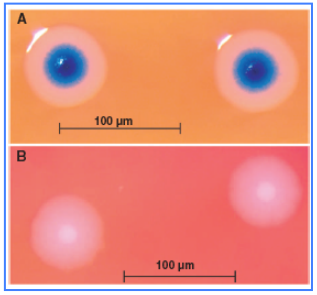
Bible quote to Morse code to DNA code to expressed protein



Gibson et al, 2010

Synthetic Genomics

Watermarks with 46 authors and James Joyce and Richard Feynman quotes



George Church, 2012

Regenesis

Next-Generation Digital Information Storage in DNA

George M. Church[1,2], Yuan Gao[3], Sriram Kosuri[1,2,*]

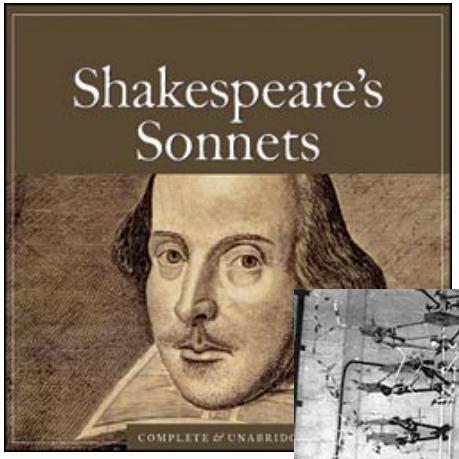
(+) storage dense (2 bits/nt)
(455 exabytes/gram)
"naturally readable"

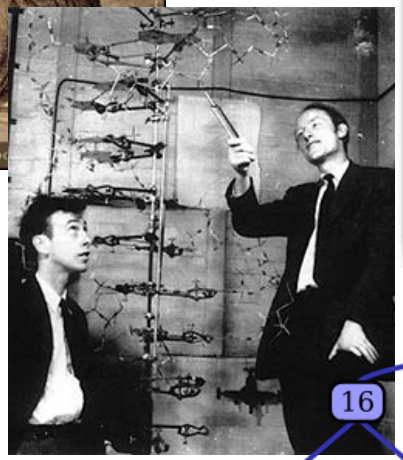
(-) living systems are unreliable, existing examples don't scale and $$$$

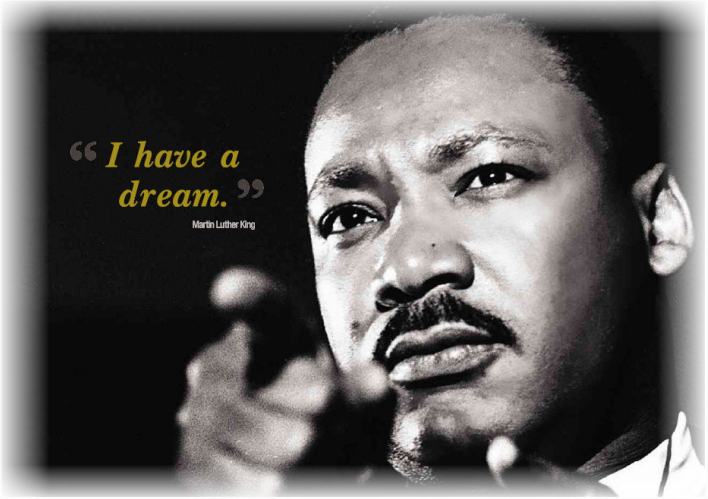
Proof of concept

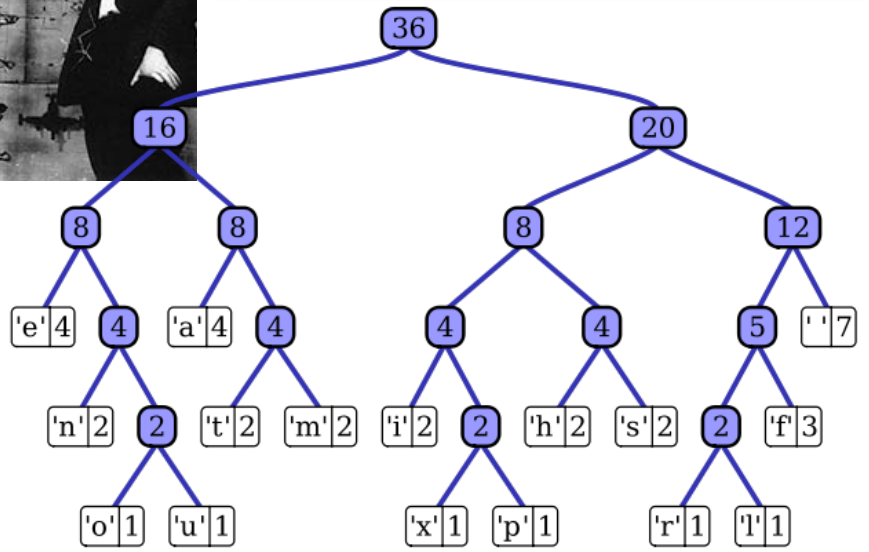
EMBL-EBI

- ASCII txt
- PDF
- JPEG 2000
- MP3
- Huffman code (way of compressing info)









TOTAL = 757,051 bytes of information

Encoding, Synthesizing, Decoding

An arbitrary computer file

a string $S_\emptyset$ of bytes, i.e. a value in the set $\{\emptyset \ldots 255\}$

$$S_\emptyset = \texttt{Birney␣and␣Goldman}$$

Encoding, Synthesizing, Decoding

An arbitrary computer file

a string $S_\emptyset$ of bytes, i.e. a value in the set $\{\emptyset \ldots 255\}$

given Huffman code, converting it to base-3.

S_1 of characters in $\{\emptyset, 1, 2\}$. 'trit'.

B	i	r	n	e	y	␣	a	n
$S_1 =$ 2ø1øø	2ø21ø	1ø1ø1	øøø21	2øøø1	222111	ø2212	ø1112	øøø21
221øø	ø2212	222212	ø211ø	ø21ø1	221øø	11ø21	ø1112	øøø21
d	␣	G	o	l	d	m	a	n

then add "indexing" and "orientation" info

Encoding, Synthesizing, Decoding

An arbitrary computer file

a string $S_\emptyset$ of bytes, i.e. a value in the set $\{\emptyset \ldots 255\}$

given Huffman code, converting it to base-3.

S_1 of characters in $\{\emptyset, 1, 2\}$. 'trit'.

converted to a DNA string

previous nt written	next trit to encode		
	∅	1	2
A	C	G	T
C	G	T	A
G	T	A	C
T	A	C	G

Encoding, Synthesizing, Decoding

An arbitrary computer file

a string $S_\emptyset$ of bytes, i.e. a value in the set $\{\emptyset \ldots 255\}$

given Huffman code, converting it to base-3.

S_1 of characters in $\{\emptyset, 1, 2\}$. 'trit'.

converted to a DNA string

previous nt written	next trit to encode		
	∅	1	2
A	C	G	T
C	G	T	A
G	T	A	C
T	A	C	G

S_5 = TAGTATATCGACTAGTACAGCGTAGCATCTCGCAGCGAGATACGCTGCTACGCAGCATGCTGT

GAGTATCGATGACGAGTGACTCTGTACAGTACGTACGTACGTACGTACGTACGTACGACTAT

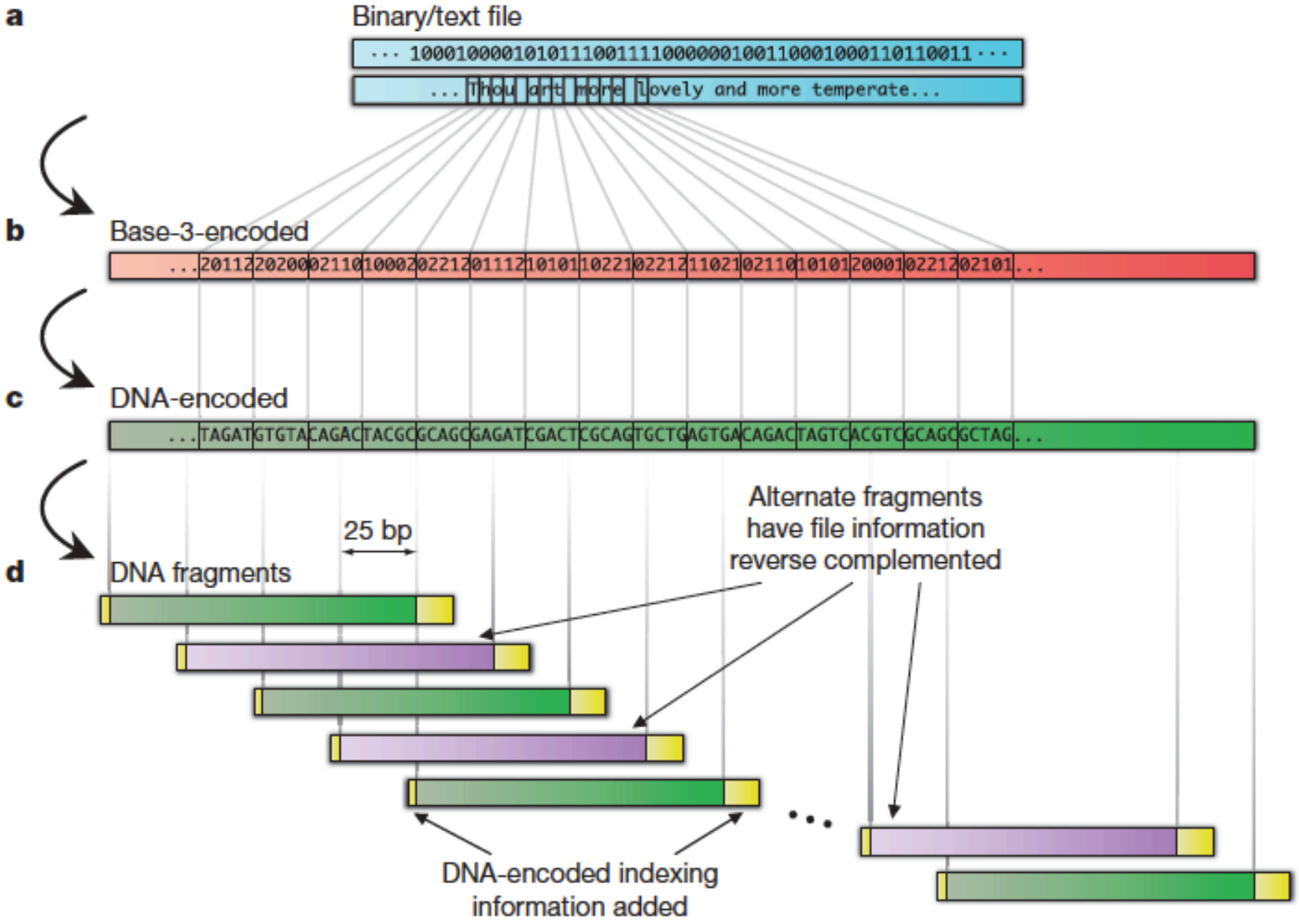
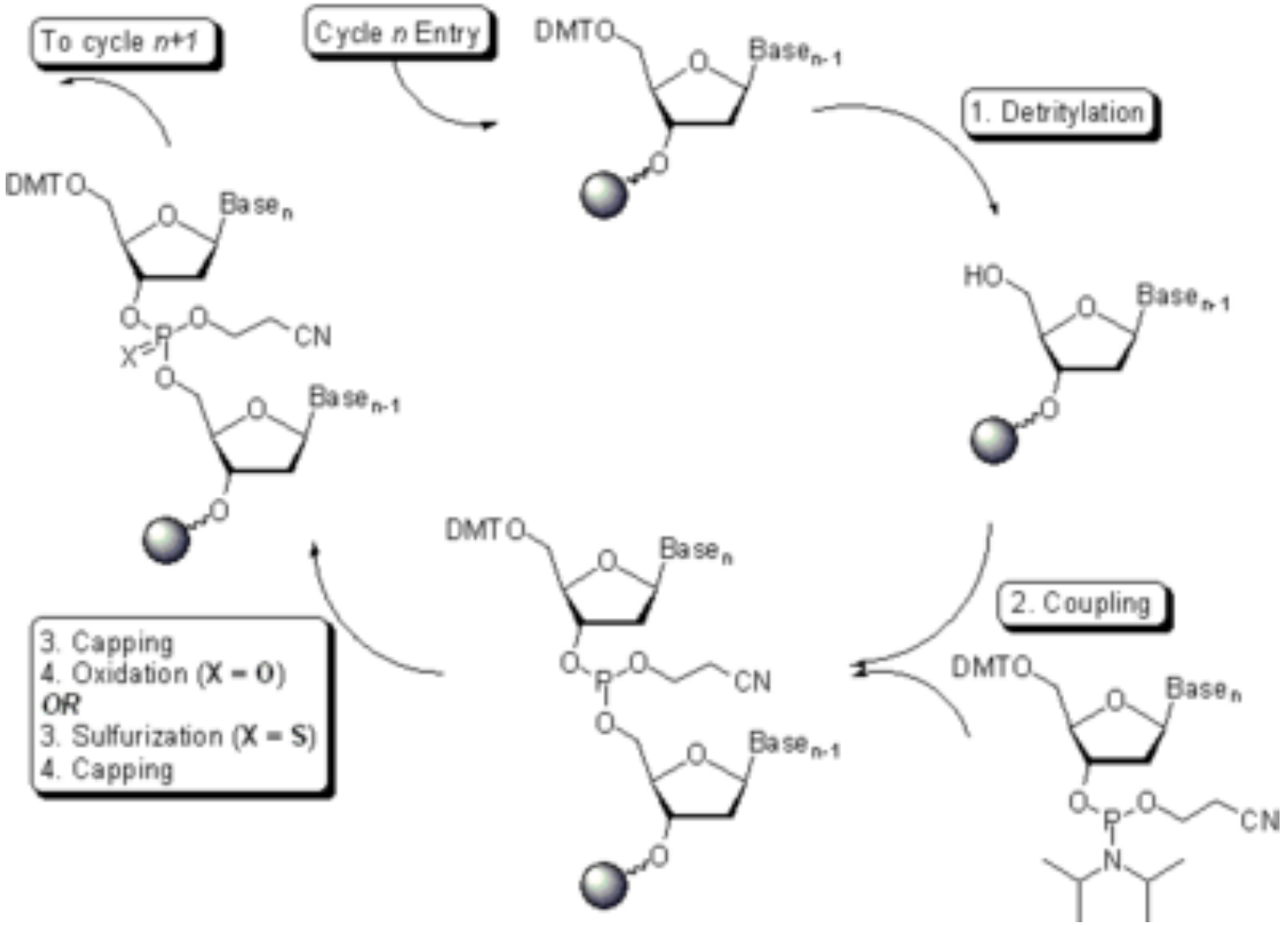

a
Binary/text file
··· 1000100001010111001111000000100110001000110110011 ···
... Thou art more lovely and more temperate...
b
Base-3-encoded
...20112202000211010002022120111210101022102212110210211010120001022120210 1...
c
DNA-encoded
...TAGATGTGTACAGACTACGCGCAGCGAGATCGACTCGCAGTGCTGAGTGACAGACTAGTCACGTCGCAGCGCTAG...
d
25 bp
DNA fragments
Alternate fragments
have file information
reverse complemented
DNA-encoded indexing
information added

Fig. 1

Encoding, Synthesizing, Decoding



Beaucage SL, Caruthers MH. (1981) Tetrahedron Lett. 22, 1859-62.

Encoding, Synthesizing, Decoding

Precise details of the decoding procedure are left as an exercise for the reader.

Errors introduced during DNA synthesis, storage or sequencing could lead to various artefacts, particularly nt insertion, deletion or substitution. Recovery of information from fragments with such errors may be possible (details left as exercise)—we have not found this to be necessary due to the large numbers of perfectly-sequenced fragments available via high-throughput sequencing.

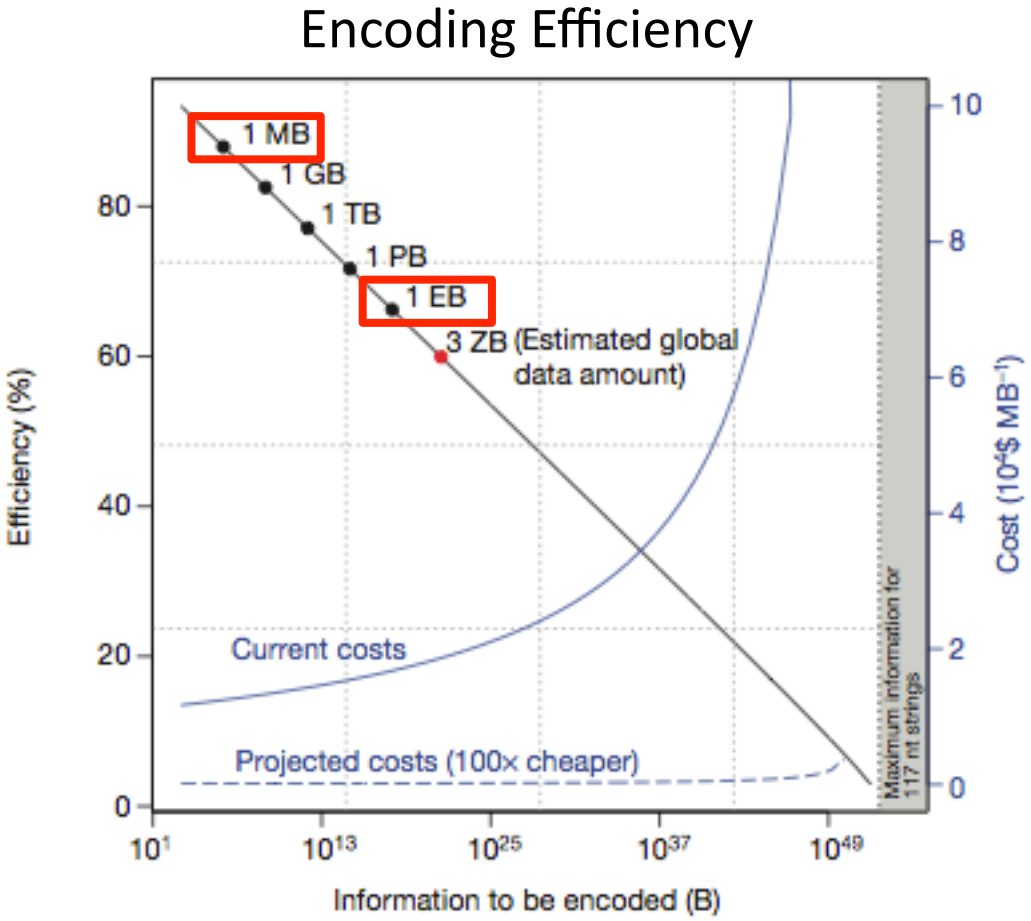
Key Questions

1. Can large amounts of data be stored using DNA?

2. Is decoding accurate?

3. Is DNA-storage cost-effective?

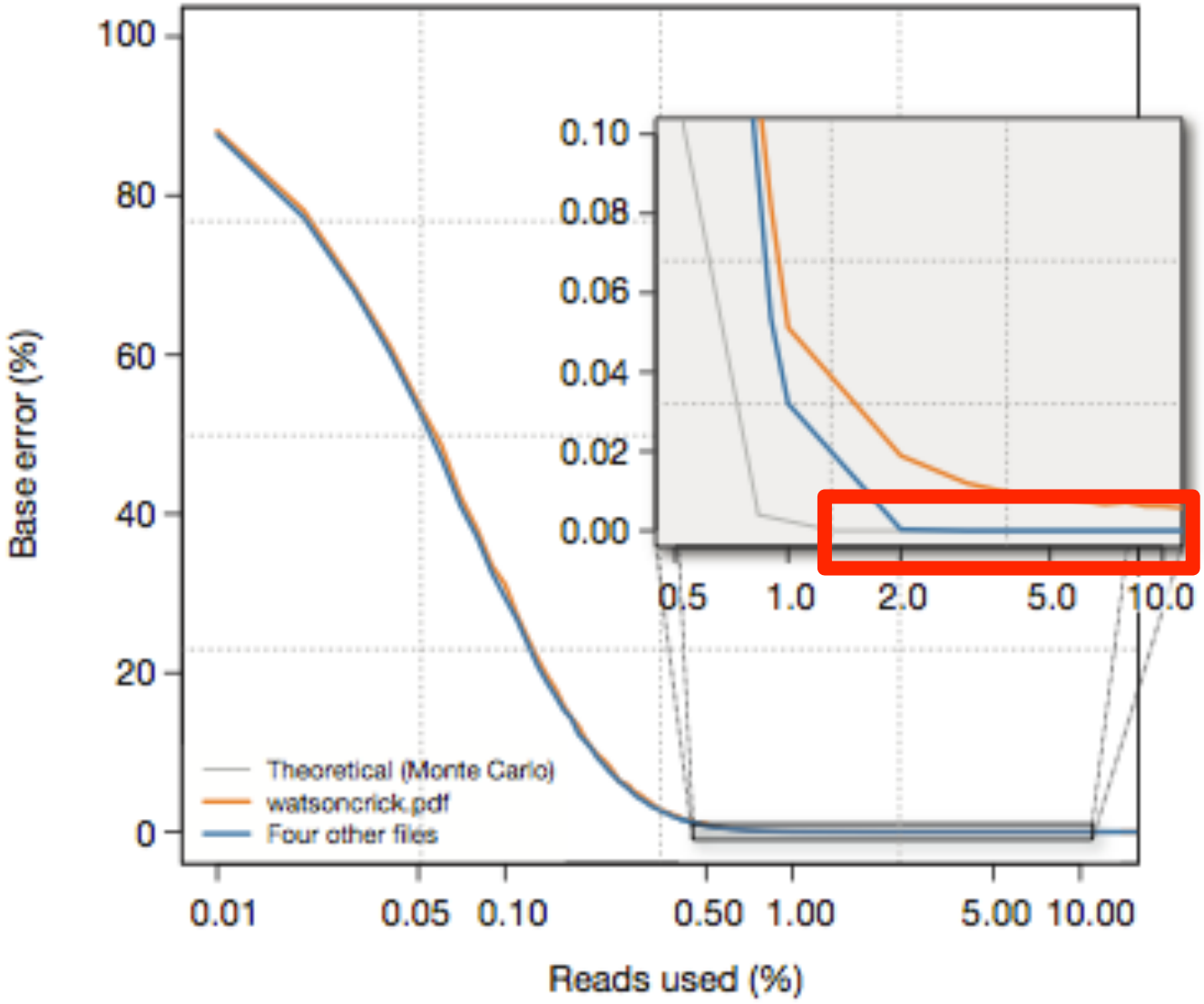
DNA-based Storage Feasible for Large Data



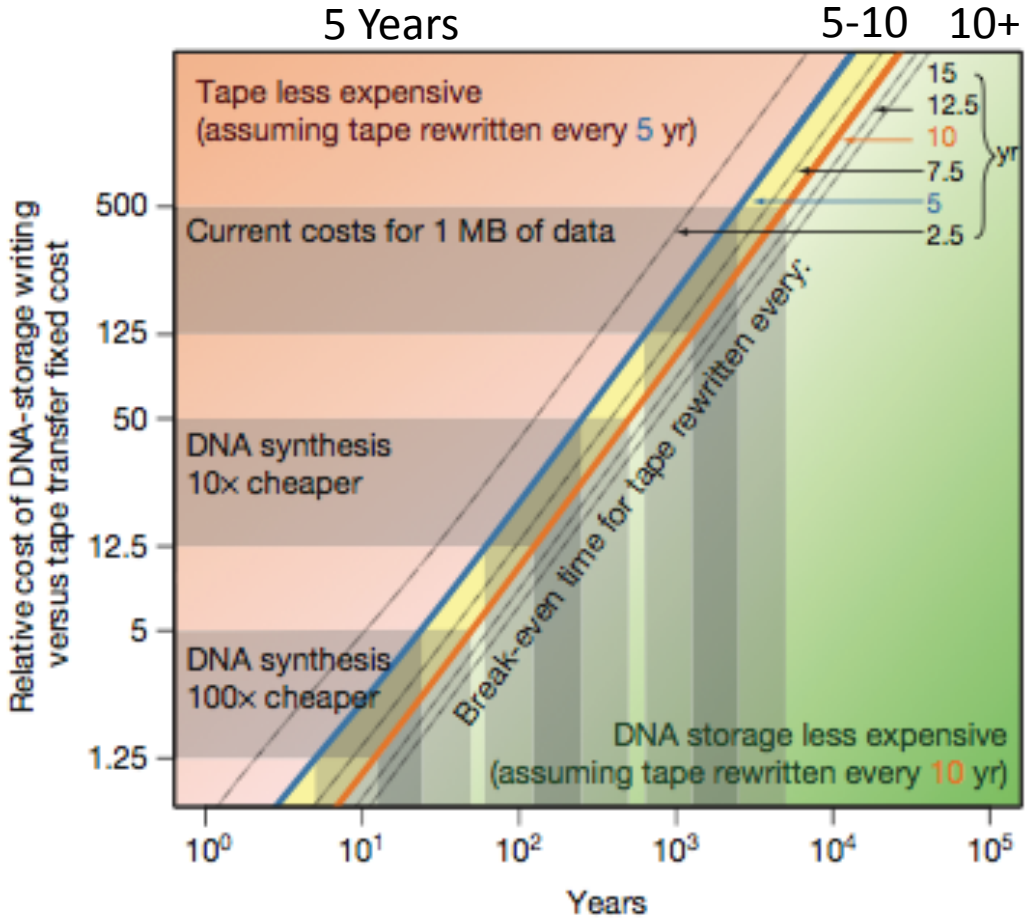
Bytes	**Efficiency**
10^{3}	88%
10^{18}	>65%

Projected 2 magnitude reduction in cost

Zero Percent Error Rate after Reading More Than 2% of Sample

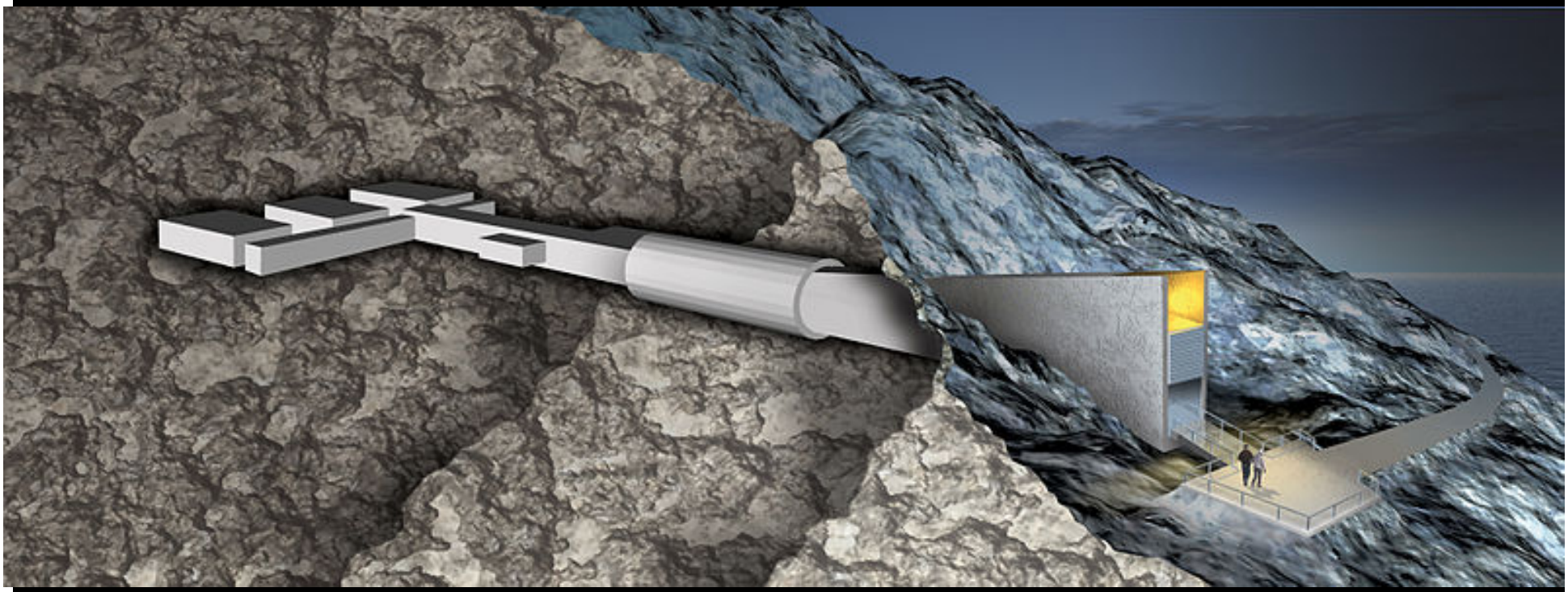


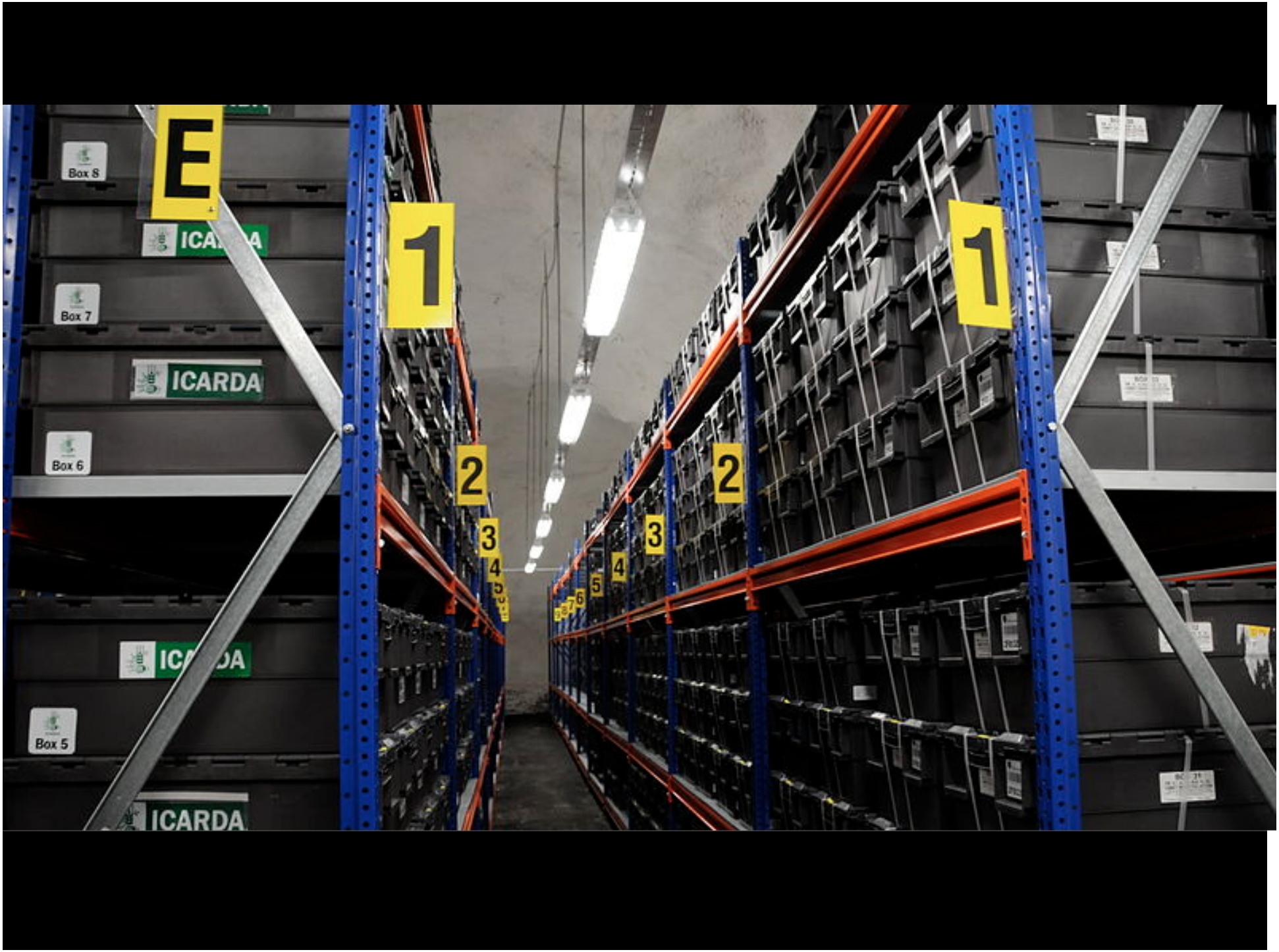
DNA-based Storage Ideal for Long-Horizon Archives with Low Access Rates



- LHC stores 90% of data on magnetic tape
- Magnetic tape needs to be rewritten periodically to counter data loss resulting from magnetic field breakdown
- Shortest breakeven time: ~100 years


SVALGARD GLOBAL SEED VAULT
810 MILES FROM NORTH POLE




E
1
1
2
2
3
3
4
4
5
6
7
8
Box 8
Box 7
ICARDA
Box 6
Box 5
ICARDA

Assumptions

- Cost of synthesis will decrease 100x
- No public wariness
- DNA will remain stable and fully readable over all periods of time
- Current decoding practices will still be in place and understandable in 500+ years
- Preferable to not maintain records

Limitations

- Accessibility
- Not widespread information storage plan
- Consume 10% of information every time you decode
- Not yet cost-effective
- Easily-susceptible to environmental damage

Significance

- Proof of concept of DNA as a dense and stable information storage system
- DNA as storage device for multiple file types
- DNA as a computer language
- Cool science!

Pros

- Low maintenance
- Low error-rate
- Addresses the problem of long-term data storage

Cons

- Cost-effectiveness
- Accessibility
- Public hesitation towards biology

Discussion Questions

- What are some concerns the general public would have about DNA-based storage?
- Would it be wise to put vital information such as government information and historical records in a relatively inaccessible place?
- How likely are people to adopt DNA-storage in the near future?