DNA as Storage Medium

20210724
Literature Seminar
Yun-wei Xue

Contents

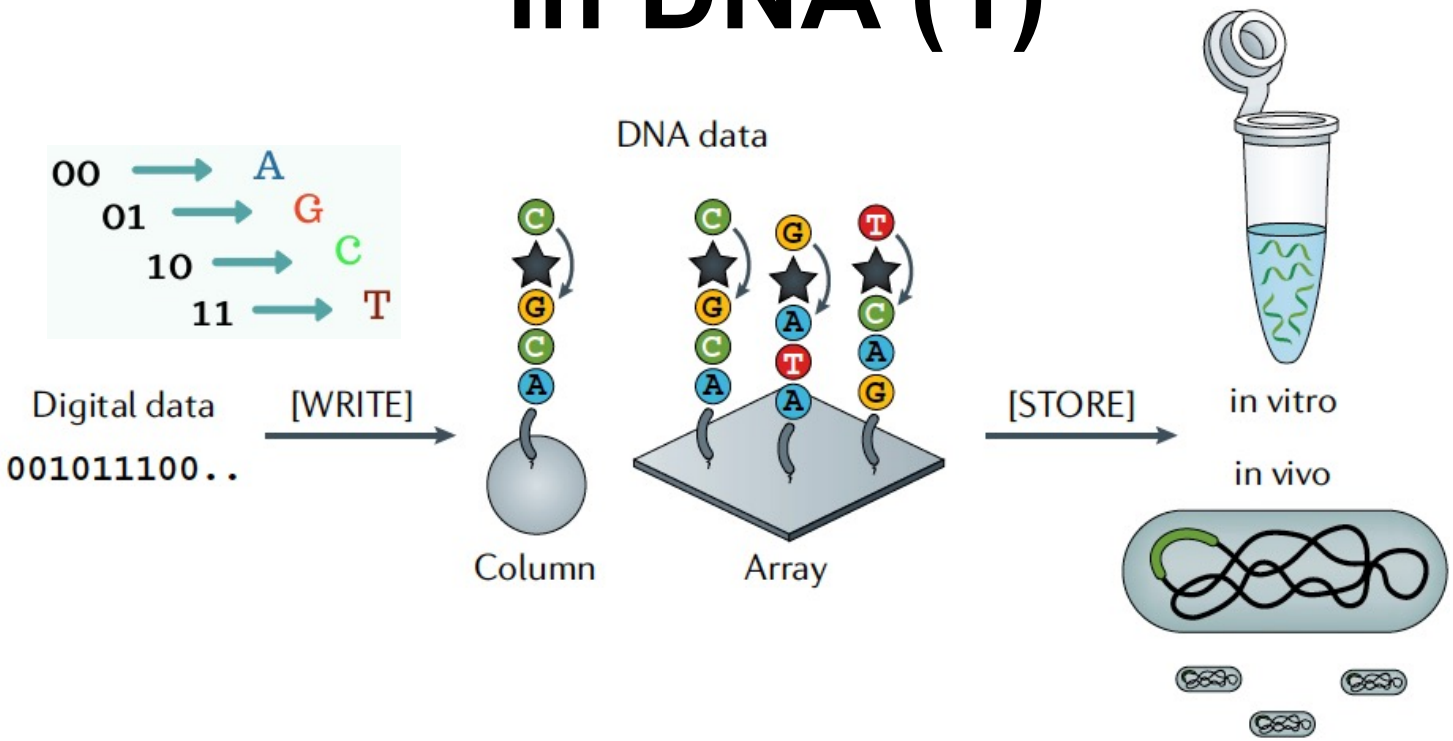
1.1 Storage Density

Classes	Examples	Achieved density	Theoretical density
Magnetic tape media		0.84 **GB**/in^2 [1]	
Optical disc media	CD, DVD, Blu-ray, *etc.*	12.5 **GB**/in^2	
Magnetic disk media	HDD (hard disk drives)	1.34 **TB**/in^2 [2]	~ 5 **TB**/in^2 [3]
Solid state media	SSD (solid-state drive)	~ 6-fold denser than HDD	
DNA		215 **PB**/g [4]	455 **EB**/g [5]

KB	10^3 byte
MB	10^6
GB	10^9
TB	10^{12}
PB	10^{15}
EB	10^{18}
ZB	10^{21}
YB	10^{24}

[1] HP LTO-6 Media Metal Particle and Barium Ferrite Archived December 22, 2015, at the Wayback Machine, HP, May 2014.
[2] Re, Mark (August 25, 2015). "Tech Talk on HDD Areal Density". Seagate.
[3] Mallary, M.; Torabi, A.; Benakli, M. *IEEE Trans. Magn.* **2002**, *38*, 1719.
[4] Erlich, Y.; Zielinski, D. *Science* **2017**, *355*, 950.
[5] Church, G.; Gao, Y.; Kosuri, S. *Science* **2012**, *337*, 1628.

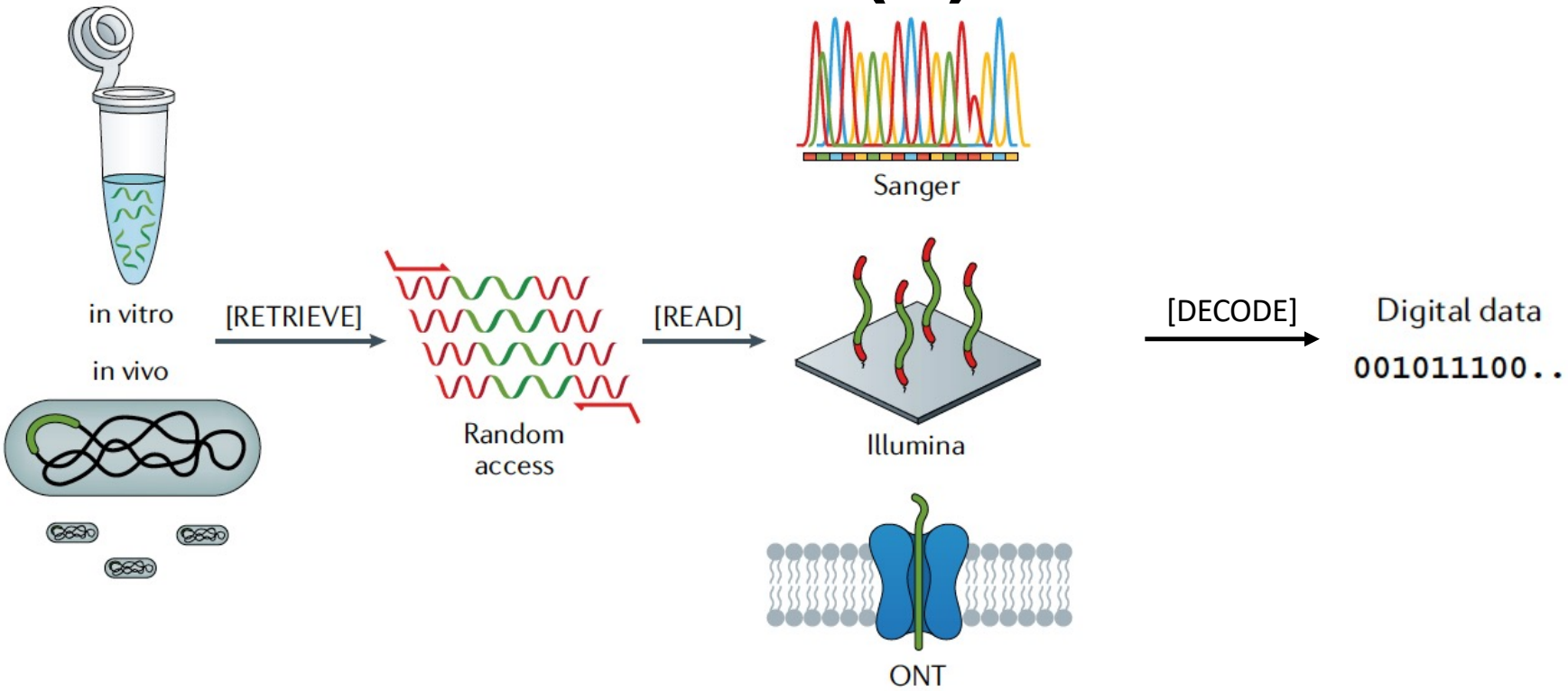
1.2.1 Major Steps of Digital Storage in DNA (1)



DNA data storage involves five major steps:
(1) **Write** (encoding): A computer algorithm maps strings of bits into DNA sequence. The resulting DNA sequences are then synthesized.
(2) **Store**: The synthesized DNA can be cloned and stored commonly in vitro, or within a biological cell (in vivo).

Edited from Ceze, L.; Nivala, J.; Strauss, K. *Nature Rev. Genet.* **2019**, *20*, 456–466.

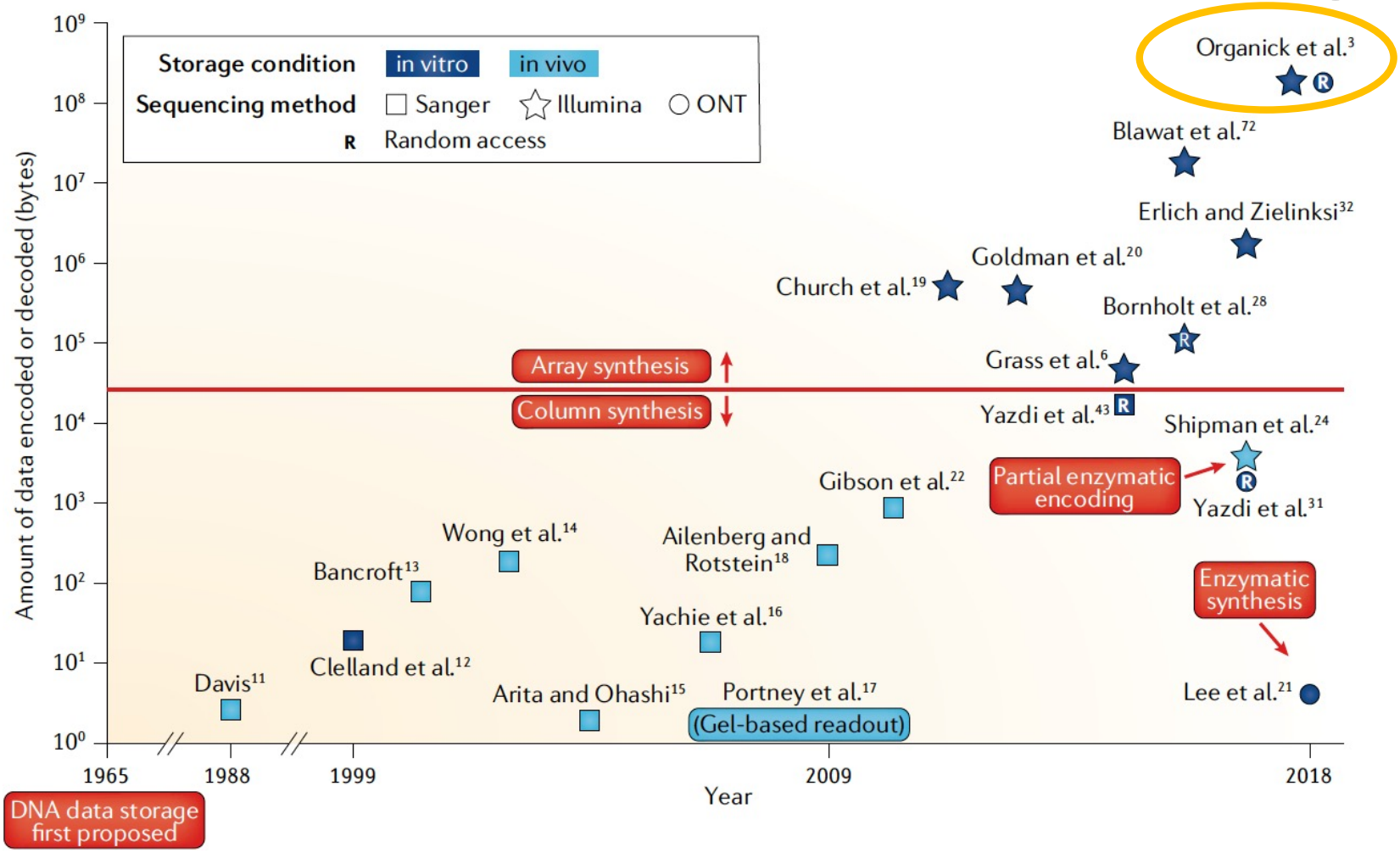
1.2.2 Major Steps of Digital Storage in DNA (2)



(3) **Retrieve**: DNA data requested to be read can be selectively retrieved from DNA pool in a process called random access (PCR enrichment).
(4) **Read**: Various sequencing machines are used to extract DNA sequence.
(5) **Decode**: Sequence detected are decoded back to binary data.

Edited from Ceze, L.; Nivala, J.; Strauss, K. *Nature Rev. Genet.* **2019**, *20*, 456–466.

1.3 Development of DNA storage



Most early work on DNA storage involved in vivo cloning. Recently, in vitro storage is the mainstream for the development of DNA synthesis.

Ceze, L.; Nivala, J.; Strauss, K. *Nature Rev. Genet.* **2019**, *20*, 456–466.

Contents

1. Introduction

2. In vitro data storage

3. Direct in vivo data storage (main paper)

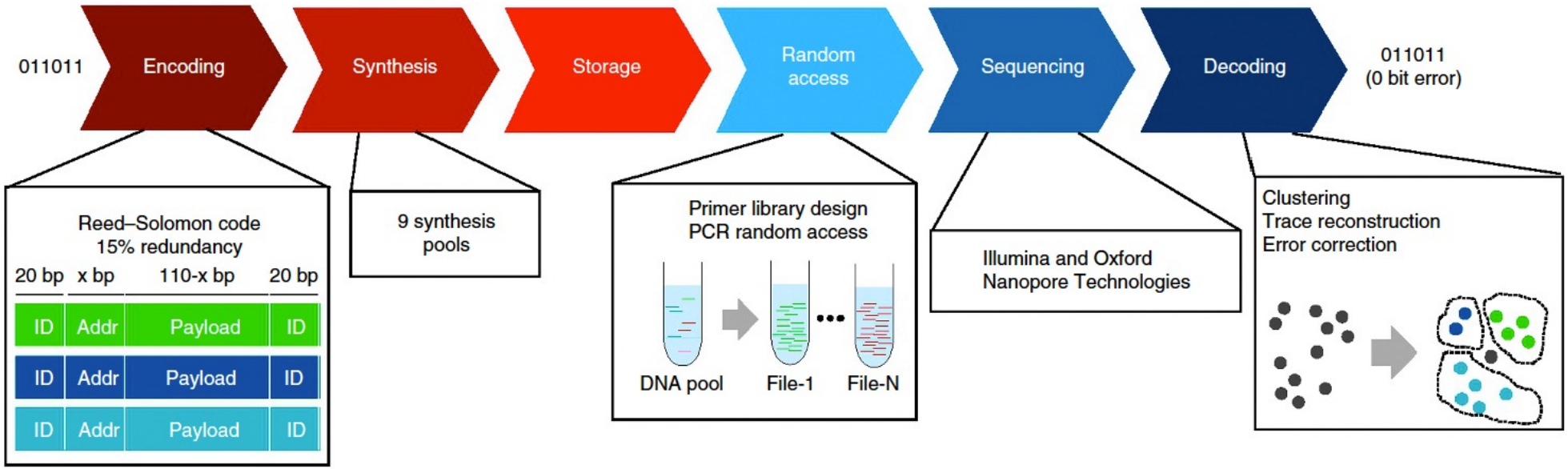
ARTICLES

nature biotechnology

Random access in large-scale DNA data storage

Lee Organick[1], Siena Dumas Ang[2], Yuan-Jyue Chen[2], Randolph Lopez[3], Sergey Yekhanin[2], Konstantin Makarychev[2,5], Miklos Z Racz[2,5], Govinda Kamath[2,5], Parikshit Gopalan[2,5], Bichlien Nguyen[2], Christopher N Takahashi[1], Sharon Newman[1,5], Hsing-Yeh Parker[2], Cyrus Rashtchian[2], Kendall Stewart[1], Gagan Gupta[2], Robert Carlson[2], John Mulligan[2], Douglas Carmean[2], Georg Seelig[1,4], Luis Ceze[1] & Karin Strauss[2]

2.1 DNA Data Storage Workflow



The workflow basically corresponds to the major steps of digital storage in DNA. To achieve random access and low error rate, primer design and redundancy were applied, respectively.

Organick, L., Ang, S., Chen, YJ. *et al. Nat. Biotech.* **2018**, 36, *242*.

2.2 Primer Library Design

Generate random 20-mer based on:
1. GC-content
2. sequence complementarity
3. long homopolymers
4. Hamming distance

19,840 sequences

→

Filter secondary structure and melting temperature

9,869 sequences

→

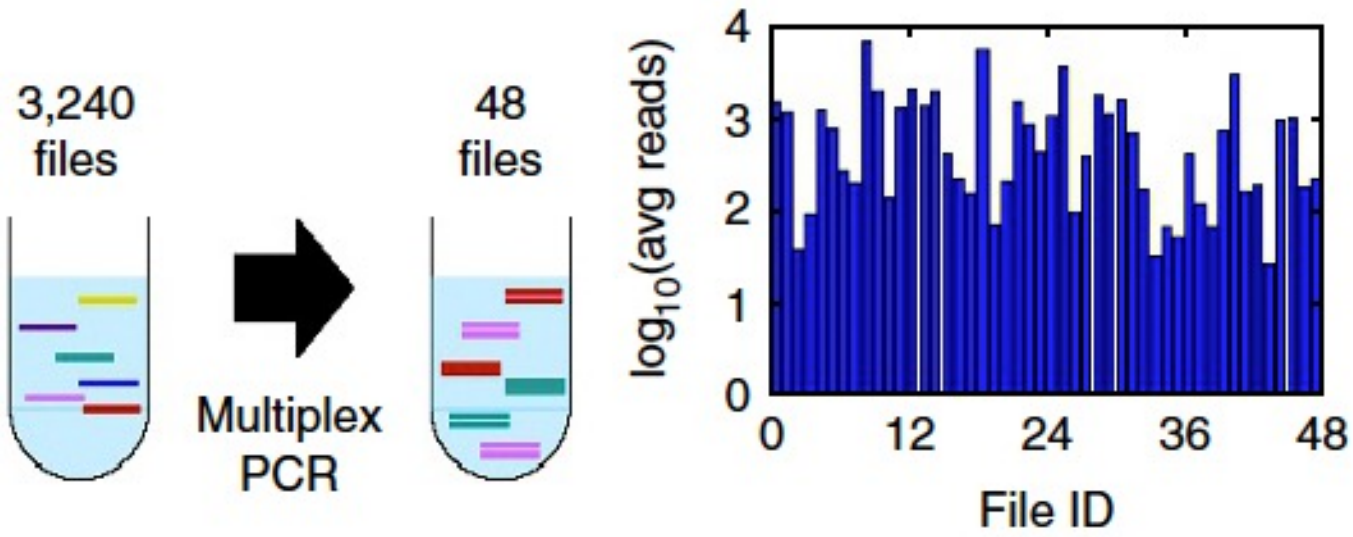
Filter similarity

5,625 sequences

Various criteria were adopted to give out unique primer sequences that meet the requirement of random access.

Organick, L., Ang, S., Chen, YJ. *et al. Nat. Biotech.* **2018**, 36, *242*.

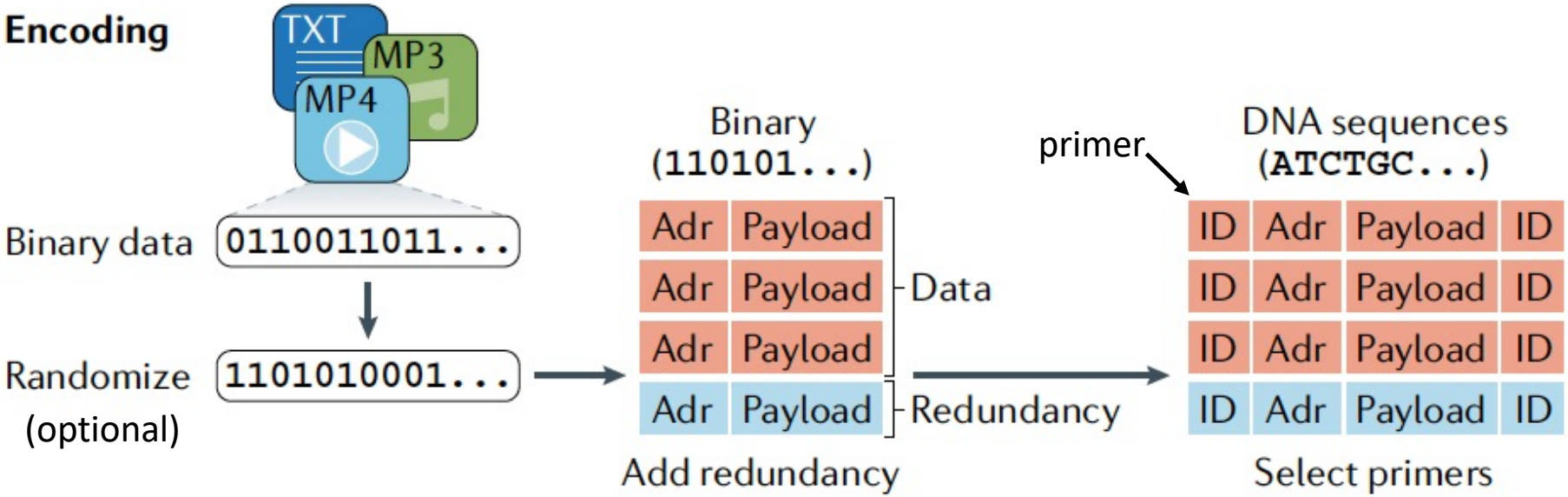
2.3 Primer Library validation



48 out of 3,240 sequences were randomly selected for amplification. The candidate primers were validated experimentally.

Organick, L., Ang, S., Chen, YJ. *et al. Nat. Biotech.* **2018**, 36, *242*.

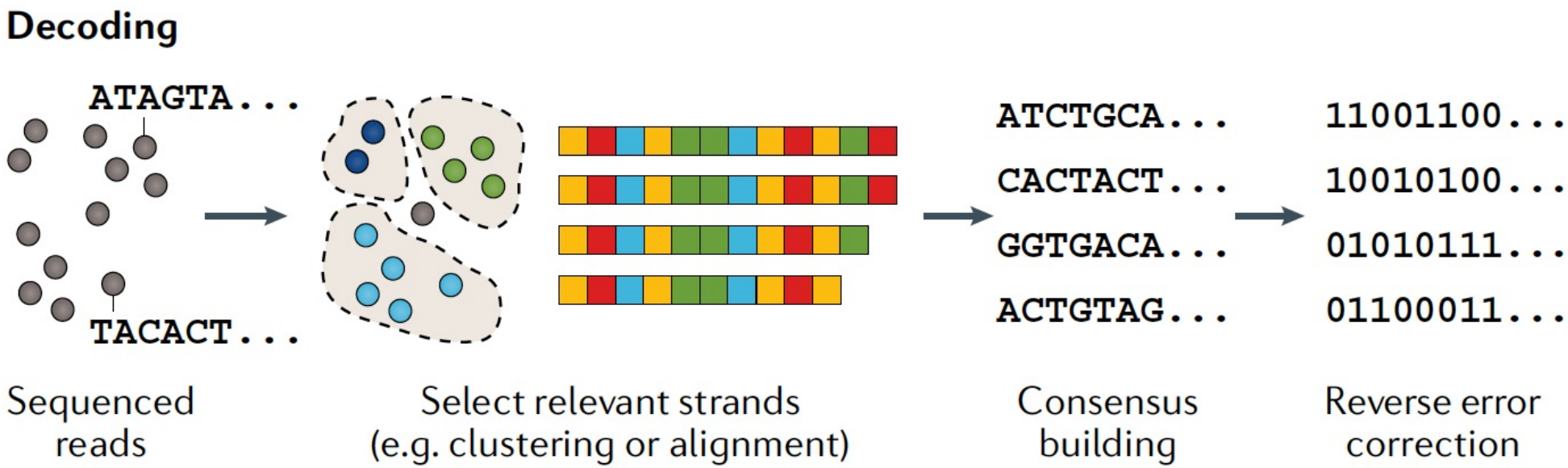
2.4 Encoding of DNA Storage



The binary data are partitioned into small bit sequences (payload) with sequence numbers (addressing information, adr). Redundancy is added for error correction. After conversion to DNA sequence, primer target sites are added for data selectivity.

Organick, L., Ang, S., Chen, YJ. *et al. Nat. Biotech.* **2018**, 36, *242*.
Ceze, L.; Nivala, J.; Strauss, K. *Nature Rev. Genet.* **2019**, *20*, 456–466.

2.5 Decoding of DNA Storage



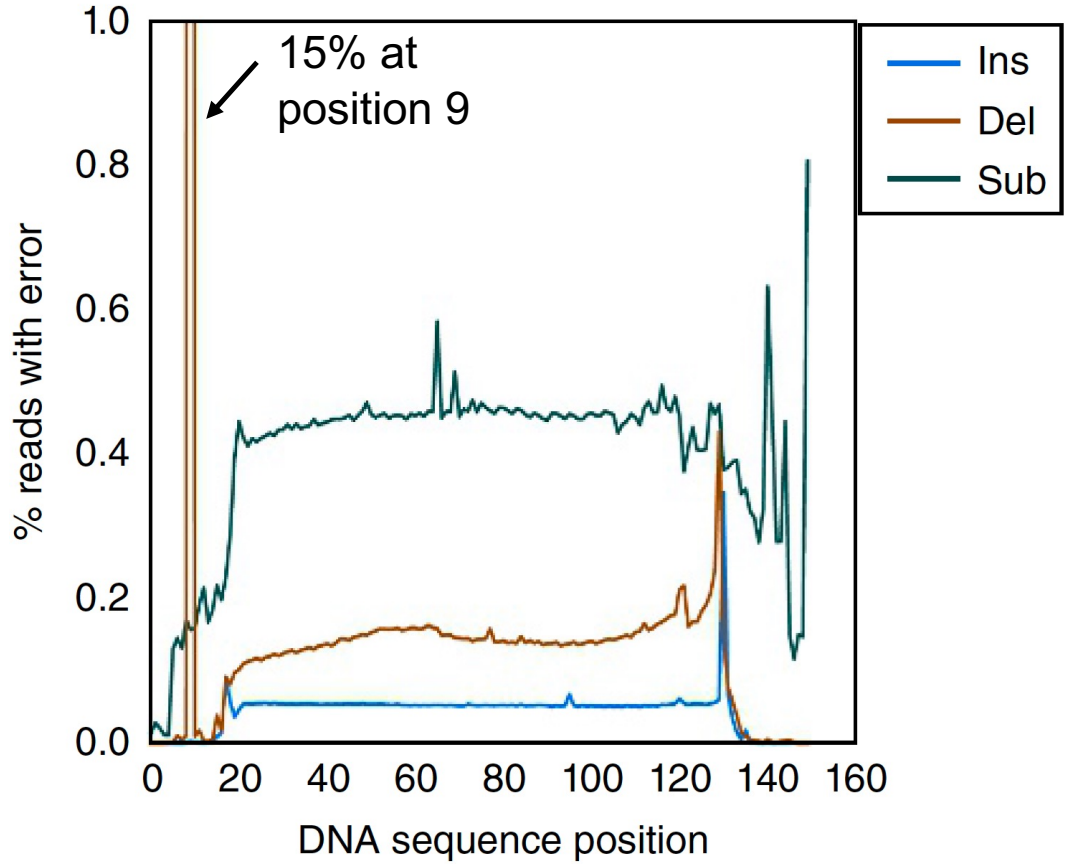
The decoding process starts by clustering reads based on similarity. Then a consensus is to be found between the sequences in each cluster to reconstruct the original sequences. Finally, the sequence read are decoded back to digital data.

Organick, L., Ang, S., Chen, YJ. *et al. Nat. Biotech.* **2018**, 36, *242*.
Ceze, L.; Nivala, J.; Strauss, K. *Nature Rev. Genet.* **2019**, *20*, 456–466.

2.6.1 Error Analysis (1)

DNA pool

PCR and sequencing

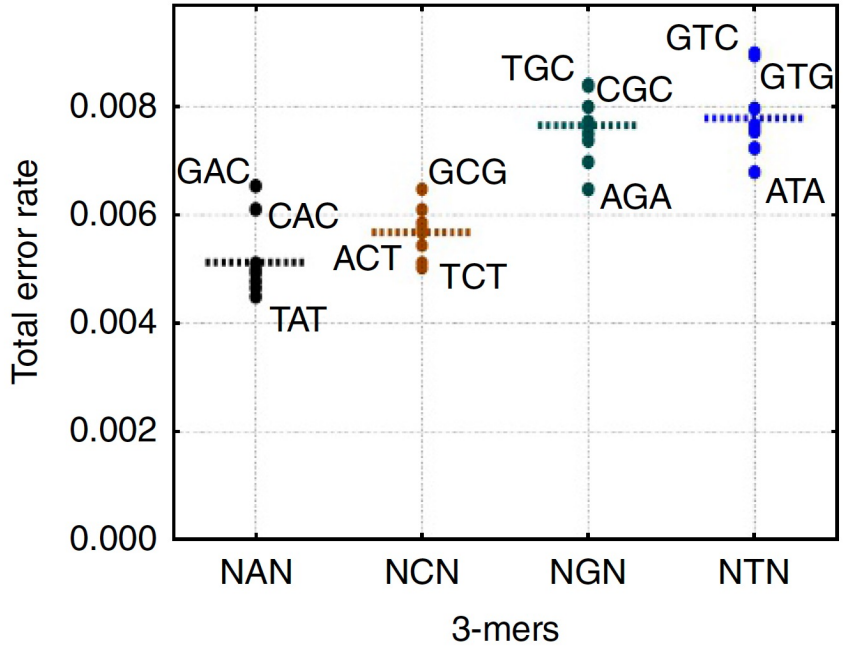


A file of interest is accessed via PCR and sequencing from a stored DNA pool. Error rates of insertion, deletion, and substitution were analyzed.

Organick, L., Ang, S., Chen, YJ. *et al. Nat. Biotech.* **2018**, 36, *242*.

2.6.2 Error Analysis (2)

	Insertion rate	Deletion rate	Substitution rate	% Total reads
A	1.1×10^{-4}	4.1×10^{-4}	7.5×10^{-4}	24.6
C	9.1×10^{-5}	3.6×10^{-4}	9.8×10^{-4}	25.1
G	2.5×10^{-4}	3.8×10^{-4}	1.3×10^{-3}	25.1
T	8.4×10^{-5}	3.7×10^{-4}	1.5×10^{-3}	25.2
Total	5.4×10^{-4}	1.5×10^{-3}	4.5×10^{-3}	100.0



Insertion and substitution errors are biased toward certain base types. Almost half of the insertions are associated with type G, and about a third of the substitution are associated with type T. Error rates concerning 3-mers were also analyzed, showing type G, T with higher rates.

Organick, L., Ang, S., Chen, YJ. *et al. Nat. Biotech.* **2018**, 36, *242*.

2.7 Short Summary

In vitro data storage in DNA

1. High density

(theoretically 455 EB/g)

2. Error tolerance

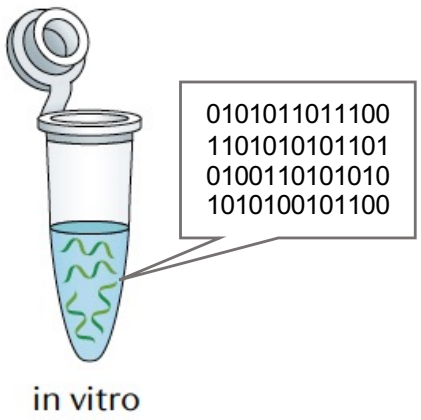
(due to redundancy, error-correcting algorithm)

3. Long-term durability

1. High cost

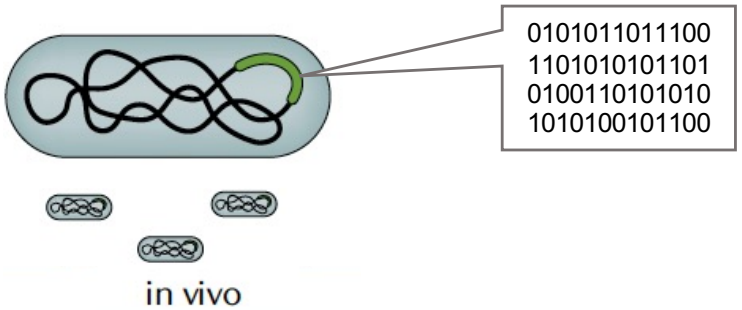
2. Low write/ read speed

(Limited by the speed of DNA synthesis and sequencing)



in vivo data storage in DNA

Digital information is stored in cells indirectly by inducing synthesized DNA segment into the genome.



Organick, L., Ang, S., Chen, YJ. *et al. Nat. Biotech.* **2018**, 36, *242*.

Contents

1. Introduction

2. In vitro data storage

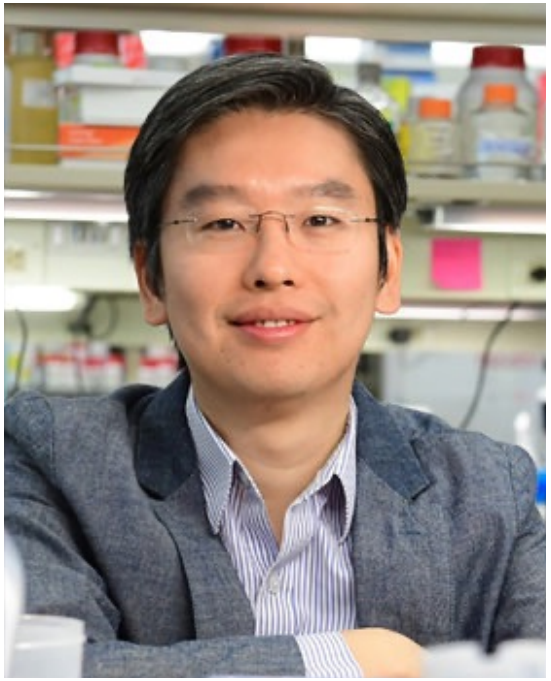
3. Direct in vivo data storage (main paper)

ARTICLES
https://doi.org/10.1038/s41589-020-00711-4
nature chemical biology

Robust direct digital-to-biological data storage in living cells

Sung Sun Yim[1], Ross M. McBee[1,2], Alan M. Song[1], Yiming Huang[1,3], Ravi U. Sheth[1,3] and Harris H. Wang[1,4]

Harris H. Wang



2001.9 - 2005.6
Massachusetts Institute of Technology
B.S. in Physics B.S. in Applied Mathematics
Minor in Biomedical Engineering

2005.9 - 2010.6
Harvard University
Ph.D. in Biophysics
Harvard-MIT Health Sciences and Technology
Joint Ph.D. in Medical Engineering Medical Physics
(Advisor: George Church)

2011.9 - 2013.2
Harvard Medical School
Instructor of Systems Biology (PI status)

2013.3 - 2020.6
Columbia University Irving Medical Center
Assistant Professor of Systems Biology

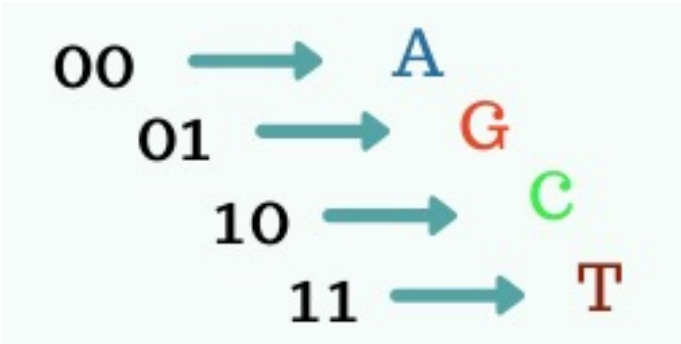
2020.7 - present
Columbia University Irving Medical Center
Associate Professor of Systems Biology (with tenure)

Research interests:
- Systems biology
- Synthetic biology

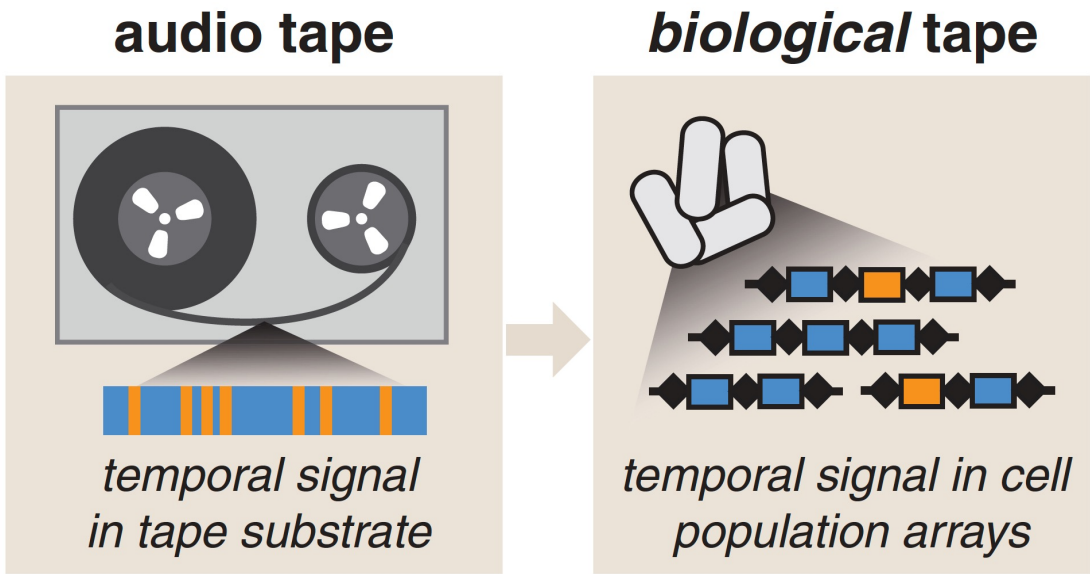
Organick, L., Ang, S., Chen, YJ. *et al. Nat. Biotech.* **2018**, 36, *242*.

3.1 Concept of Direct in vivo Storage

in vitro storage in DNA
(base pair -> information)



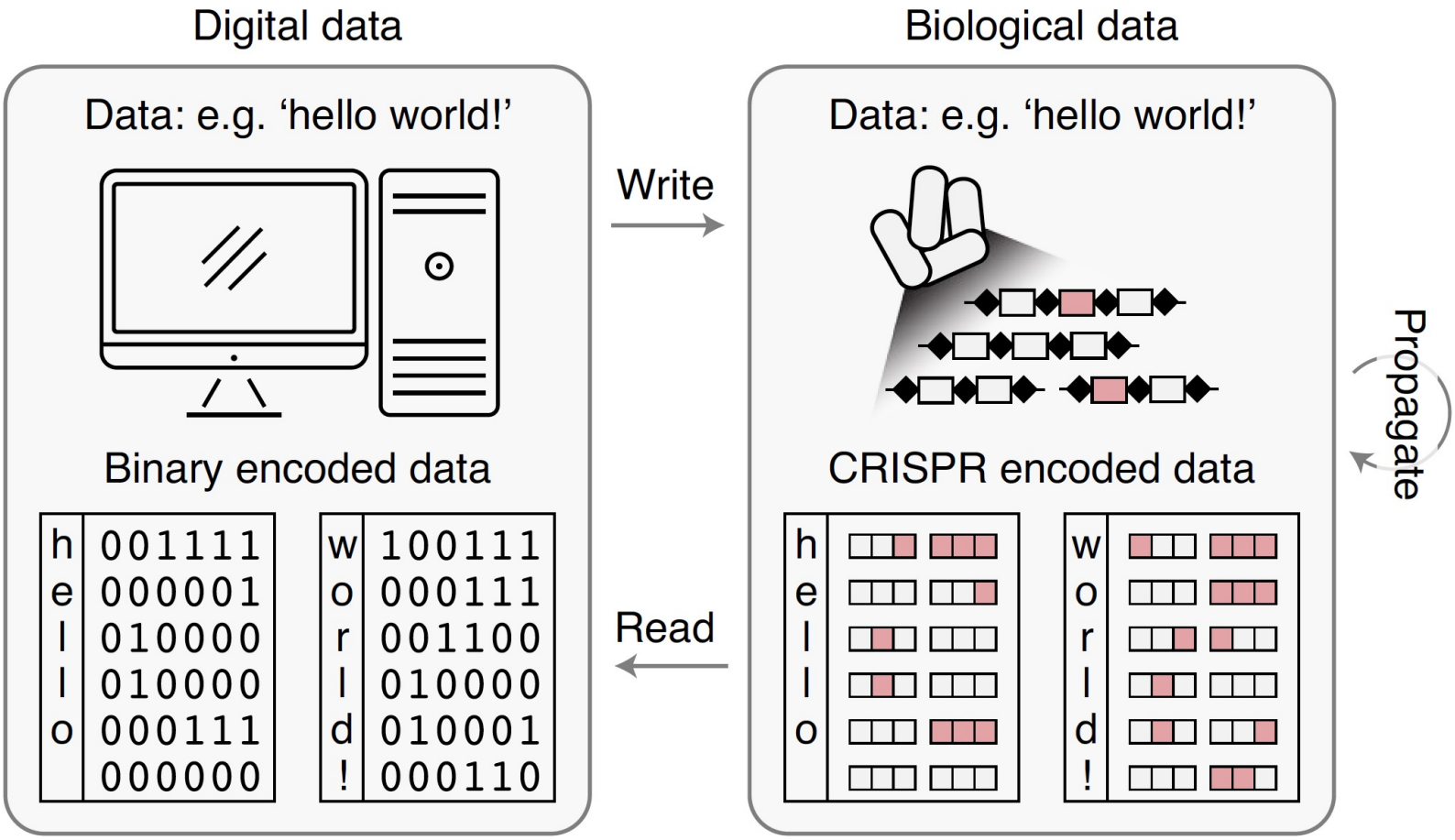
direct in vivo storage in DNA
(genome sequence -> information)



Due to the difficulty of editing every single oligonucleotide in the genome of one cell, the concept of direct in vivo storage in DNA mimics an audio tape in which the induced spacer works as signals. This concept achieved direct induction of information but resulted much lower density.

Sheth, R.; Yim, S.; Wu, F.; Wang, H. *Science* **2017**, *358*, 1457.

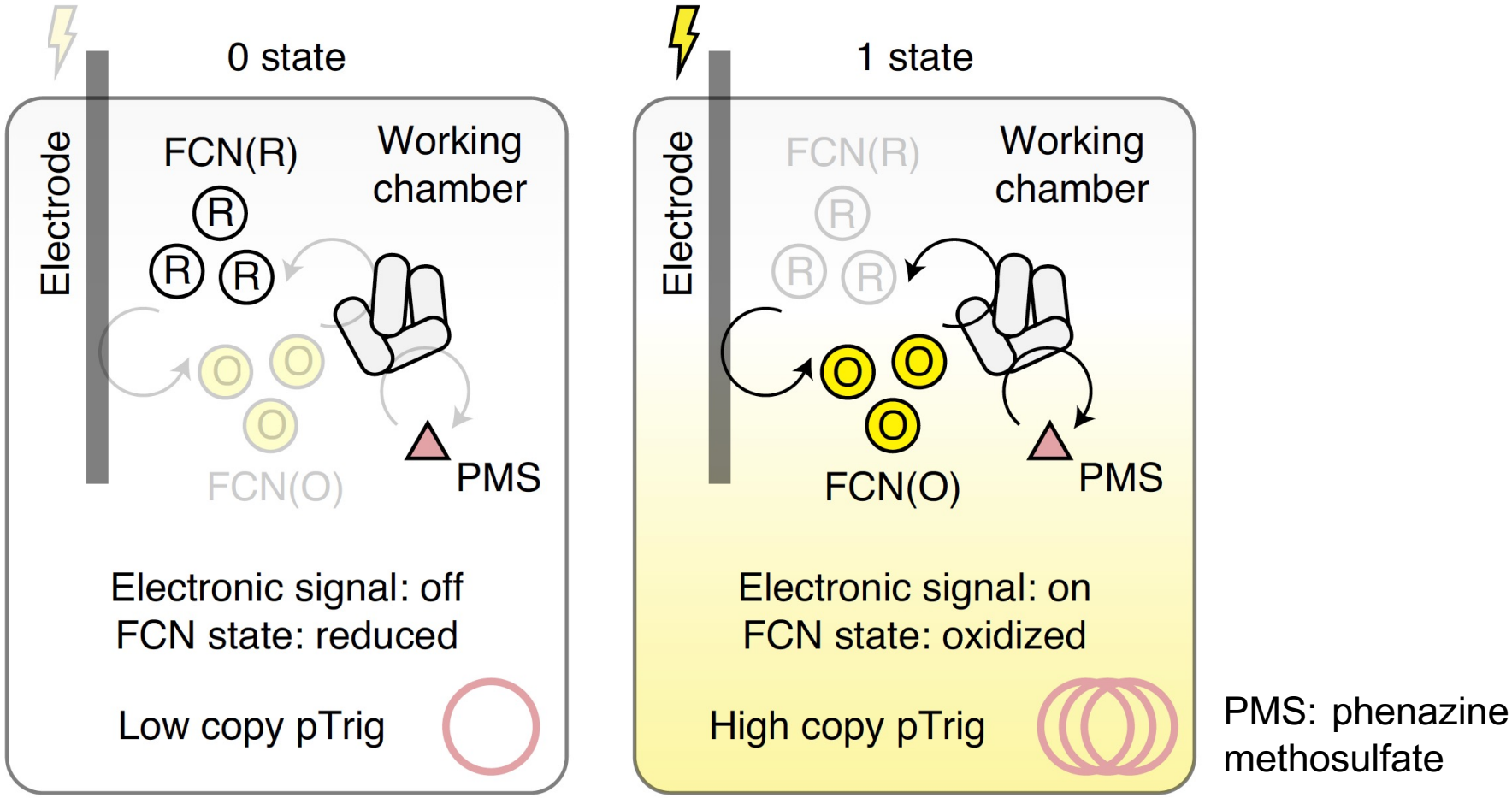
3.2 From Digital to Biological Data



Digital data (3-bit binary data) can be encoded into bacteria genome in an 'audio tape' manner.

Yim, S.; McBee, R.; Song, A.; Huang, Y.; Sheth, R.; Wang, H. *Nat. Chem. Biol.* **2021**, *17*, 246.

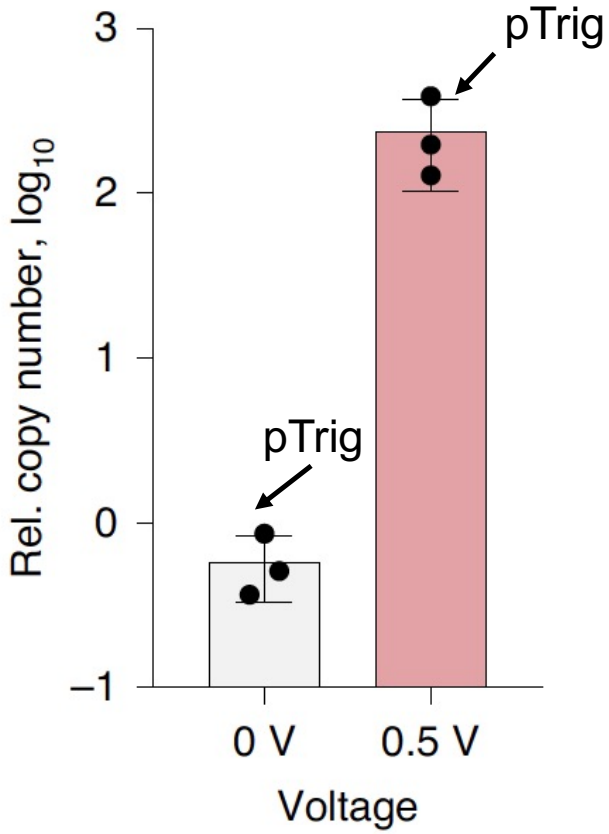
3.3 Data Recording by Electrical Stimulation

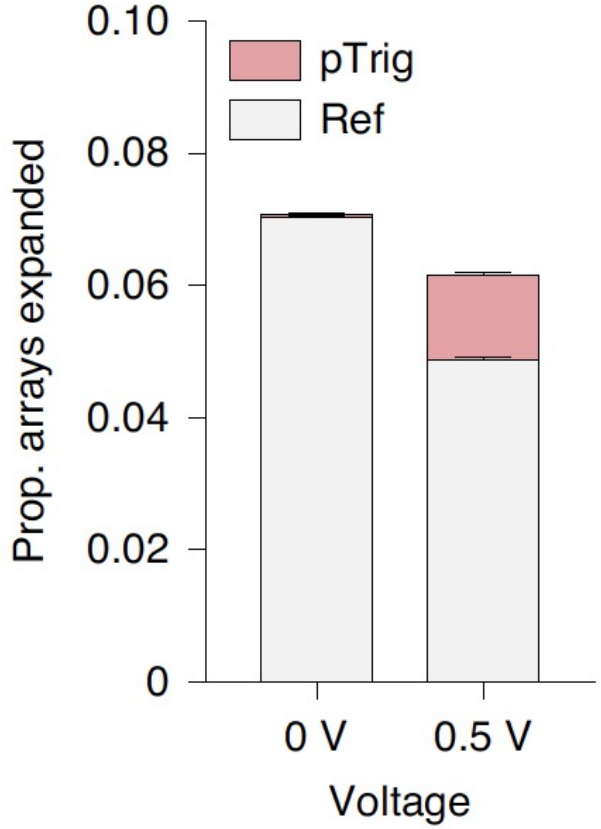


The concept of DRIVES (Data Recording In Vivo by Electrical Stimulation) involves the gene expression of trigger DNA (pTrig) regulated through electrical stimuli.

Yim, S.; McBee, R.; Song, A.; Huang, Y.; Sheth, R.; Wang, H. *Nat. Chem. Biol.* **2021**, *17*, 246.

3.4 Copy Number of pTrig

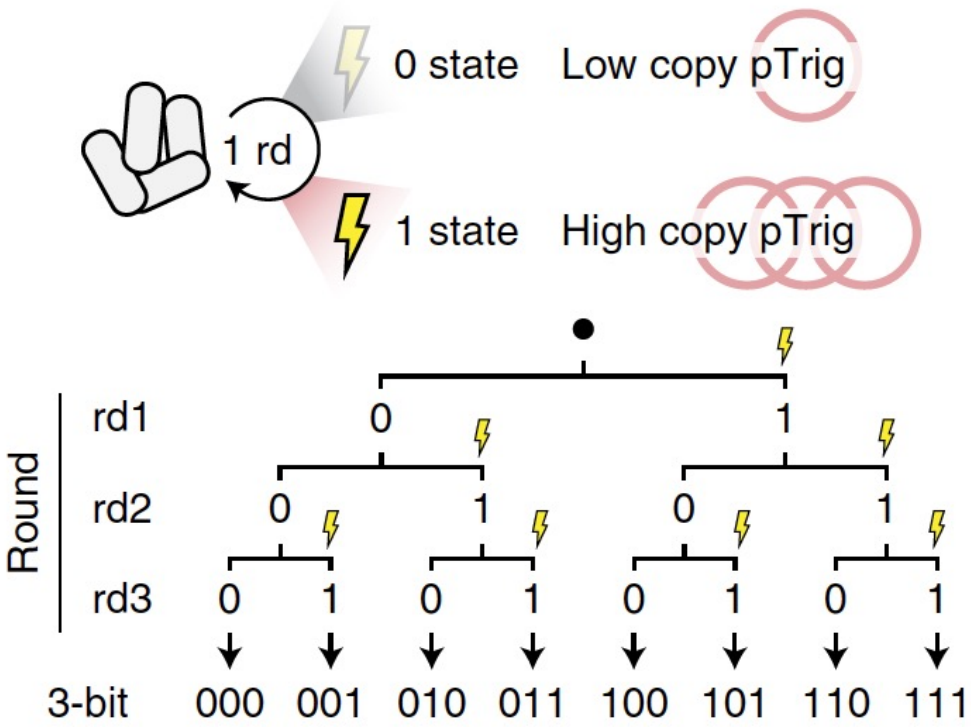


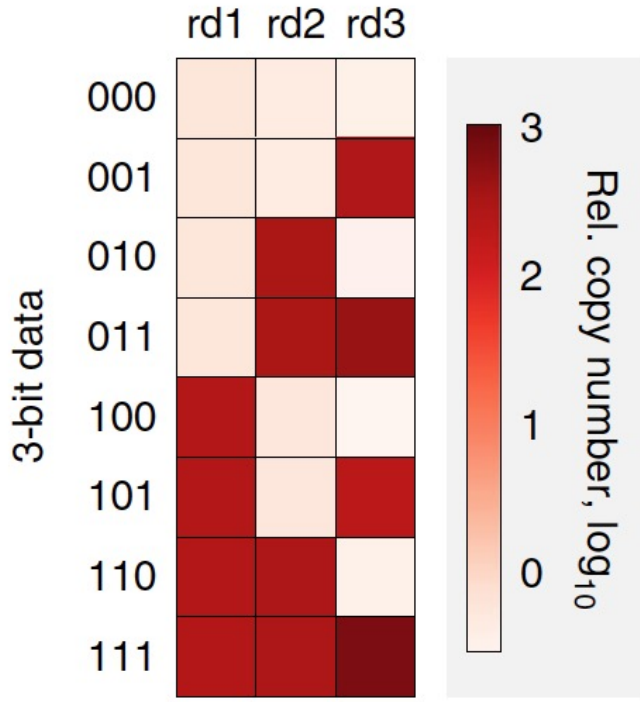


Over 400-fold increase of pTrig copy number was observed in the presence of electrical signal (left).
The significant difference in the proportion was also demonstrated (right).

Yim, S.; McBee, R.; Song, A.; Huang, Y.; Sheth, R.; Wang, H. *Nat. Chem. Biol.* **2021**, *17*, 246.

3.5 3-bit Binary Profile

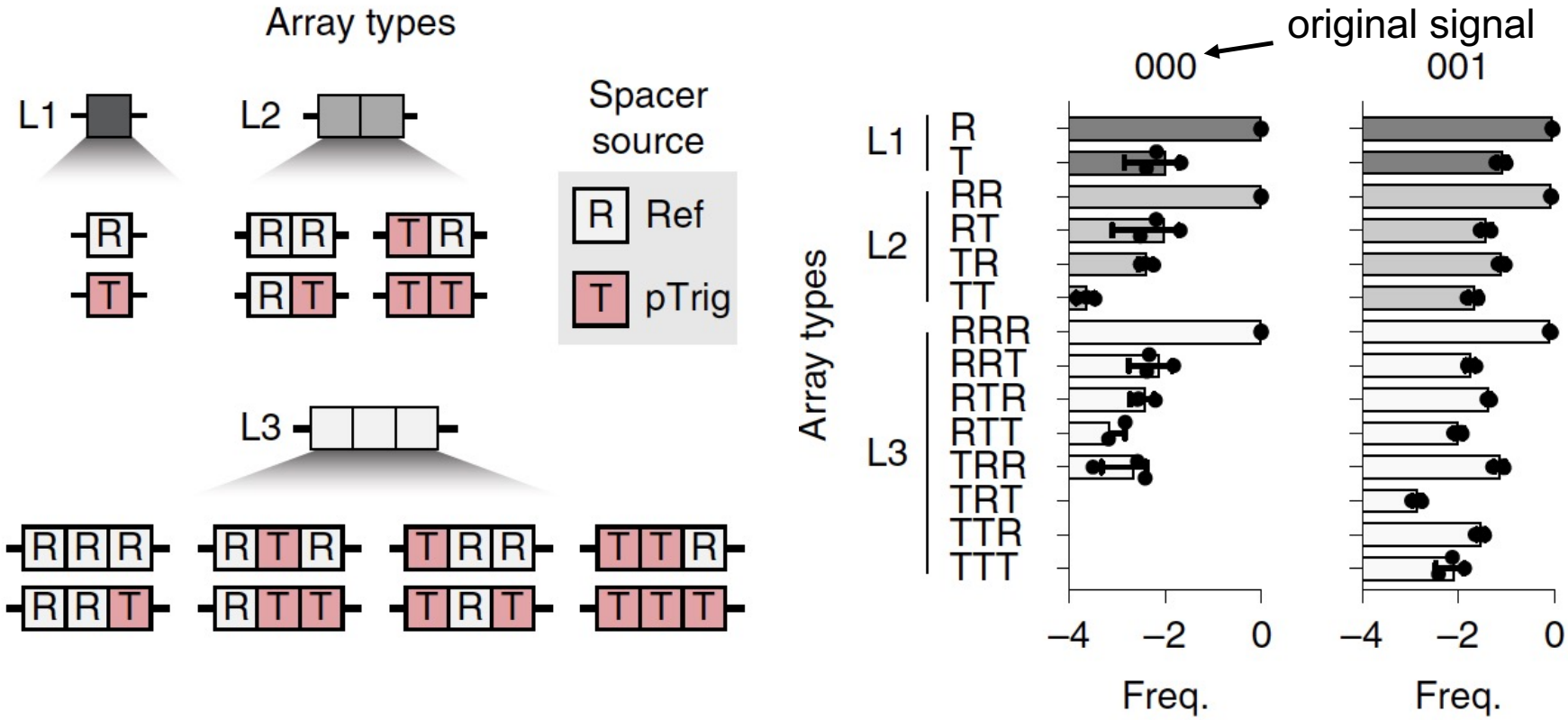




Cells were subjected to electrical signals over three sequential rounds in order to constitute all eight possible 3-bit binary profiles marked by the copy numbers of pTrig.

Yim, S.; McBee, R.; Song, A.; Huang, Y.; Sheth, R.; Wang, H. *Nat. Chem. Biol.* **2021**, *17*, 246.

3.6 Frequency Analysis of Array Types

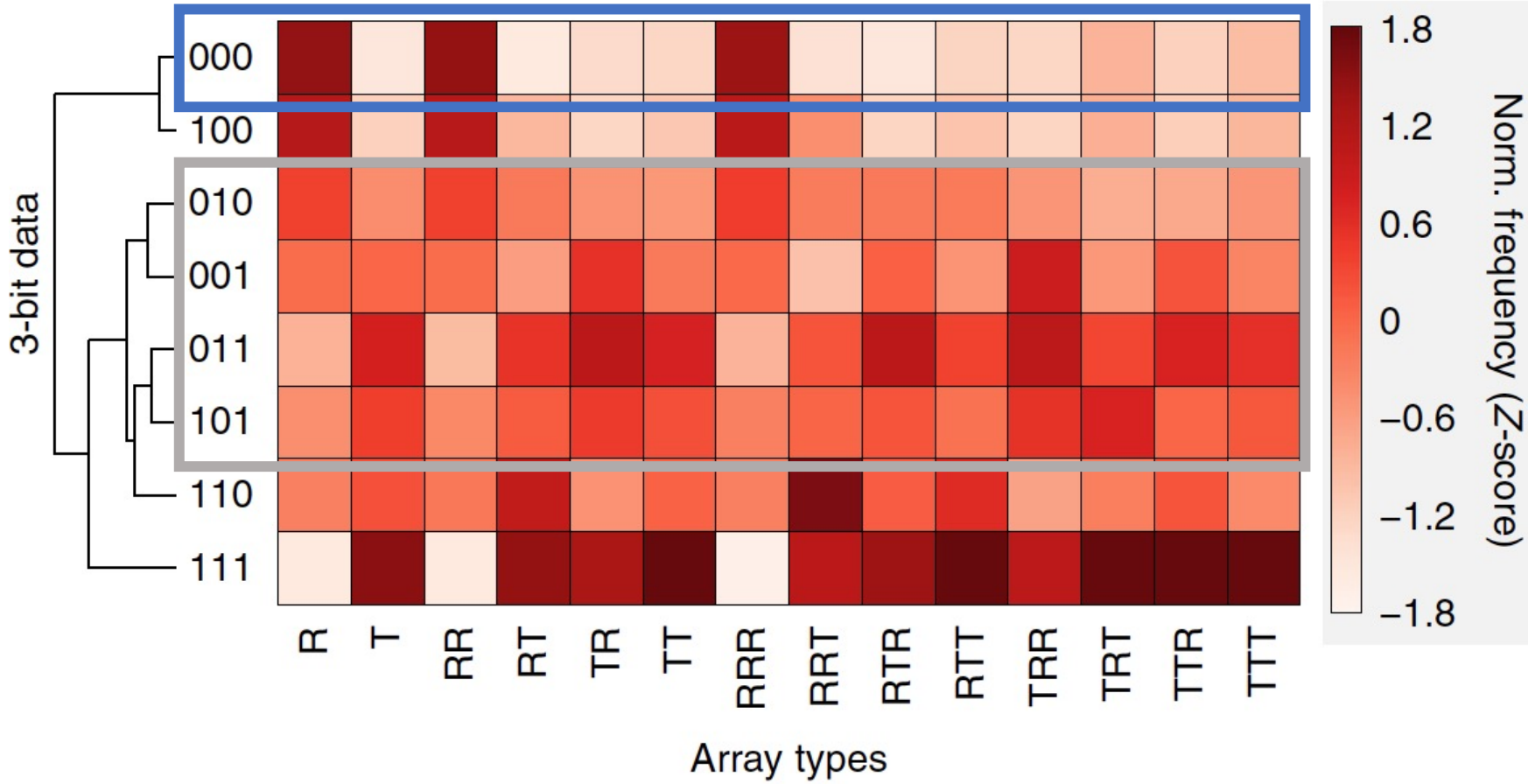


Sequencing of the genome DNA alone is also not appliable to tell the original electrical signals. For a given array length (L1, L2, and L3), a frequency distribution of reference spacers (R) and trigger spacers (T) was analyzed. The input signals resulted different frequency distribution upon the array types.

Yim, S.; McBee, R.; Song, A.; Huang, Y.; Sheth, R.; Wang, H. *Nat. Chem. Biol.* **2021**, *17*, 246.

3.7 Array-Type Frequency Profile

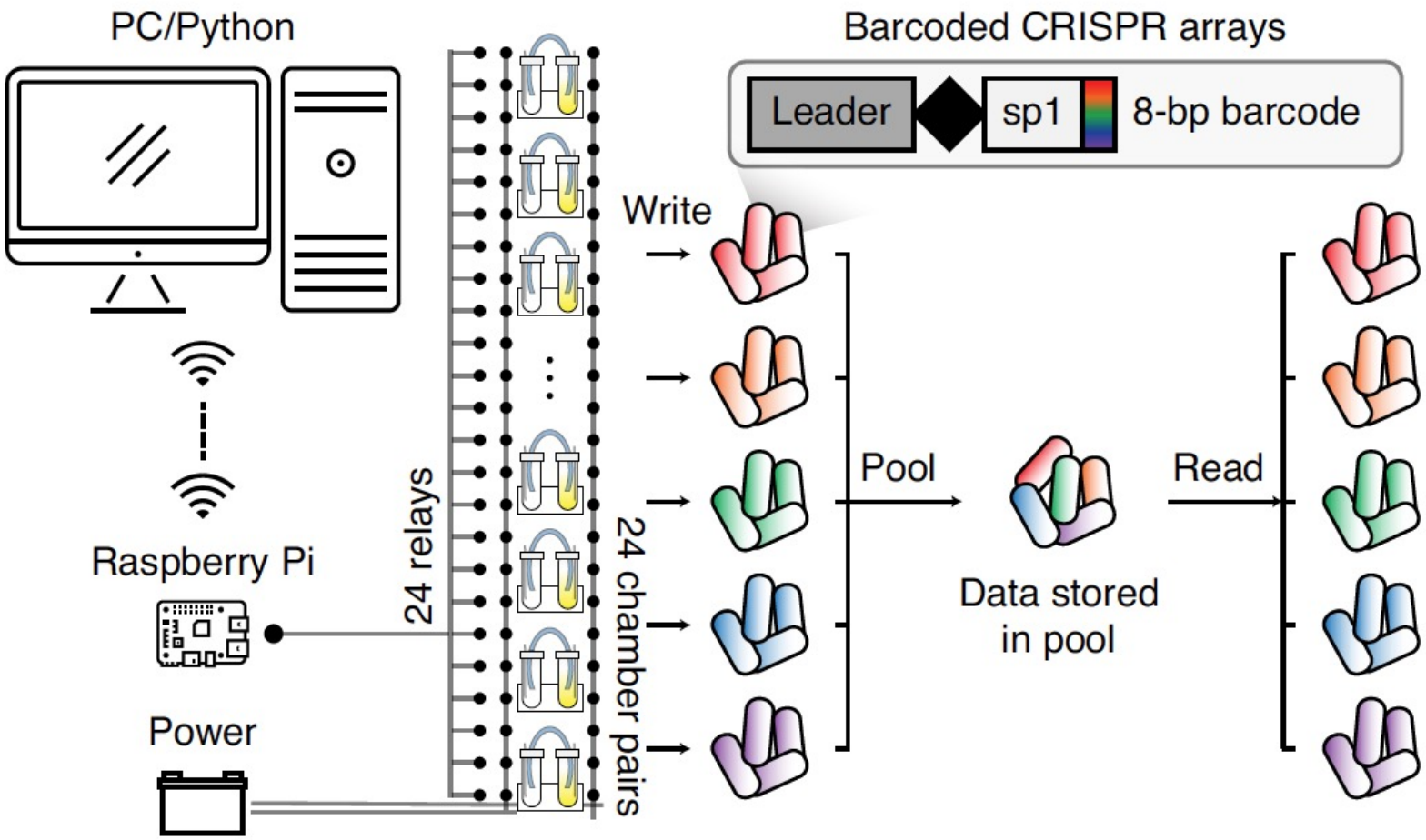
signal – distribution pattern



Whether these array-type frequencies could differentiate between different input signals was tested. The authors demonstrated that digital data can be stored directly through electrical stimulation and the resulting frequency profiles can be used to recover the stored data.

Yim, S.; McBee, R.; Song, A.; Huang, Y.; Sheth, R.; Wang, H. *Nat. Chem. Biol.* **2021**, *17*, 246.

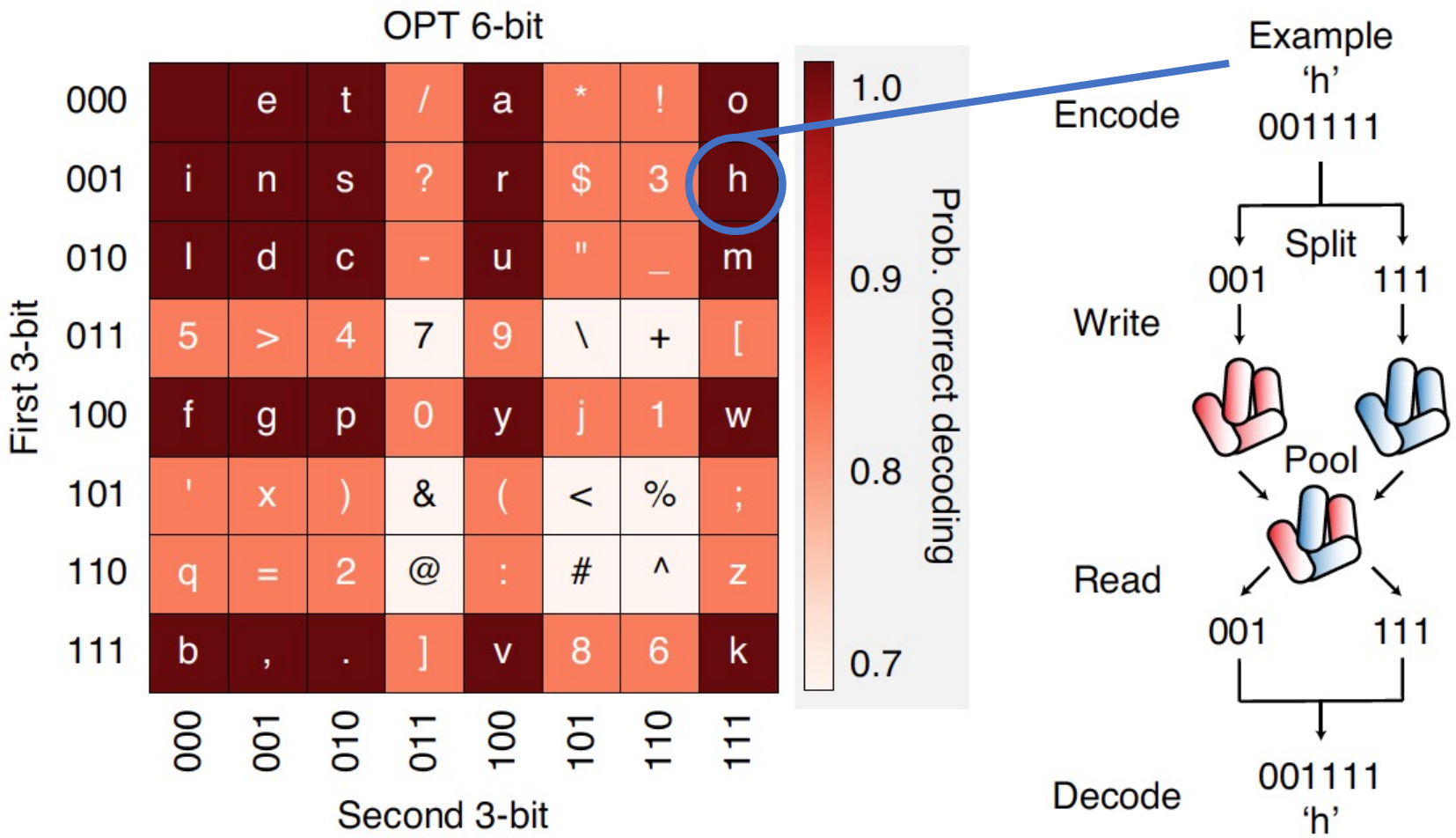
3.8 DRIVES Set-up



A multiplexing strategy to write binary data across multiple barcoded cell populations in parallel was devised.

Yim, S.; McBee, R.; Song, A.; Huang, Y.; Sheth, R.; Wang, H. *Nat. Chem. Biol.* **2021**, *17*, 246.

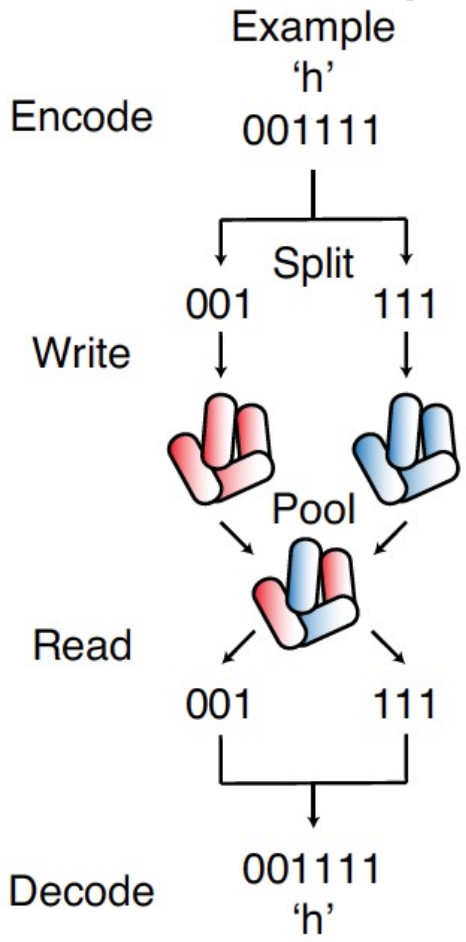
3.9 6-bit Character Table

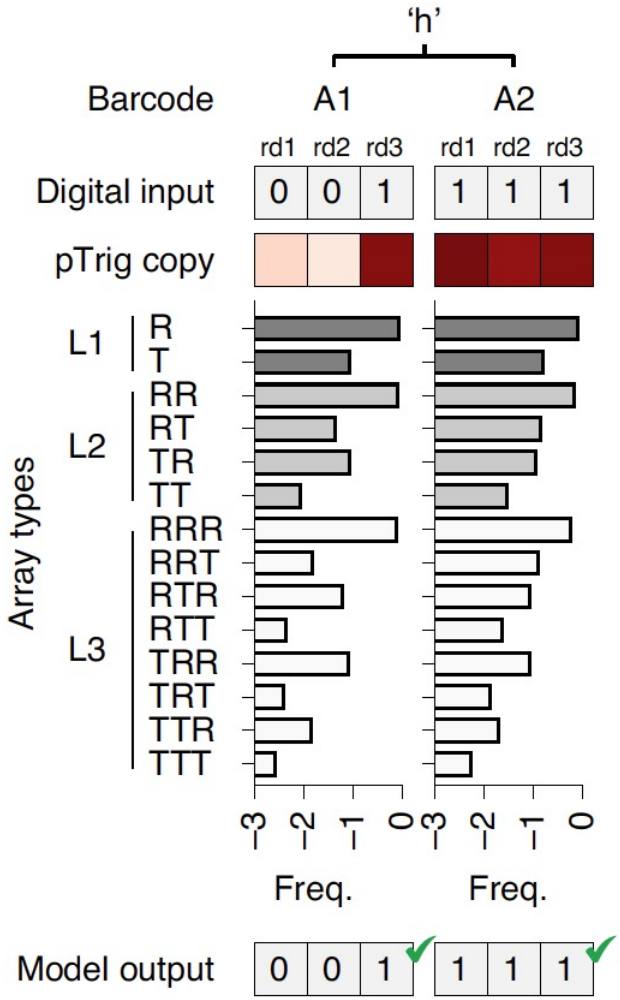


Characters are encoded into 2 sets of 3-bit binary data according to the character table. One character is thus split into two barcoded cell populations.

Yim, S.; McBee, R.; Song, A.; Huang, Y.; Sheth, R.; Wang, H. *Nat. Chem. Biol.* **2021**, *17*, 246.

3.10 Decoding of Stored Character

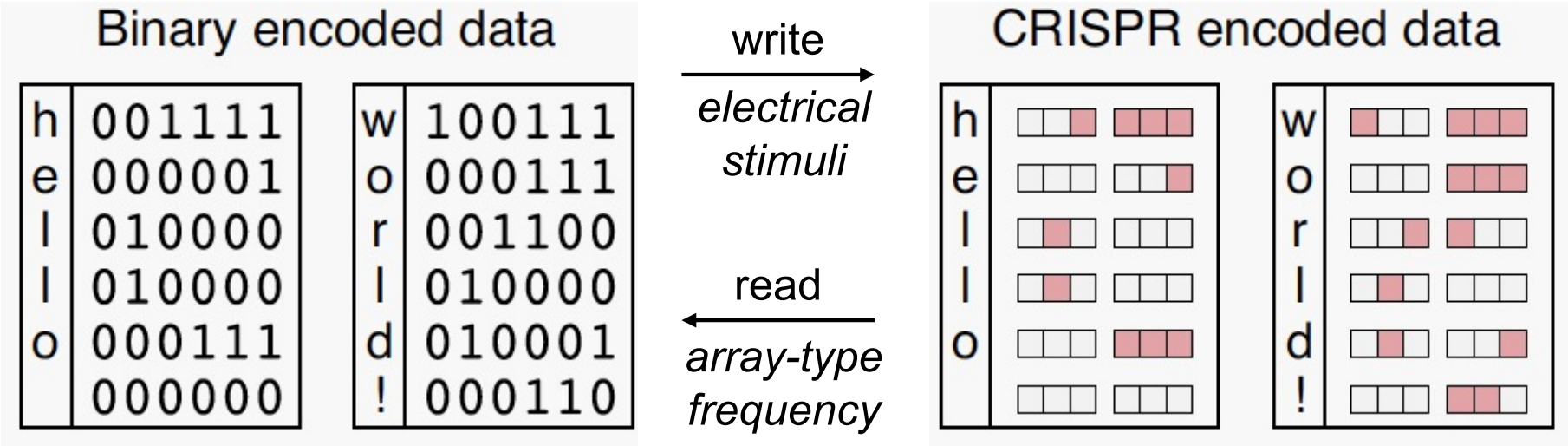




In experiment, two sets of cells are sequenced and analyzed by frequency distribution which in return output the original signals. Then, the signals were decoded to the character "h". This result successfully proved the direct in vivo storage method by electrical stimulation.

Yim, S.; McBee, R.; Song, A.; Huang, Y.; Sheth, R.; Wang, H. *Nat. Chem. Biol.* **2021**, *17*, 246.

3.11 Summary



The 'DRIVES' (data recording in vivo by electrical stimulation) first managed to encode digital data directly into the living cells without the need to synthesize DNA in vitro.

1. High cost

2. Hard to write/ read

3. Much lower density

1. Protection from degradation

2. New direction for DNA storage

.....

Yim, S.; McBee, R.; Song, A.; Huang, Y.; Sheth, R.; Wang, H. *Nat. Chem. Biol.* **2021**, *17*, 246.