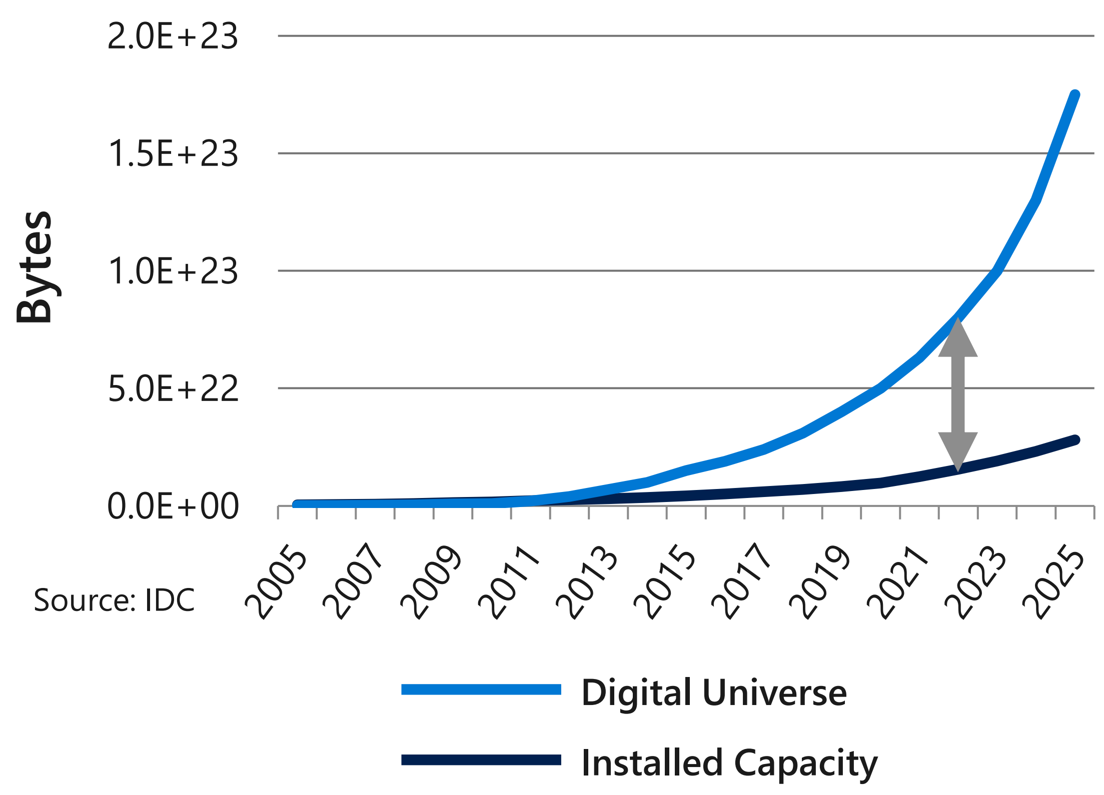
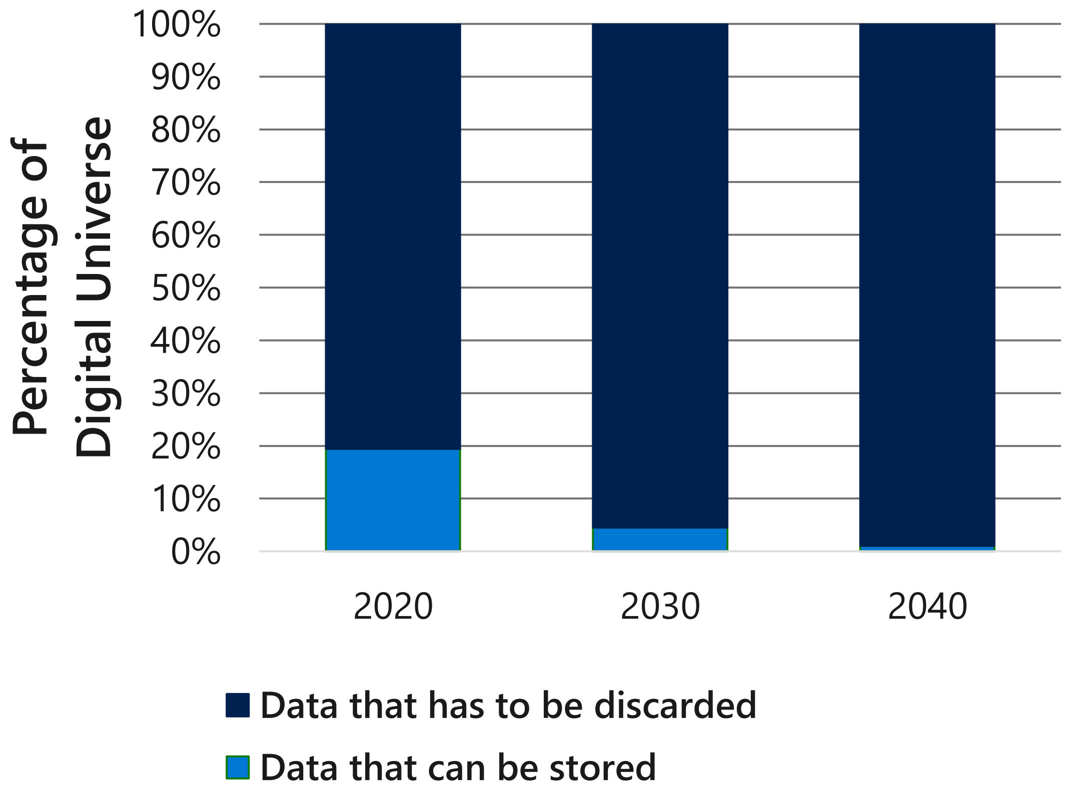
# DNA Data Storage and Near-Molecule Processing for the Yottabyte Era

Karin Strauss, Microsoft Research
Luis Ceze, University of Washington

# Storage capacity is growing too slowly



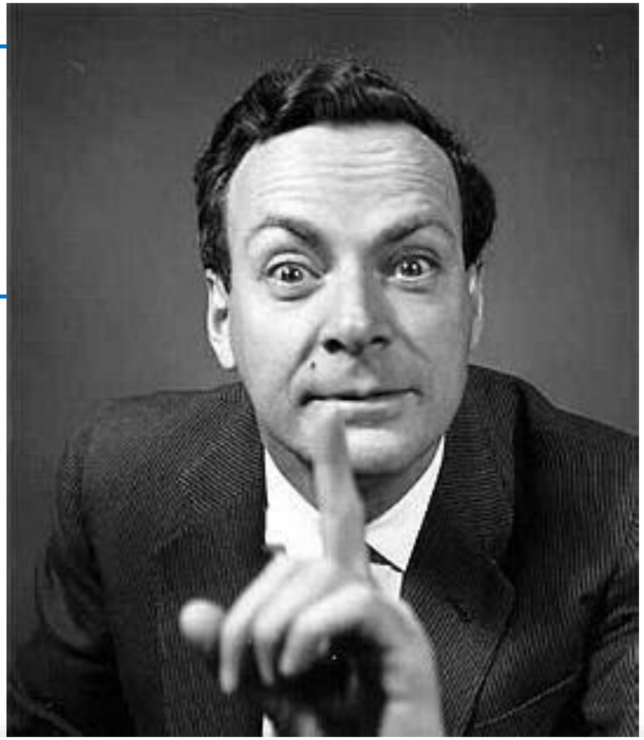
# When Moore met Feynman

The number of transistors doubles every 18-24 months

The industry roadmaps are based on that continued rate of improvement

Arrange the atoms the way we want

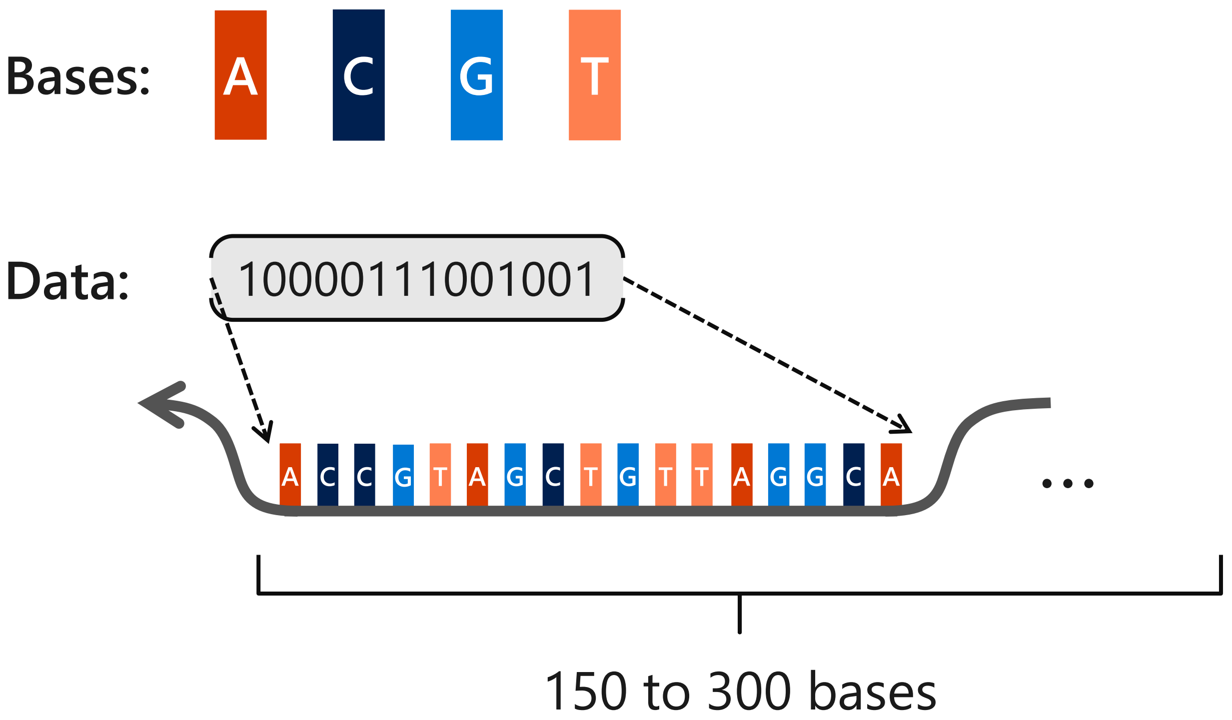
DNA molecules use approximately 50 atoms for one bit



## Let's store data in DNA!

Norbert Wiener, U.S. News and World Rep, 1964; Mikhail Neiman, Radiotechnika, 1965; Church et al, Science, 2012; Goldman et al, Nature, 2013

# DNA data storage basics




**Store data in synthetic DNA strands**

**Simple mapping:**

| Bits | Base |
|---|---|
| 00 | A |
| 01 | C |
| 10 | G |
| 11 | T |


0.5
0
Credit: Tara Brown Photography/University of Washington

# Dense, really dense

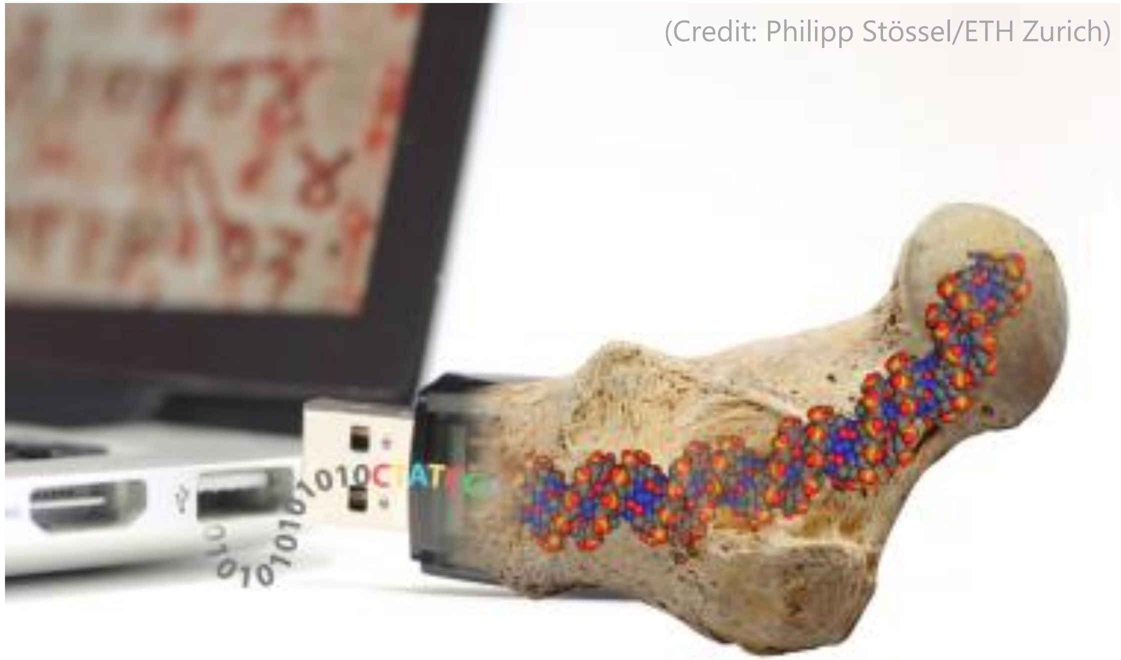
**Cold Storage**: 1EB
**Size**: Two Walmart Supercenters

vs.

It's here!

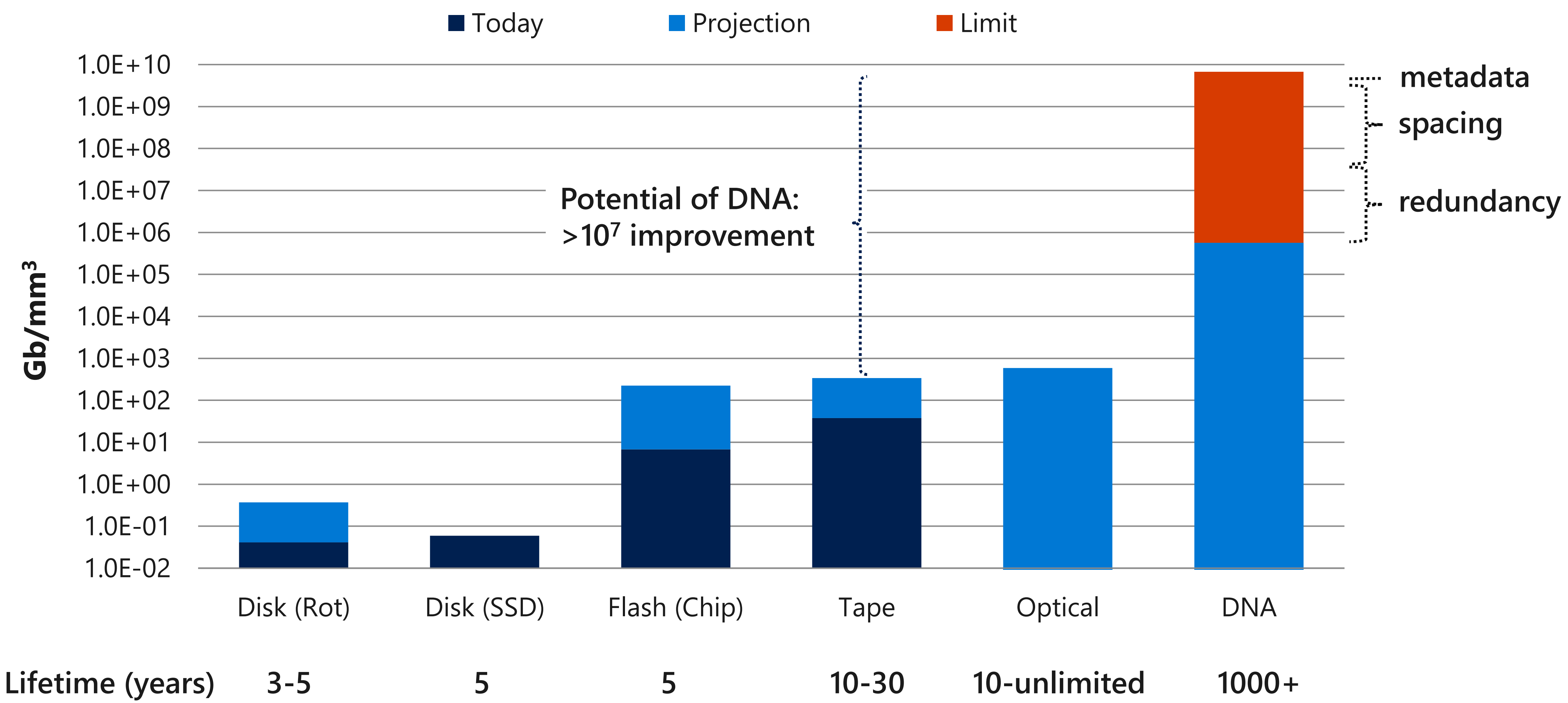
# Information durability

## DNA “synthetic fossils” last 2,000 – 2,000,000 years

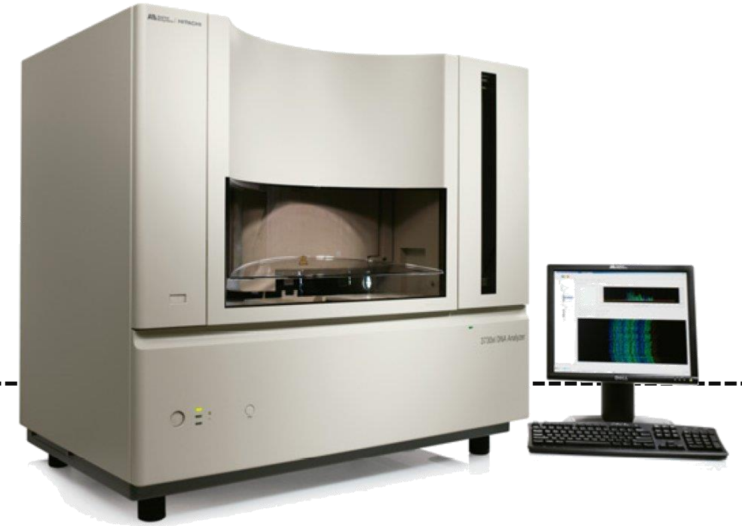



## Extreme density makes these conditions cheap and easy to keep

# Comparison with other media



# No obsolescence



*Size of a mainframe*

800 Kbases/day

*Size of a workstation*

80 Gbases/day

*Size of a portable SSD*

10 Gbases/day

**Same medium as read technology improves:**

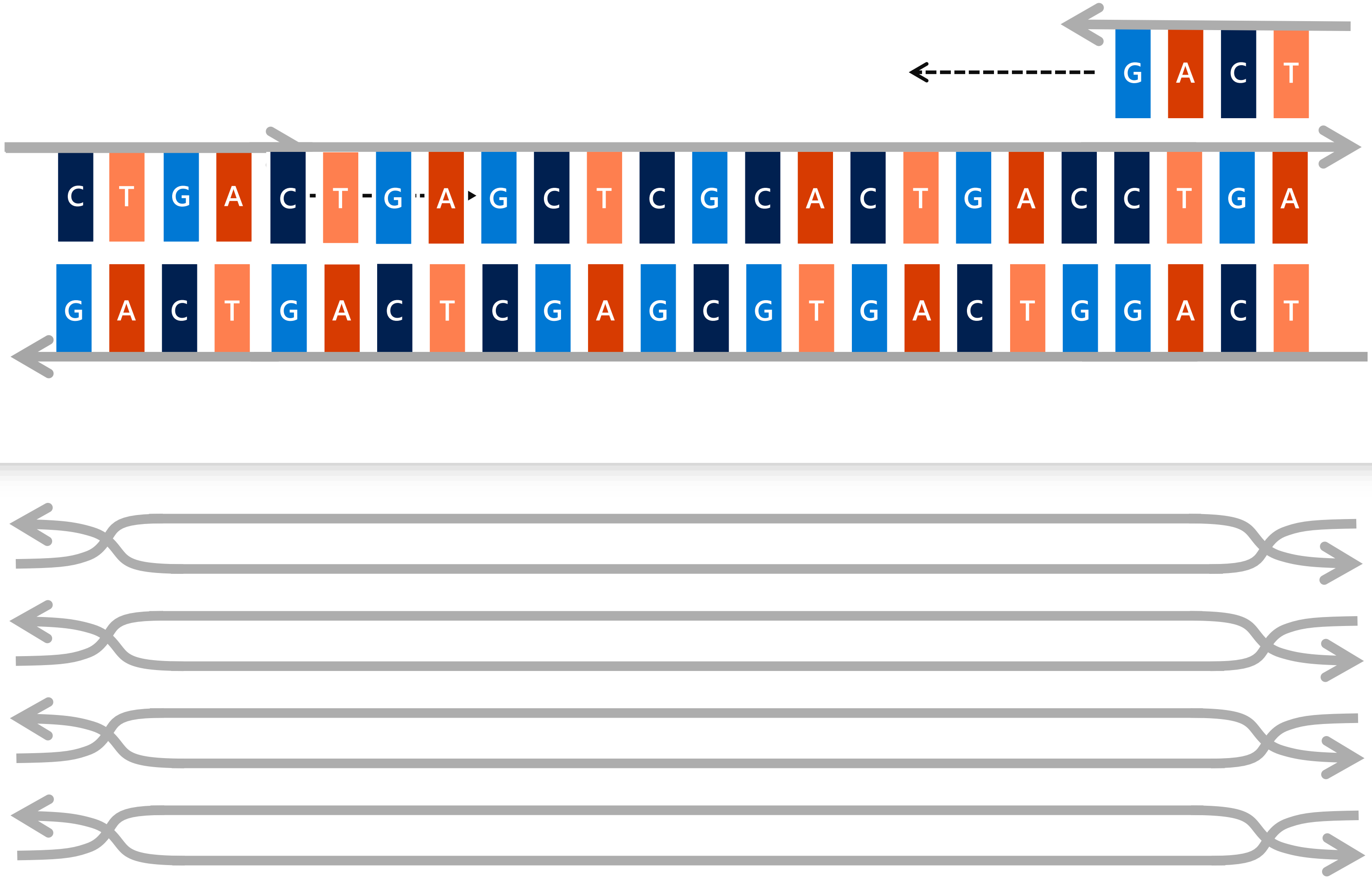
# No obsolescence issue, DNA will always be relevant

Same medium as read technology improves:

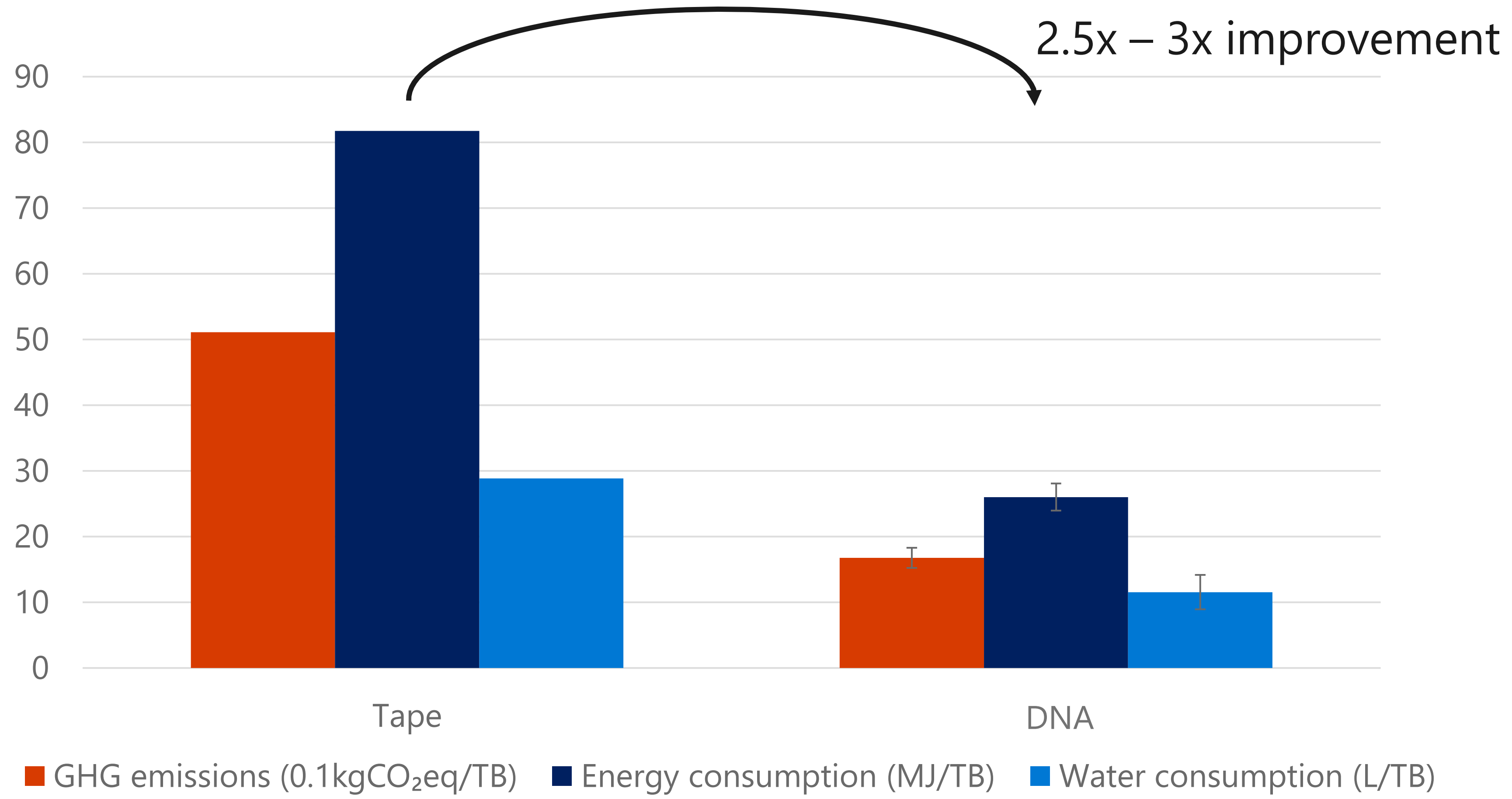

Medium changes as read and write technology improves:

# Ability to make copies

Polymerase Chain Reactions (PCR) create copies exponentially



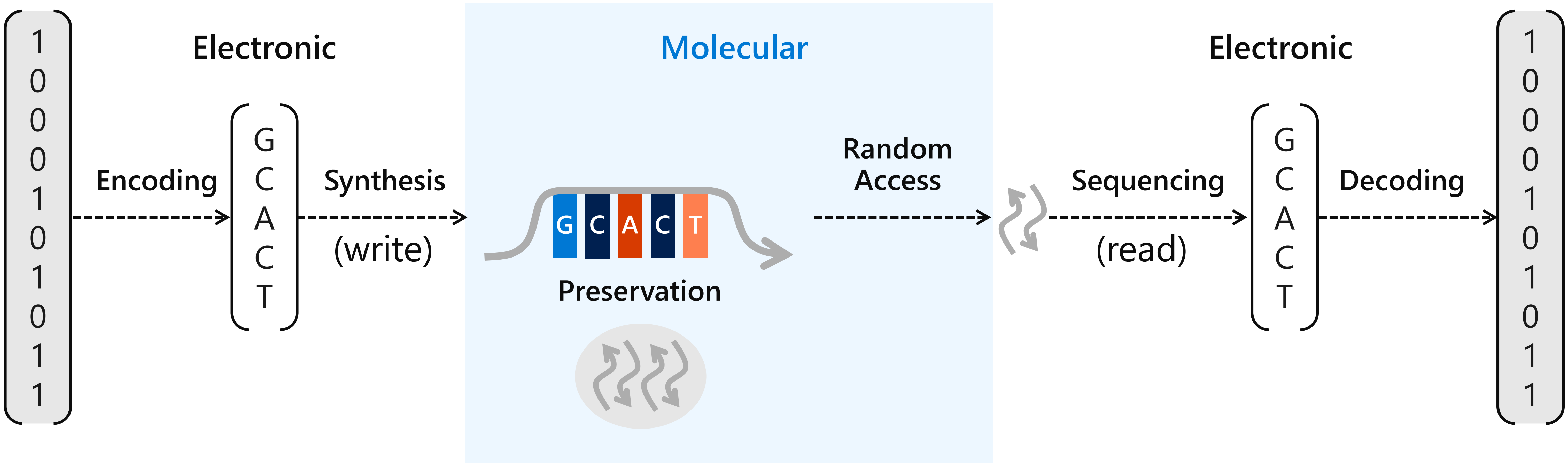
# Sustainability




▶ DNA promises to be significantly more sustainable than tape

Nguyen et al., Electronics Goes Green, 2020

# DNA storage end-to-end system



# Our results so far

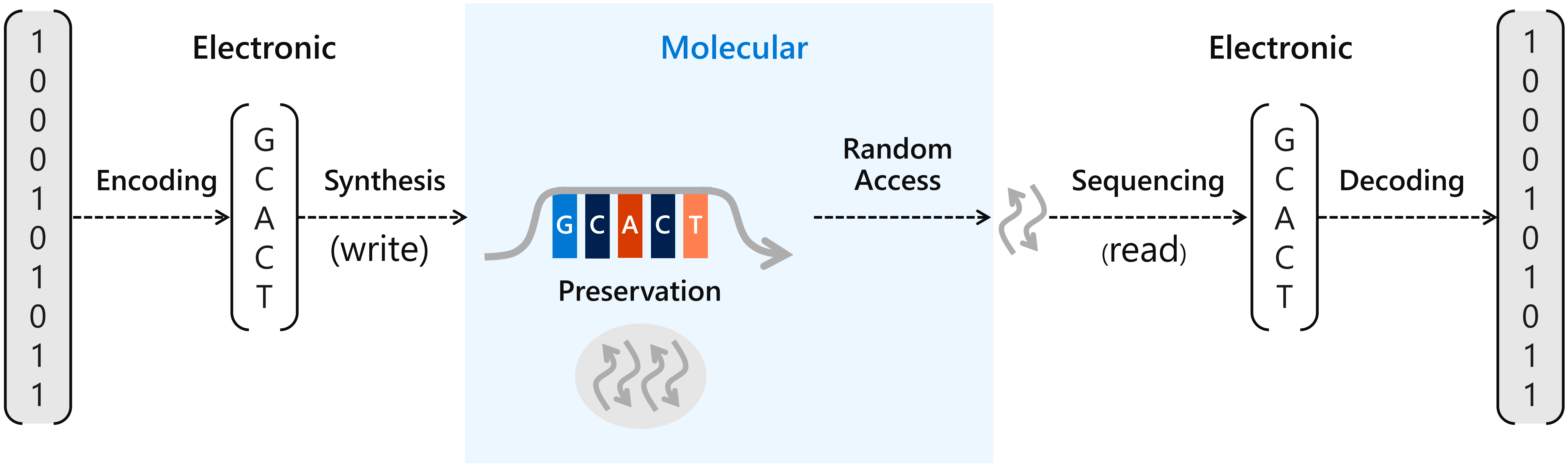
## 1GB of data stored and fully recovered

## Published


# Most data in DNA in peer reviewed publication

Organick et al., Nature Biotechnology, 2018

# DNA storage end-to-end system



# DNA encoding

0101000101011100 1101100001111000 0111010111001000 1001010001001001

Photo credit: OK Go.

1 of 4

2 of 4

3 of 4

4 of 4

ACGT ACG AGTCGATCAGTCGA TGCA

ACGT ACT GTCATACGTAGCGA TGCA

ACGT AGA CTATCGTGACTAGTC TGCA

ACGT AGC ATCGCTCGATCGAT TGCA

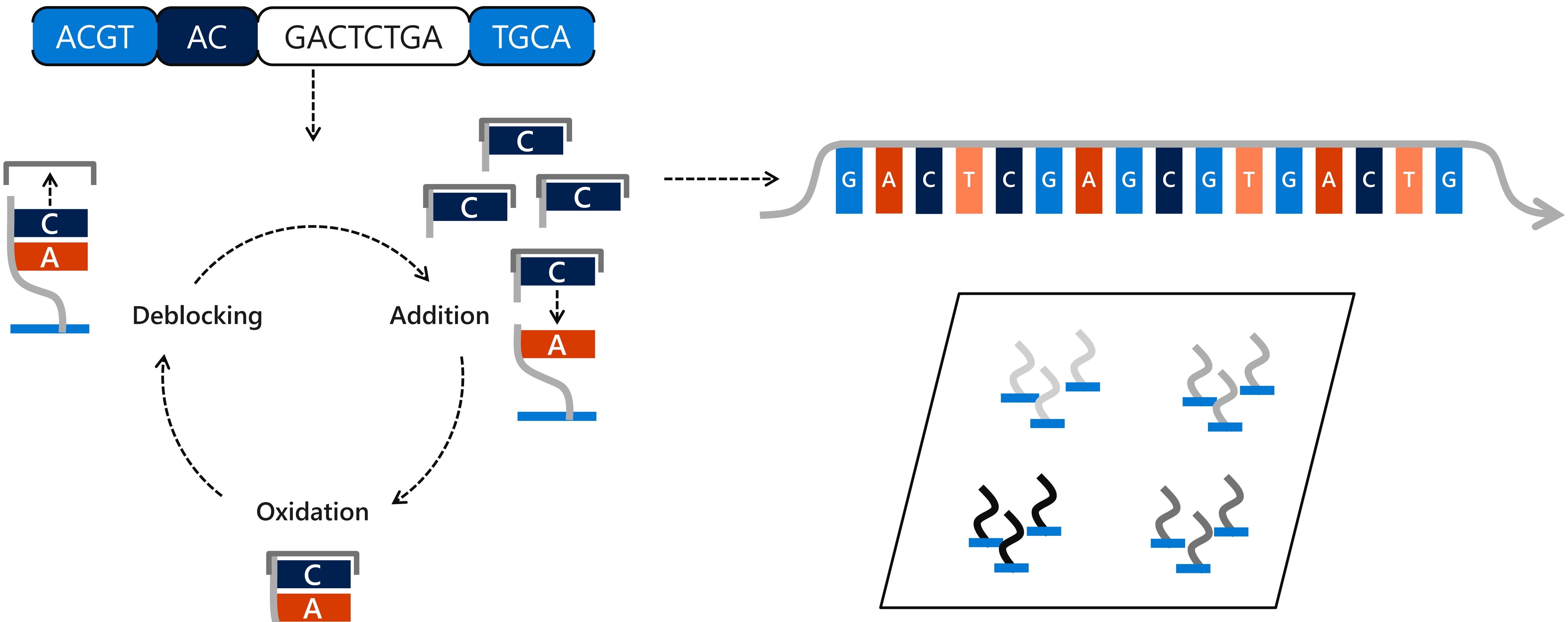
Redundant data for later error correction*

ACGT AGT GCTAGCTAGATCGA TGCA

ACGT CAC TCGATCGATAGCTA TGCA

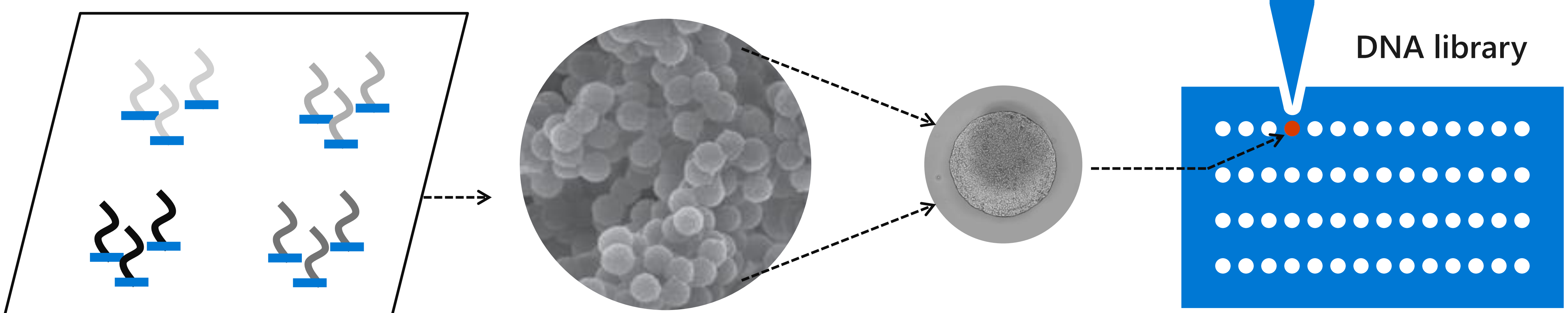
Grass et al., Angewandte Chemie, 2015; Bornholt et al., ASPLOS, 2016

# DNA synthesis




Array synthesis

# DNA preservation

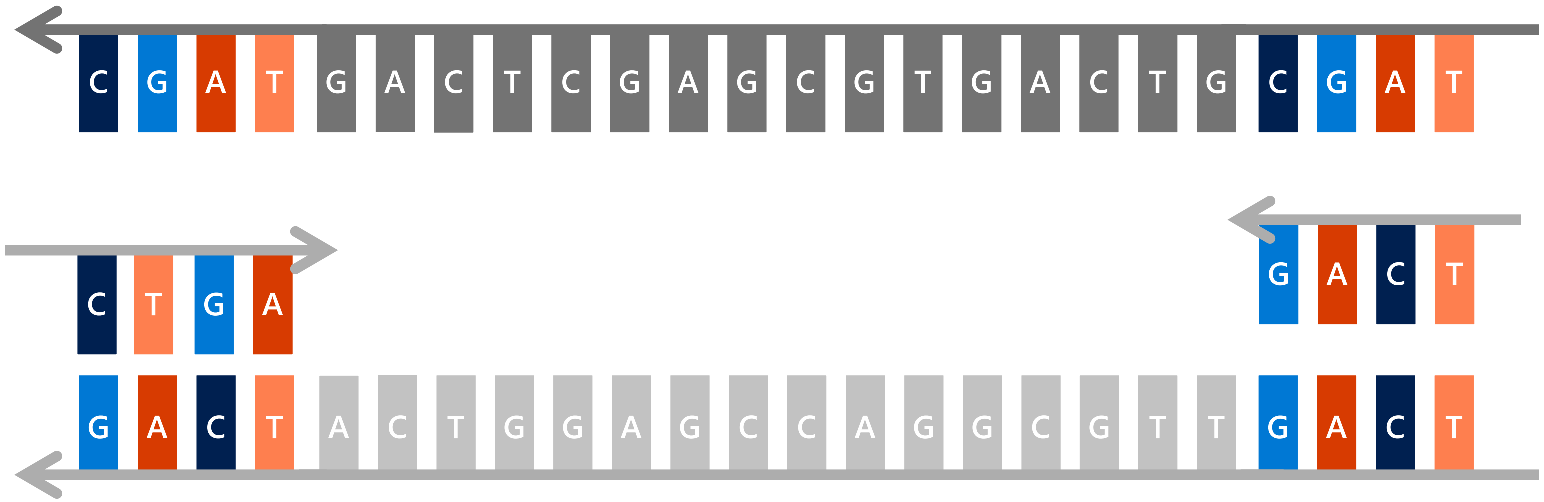



(Credit: Grass et al./ETH Zurich)

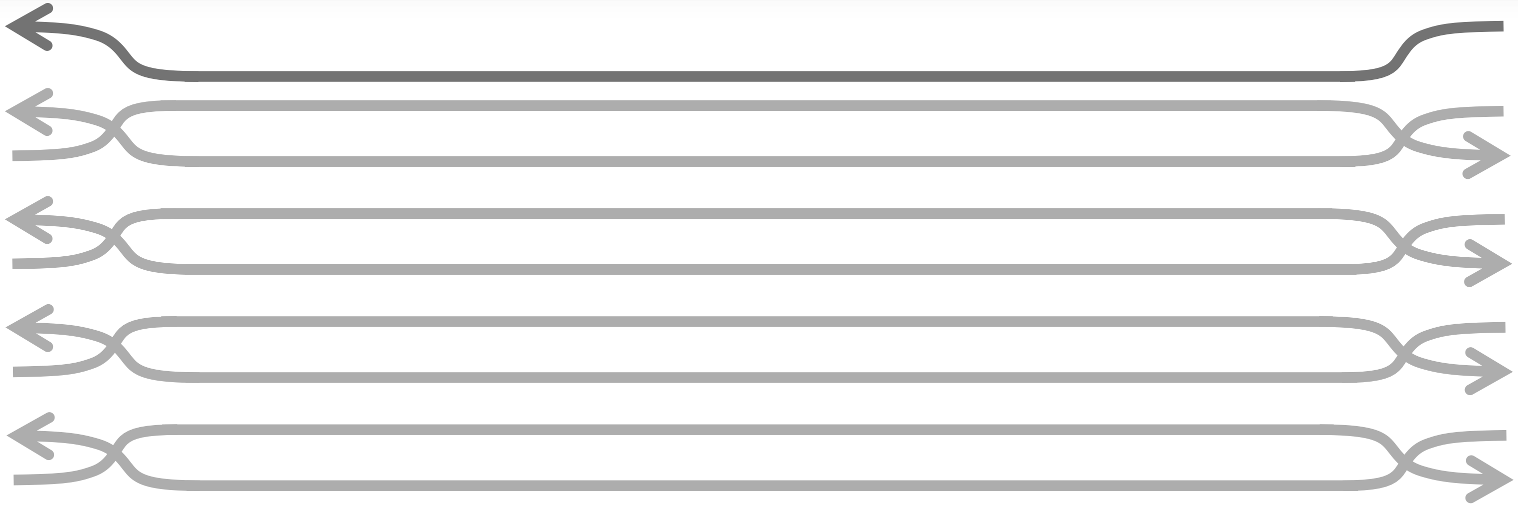
Grass et al., Angewandte Chemie, 2015; Chen et al., Advanced Functional Materials, 2019 ; Kohll et al. Chemical Communications, 2020

# DNA storage random access with PCR

Selecting one item out of two

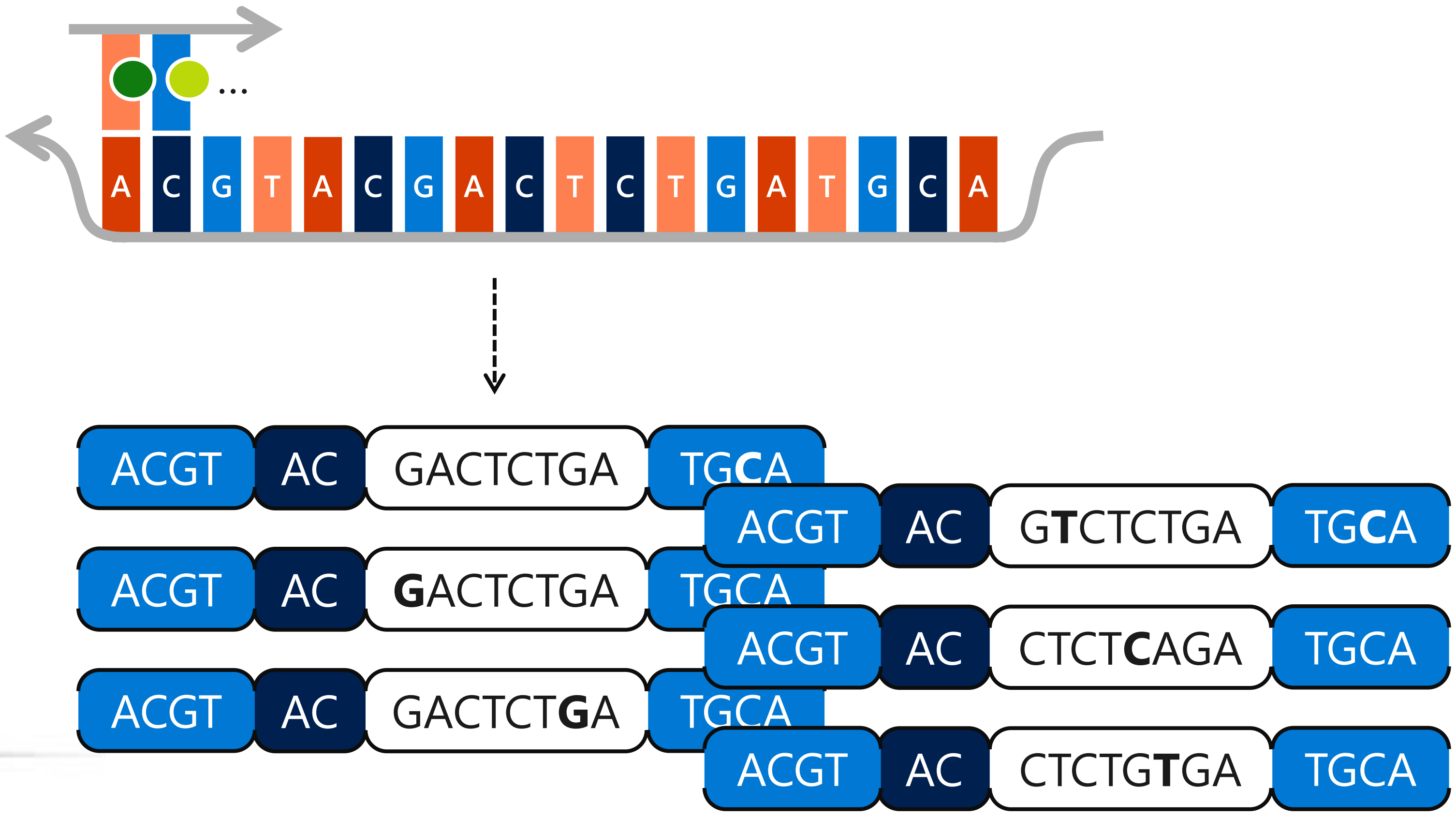



PCR primers are used to access units of storage individually

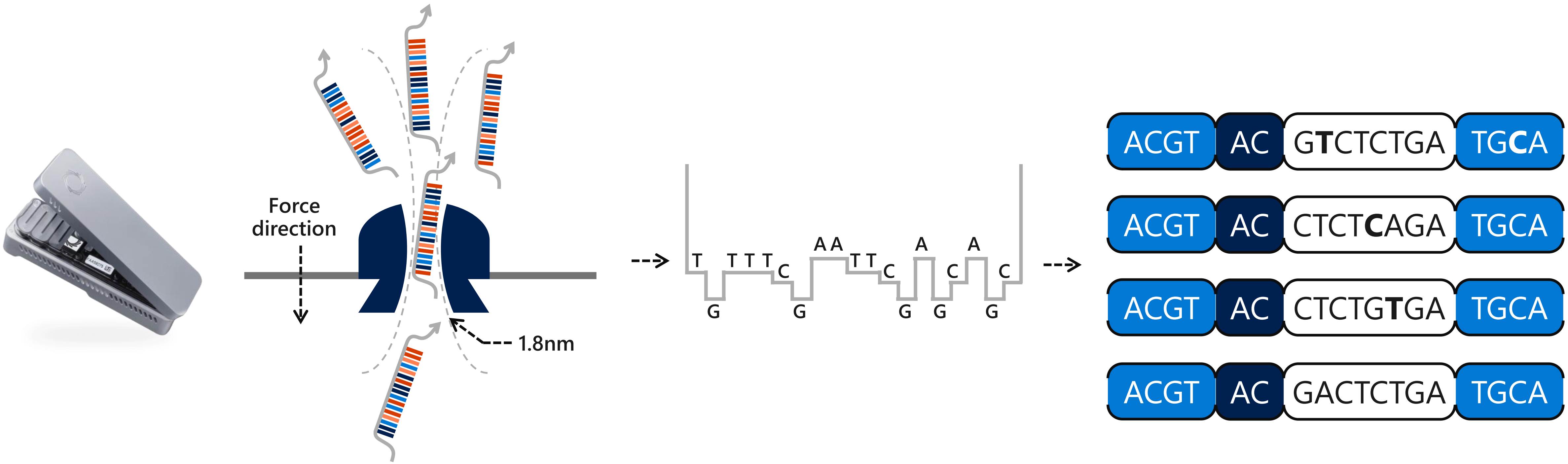


Yazdi et al., Scientific Reports, 2015; Bornholt et al., ASPLOS, 2016; Organick et al., Nature Biotechnology, 2018; Organick et al., Nature Communications, 2020; Chen et al., Nature Communications, 2020

# Reading DNA with sequencing by synthesis



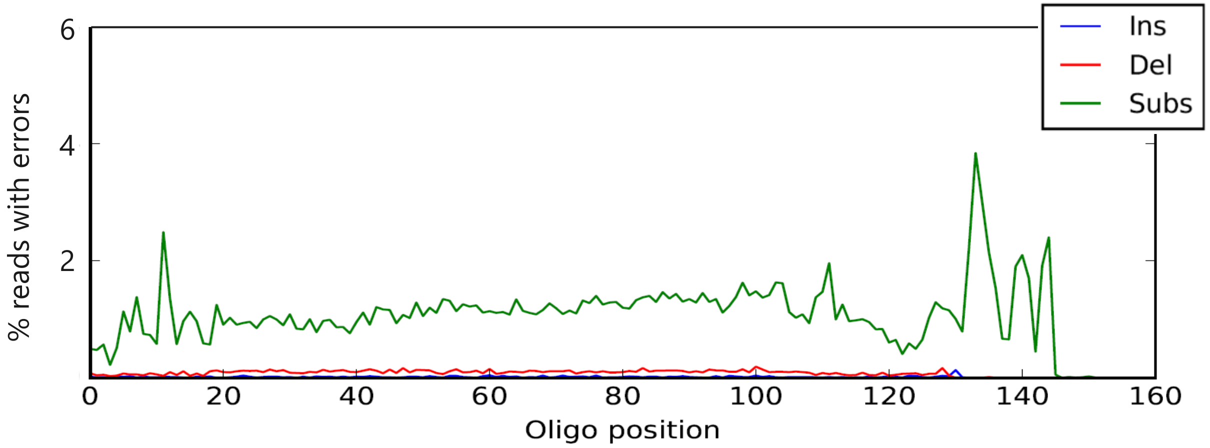
# Reading DNA with nanopores



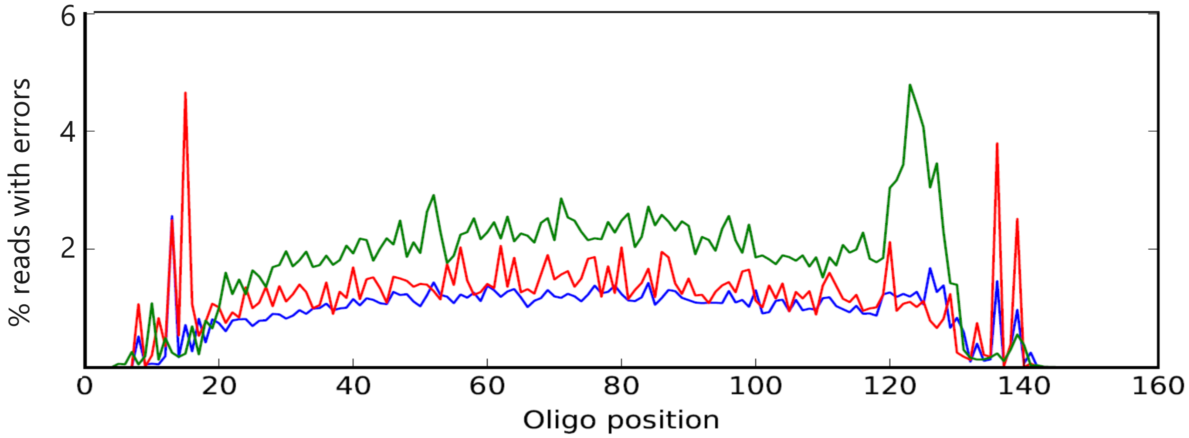

Yazdi et al., Scientific Reports, 2017; Organick et al., Nature Biotechnology, 2018; Lopez et al., Nature Communications, 2019

# Different error profile across platforms

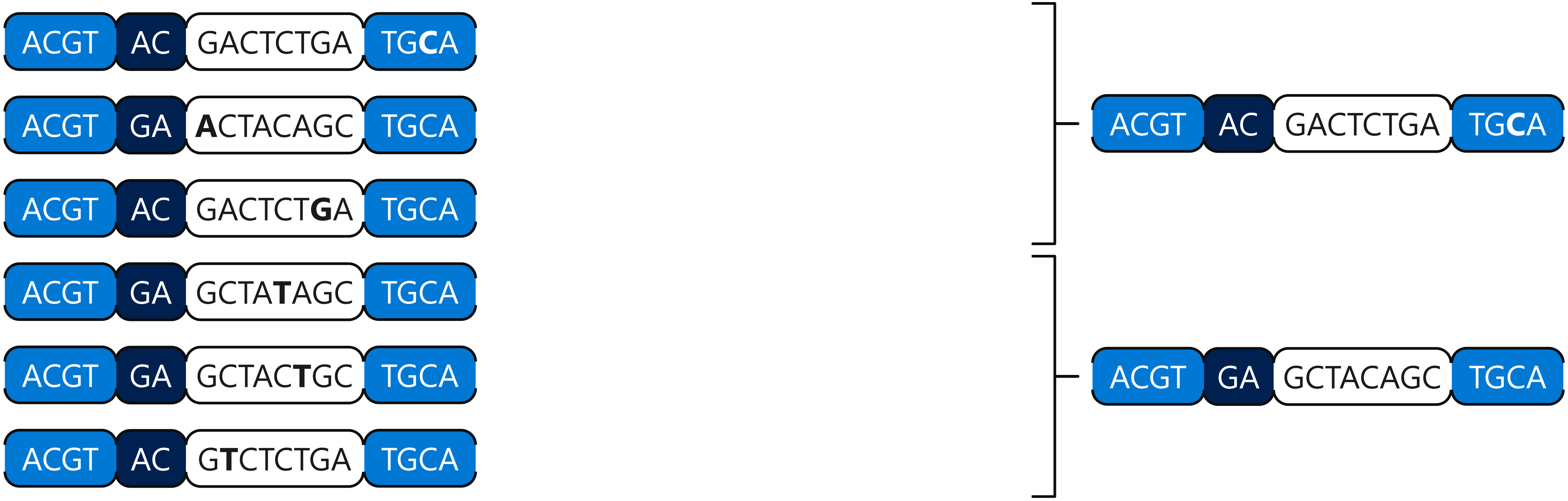
## Illumina NextSeq


## ONT MinION

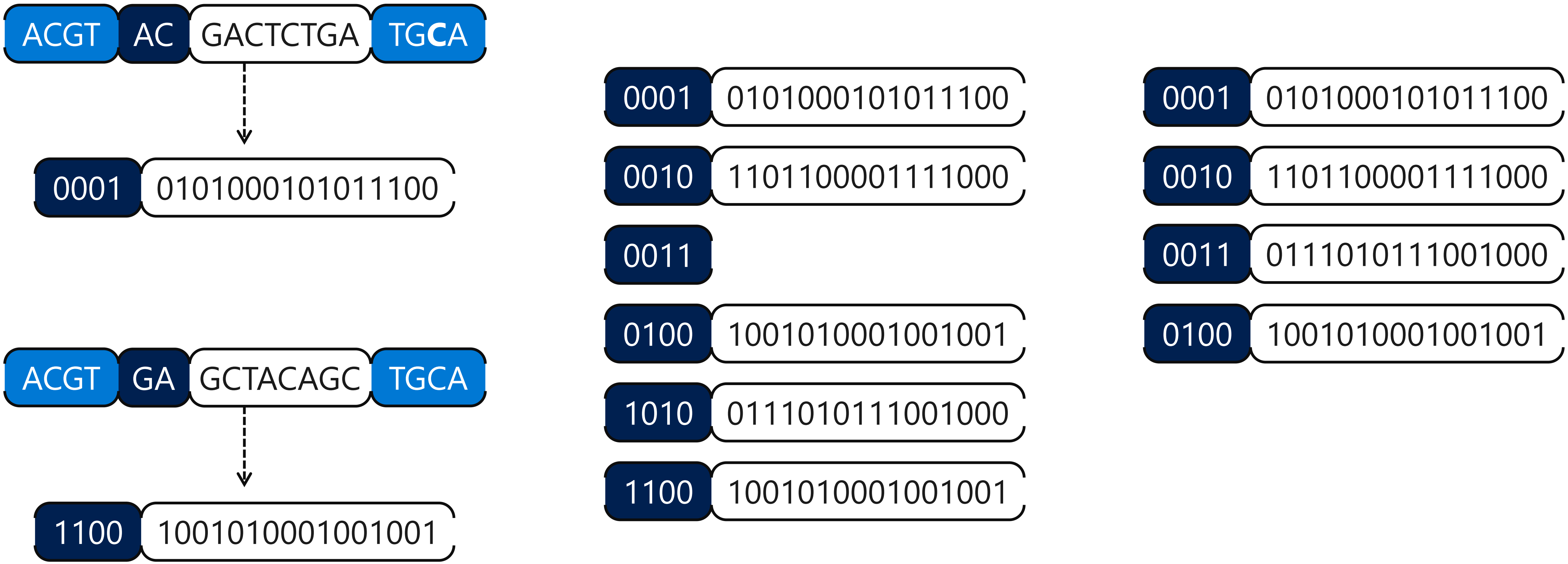


# DNA decoding and error correction




Rashtchian et al., NeurIPS, 2017; Organick et al., Nature Biotechnology, 2018

# DNA decoding and error correction



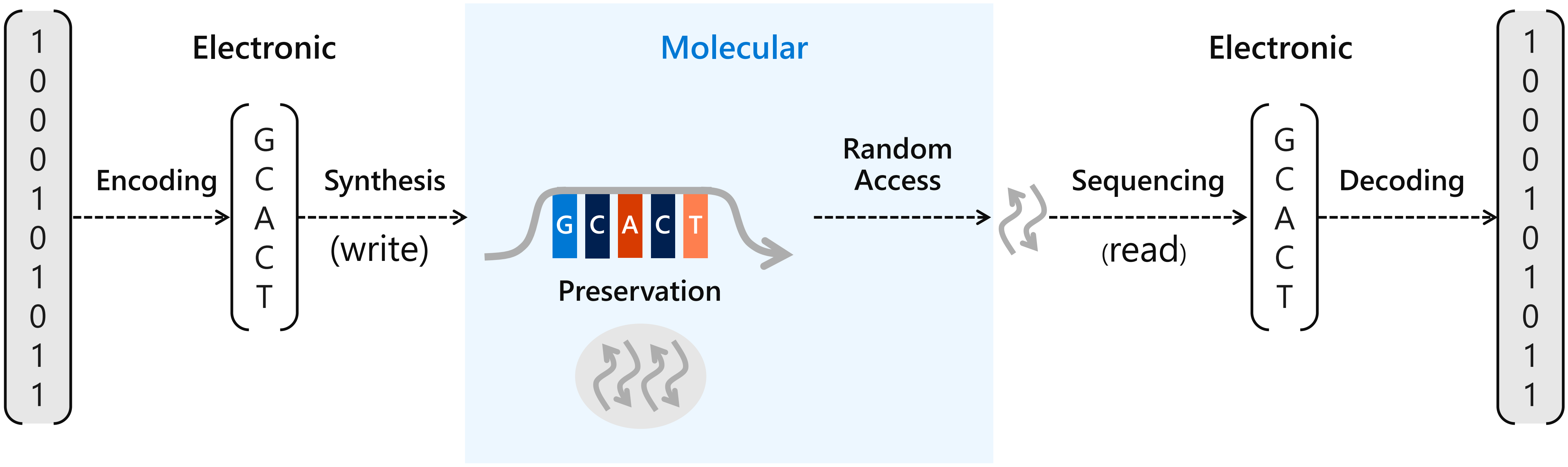
# DNA decoding and error correction

| | |
|---|---|
| 0001 | 0101000101011100 |
| 0010 | 1101100001111000 |
| 0011 | 0111010111001000 |
| 0100 | 1001010001001001 |

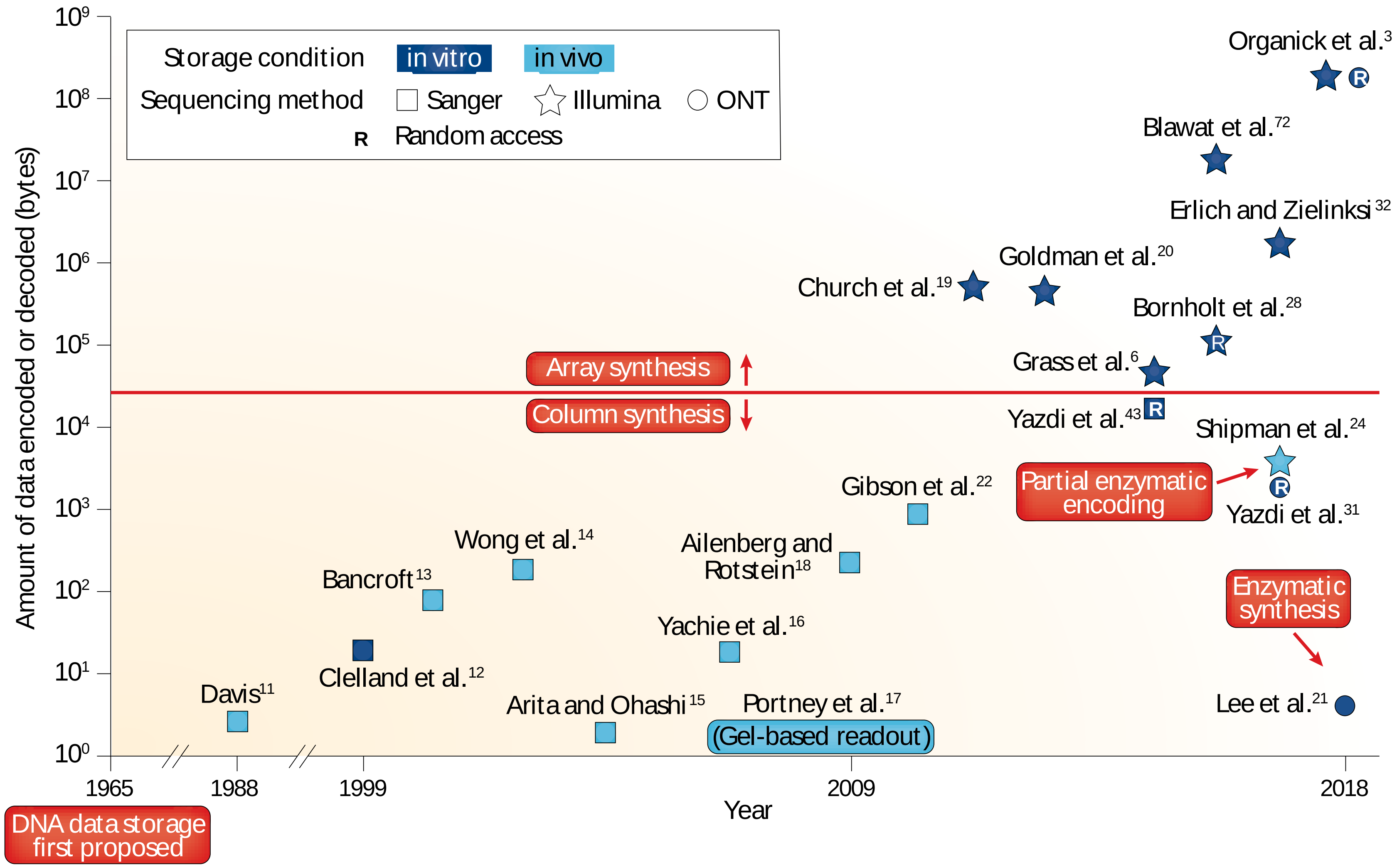
0101000101011100 1101100001111000 0111010111001000 1001010001001001

*Photo credit: OK Go.*

# DNA storage end-to-end system

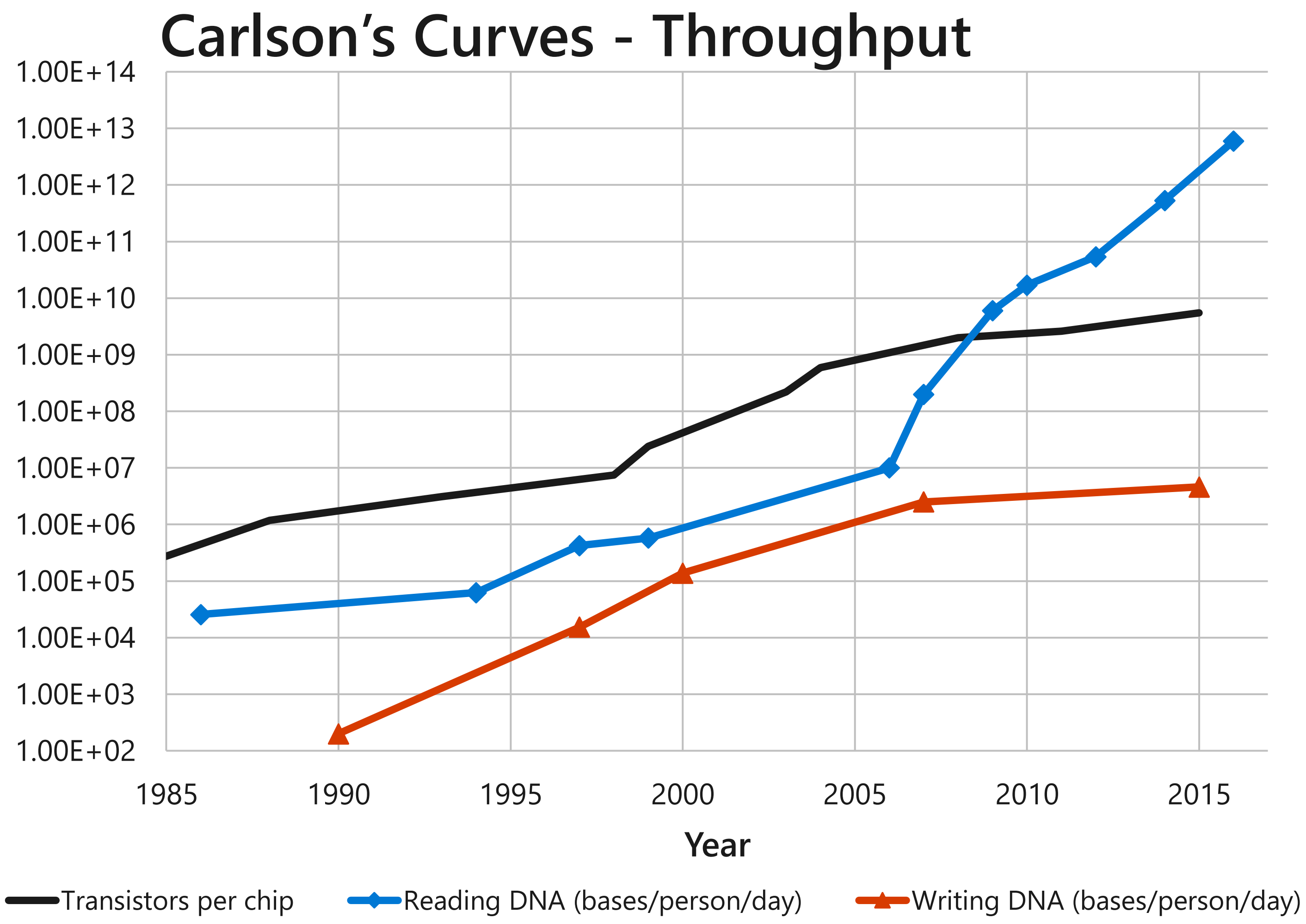


# Exponential improvements in DNA data storage




[Ceze et al., Nature Genetics Reviews'19]

# Performance of reading and writing DNA




Source: Robert Carlson

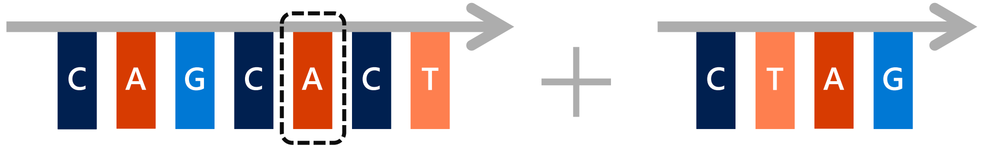
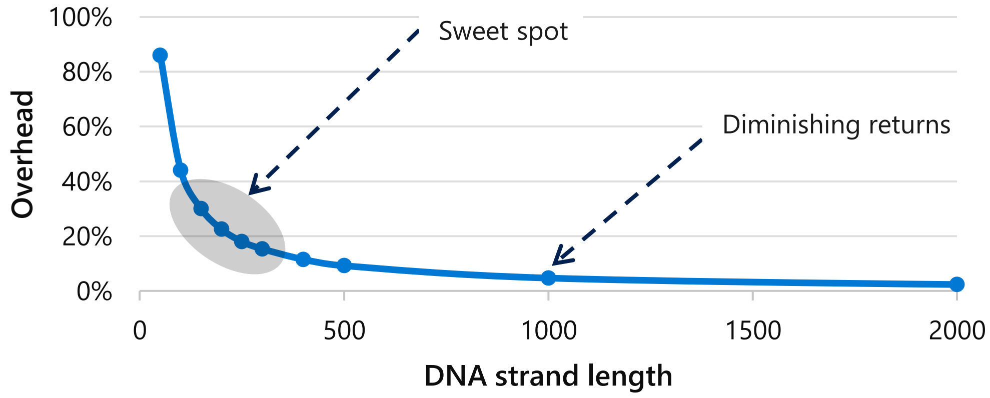
## Latency

Synthesis and sequencing are currently batch processes, matches archival storage SLAs (~hours).

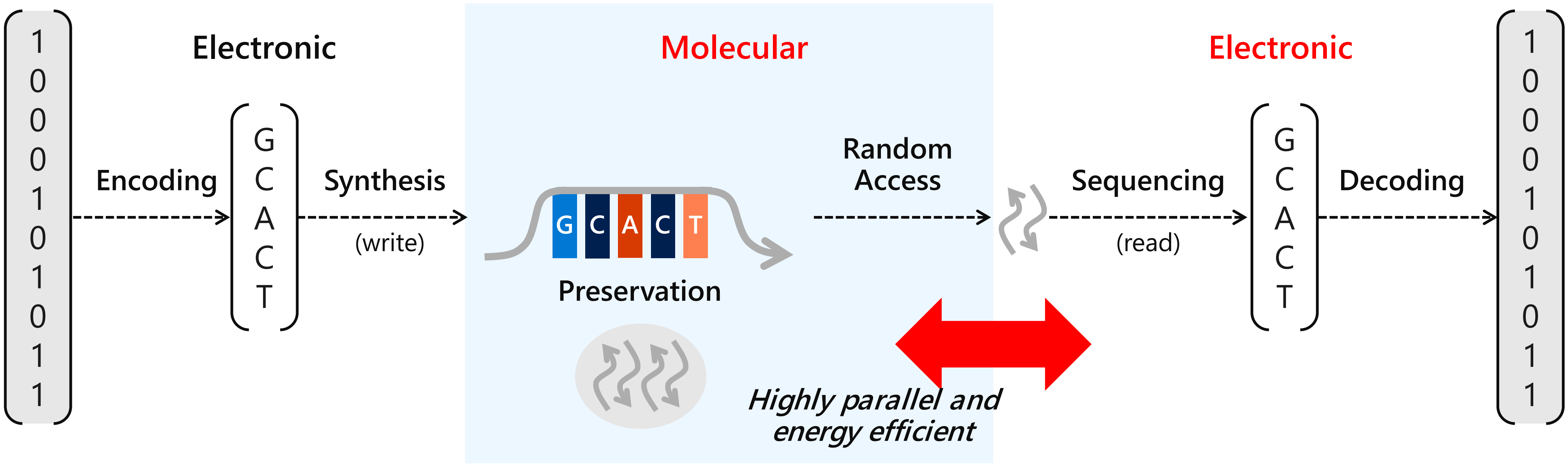
Emerging technologies, like nanopore devices, provide closer to real time latency

# Write and read mechanisms

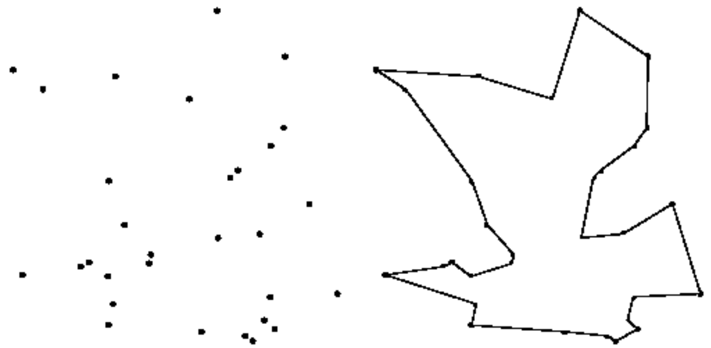
Opportunities for improvements in write and read throughput, latency and cost

| Opportunities | Life sciences | Data storage |
|---|---|---|
| Error rate | Single base mutations affect function | Error correcting codes allow data recovery even in the presence of multiple errors   **Error types:** substitutions, deletions and insertions |
| Length ("block size") | Longer sequences have more function | Shorter sequences are faster and easier to make   |

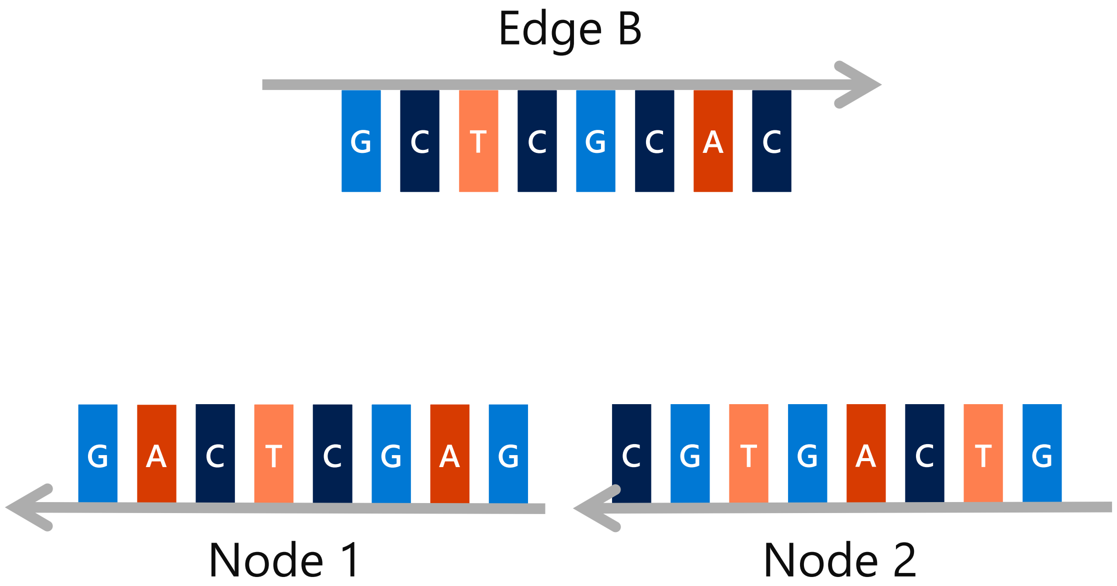
# DNA storage end-to-end system w/ integrated computing



# DNA computing in the 80s



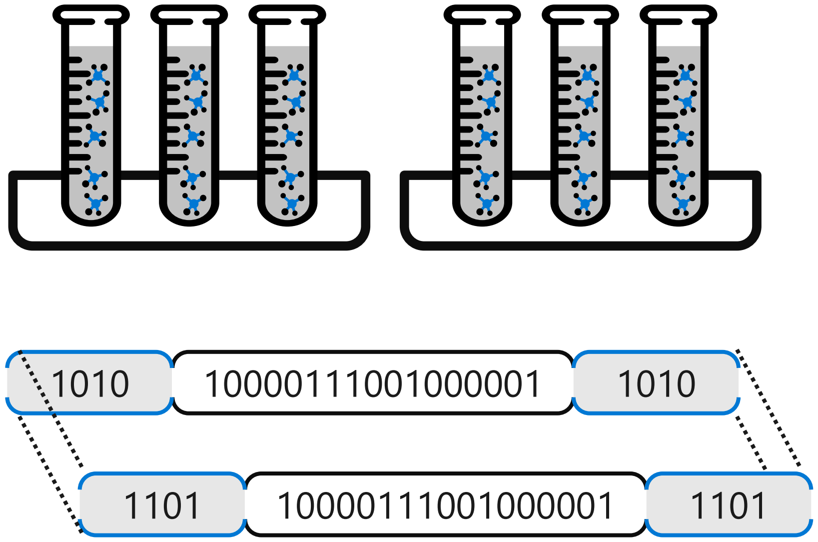
Hamiltonian path problem




**Problem:** shifts complexity from time to amount of material

Adleman, DNA1, 1994

# DNA "computing" in the age of big data

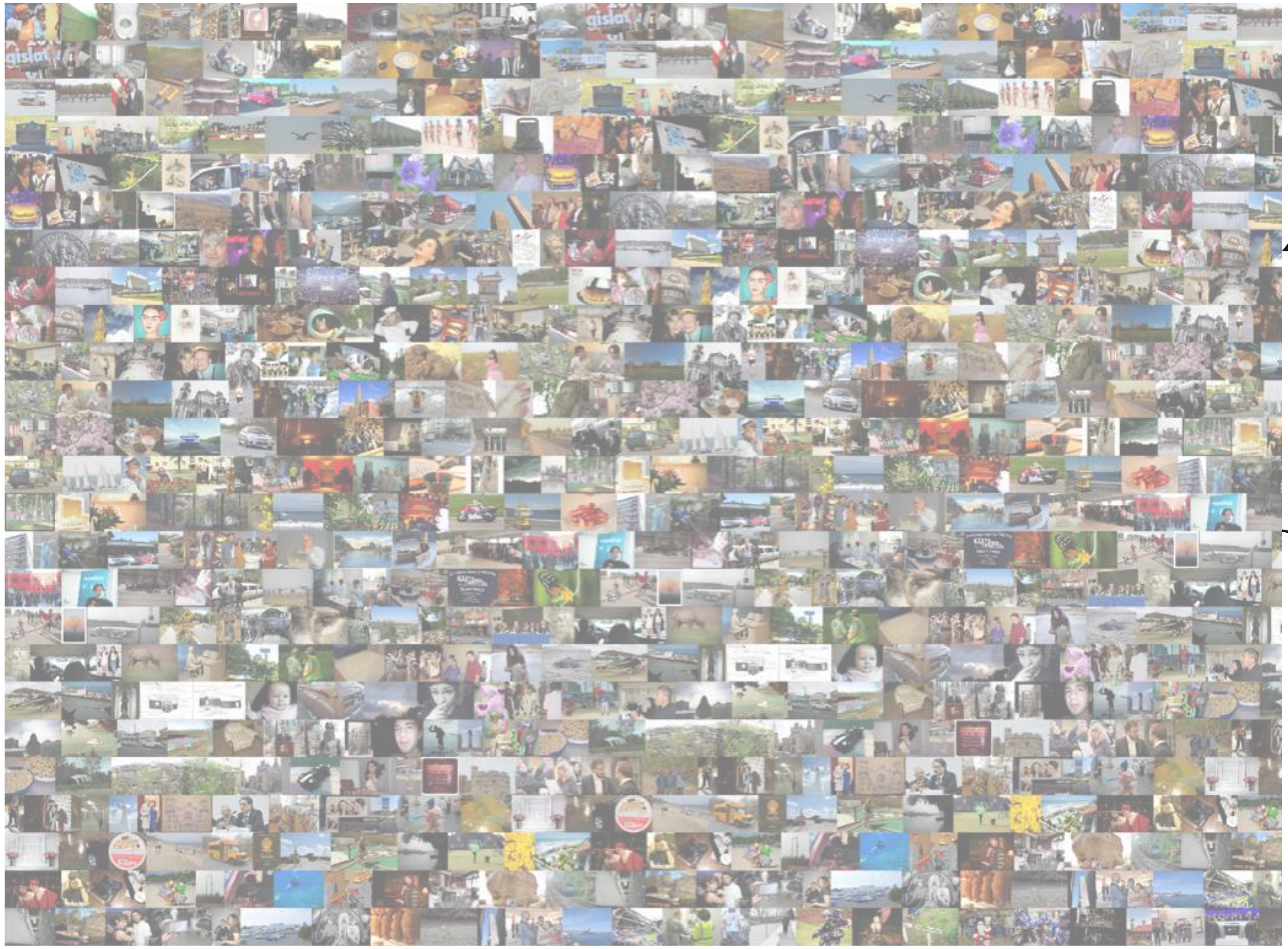



Operate over data already stored in DNA

Target polynomial time algorithms

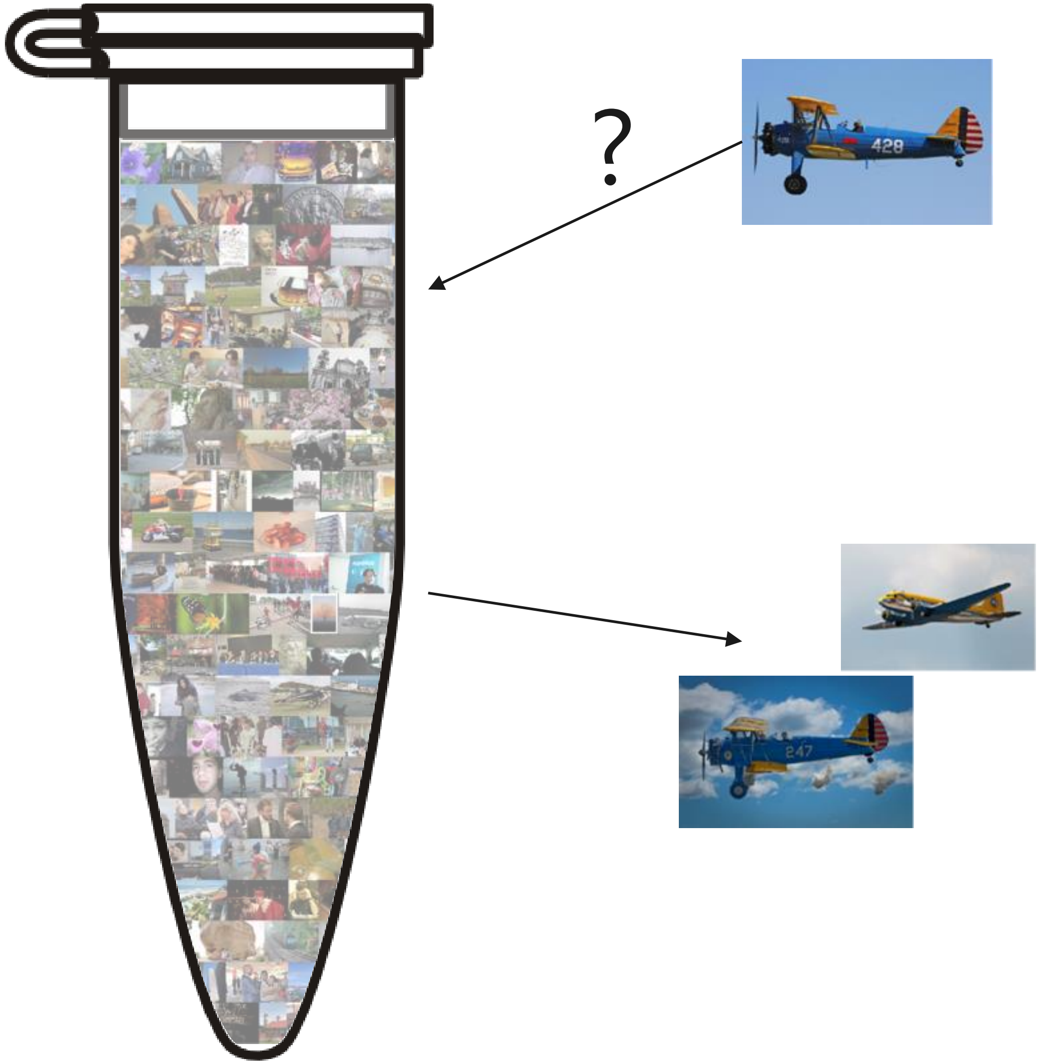
Extremely parallel and energy efficient

# Content-based image/video search



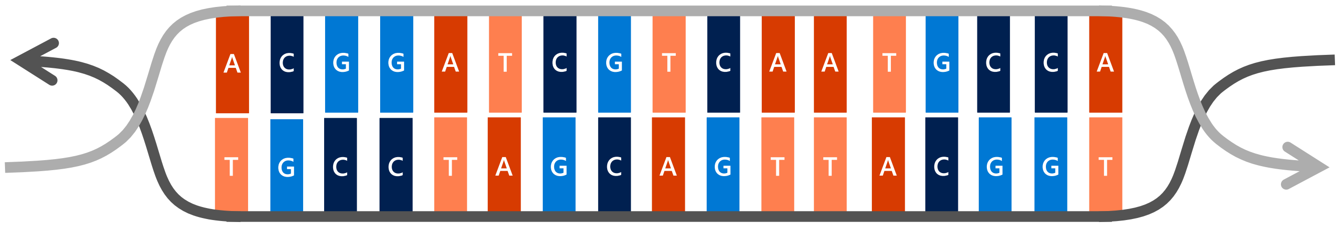
?

# Content-based image/video search in DNA

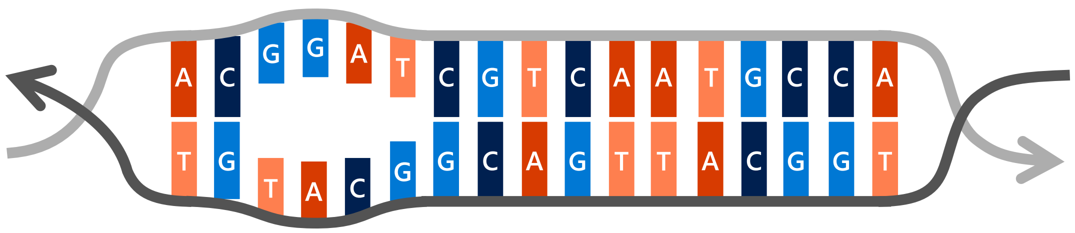


# Exploiting matches for exact and approximate search

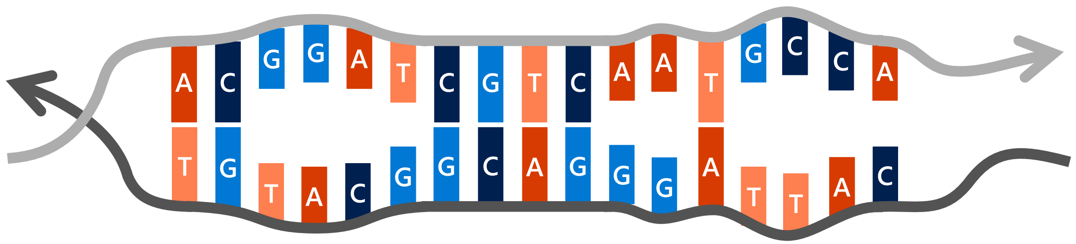
## Double helix: complete match



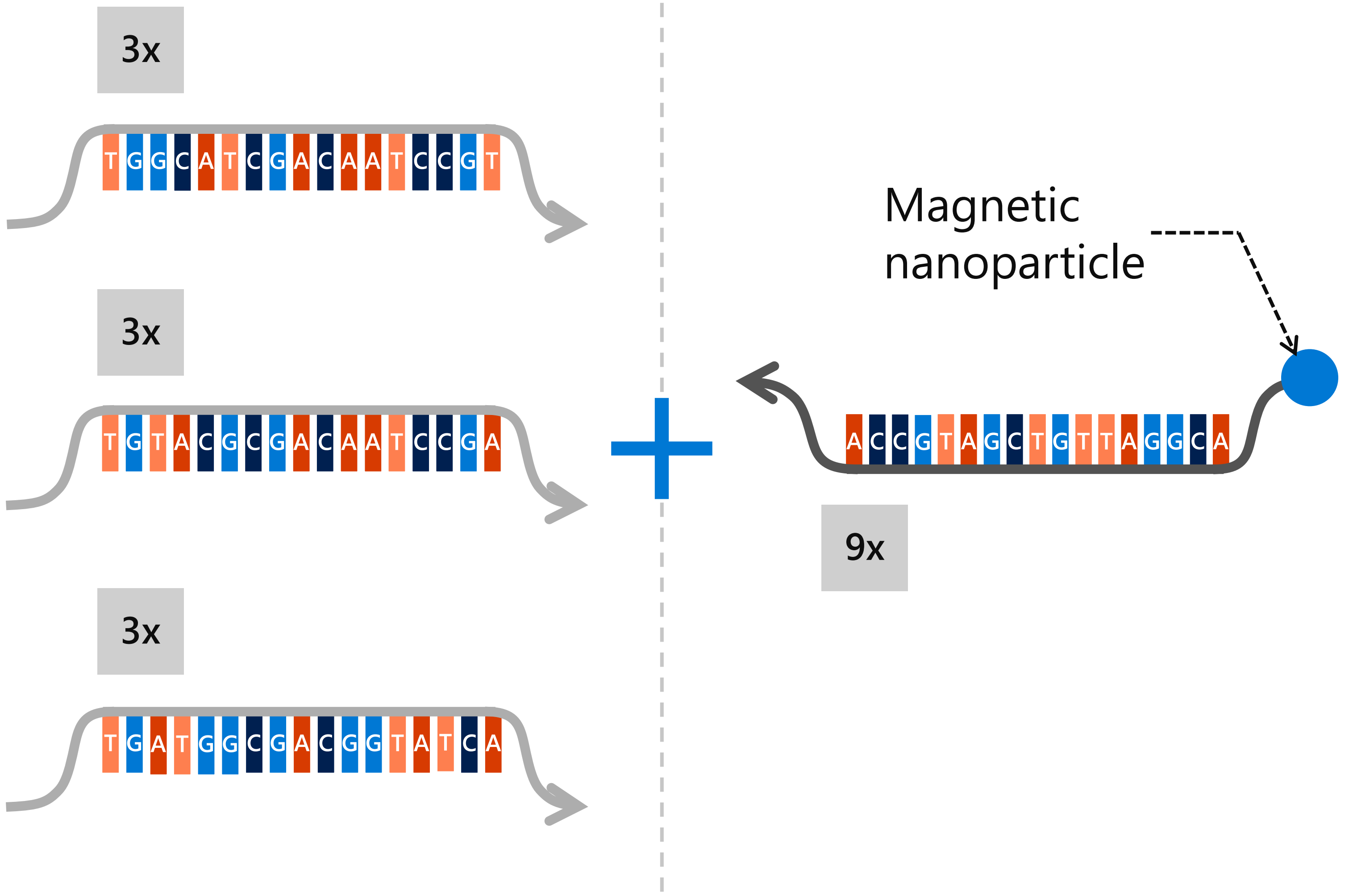

## Good partial match



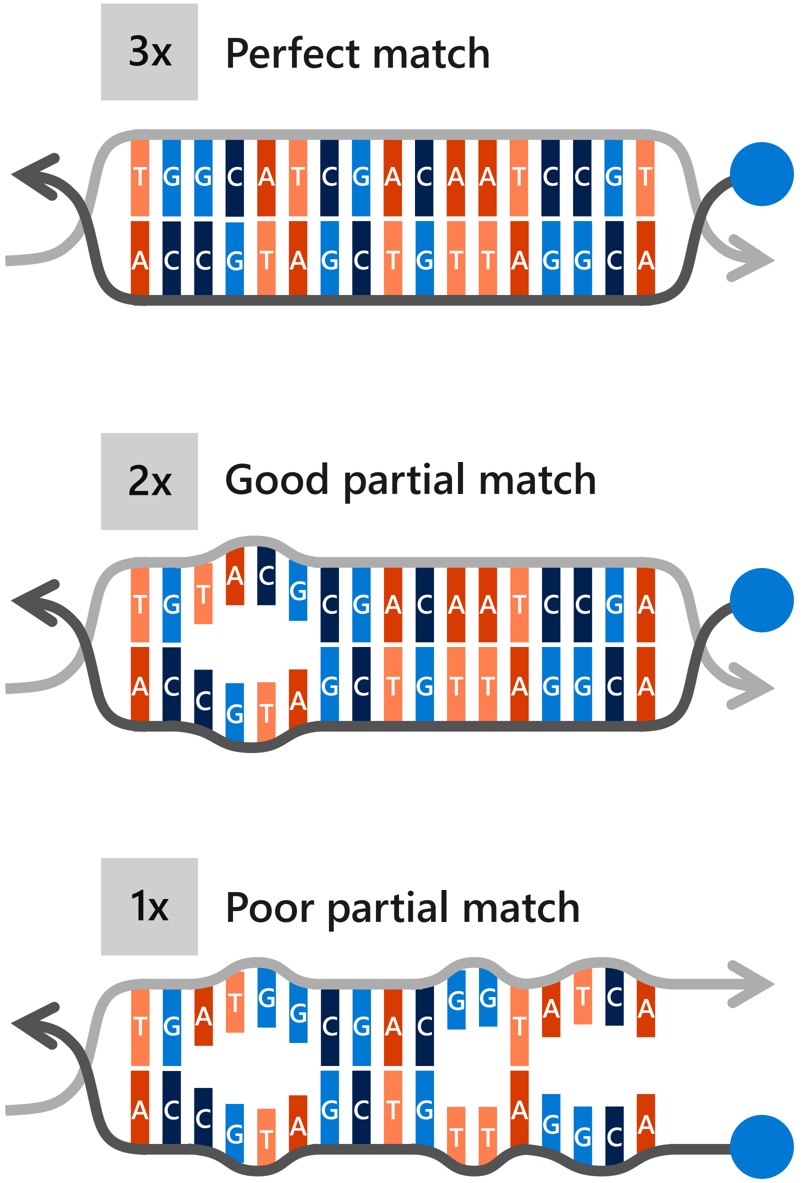

## Poor partial match



# Searching with DNA

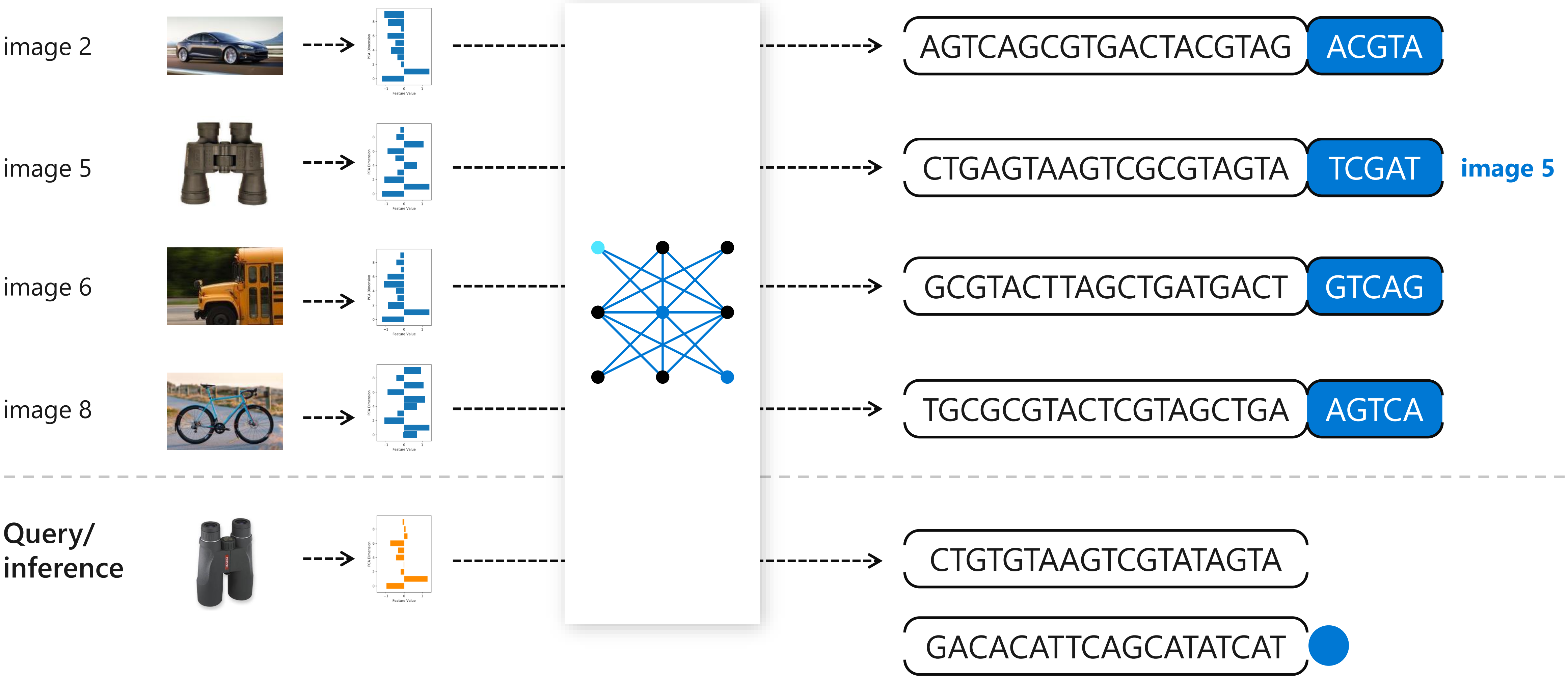



## Match-dependent yield



# Content based media search

## Database/ training




Stewart et al., DNA24, 2018; Nature Communications, 2021

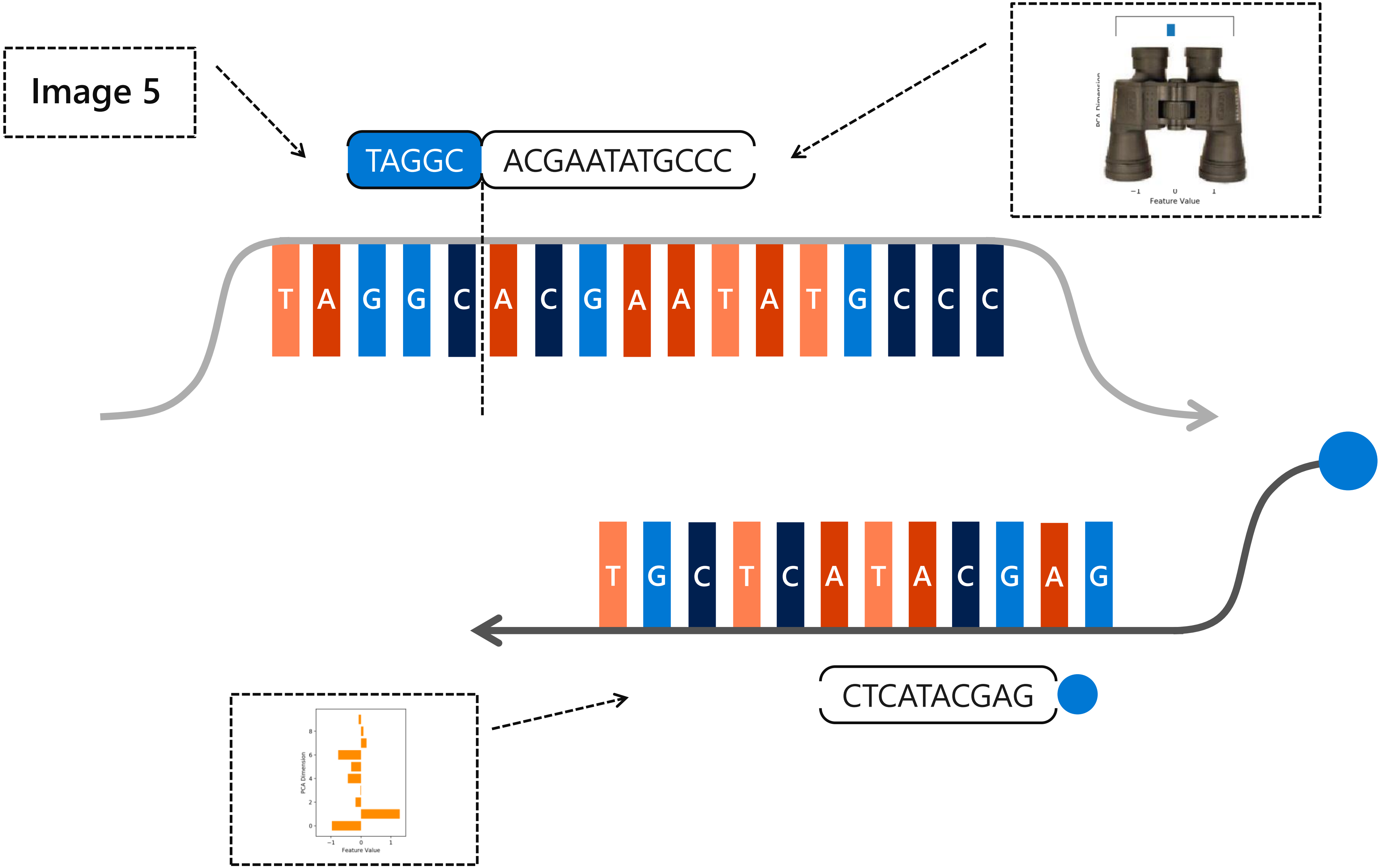


Image 5
TAGGC
ACGAATATGCCC
T A G G C A C G A A T A T G C C C
Feature Value
T G C T C A T A C G A G
CTCATACGAG
PCA Dimension
Feature Value

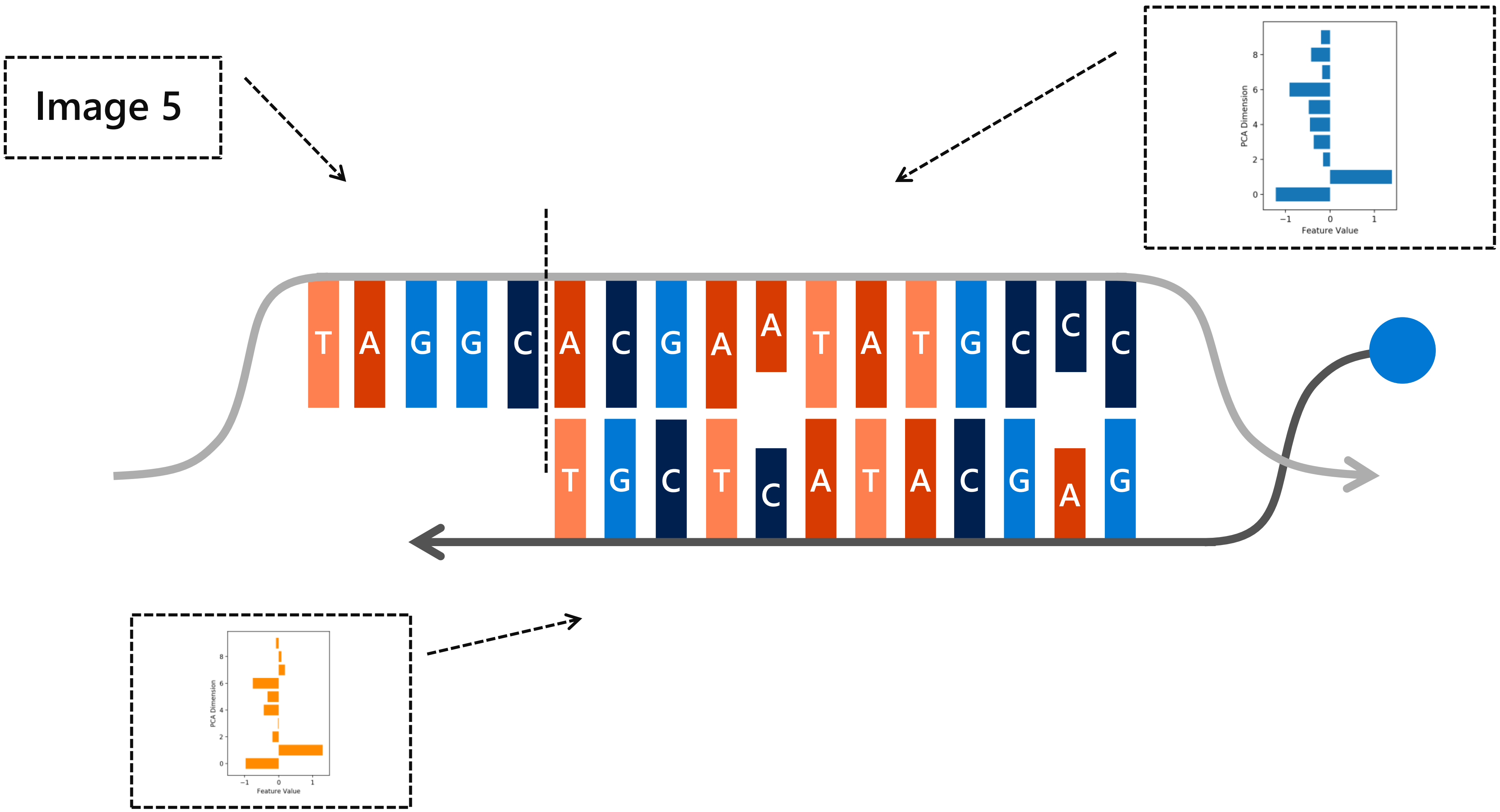
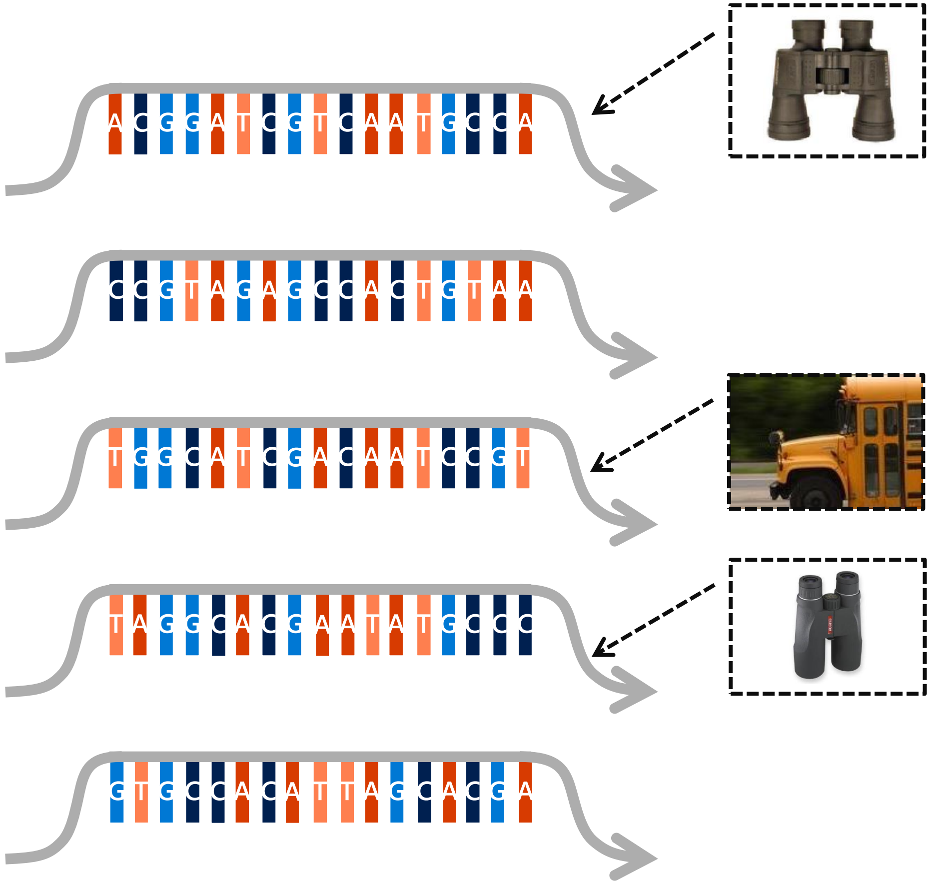
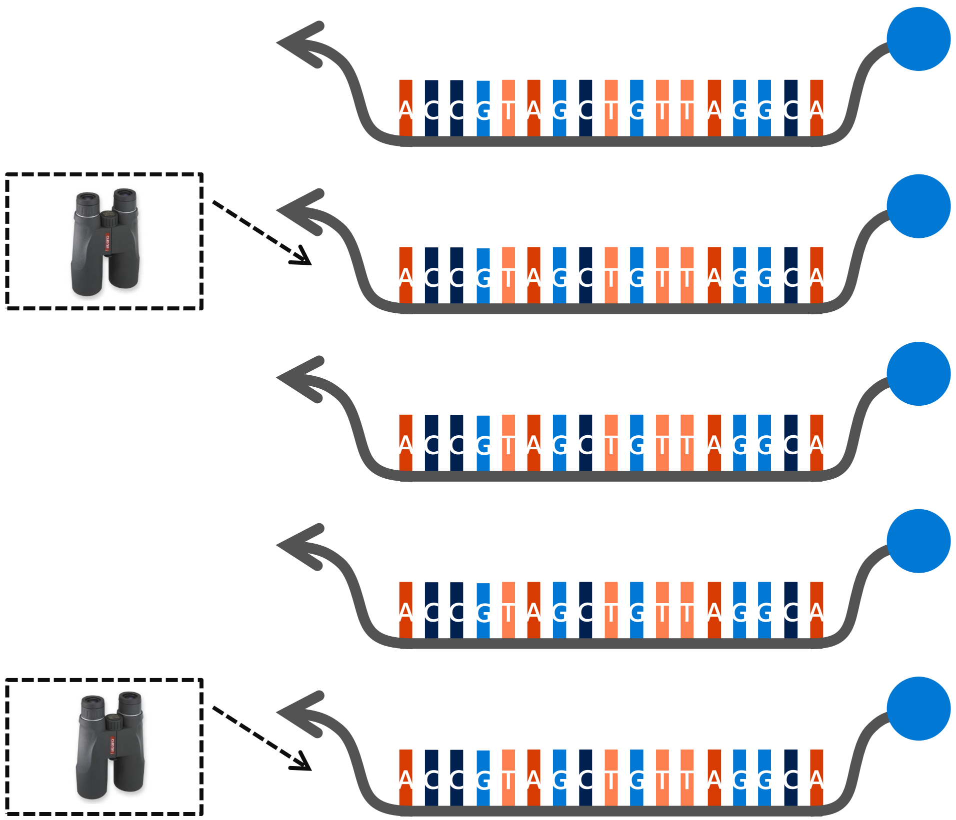


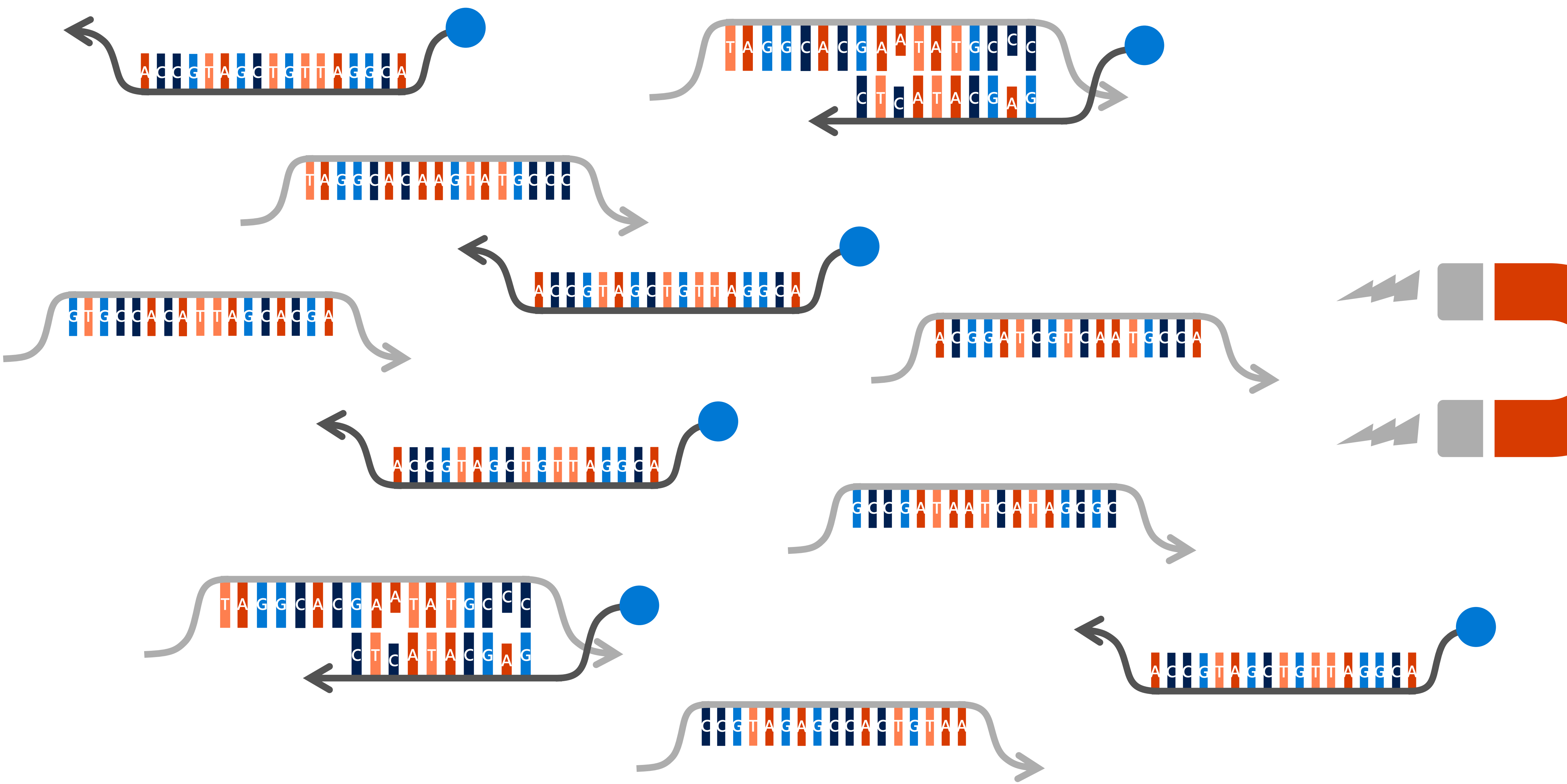
Image 5
T A G G C A C G A A T A T G C C C
T G C T C A T A C G A G
PCA Dimension
Feature Value
PCA Dimension
Feature Value

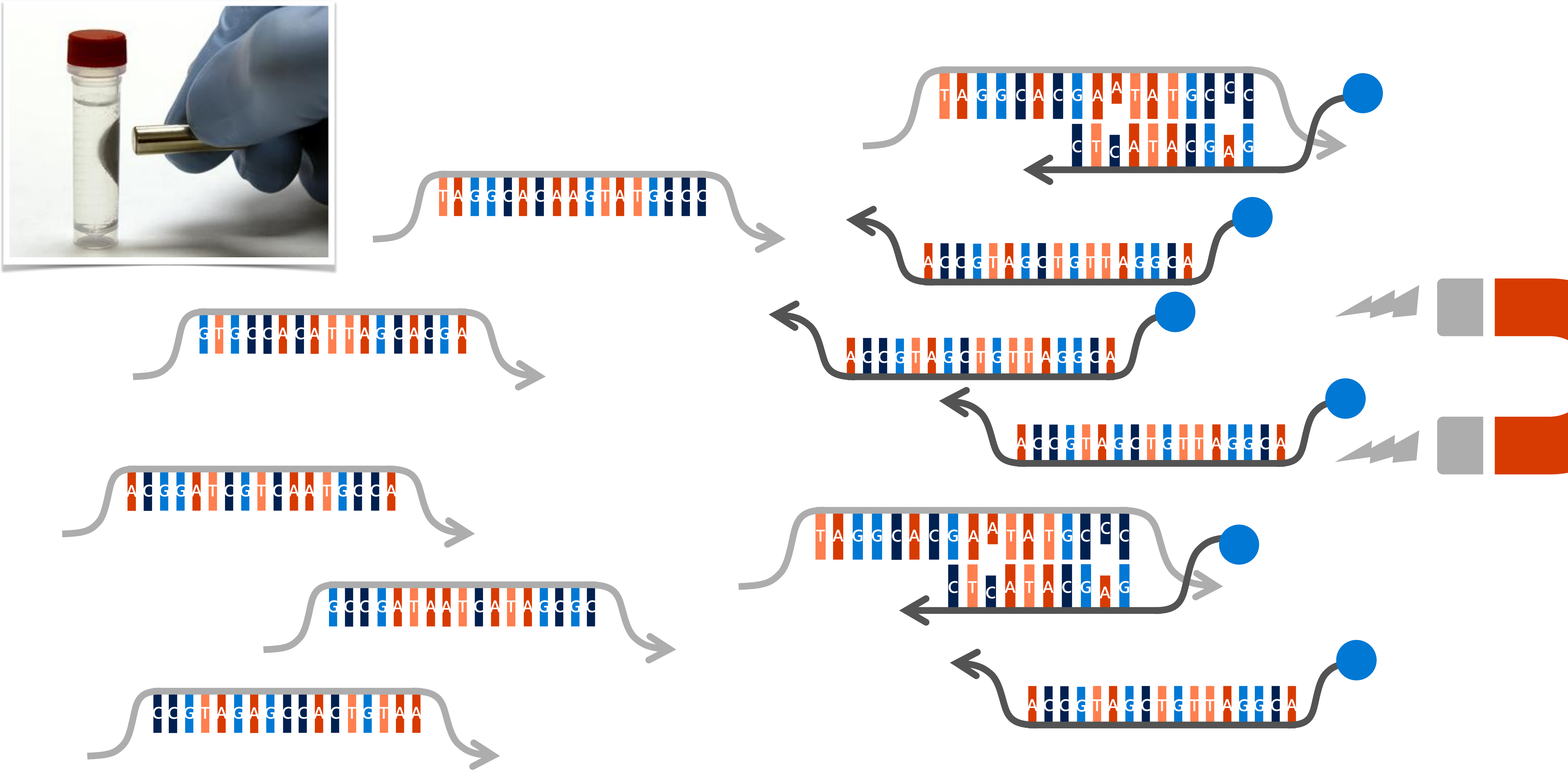
# “Database” strands
*(many)*

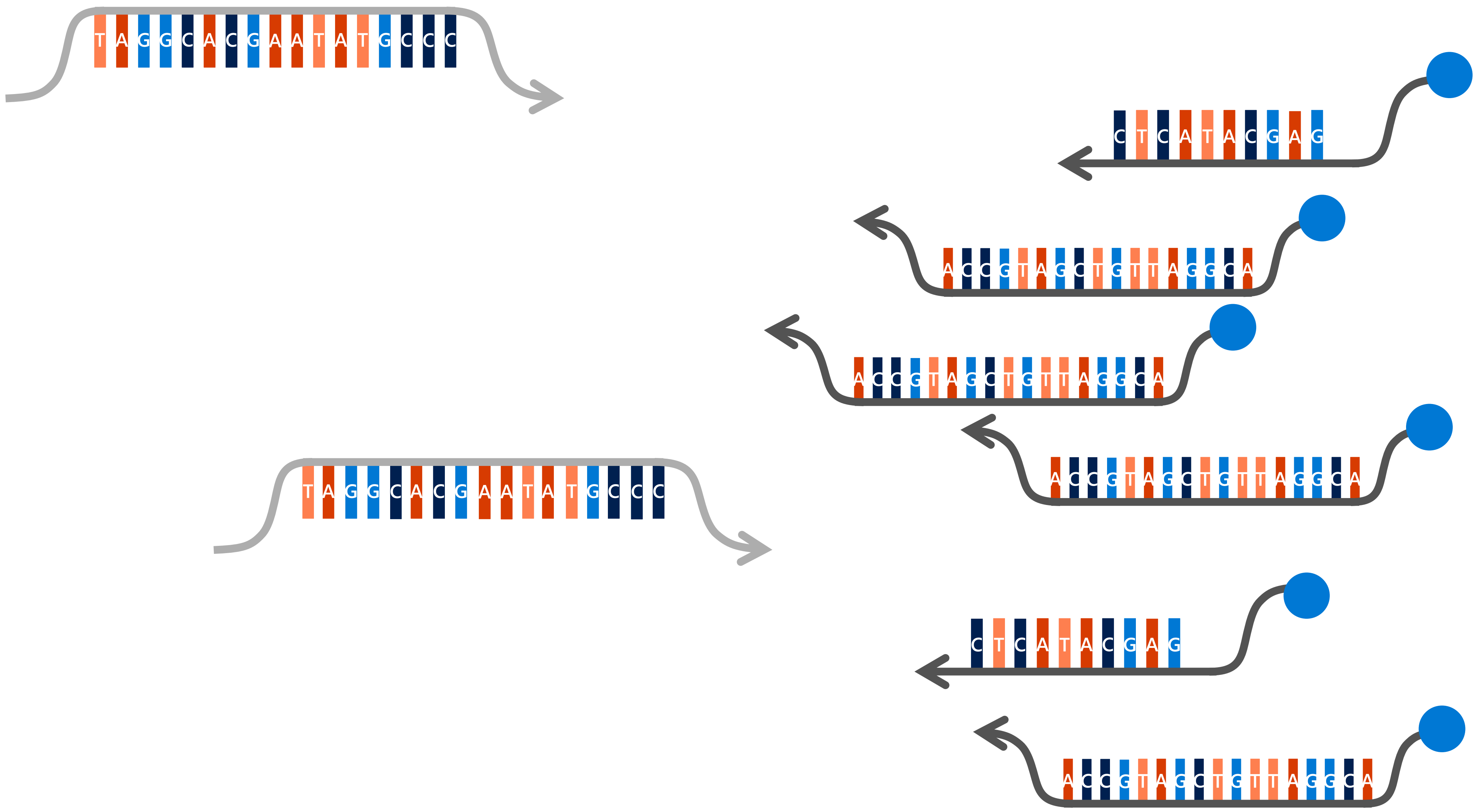


# “Query” strands
*(one type)*

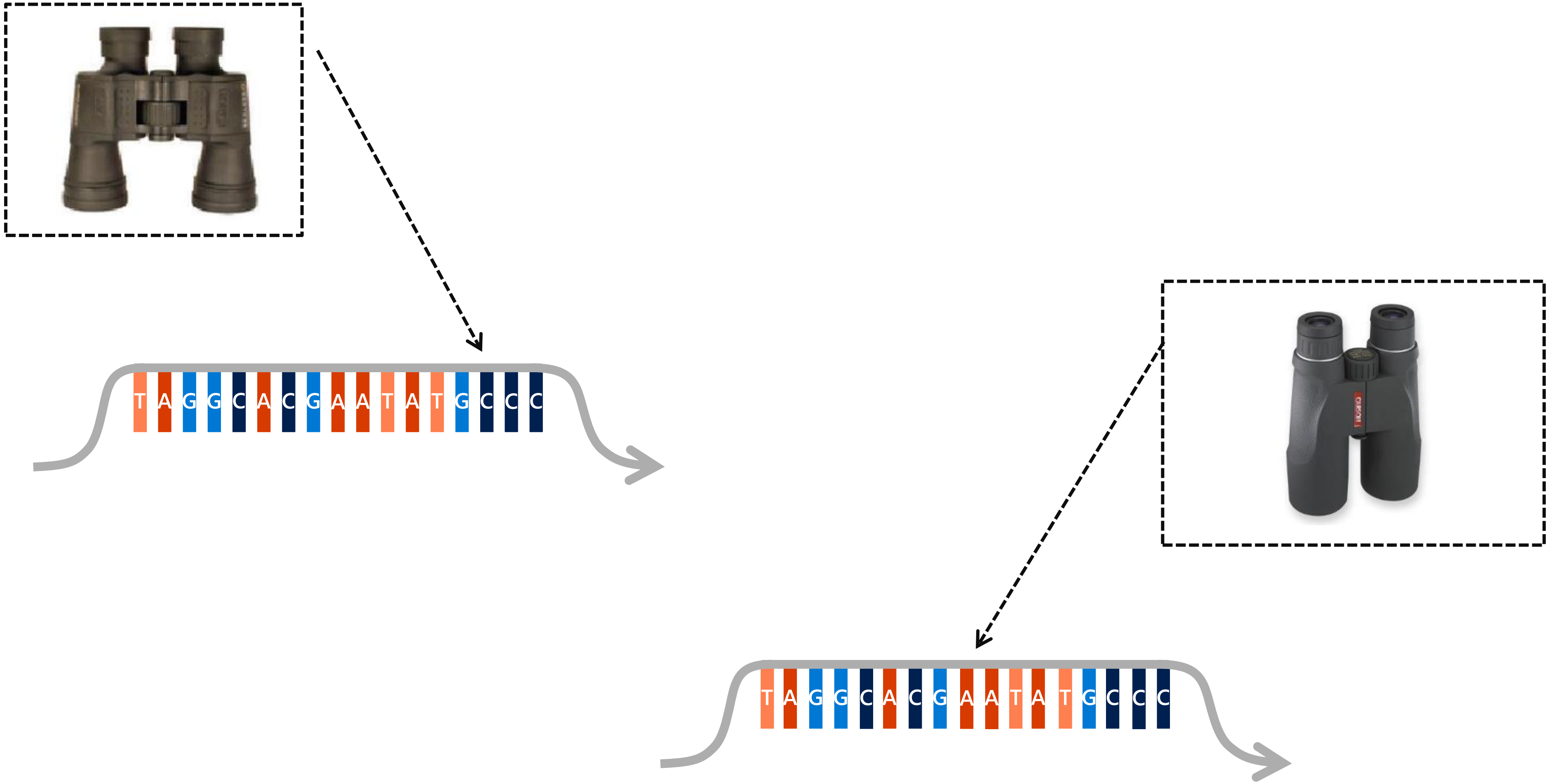





TAGGCACAAGTATGCCC
GTGCCACATTAGCACGA
ACGGATCGTCAATGCCA
GCCGATAATCATAGCGC
CCGTAGAGCCACTGTAA
TAGGCACGAATATGCCC
CTCATACGAG
ACCGTAGCTGTTAGGCA
ACCGTAGCTGTTAGGCA
ACCGTAGCTGTTAGGCA
TAGGCACGAATATGCCC
CTCATACGAG
ACCGTAGCTGTTAGGCA

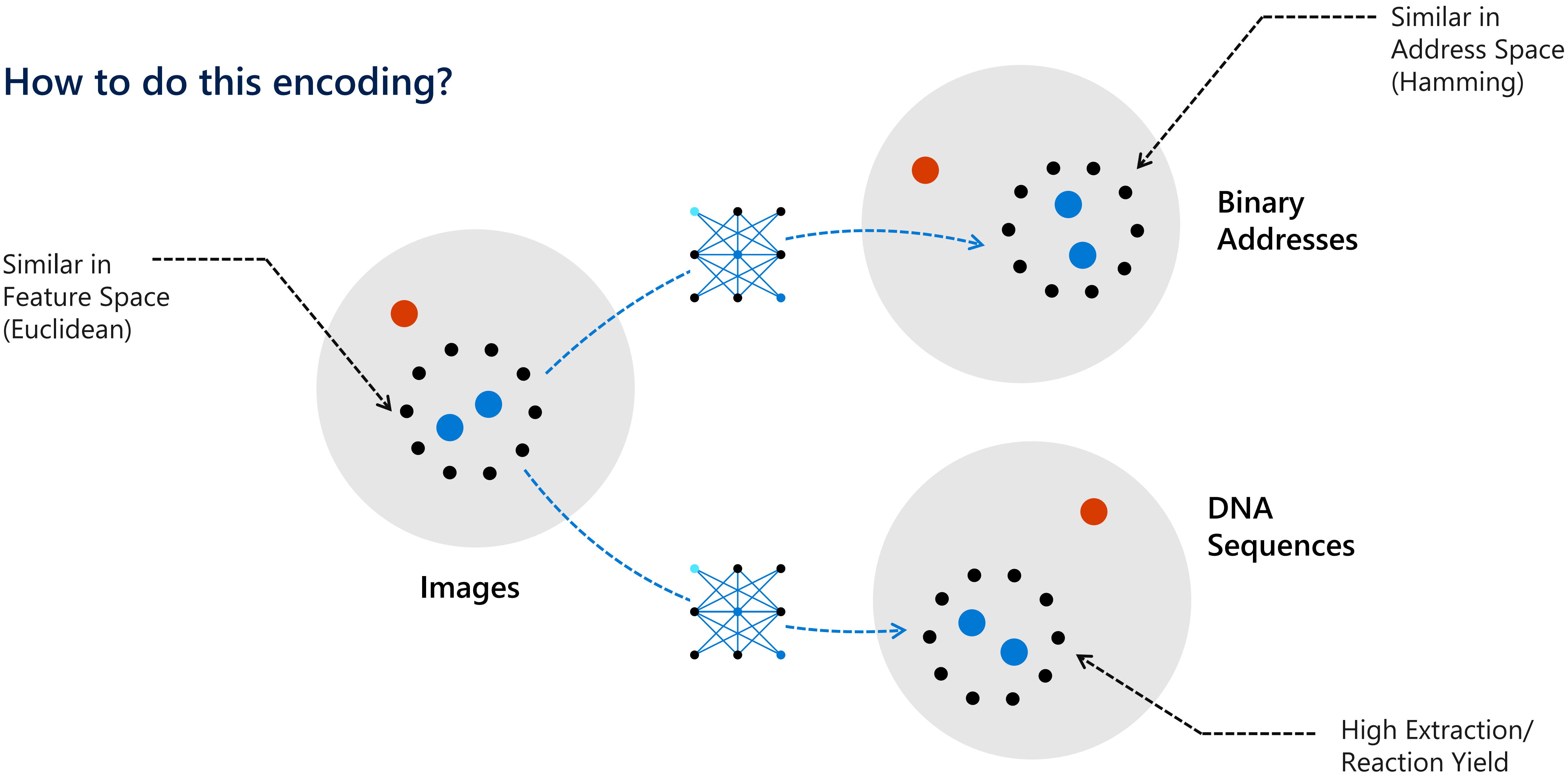


T A G G C A C G A A T A T G C C C
C T C A T A C G A G
A C C G T A G C T G T T A G G C A
A C C G T A G C T G T T A G G C A
A C C G T A G C T G T T A G G C A
T A G G C A C G A A T A T G C C C
C T C A T A C G A G
A C C G T A G C T G T T A G G C A


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

# “Semantic” Hashing

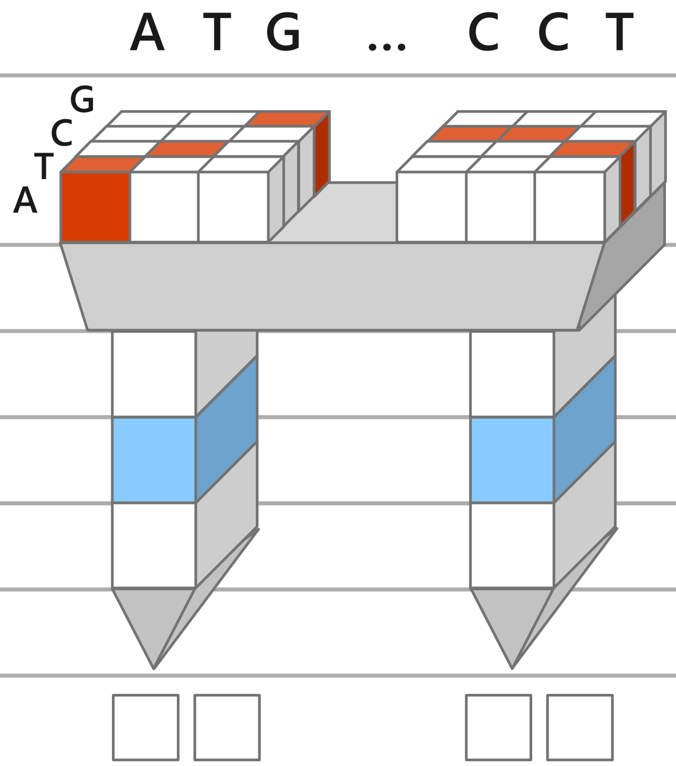
## How to do this encoding?



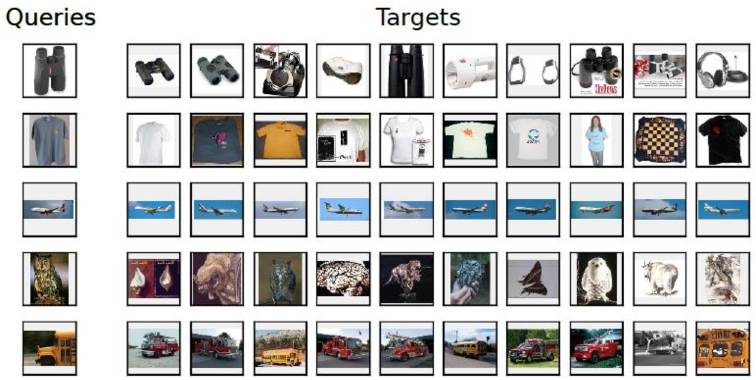

Adapted from Salakhutdinov et al. 2007

# Learning-based encoding

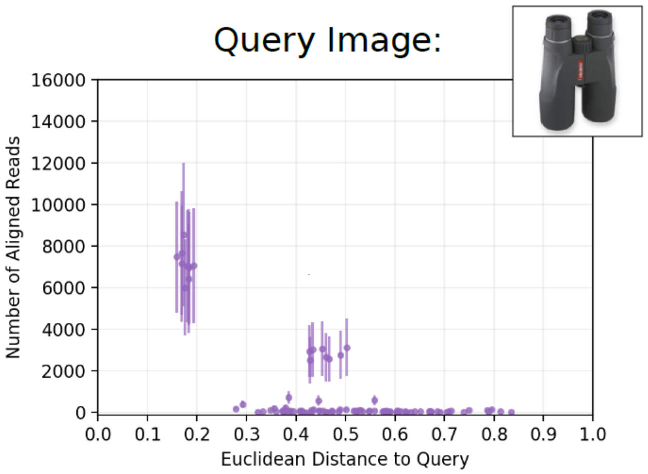
| Layer | Size |
|---|---|
| Sequence Output | 30 x 1 |
| ReLU + Softmax Activations | 30 x 4 |
| Fully Connected Weights | 10 x 128 x 30 x 4 |
| ReLU Activations | 10 x 128 |
| Convolutional Weights 2 | 1 x 128 x 128 |
| Sine Activations | 10 x 128 |
| Convolutional Weights 1 | 1 x 128 |
| Input Features | 10 x 1 |

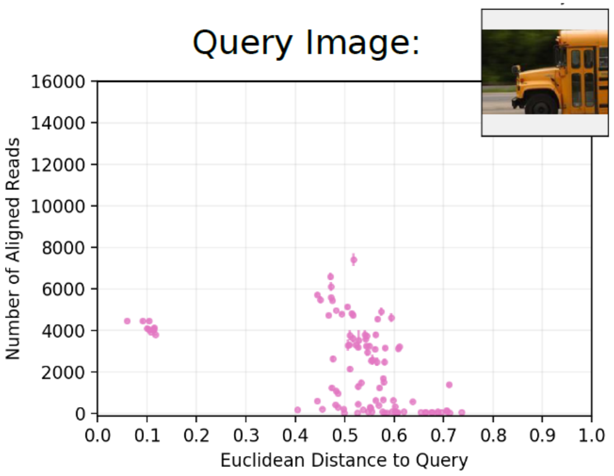


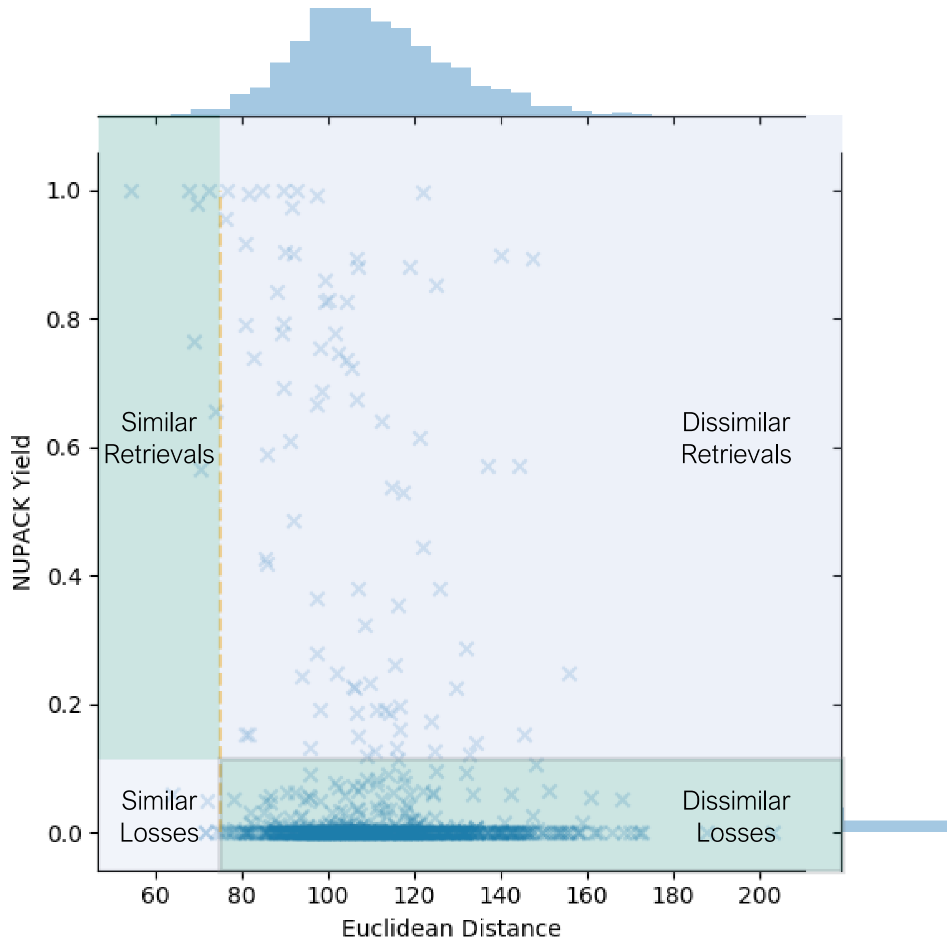
# Experiments show encouraging results



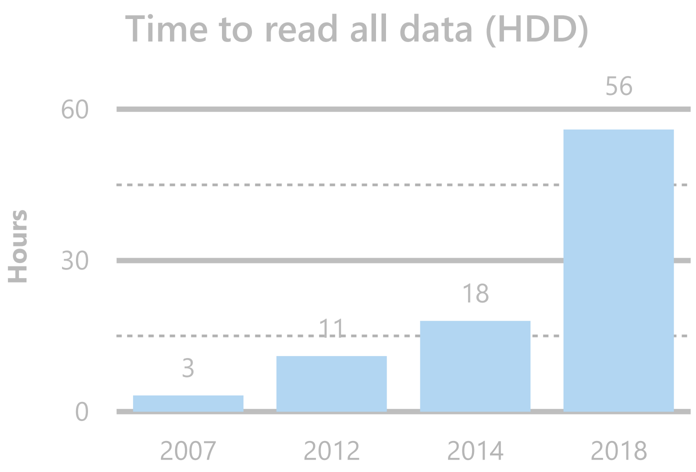

1.6M images, 3 queries
Magnetic bead extraction
Illumina-based sequencing



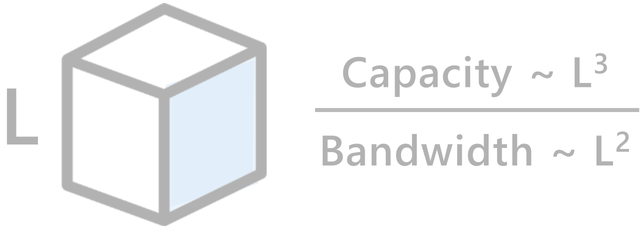



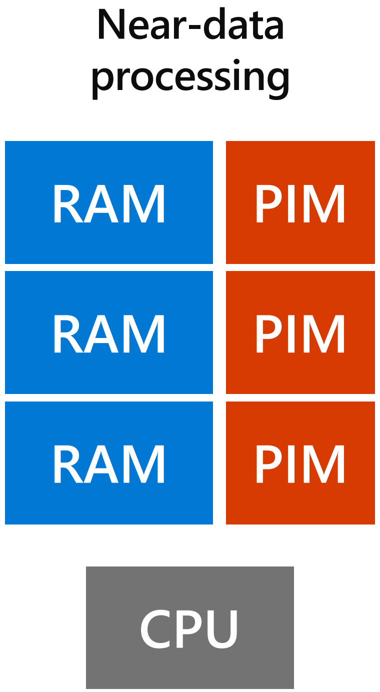
# Yottabyte-scale near-molecule computing?

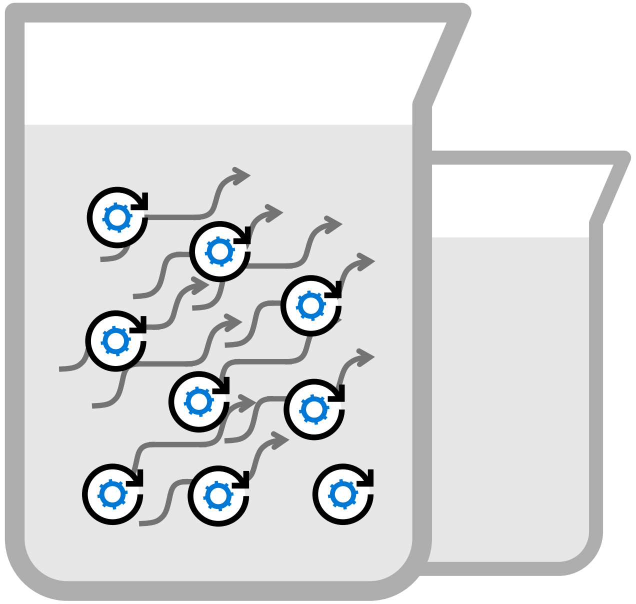



Capacity/bandwidth going up



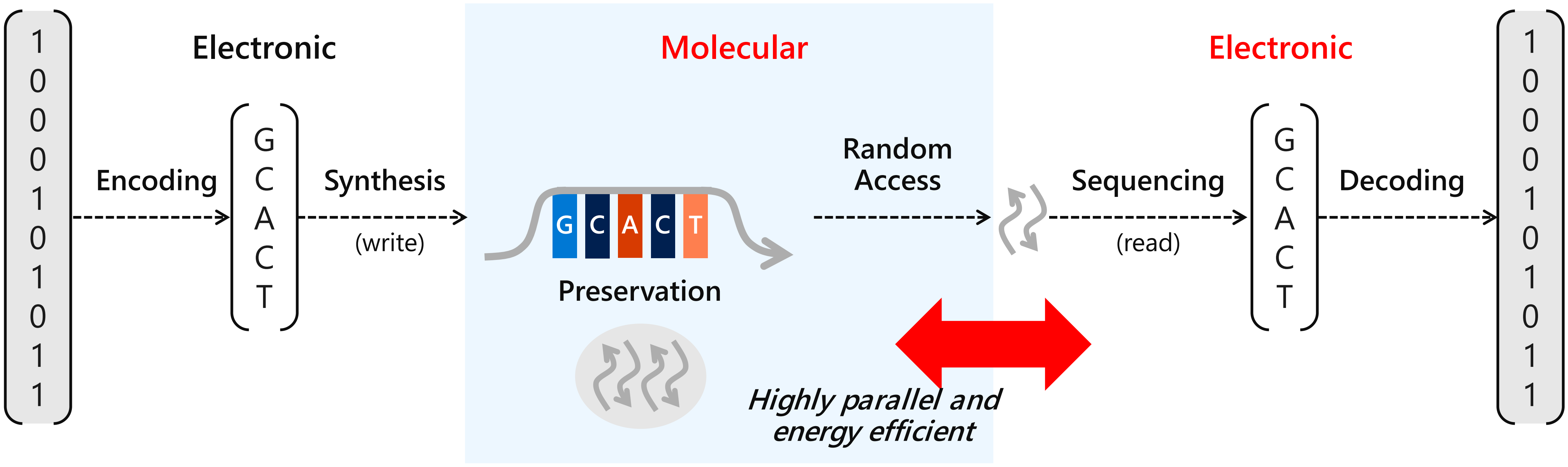

Carmean et al., IEEE Proceedings, 2018



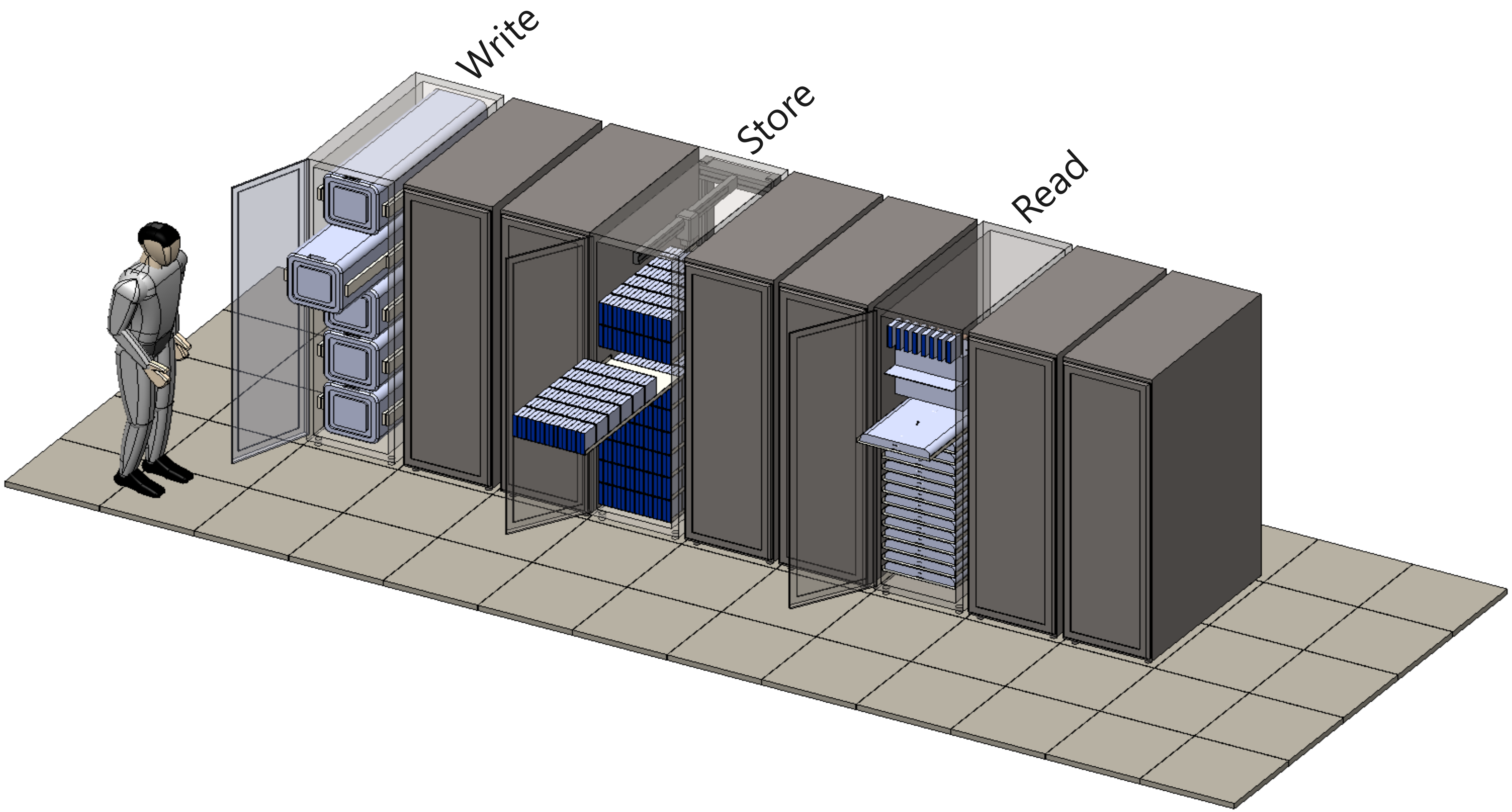



Physically "diffusing" computation through data offers parallelism and virtually unlimited access bandwidth. Yes, at a higher latency.

# DNA storage end-to-end system w/ integrated computing

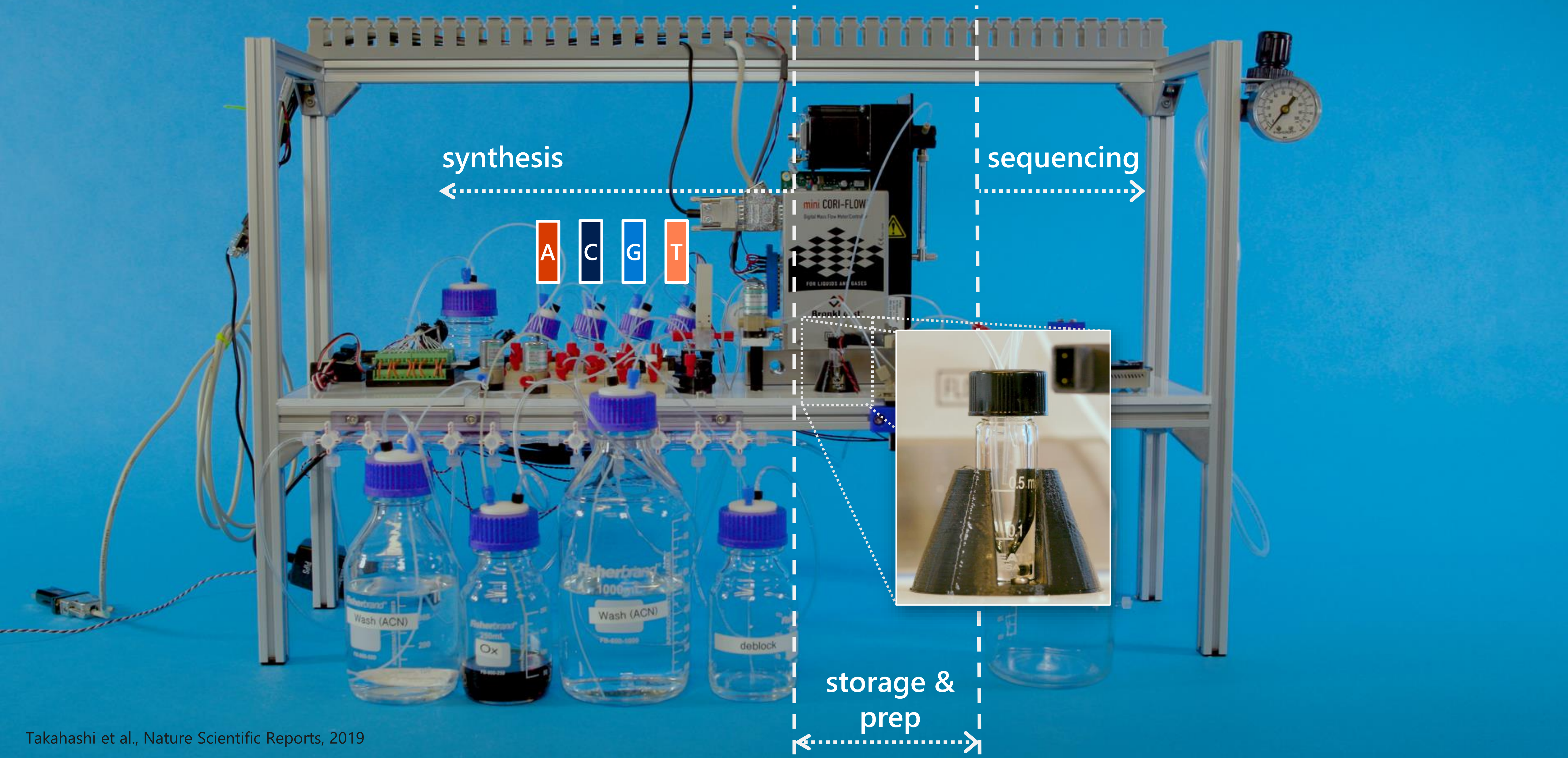


# End-to-end system in a datacenter



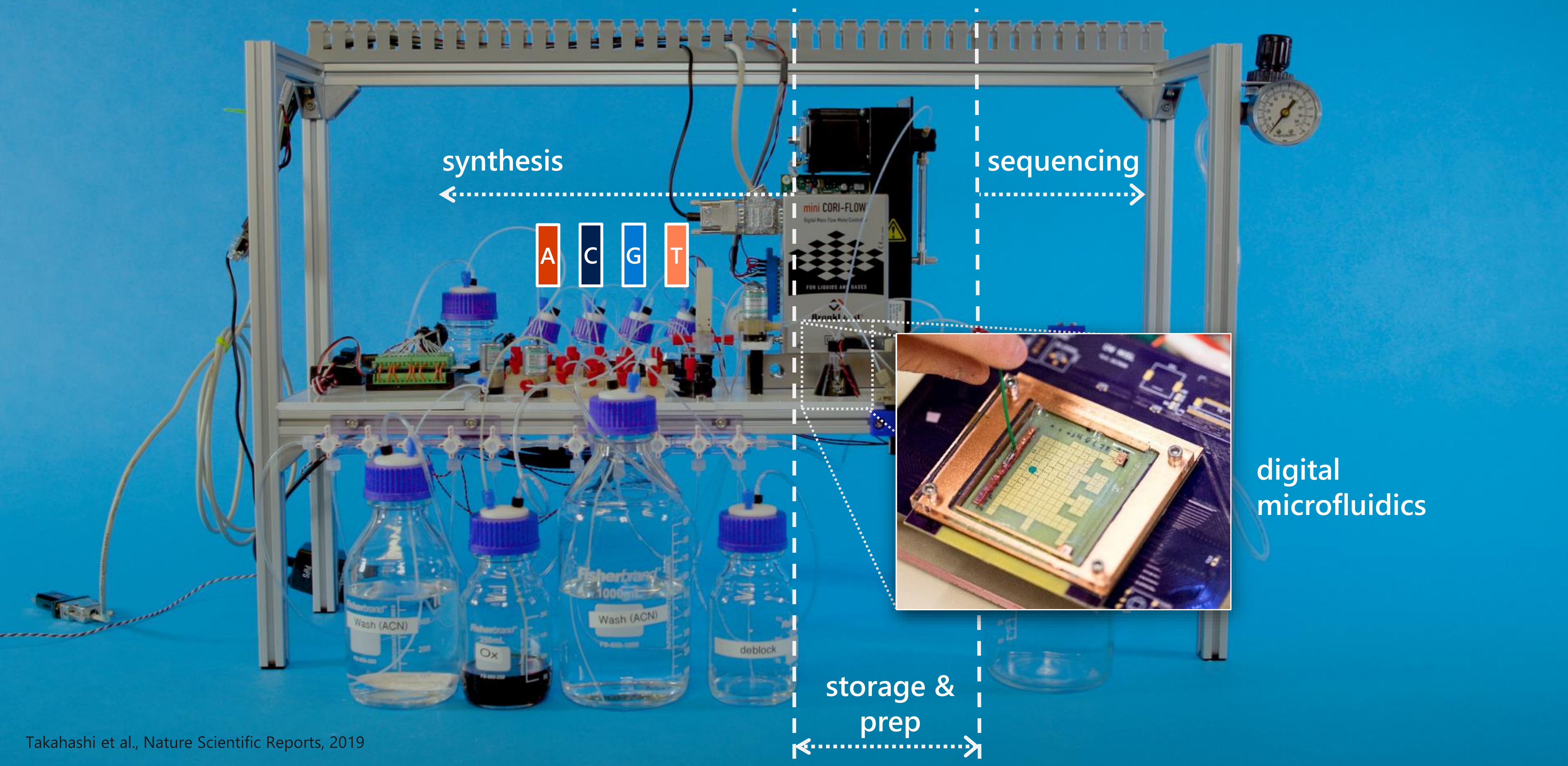

Mate
HAND WASH

# First fully automated DNA data storage system




Takahashi et al., Nature Scientific Reports, 2019

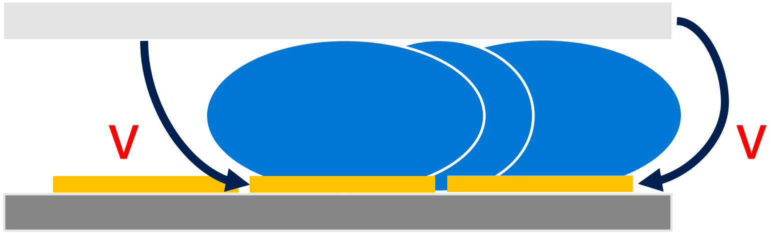
# First fully automated DNA data storage system

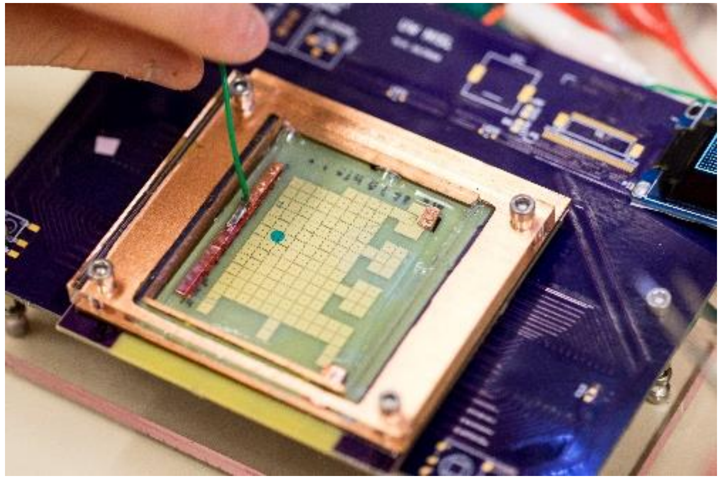


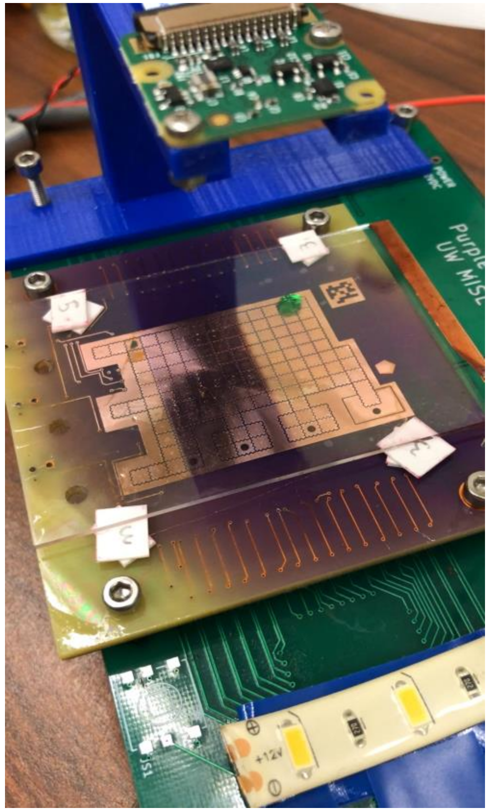

Takahashi et al., Nature Scientific Reports, 2019

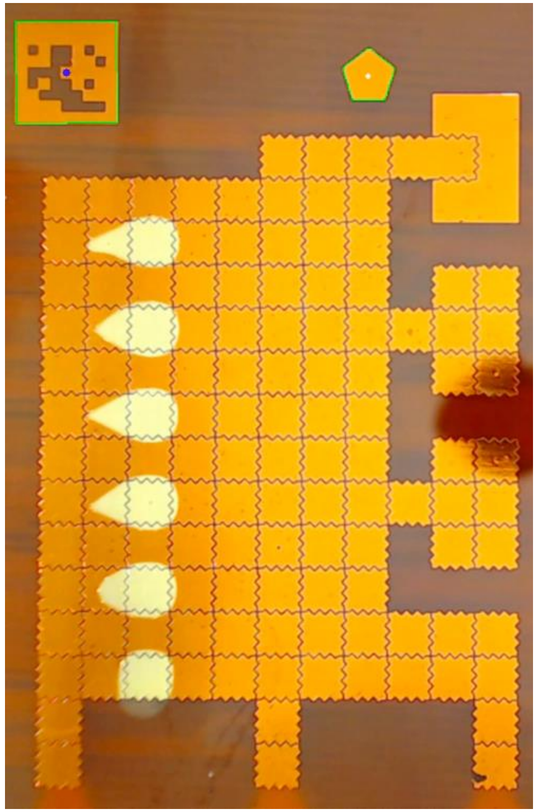
# Digital microfluidics

Versatile platform to implement wet lab preparation protocols



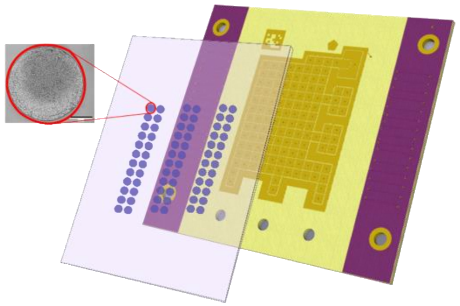



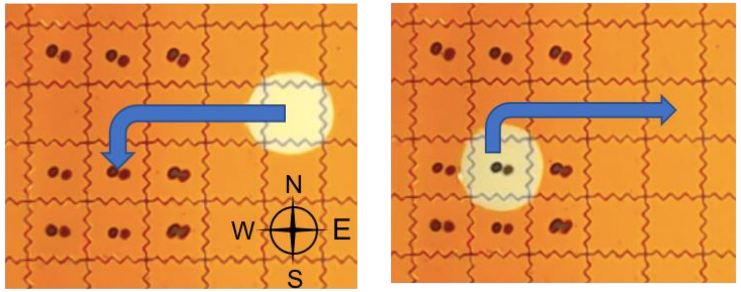




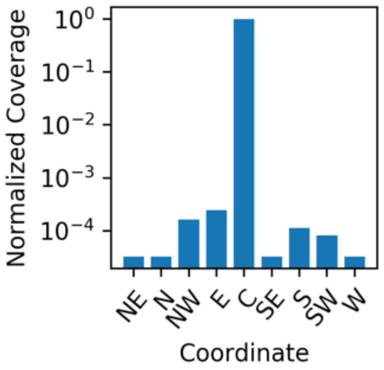
Newman et al., Nature Communications, 2019; Willsey, ASPLOS, 2019; Stephenson et al., IEEE Micro, 2020

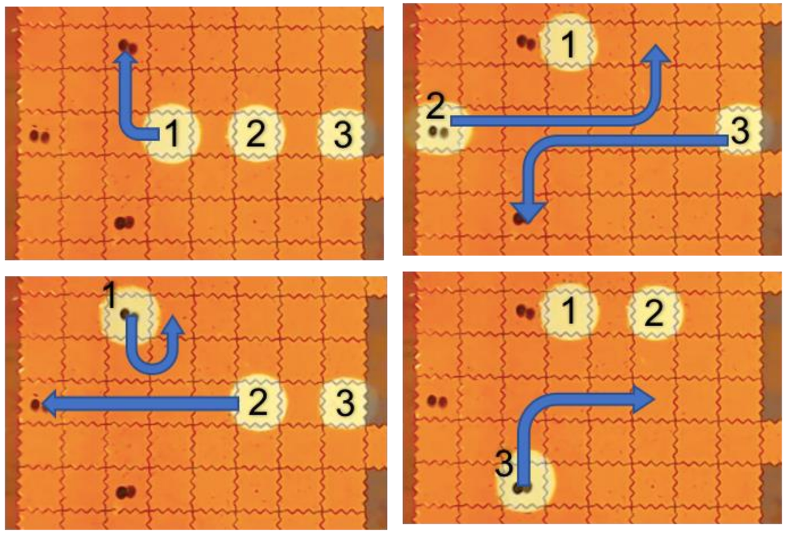
# Random access with spots+digital microfluidics






60s dwell time
33ng mass




No measurable contamination

[Stephenson, Takahashi, Nguyen, et al.., Nature Communications'19]

# Affordable full-stack SW/HW digital microfluidics platform

**High-level programming with *Puddle***

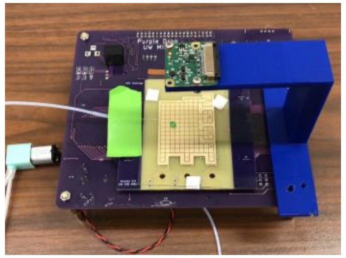
```
def thermocycle(droplet, temps_and_times):
    for temp, time in temps_and_times:
        heat(droplet, temp, time)
    if droplet.volume < MIN_VOLUME:
        droplet += input("water", min_volume)

def pcr(droplet, n_iter):
    thermocycle(droplet, n_iter * [
        (95,  3 * minutes),
        (62, 30 * seconds),
        (72, 20 * seconds),
    ])
```

**"Assembly code"**

```
activate(3,0)
activate(3,1)
activate(3,2)
...
```

**Hardware**



Willsey et al., ASPLOS, 2019; Stephenson et al., IEEE MICRO 2020.

| Molecular Computing | Experiment | Medical Diagnostics |
|---|---|---|
| Synthetic DNA | Chemistry | Medicine |

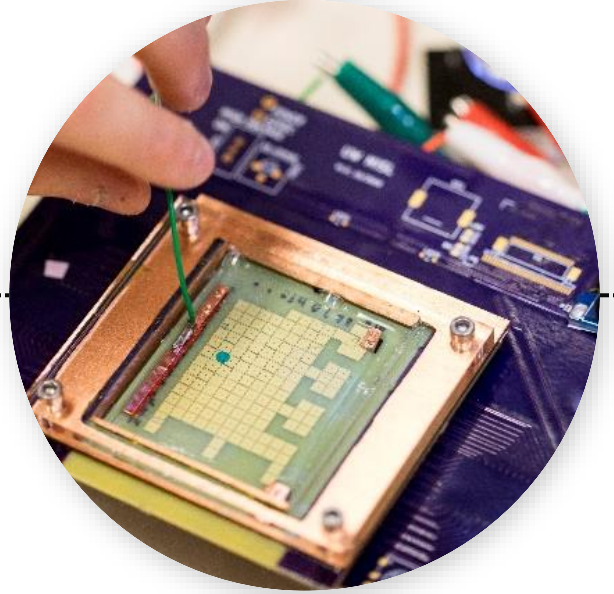
Domain Specific Reasoning

Core Fluidic Semantics

Hardware Abstraction

Microfluidic Chips

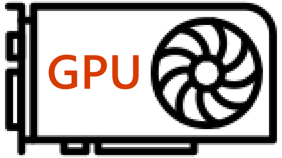
# Hardware, software, wetware

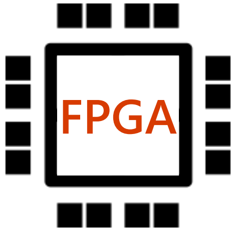


Molecular domain

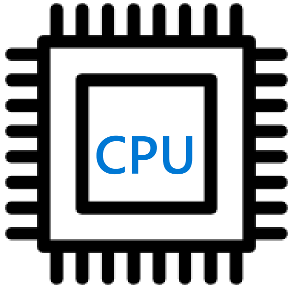
Electronic domain

# Future hybrid systems



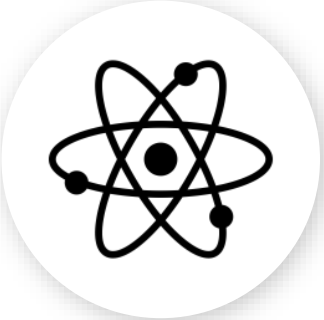

**Special purpose**




**General purpose**

**Electronics:**
Ultra low latency,
engineerable, perfect control

**Biomolecules:**
Self assembly, massive
data and efficiency



**Quantum:**
Massive specialized parallel
computing, little data

# Questions?

https://misl.cs.washington.edu

https://www.microsoft.com/en-us/research/project/dna-storage/