A DNA-Based Archival Storage System

Luis Ceze and Karin Strauss

University of Washington
Microsoft Research

MISL

W

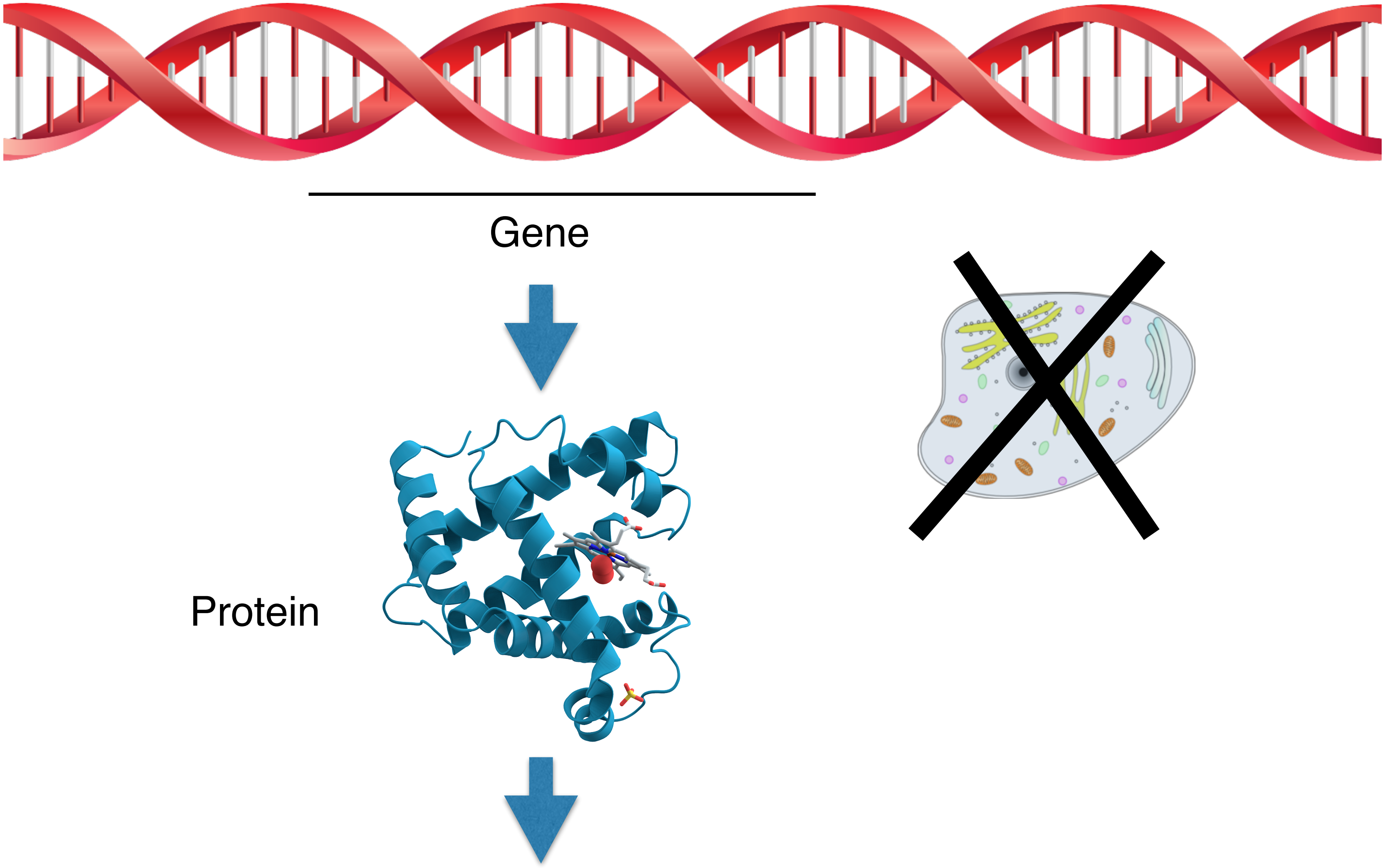
Microsoft® Research

joint work with Doug Carmean, Georg Seelig, James Bornholt, Randolph Lopez, Lee Organick, Rob Carlson, Hsing-Yeh Parker, Yuan Chen, Chris Takahashi, Bichlien Nguyen, Sergey Yekhanin, Siena Dumas Ang, Sharon Newman.

Library of Congress, Sep 2016.

© University of Washington and Microsoft Research. All rights reserved.

DNA is the information storage medium for life



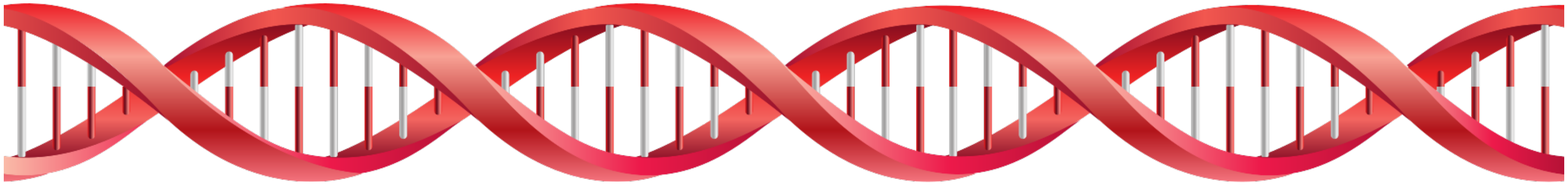
Gene

Protein

Function/Characteristic

© University of Washington and Microsoft Research. All rights reserved.

Using *synthetic* DNA for data storage



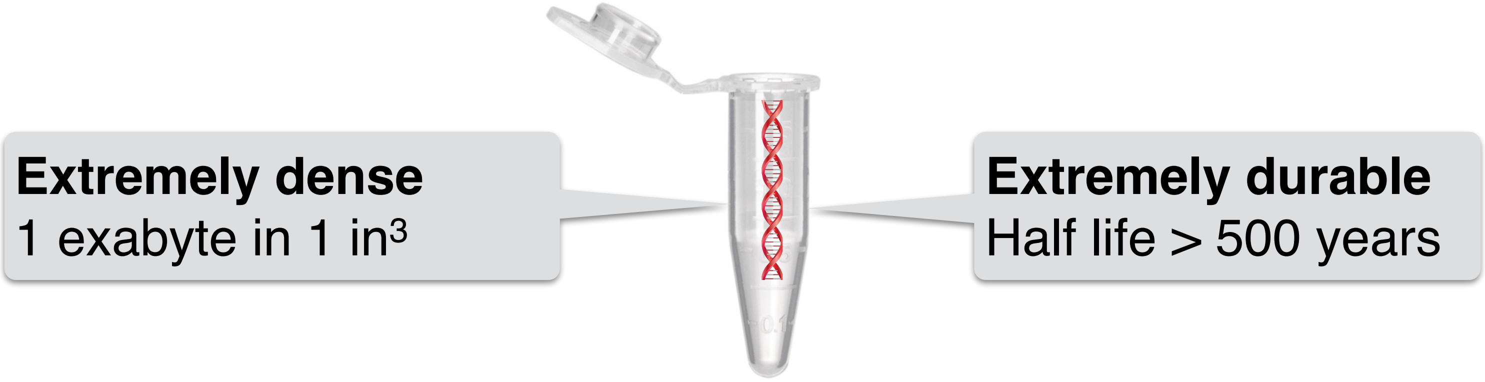
Manufacture DNA
Dehydrate & store
Read DNA

100101010

But why?

© University of Washington and Microsoft Research. All rights reserved.

DNA molecules for digital data



Extremely dense
1 exabyte in 1 in^3

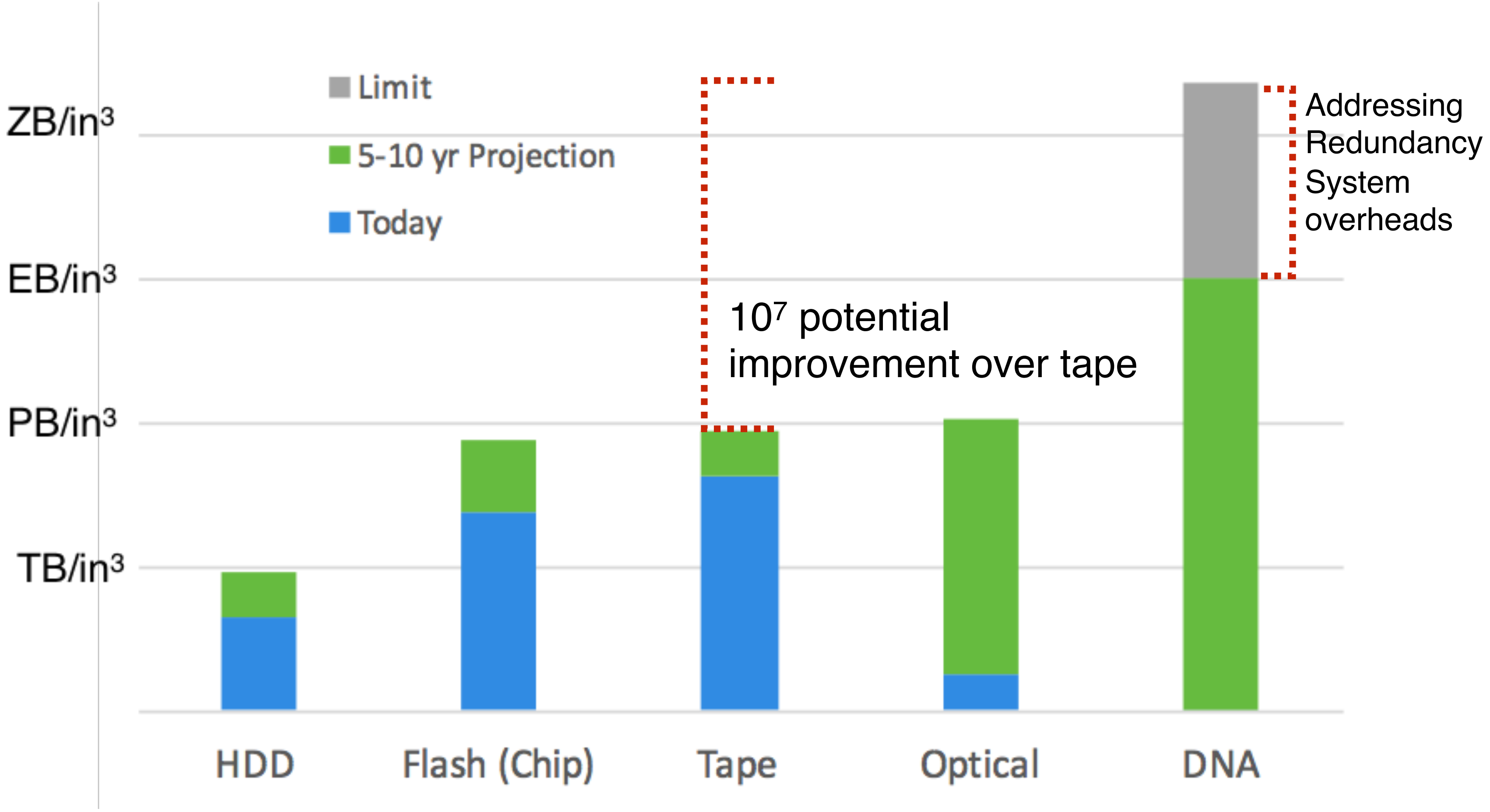
Extremely durable
Half life > 500 years

Readers never become obsolete!
(no migration :)

And consumes very little power at rest.

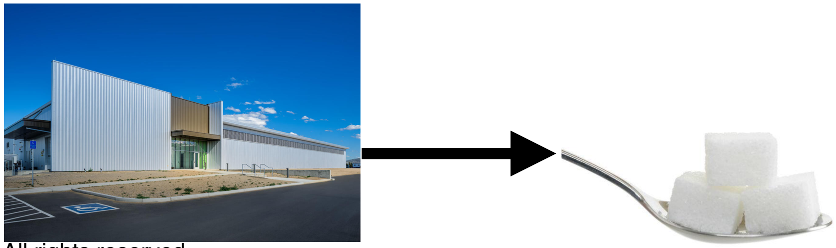
© University of Washington and Microsoft Research. All rights reserved.

Comparing storage density



10^7 potential improvement over tape

Addressing
Redundancy
System
overheads



© University of Washington and Microsoft Research. All rights reserved.

The ultimate storage hierarchy

	Access Time	*Capacity*	*Durability*
Flash	μs-ms	TBs	~5 yrs
HDD	10s ms	100s TBs	~5 yrs
Tape	minutes	PBs	~10s yrs
DNA-based Archival	hours	ZBs	~100s yrs

Our goal: build an integrated DNA storage system.

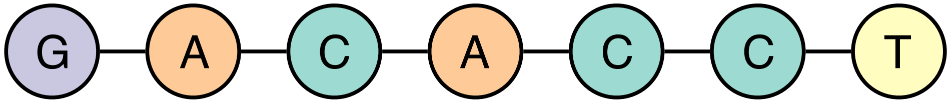
© University of Washington and Microsoft Research. All rights reserved.

DNA molecules

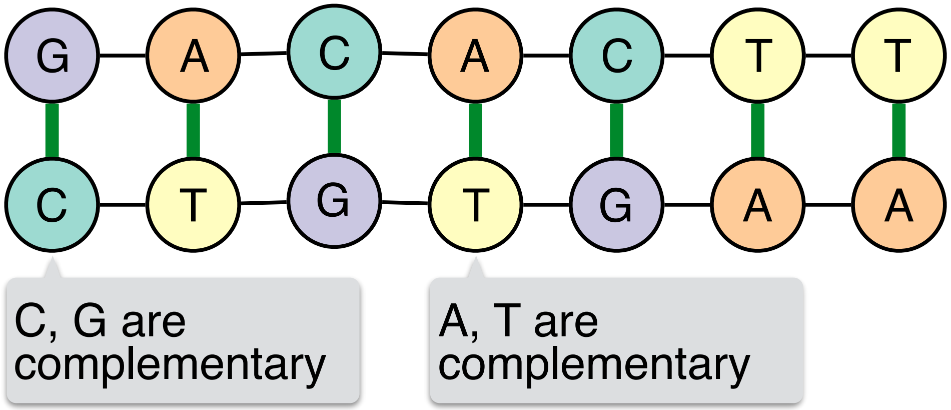
Four nucleotides:

- A Adenine
- C Cytosine
- G Guanine
- T Thymine

DNA strand (oligonucleotide) is a linear sequence of these nucleotides



Two strands can bind to each other if they are complementary:

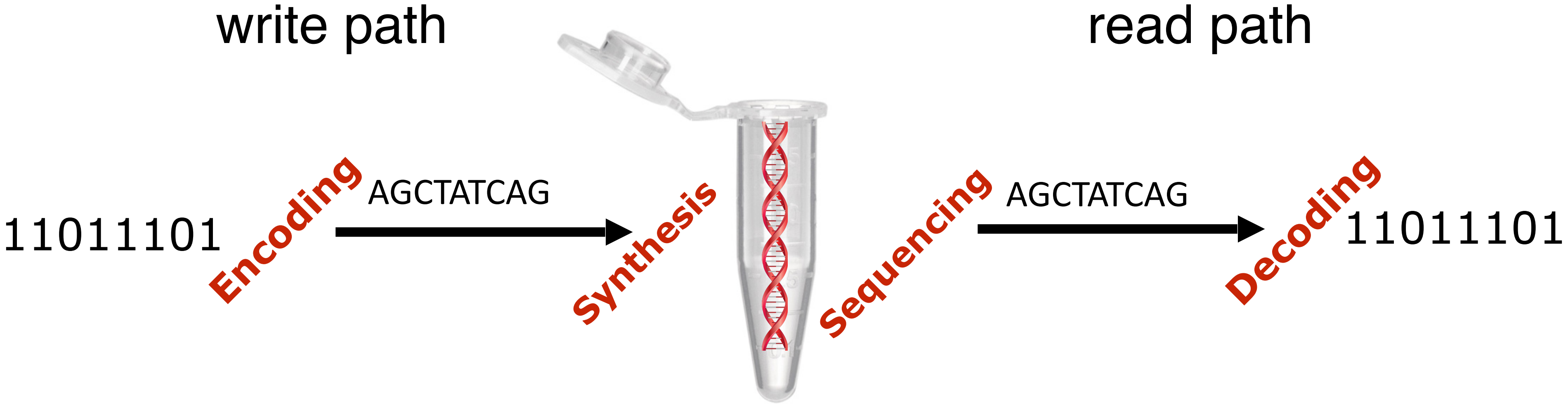


C, G are complementary

A, T are complementary

© University of Washington and Microsoft Research. All rights reserved.

DNA data storage at 30,000 feet



write path

read path

11011101

Encoding

AGCTATCAG

Synthesis

Sequencing

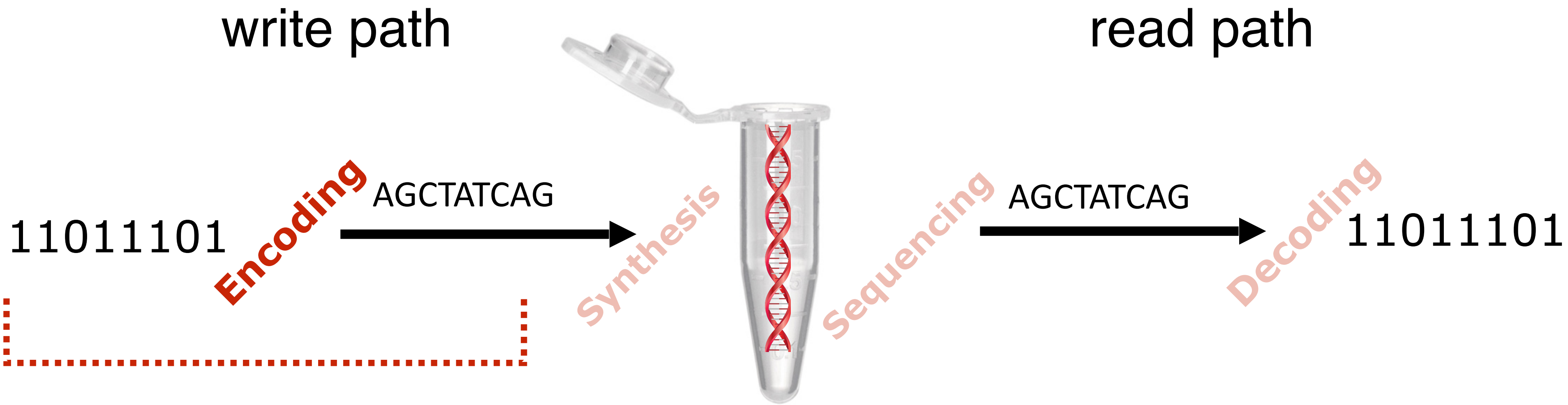
AGCTATCAG

Decoding

11011101

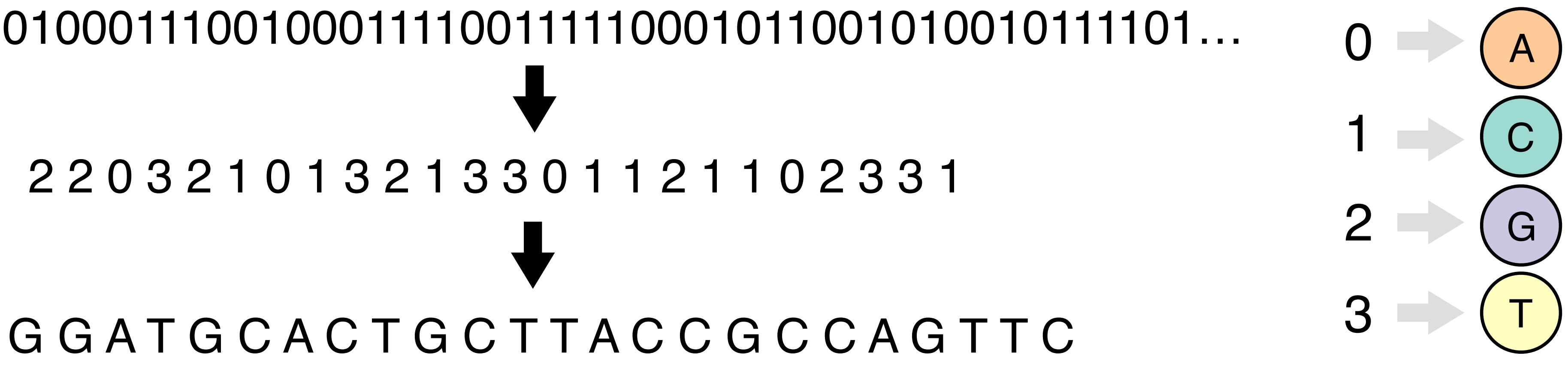
© University of Washington and Microsoft Research. All rights reserved.

DNA data storage at 30,000 feet



© University of Washington and Microsoft Research. All rights reserved.

Encoding digital data in DNA

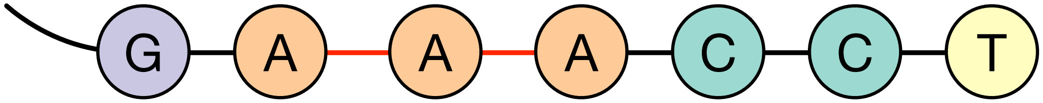


101000111001000111100111110001011001010010111101…

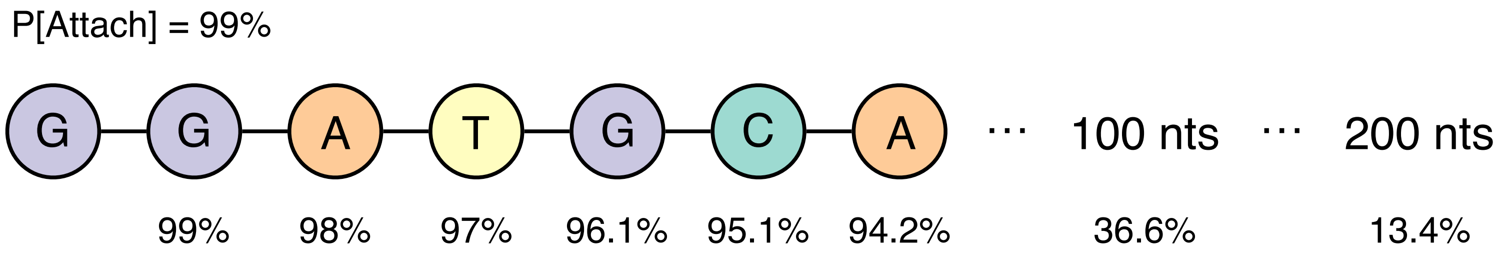
2 2 0 3 2 1 0 1 3 2 1 3 3 0 1 1 2 1 1 0 2 3 3 1

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

Repeated letters are bad: Use base 3 and “rotate” mapping.



Synthetic DNA sequences have limited length: Break it up



P[Attach] = 99%

© University of Washington and Microsoft Research. All rights reserved.

MISL UW MSR

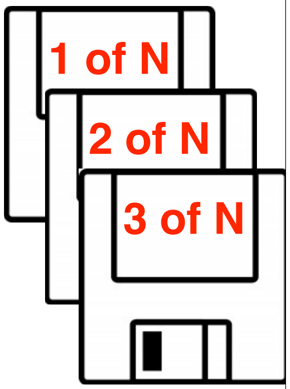
Breaking up data into chunks (~150nts)

CGAT GCAC TGCT GACG GCTA GCTC

ATGTT	AAAA	CATCC
ATGTT	AAAC	CATCC
ATGTT	AAAG	CATCC

File identifiers (“primers”)

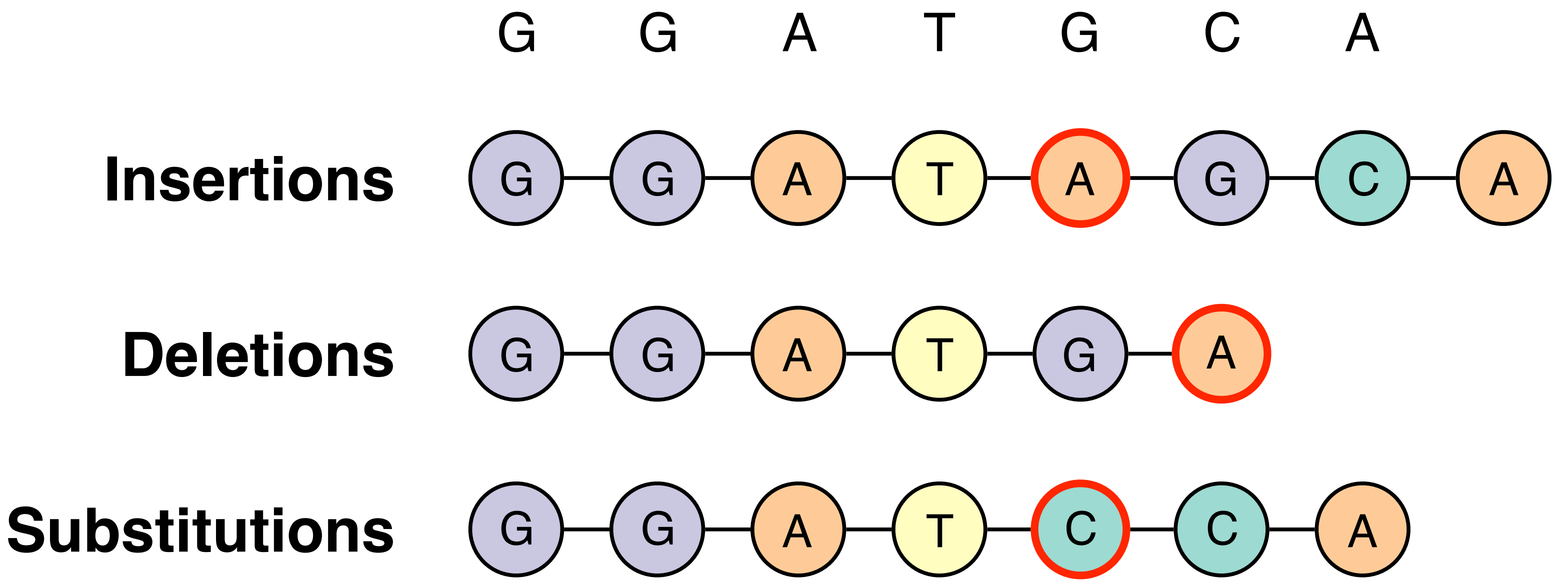
Addresses within the file



~ 20 bytes per DNA strand. Many strands per file.

© University of Washington and Microsoft Research. All rights reserved.

Errors in writing/reading DNA

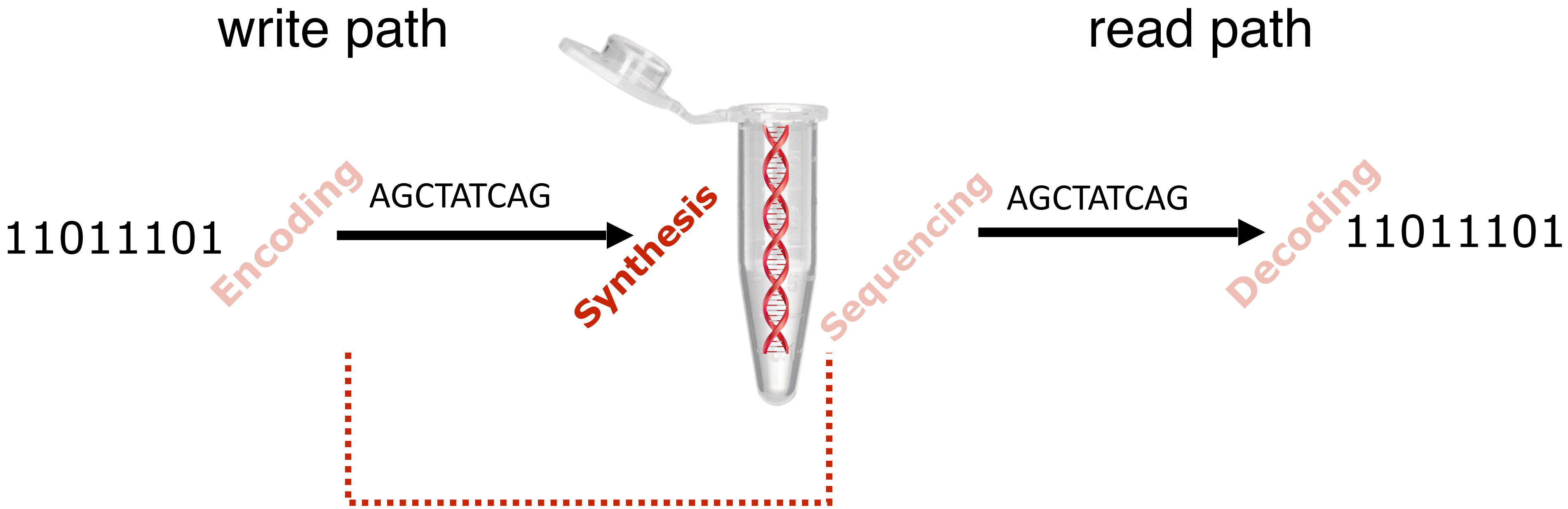


Aggregate error rates ~1%

Encode redundant data in additional DNA strands. Many possibilities: parity, Reed Solomon, LDPC, …

© University of Washington and Microsoft Research. All rights reserved.

DNA data storage at 30,000 feet

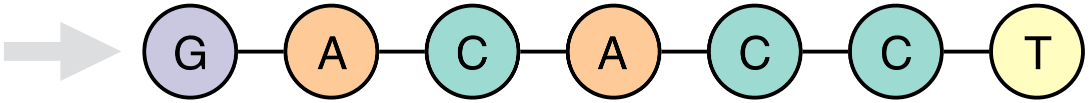


© University of Washington and Microsoft Research. All rights reserved.

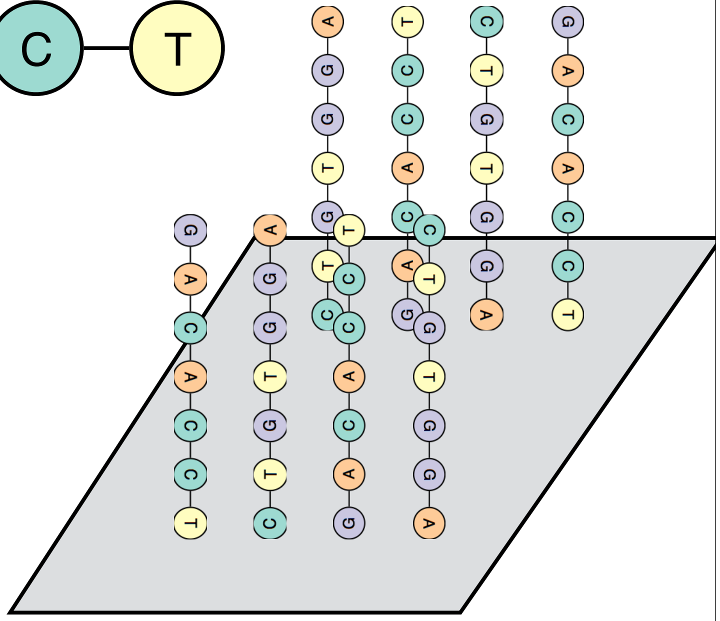
DNA Synthesis

Manufacturing DNA strands

GACACCT

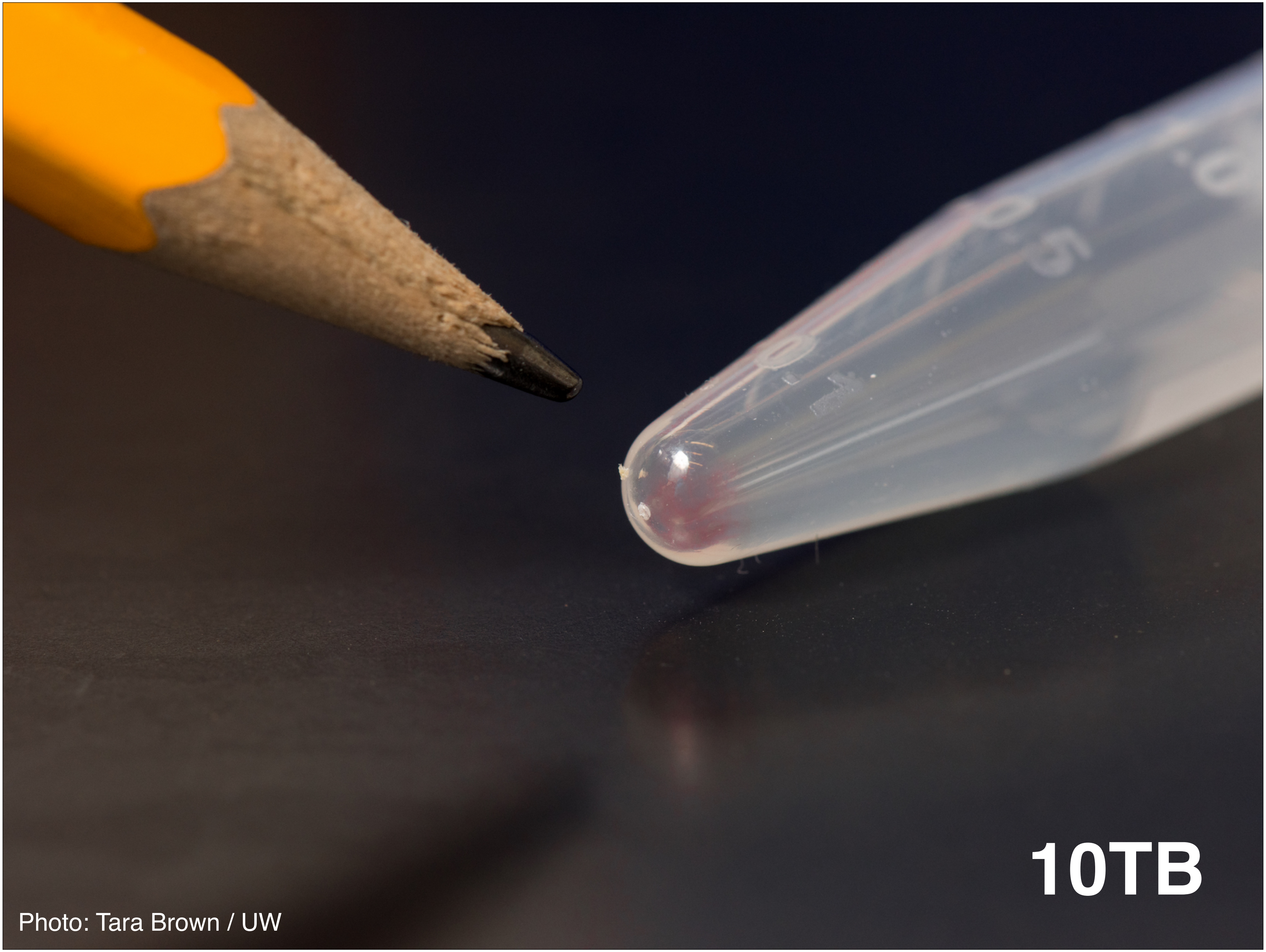


- Normally used for life sciences and medicine
- Millions of copies of each sequence
- Can make many different sequences in parallel



Twist Bioscience

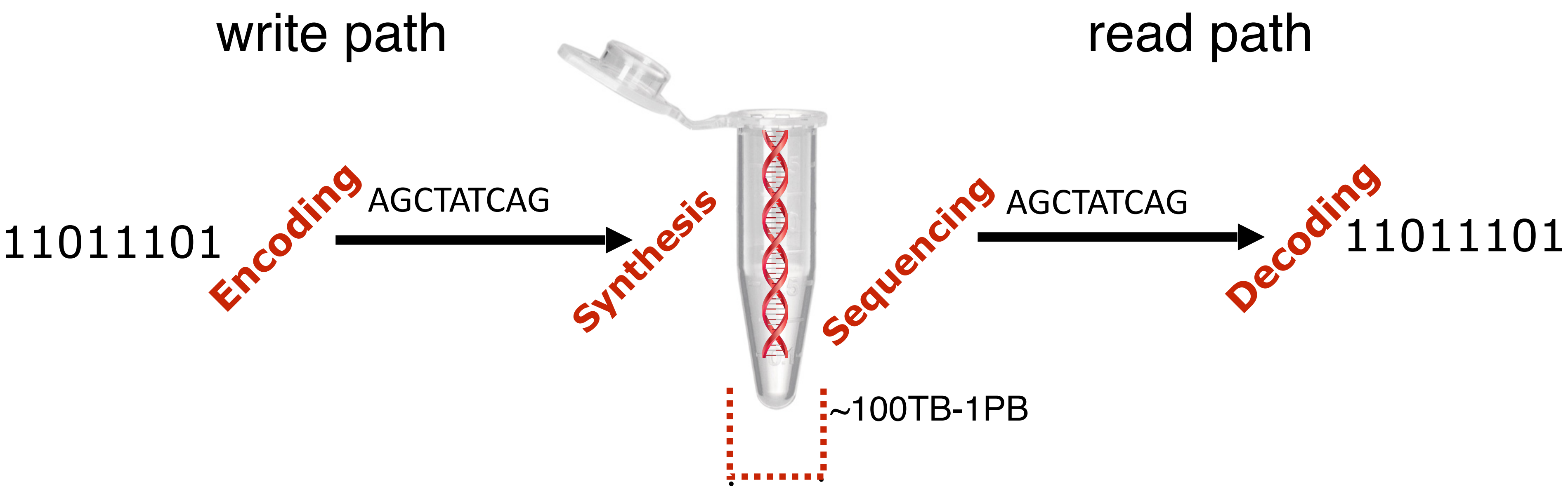
© University of Washington and Microsoft Research. All rights reserved.



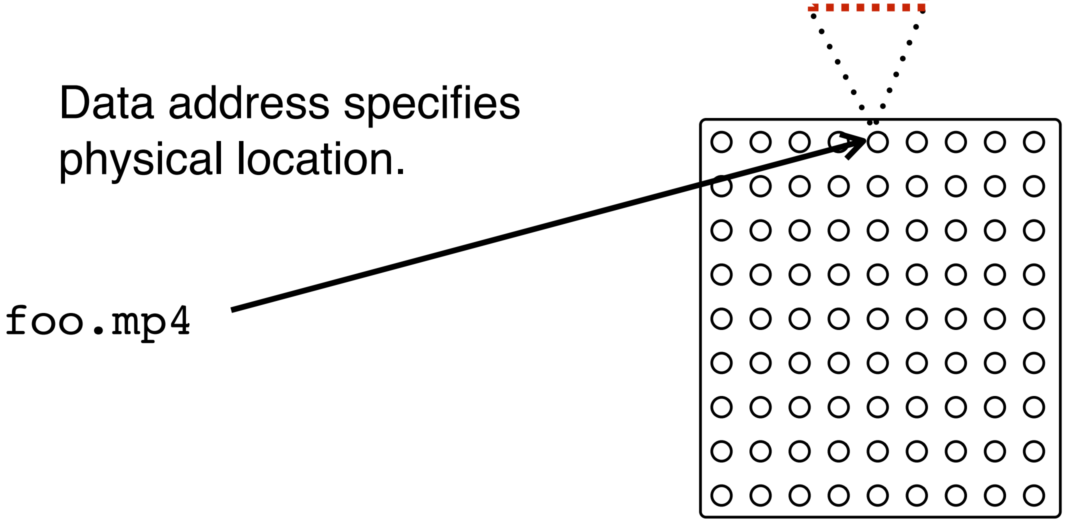
10TB

Photo: Tara Brown / UW

DNA Storage “Library”



Data address specifies physical location.

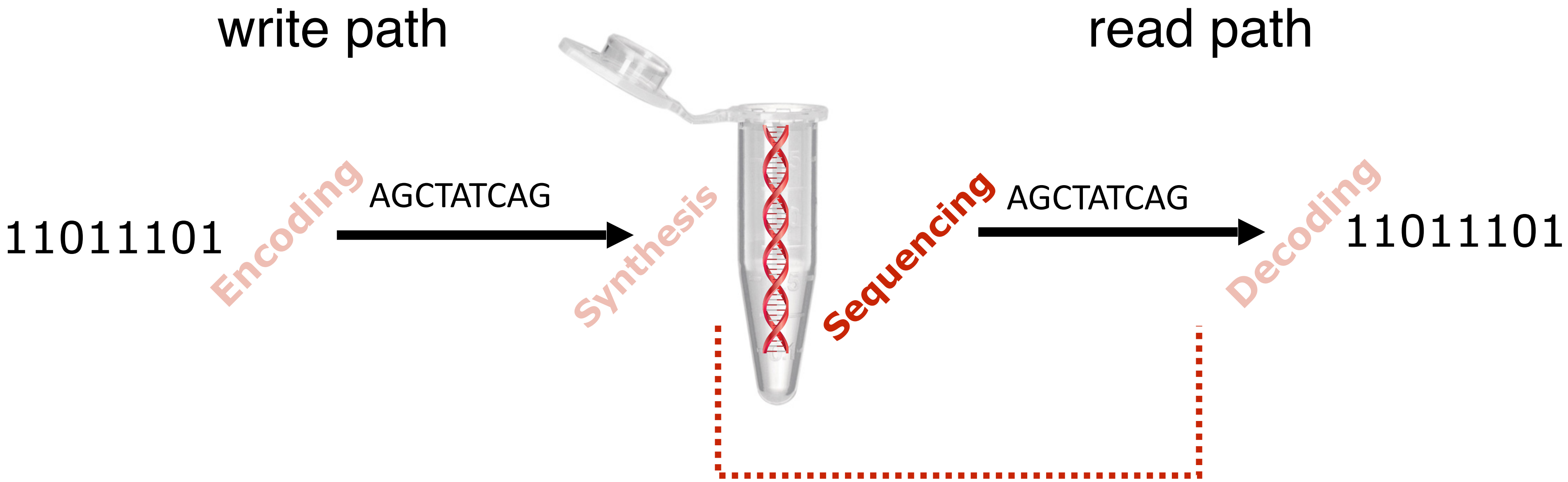


DNA storage (physical) library

© University of Washington and Microsoft Research. All rights reserved.

MISL UW MSR

DNA data storage at 30,000 feet



write path

read path

11011101

Encoding

AGCTATCAG

Synthesis

Sequencing

AGCTATCAG

Decoding

11011101

© University of Washington and Microsoft Research. All rights reserved.

Random access?

© University of Washington and Microsoft Research. All rights reserved.

Random access!

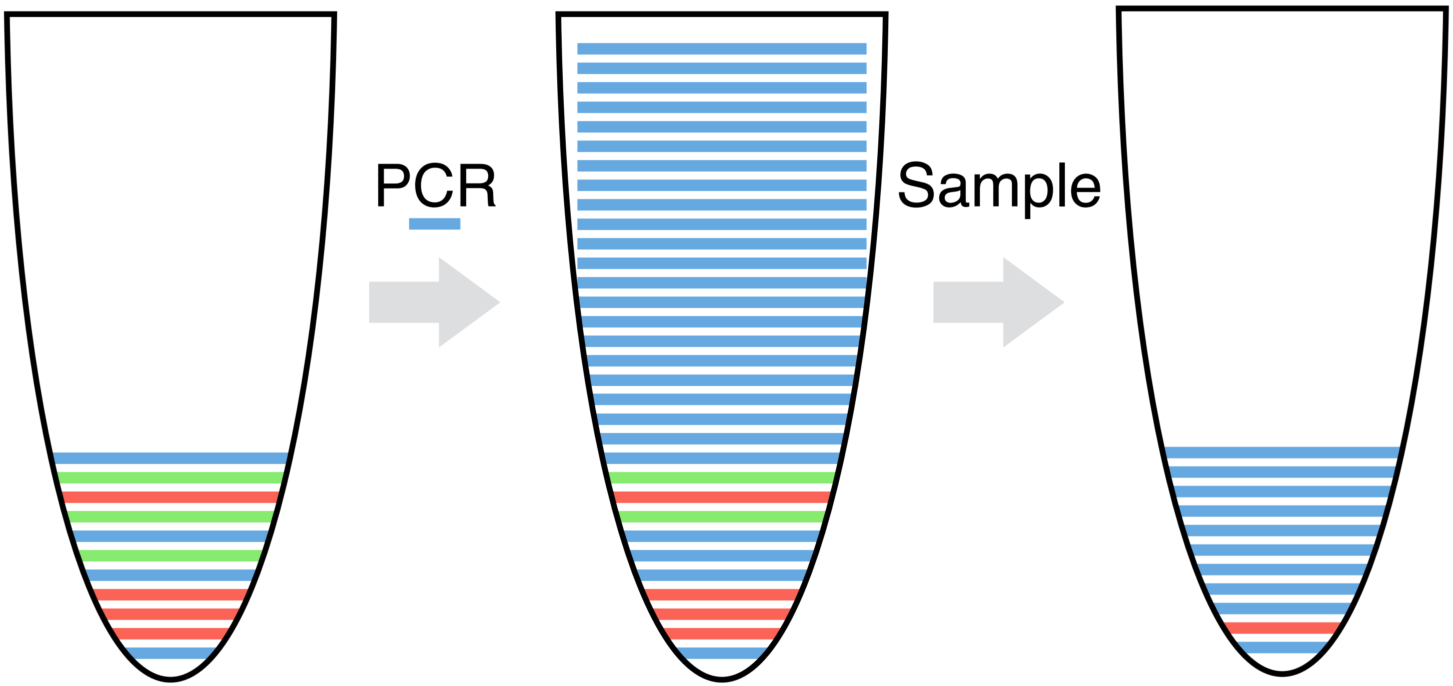
File identifiers ("primers")

ATGTT GGAT GCAC ACTA CATCC

CCACT TGCT TACC GATC GATAC

ATGTT GCCAGTTCAGCG TATCT

TACAA ATAGA

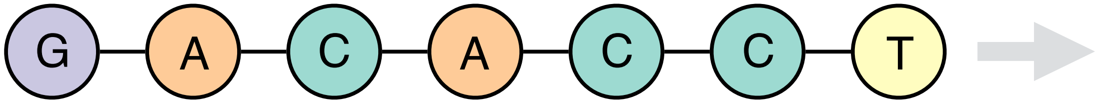


© University of Washington and Microsoft Research. All rights reserved.

MISL UW MSR

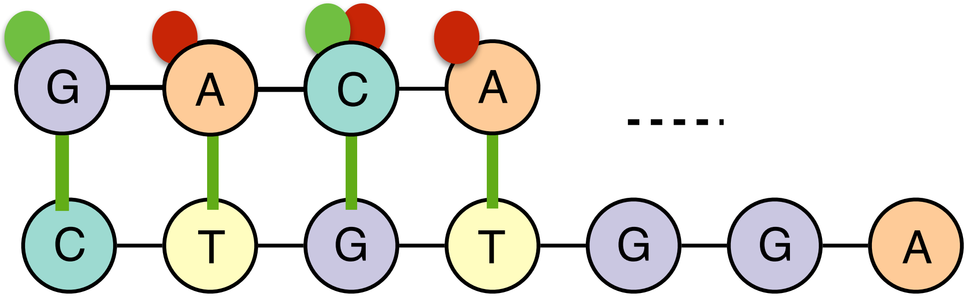
DNA Sequencing

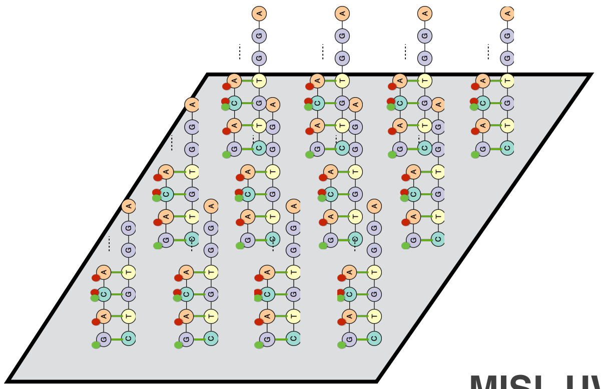
Reading DNA strands



GACACCT

- Normally used for genome sequencing
- Reads many copies of millions of DNA strands at a time
- Currently much higher throughput than synthesis

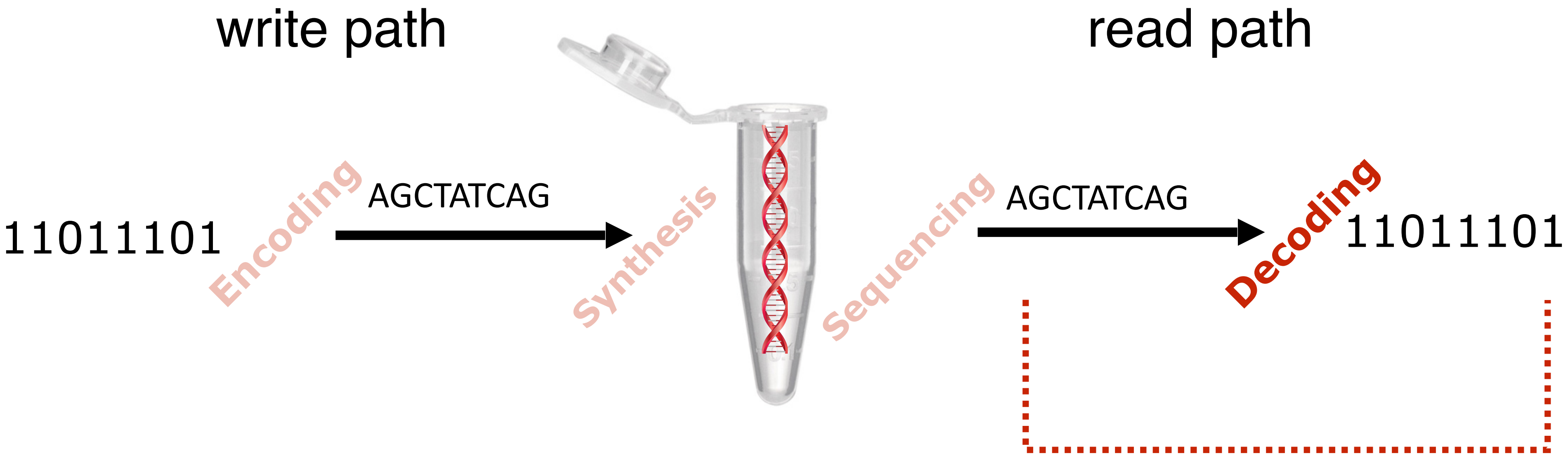




© University of Washington and Microsoft Research. All rights reserved.

MISL UW MSR

DNA data storage at 30,000 feet



write path

read path

11011101

Encoding

AGCTATCAG

Synthesis

Sequencing

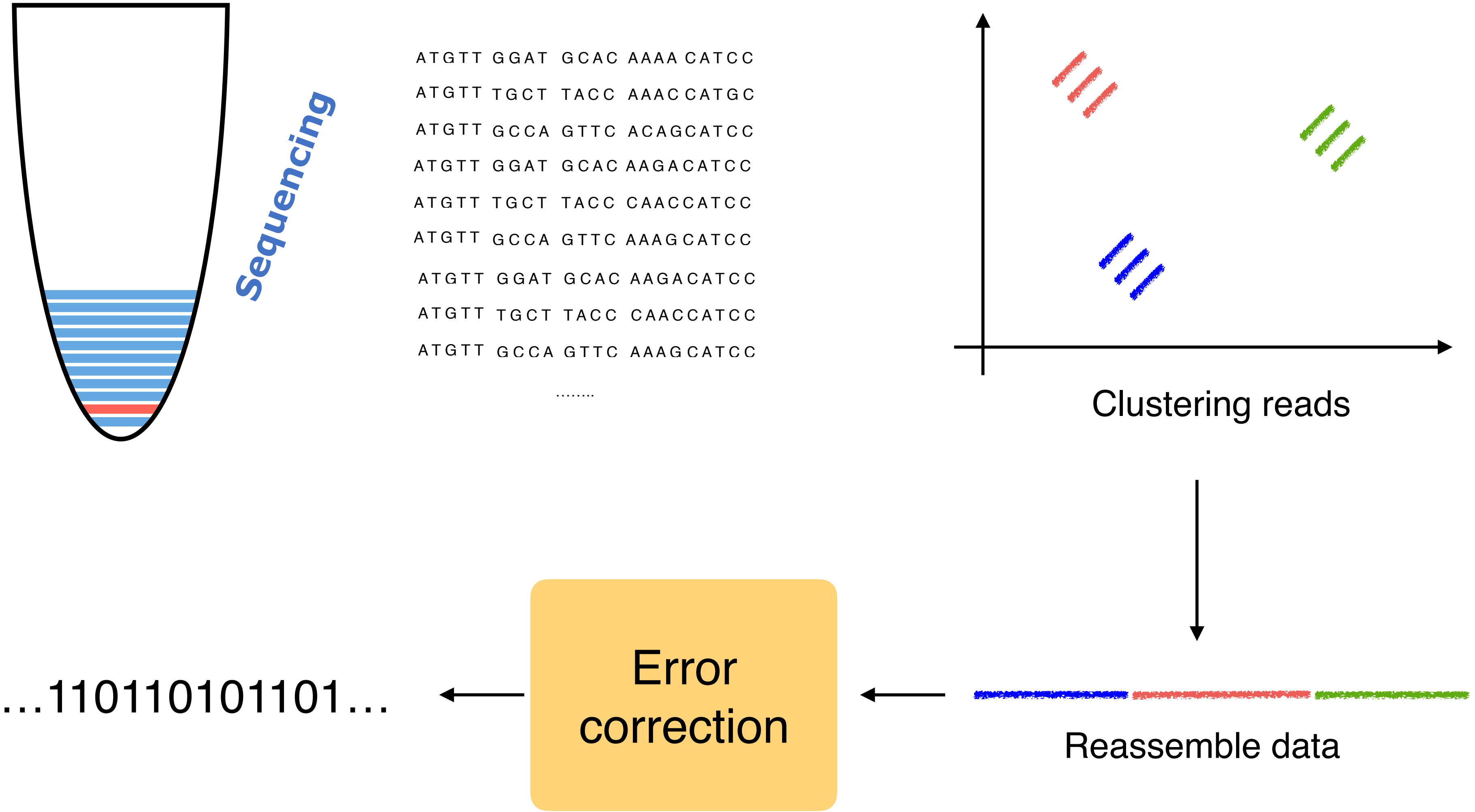
AGCTATCAG

Decoding

11011101

© University of Washington and Microsoft Research. All rights reserved.

Decoding back to digital data



Sequencing

ATGTT GGAT GCAC AAAA CATCC
ATGTT TGCT TACC AAAC CATGC
ATGTT GCCA GTTC ACAGCATCC
ATGTT GGAT GCAC AAGACATCC
ATGTT TGCT TACC CAACCATCC
ATGTT GCCA GTTC AAAG CATCC
ATGTT GGAT GCAC AAGACATCC
ATGTT TGCT TACC CAACCATCC
ATGTT GCCA GTTC AAAG CATCC
....

Clustering reads

Reassemble data

Error correction

…110110101101…

© University of Washington and Microsoft Research. All rights reserved.

MISL UW MSR

Results

200MB as of July’16. last year: 1MB

1.5 B nucleotides

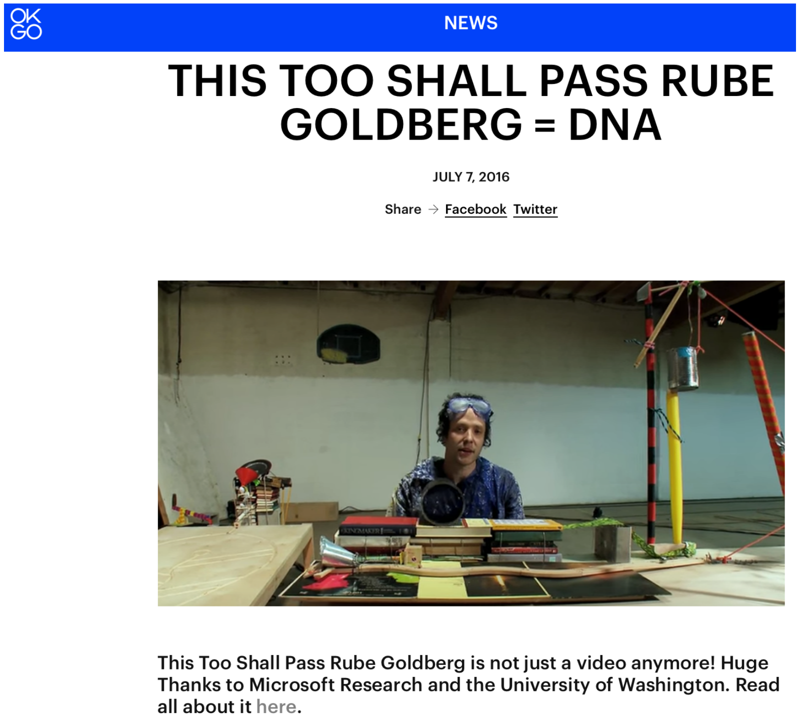
10M DNA strands

UNIVERSAL DECLARATION OF HUMAN RIGHTS

United Nations

CROP TRUST

SECURING OUR FOOD, FOREVER



OK GO NEWS

THIS TOO SHALL PASS RUBE GOLDBERG = DNA

JULY 7, 2016

Share → Facebook Twitter

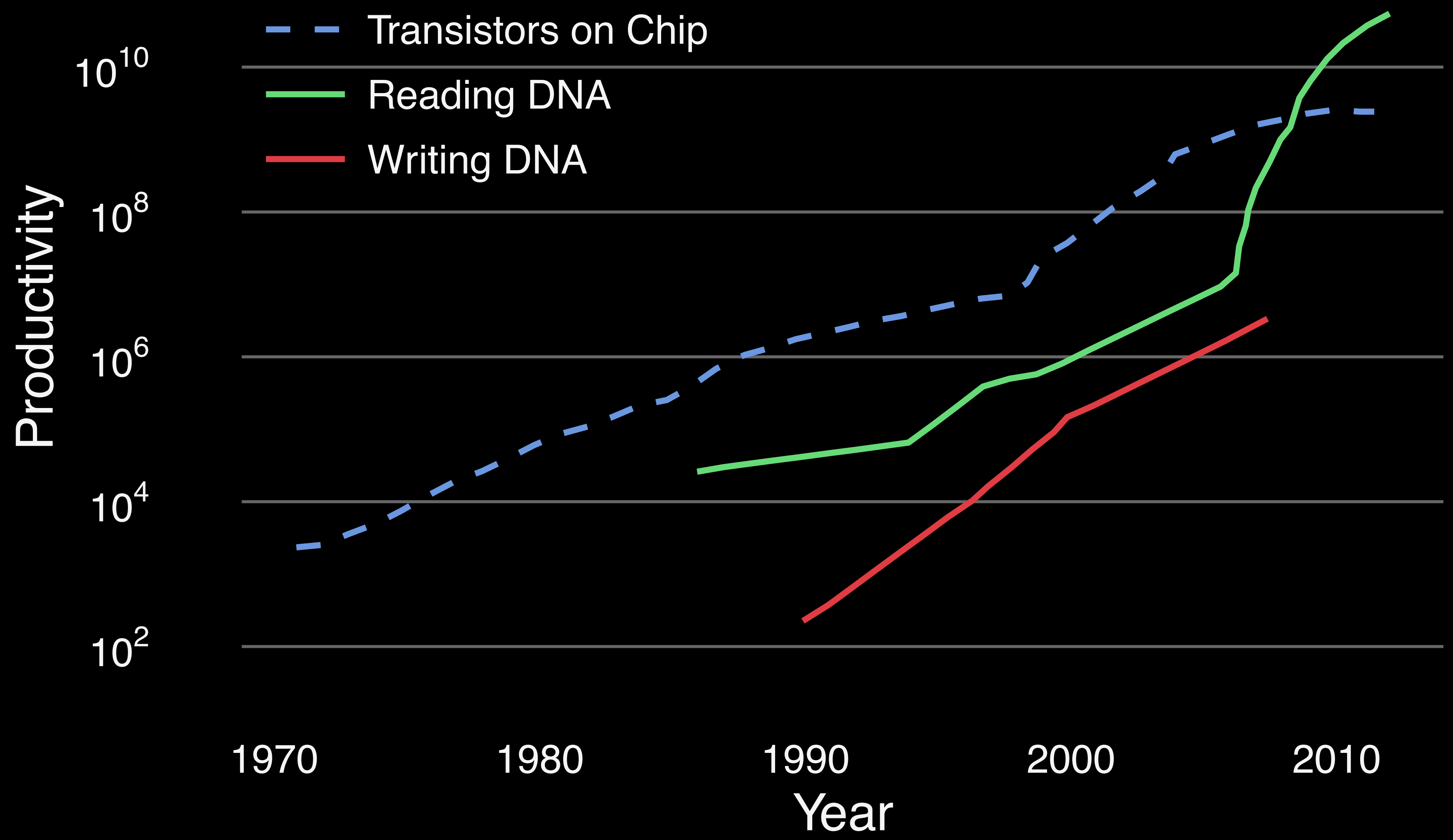
This Too Shall Pass Rube Goldberg is not just a video anymore! Huge Thanks to Microsoft Research and the University of Washington. Read all about it here.

© University of Washington and Microsoft Research. All rights reserved.

10MBs/week → 100GBs/second

DNA manipulation productivity is growing

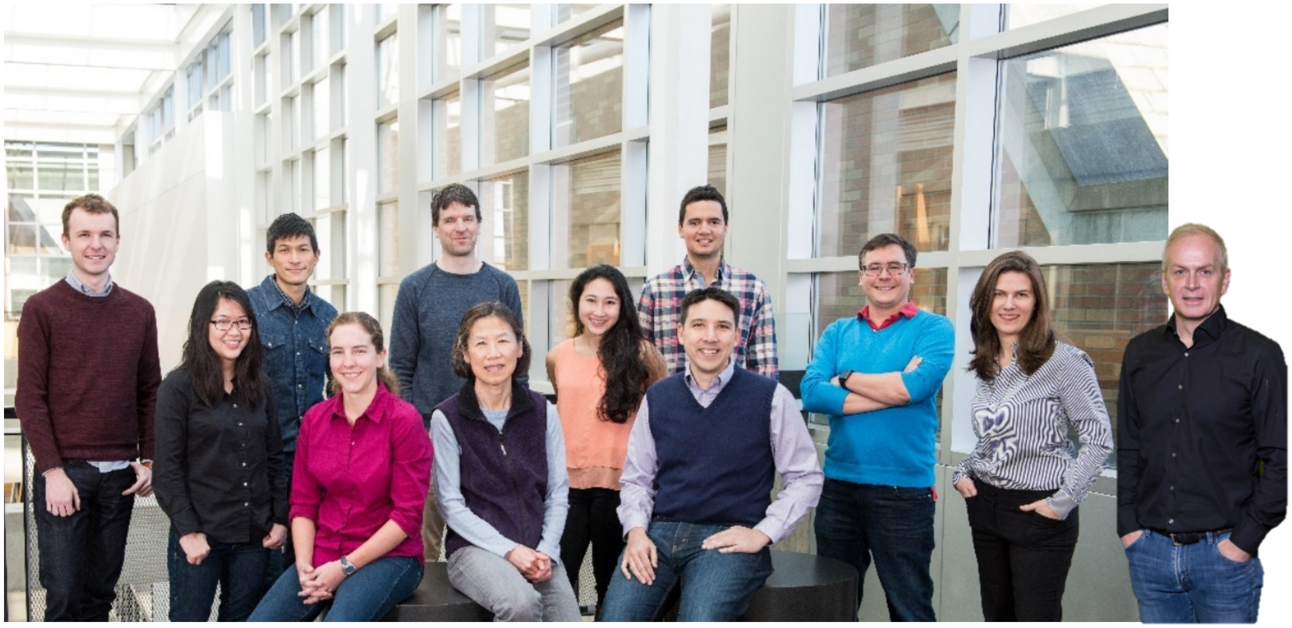
And cost is decreasing…



Source: Robert Carlson

MISL

Molecular Information Systems Lab



Computer architects, coding theorists, molecular biologists, fluidics, algorithms, …

Microsoft Research

© University of Washington and Microsoft Research. All rights reserved.