STORAGE DEVELOPER CONFERENCE

SDC22 | Fremont, CA
September 12-15, 2022

BY Developers FOR Developers

A SNIA Event

DNA Data Storage Alliance: Building a DNA Data Storage Ecosystem

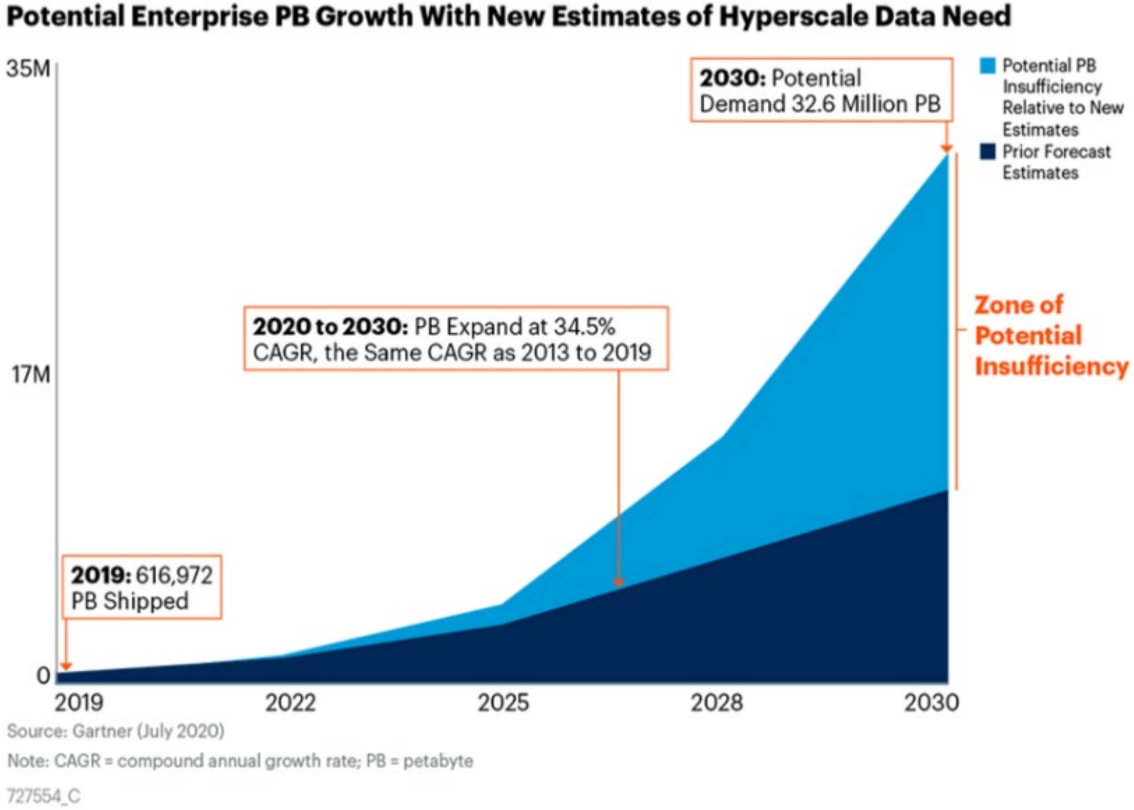
Dave Landsman
Senior Director Industry Standards
Distinguished Engineer
Western Digital

DNA DATA STORAGE ALLIANCE

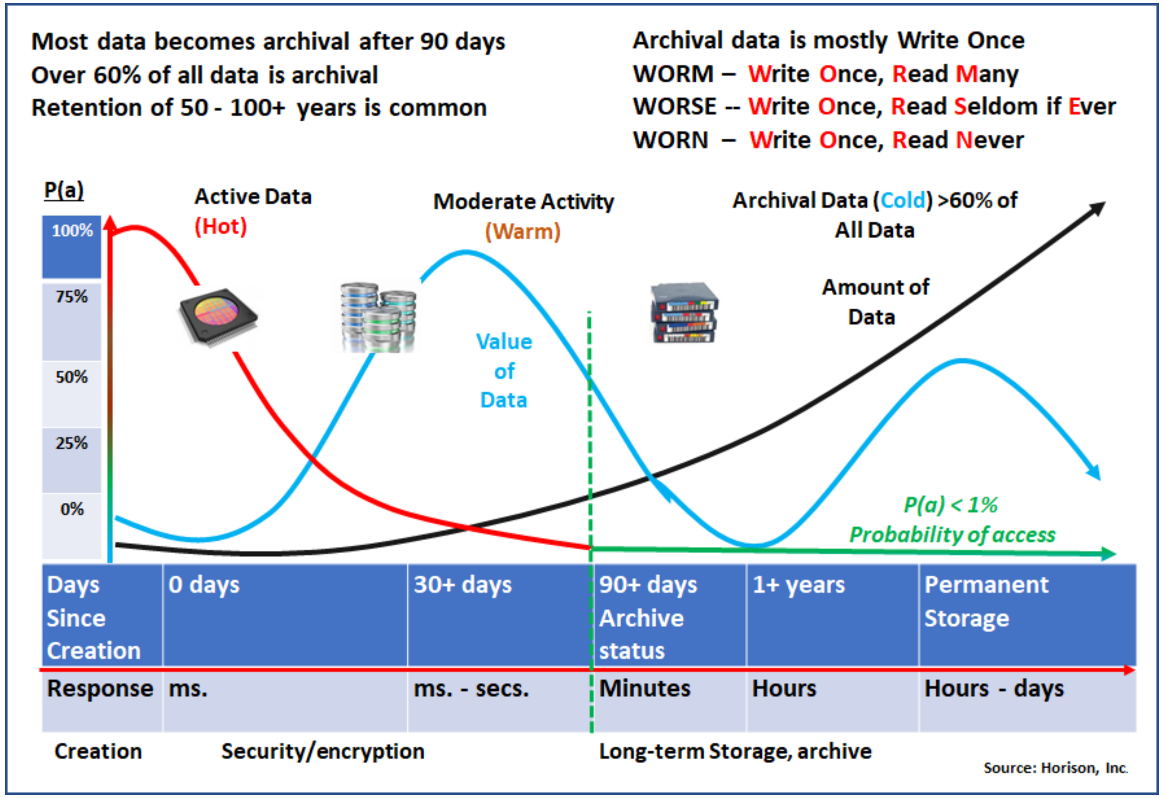
A SNIA Technology Affiliate

What is the problem?

There is too much data to save



And value of saved data is growing

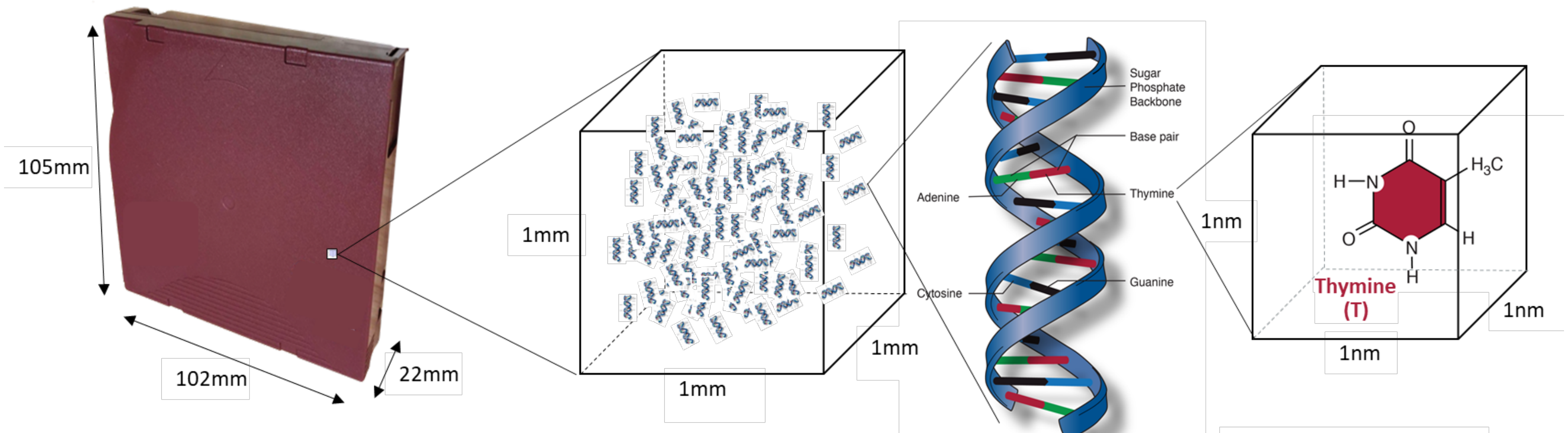


DNA DATA STORAGE ALLIANCE
A SNIA Technology Affiliate

STORAGE DEVELOPER CONFERENCE
SDC 22

Why DNA?

DNA bits are very small



Volume of an LTO cassette filled with DNA bits would hold ~2,000,000 TB; this would fill ~115,000 LTO-9 tapes

1mm^3 holds ~9TB of encoded DNA bits (1/2 LTO-9 tape capacity)

Single DNA Base fits in ~1nm^3

Source: An Introduction to DNA Data Storage - DNA Data Storage Alliance

… and they last a long time, don't need much care and feeding, and don't need migration (= TCO)

… and they'll benefit from, and accelerate, established investment in DNA technology

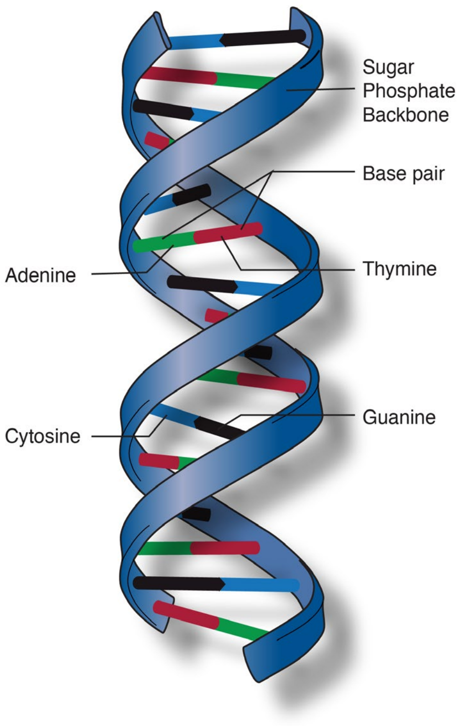
But there are questions

- Is there really a need for a medium as dense as DNA?
- Can we scale the underlying technologies?
- How do we create an interoperable DNA data storage ecosystem?

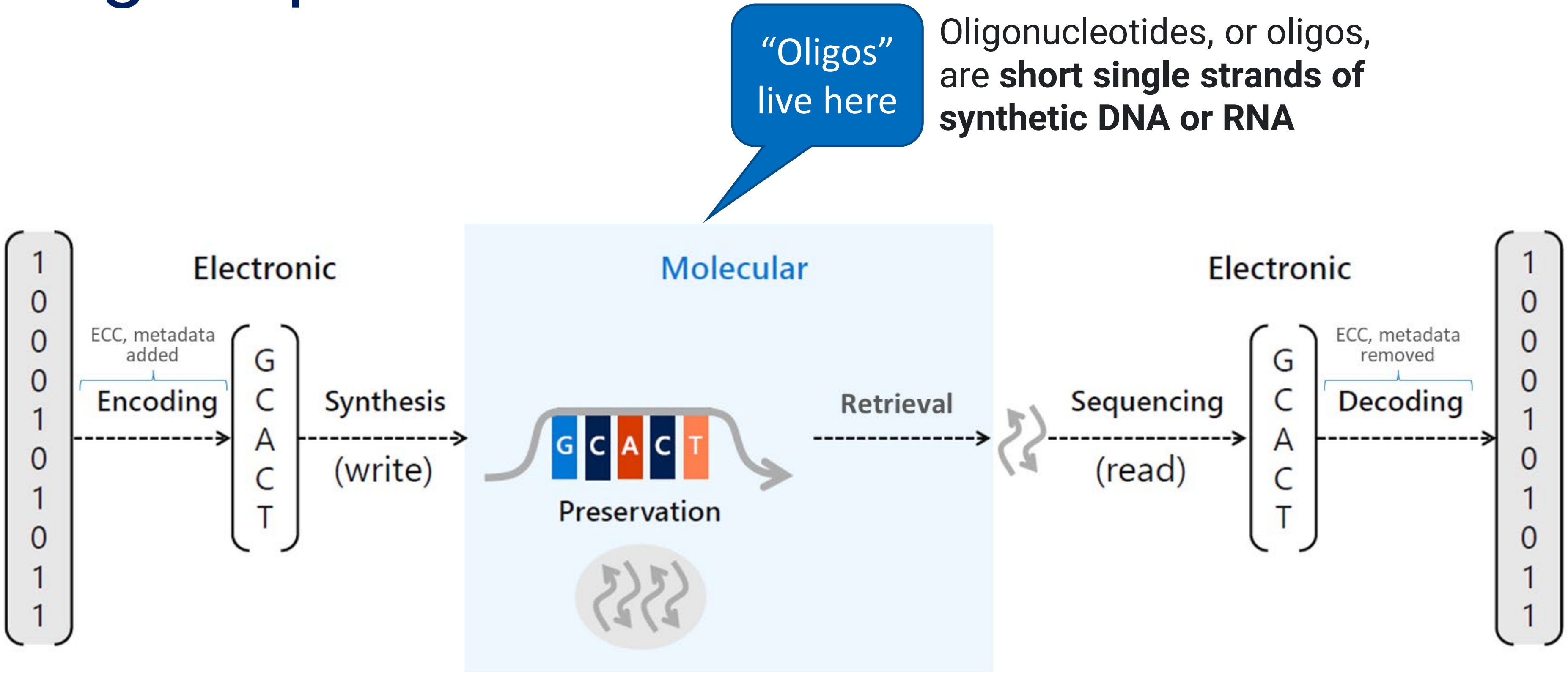
Don't miss Aaron Ogus keynote tomorrow, 11:20am, The Looming Need for Molecular Storage

DNA Data Storage Track			
8:30am-9:20am	Dave Landsman	Distinguished Engineer Western Digital	Creating a DNA Data Storage Ecosystem
9:30am-10:20am	Joel Christner Alessia Marelli Mark Wilcox	Distinguished Engineer - Dell Technologies CTO - DNA Algo CEO - 21e8	Rosetta Stone: Standard to enable discovery and decode of info in a DNA data archive
10:35am-11:25am	Alessia Marelli Rino Micheloni	CTO and COO DNAalgo	DNAssim: A Full System Simulator for DNA Storage
11:35am-12:25pm	João Reis Marília Menossi	Researchers Lenovo / Instituto de Pesquisas Tecnologicas	End-to-End DNA data storage system study
1:30pm-2:30pm	João Gervasio Adriano Galindo Leal	Researchers Lenovo / Instituto de Pesquisas Tecnologicas	DNA Coding Overview
2:30pm-3:20pm	Luca Piantanida	Research Scholar Boise State University	Nucleic Acid Memory
3:35pm-4:25pm	**Informal Q&A w/ Presenters**		

DNA Data Storage Pipeline

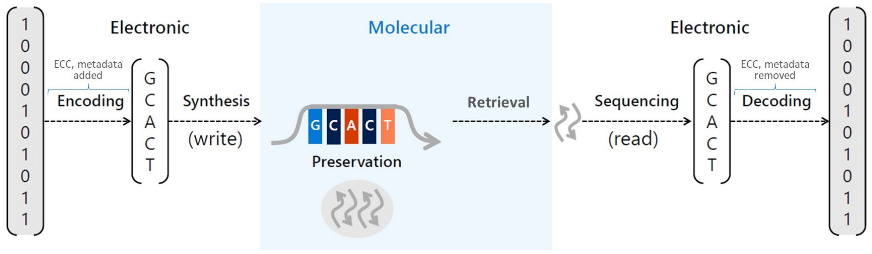


A⬄T C⬄G



Oligonucleotides, or oligos, are **short single strands of synthetic DNA or RNA**

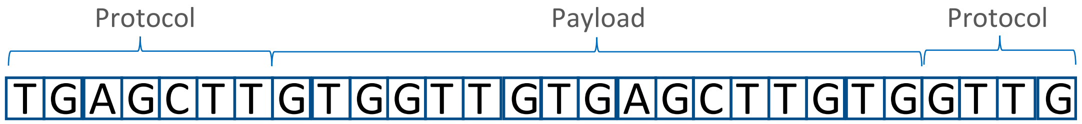
The DNA Data Storage "Channel"



Some (hopefully) illustrative analogies between Electrical and DNA

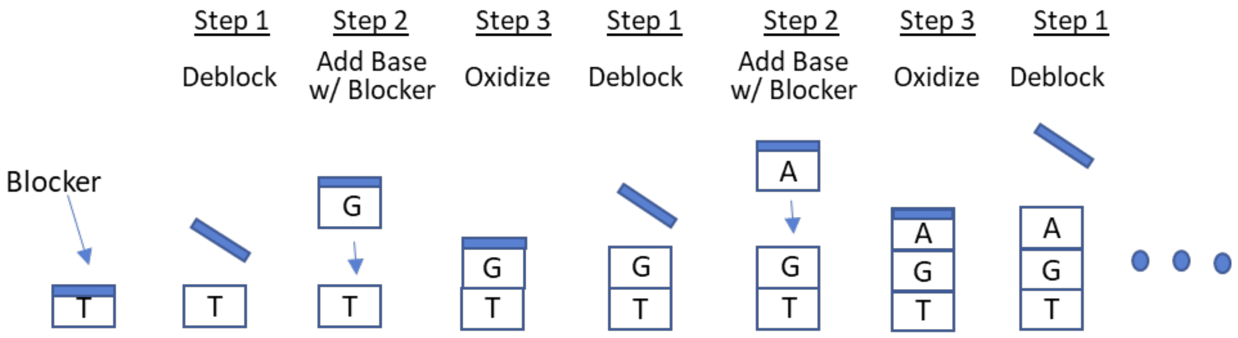
Electrical Channel	DNA "Channel"
• 1's and 0's converted to analog wave forms at transmitter, back to 1's and 0's at receiver	• 1's and 0's converted to DNA bases; that is, "wave forms" in electrical case become DNA bases in DNA case
• ECC bits added to digital bit stream before transmission, checked/stripped at receiver to check/correct data errors	• ECC bits added to digital bit stream (by codec) before synthesis (transmitter) and checked/stripped after sequencing (receiver) • DNA errors: insertions/deletions (indels), substitutions (SNVs), …
• Scrambling patterns (reordering 1's, 0's) added at transmitter to avoid analog effects which can cause errors on wire	• Some patterns of bases problematic to synthesize/sequence, so we may alter bases (ACCG…) after ECC/metadata encoded

- As with electrical, the "line protocol" for DNA data storage is critical to overall channel efficiency
- There are also "logical" protocol layers above line protocol, e.g., file tagging, packetization

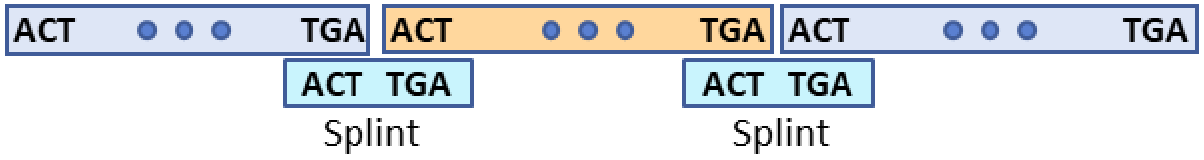


Synthesis Techniques - Snapshot

- Base-by-base
  - Two underlying base-by-base techniques: Phosphoramidite and Enzymatic
  - Both methods use similar cyclic process
  - Limit of 200-300 bases per oligo today

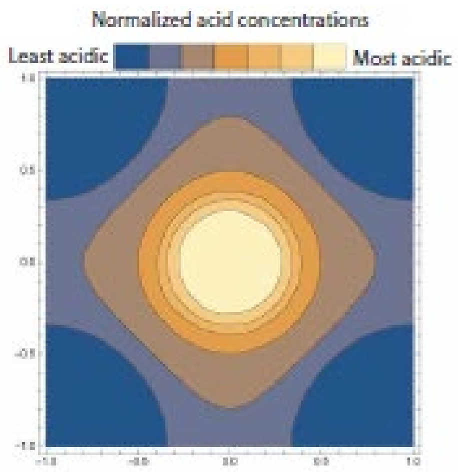


- Ligation
  - Can enable strands of many hundreds of bases or longer
  - More bits in payload means, in general, less protocol overhead



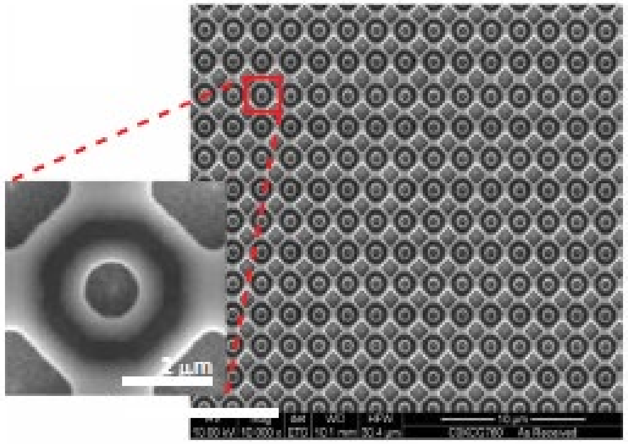
Synthesis - Scaling

Scaling DNA data storage w/ nanoscale electrode wells
Sci. Adv. 7, eabi6714 (2021) 24 November 2021 Nguyen et al



- Study results
  - Electrochemical DNA synthesis w/ 650nm wide electrodes in 200nm wells
  - Acid containment at 2um pitch with these feature sizes
  - 100-base long DNA data storage with these feature sizes

- Some throughput implications from the study
  - The chip in this study reached synthesis density of 25M synthesis sites/cm^2; nearly 3 orders of magnitude greater than previous work
  - An array with this synthesis density could achieve write speed of >2.8 KB/s/cm^2* which could be practical minimum for some archival uses
  - But at this density, it would require a 360cm^2 chip (not manufacturable), or many chips, to achieve, say, 1MB/s, and it would use lots of reagents



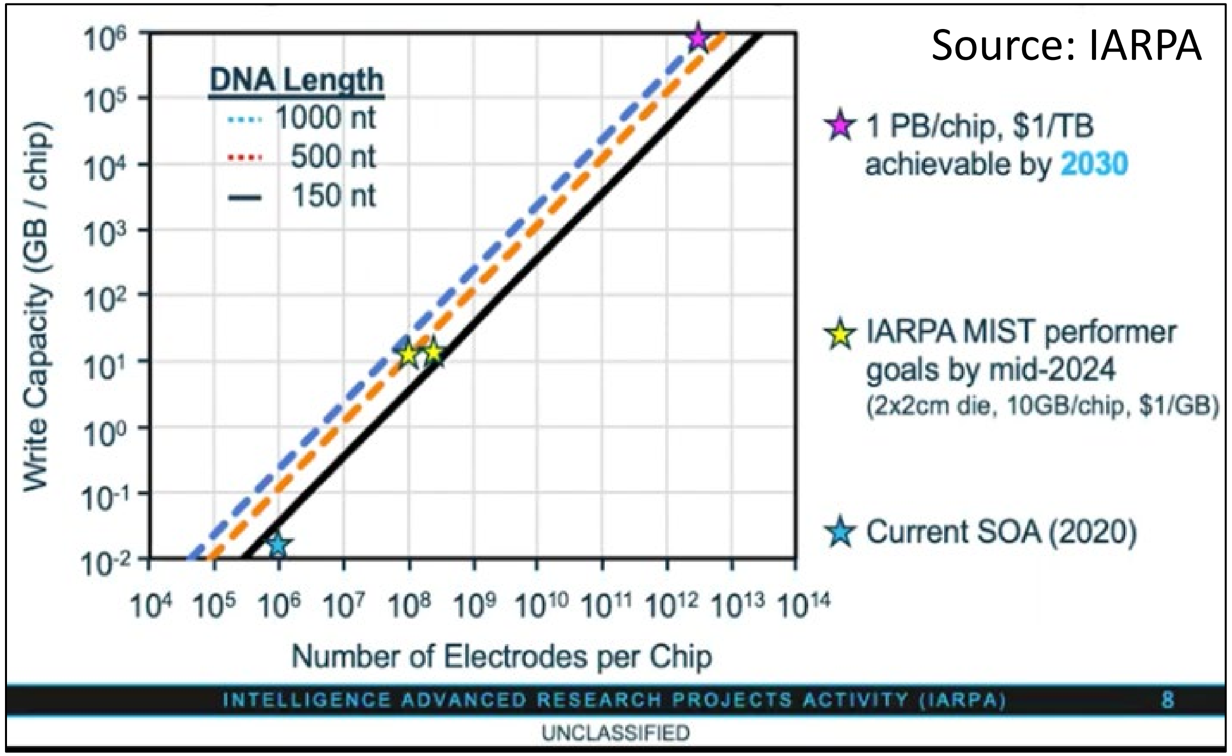
* assuming each unique oligo encodes 10 bytes of data and is written over 24 hrs

DNA DATA STORAGE ALLIANCE
A SNIA Technology Affiliate

STORAGE DEVELOPER CONFERENCE
SDC 22

Synthesis - Scaling

- There is continued progress however
  - Twist Bioscience just announced chip w/ synthesis density of 100M synthesis sites/cm^2 and ability to write 1GB per run
  - Moving toward IARPA Molecular Information Storage (MIST) synthesis goals on capacity, scale, and cost

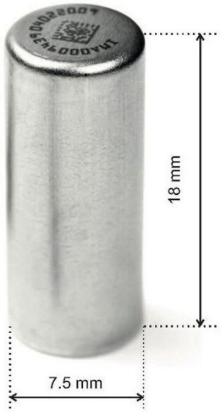


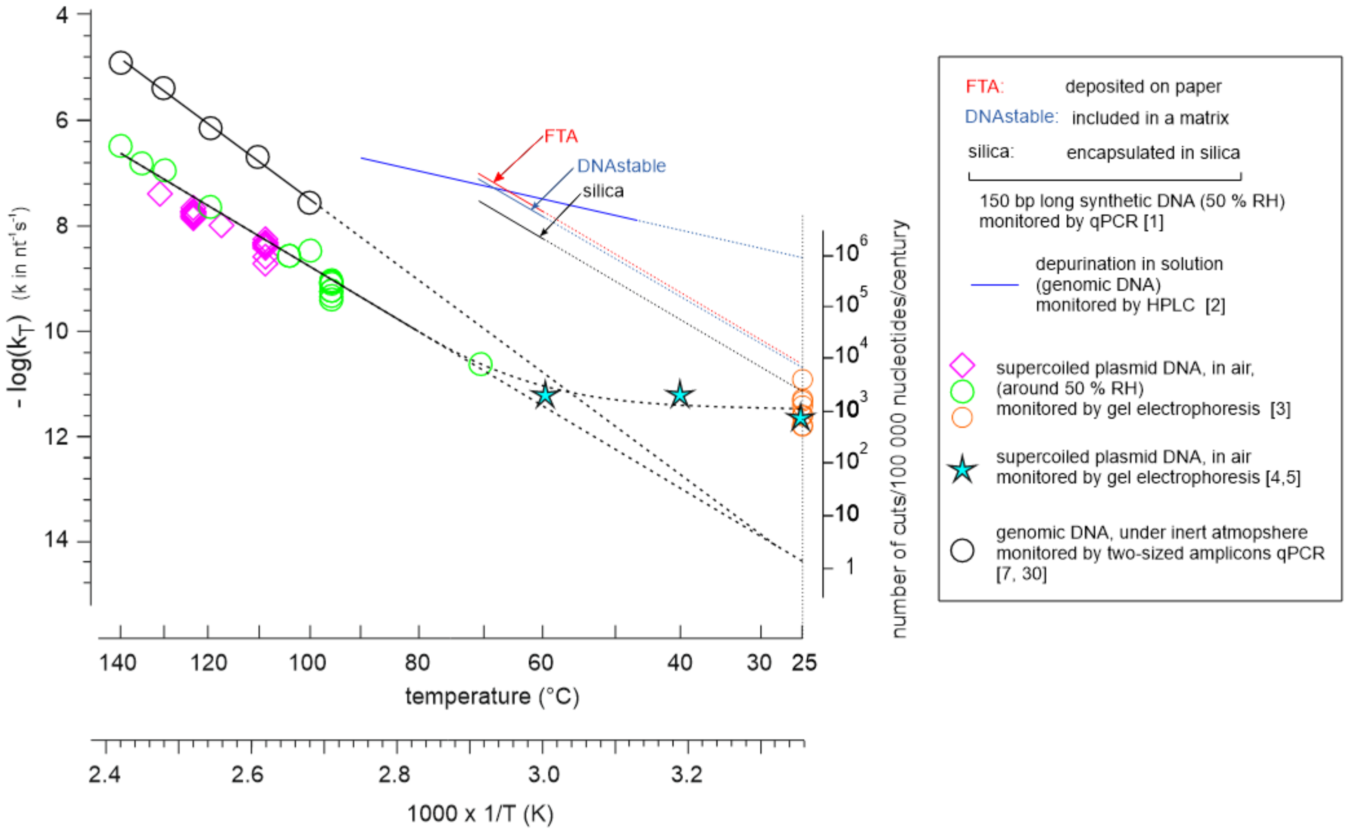
Much more scaling needed, but foundations are established

Storage of DNA based bits

- Many preservation methods being explored
- As well as verification methods to enable comparison of data retention capabilities

Principle	Procedure	
Chemical encapsulation	Moha bones	[8]
	Fox teeth	[9]
	Encapsulation in salts	[12, 16]
Physical encapsulation	Silica nanoparticles	[1]
	Stainless steel capsules	[3]
	Magnetic silica nanoparticles	[13]
Inclusion in a matrix	DNAstable	[1, 21]
	Gentegra DNA	[1, 22]
	Pullulan	[14]
	Silk	[15]
	300K matrix inclusion	[25]
Absorption on paper	FTA paper	[1, 23, 24]
	Chitosan treated paper	[17]
Dehydration on solid supports	Capillaries	[20]
	Glass	[26, 27]
	Tube walls	[28]
Dissolution in liquid salts	Imidazolium cations	[18]
	Imidazolium cations	[19]
Living organism	Bacteria	[29]

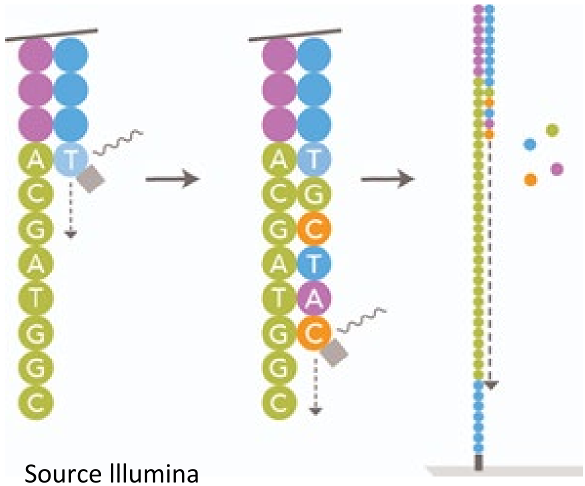




Source: Coudy et al - Long-term conservation of DNA at ambient temperature. Implications for DNA data storage. PLoSOne. 2021; 16(11): e0259868)

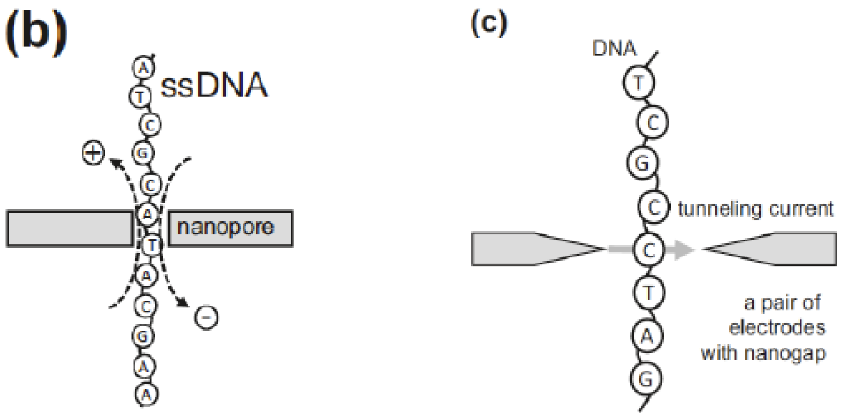
Sequencing - Snapshot

- Sequencing by Synthesis (SBS)
  - Start with a ssDNA (template)
  - Build a complementary strand (synthesis)
  - Each binding event is detectable



Source Illumina

- Nanopore
  - Guide DNA strand through very small channel: nanopore
  - As strand traverses nanopore, ionic current or tunneling current are detected and the bases are directly read

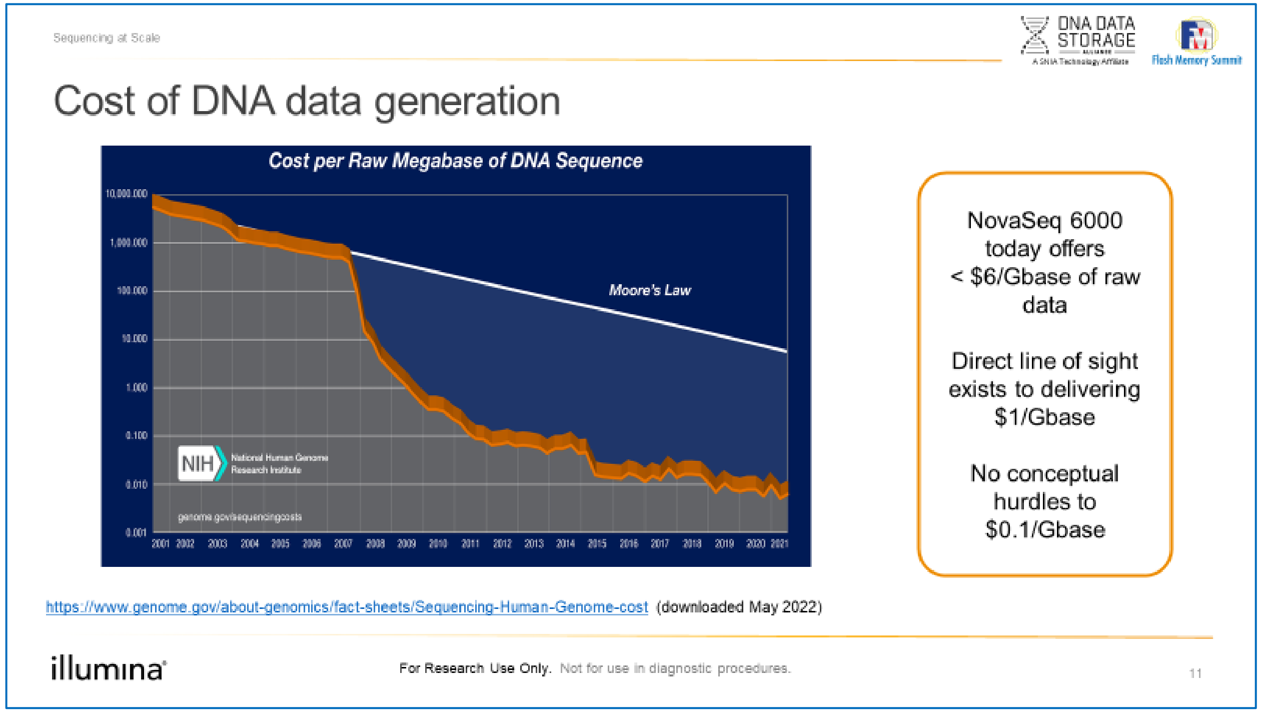


Goto, Y., Akahori, R., Yanagi, I. and Takeda, K.I., 2020. Solidstate nanopores towards single-molecule DNA sequencing. Journal of human genetics, 65(1), pp.69-77.

- In general, today, per base, SBS more accurate/slower and nanopore less accurate/faster
- Throughput battles in commercial systems

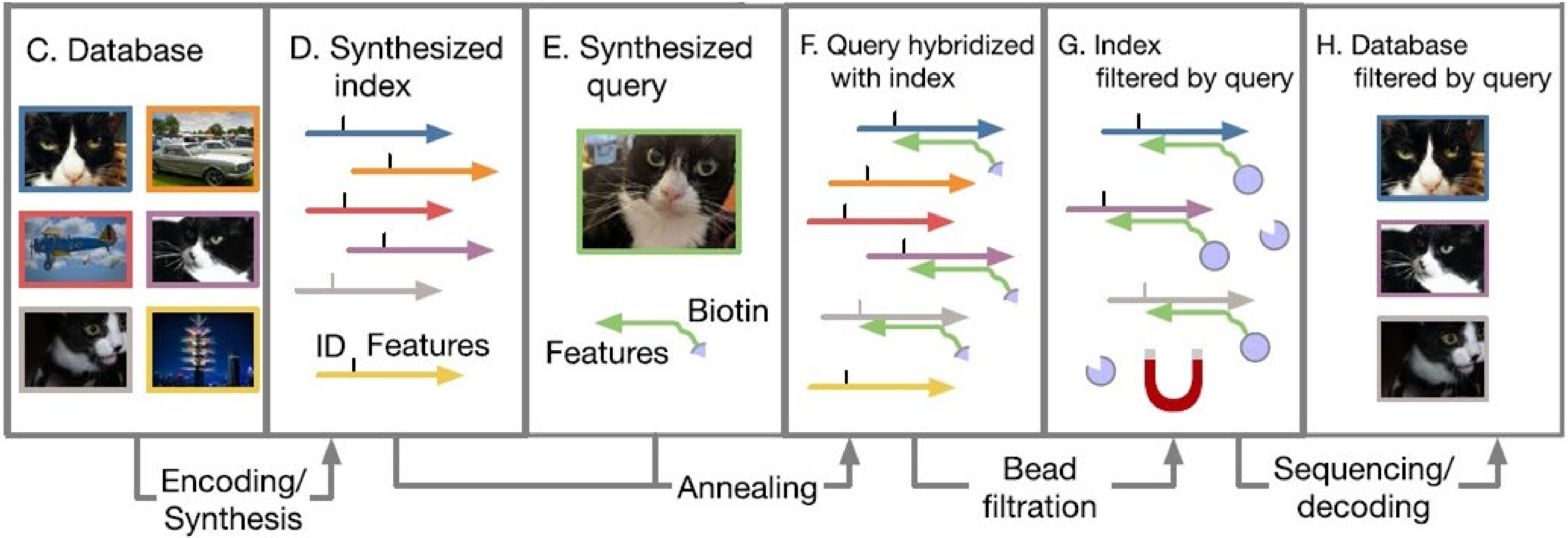
Sequencing - Scaling

- Analysis
  - **Throughput:** We are now at ~10s of GBytes/day (per machine); probably need to get to 100s of TBytes/day
  - **Cost:** Using the $/base numbers (assume 1-bit/base)
    - at $48000/TB ($6/Gbit) now
    - with **direct line of site** to $8000/TB ($1/Gbit)
    - and **no conceptual hurdles** to $800/TBytes ($0.1/Gbit)
  - This is based on list pricing and requirements for medical/genomic markets
  - DNA storage can tolerate higher error rates



- Conclusion
  - We have an apparent 3 orders of magnitude to go on both cost/price & throughput for sequencing
  - That said, there are many ways to manipulate all phases (codecs, synthesis, sequencing) to balance error tolerance and performance in the whole pipeline

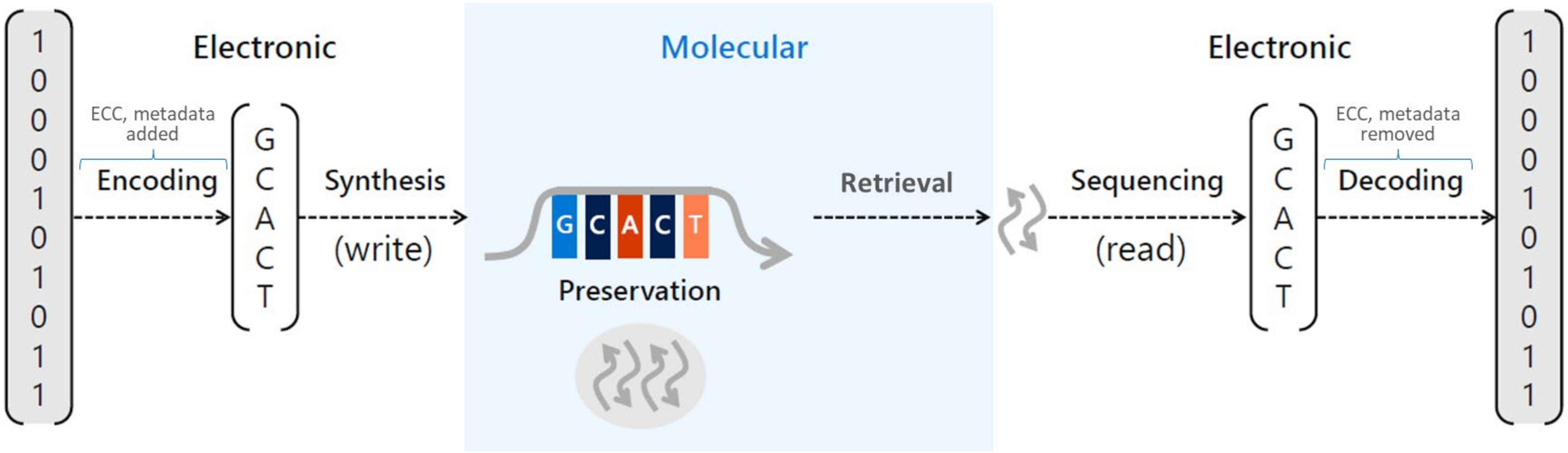
Selective retrieval of digital data encoded in DNA molecules



Bee C, Chen YJ, Queen M, Ward D, Liu X, Organick L, Seelig G, Strauss K, Ceze L. Molecular-level similarity search brings computing to DNA data storage. Nat Commun. 2021 Aug 6;12(1):4764. doi: 10.1038/s41467-021-24991-z. PMID: 34362913; PMCID: PMC8346626.

Also see [Scalable and Dynamic File Operations for DNA-based Data Storage]() (James Tuck, North Carolina State, DNAli Data Technologies) from SDC 2021.

In conclusion:
DNA data storage is now resting on a solid foundation



If we build it, who wants to use it and why?

ADAS

Digital Art

Media/Entertainment

Hyperscale

Preservation

Genomics/Omics

Smart Video

Governmental

Data Retention - Key to discovery, monetization, …

- Healthcare, astronomy, climate science, sports, smart cities and vehicles, governments/municipalities, etc. seeking to save ever larger data sets
- We cannot know today what data will become relevant to new discoveries or required information tomorrow but storing too much data imposes undue costs
- If we can store more data for a lower total cost, the tradeoff between saving or discarding data can be shifted in favor of saving vs. discarding

Source: Karl G. Jansky Very Large Array - NRAO/AUI/NSF

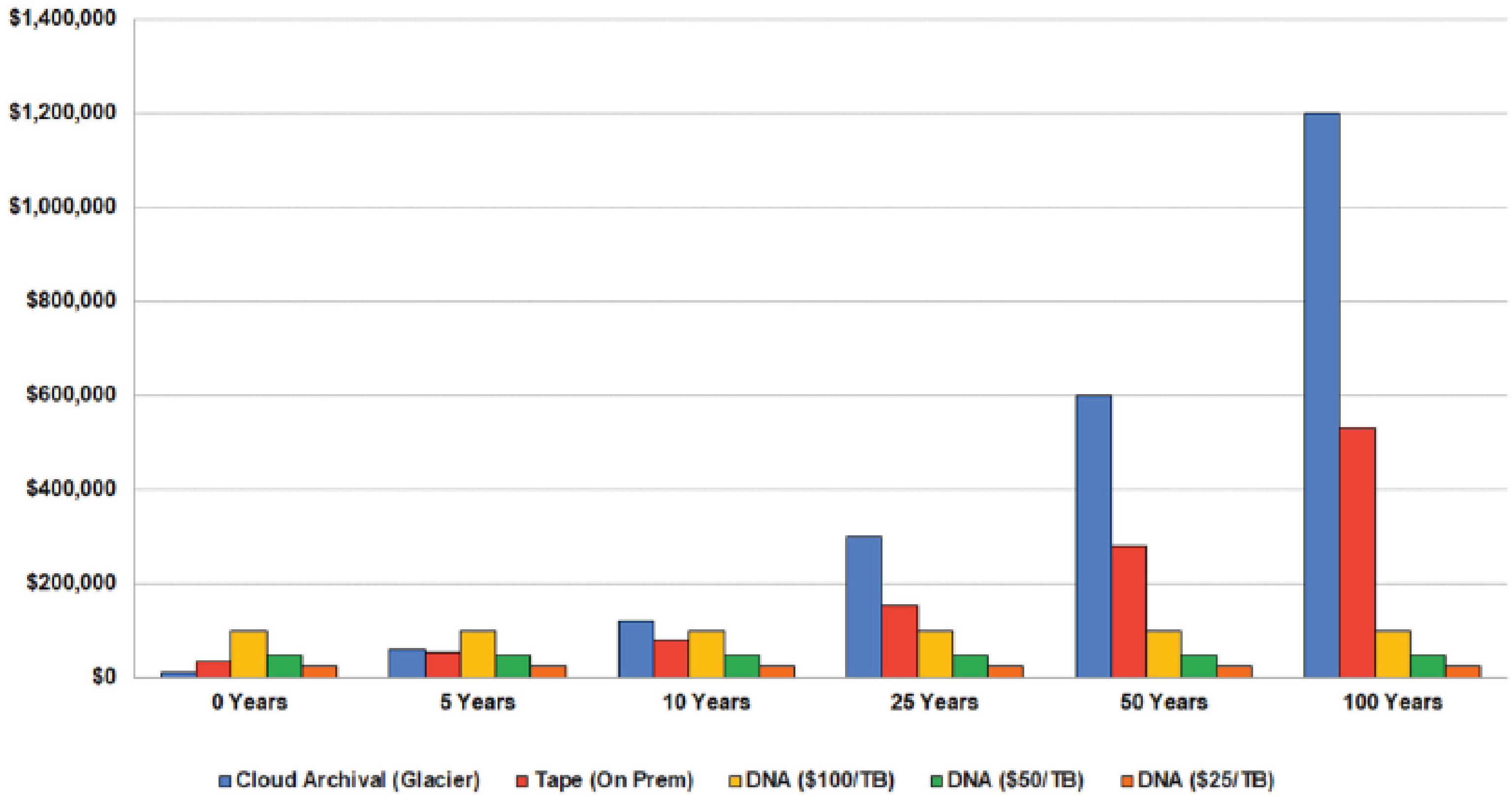
And data retention puts emphasis on TCO

- Because DNA-based storage promises no data migration and nominal-to-no fixity checks, TCO looks better vs. traditional storage as retention times lengthen
- DNA-based storage minimizes energy consumption and improves sustainability

This trend example holds even for <u>one</u> copy of the database; there will be multiple copies

Estimated Total Cost of Writing and Storing - Legacy vs. DNA

- Tape price calculated using Fujifilm TCO calculator
- Cloud prices are taken from Amazon AWS public pricing (2/1/2021).
- DNA storage prices based on selected cost scenarios for comparison only



DNA DATA STORAGE ALLIANCE
A SNIA Technology Affiliate

STORAGE DEVELOPER CONFERENCE
SDC 22

So how do we build the ecosystem?

DNA DATA STORAGE ALLIANCE

TWIST BIOSCIENCE
Microsoft
illumina
Western Digital
imagene
MOLECULAR ASSEMBLIES
FUJIFILM
CATALOG
DELL Technologies
SEAGATE
Los Alamos NATIONAL LABORATORY
Quantum
OLIGO ARCHIVE
DIGITAL BEDROCK
Keeping Your Digital Assets Safe and Evergreen
IBM
ISM Information Storage & Memories
Technion
OIOS SWISS MADE
UNIVERSITÉ CÔTE D'AZUR
cnrs
i3S
eureKARE
cacheDNA
STORING NUCLEIC ACIDS FOREVER
DNAalgo
MISL
University of WA
cinémathèque suisse
dpc DigitalPreservationCoalition
MilliporeSigma
BIO ECHO Life Sciences
Lab4ISC
Lab for Intelligent Storage and Computing
Oklahoma State University
ICMS
INSTITUTE FOR COMPLEX MOLECULAR SYSTEMS
TU/e
Philipps Universität Marburg
BIO MEMORY
Sustainable Storage Solutions
PFU
DNALI
A MOUNTAIN OF DATA IN A DROP
The University of Arizona
ANBM CENTER FOR APPLIED NANOBIOSCIENCE & MEDICINE
21e8
I(CO)$_2$S
Newcastle University
Reichman University

DNA DATA STORAGE ALLIANCE
A SNIA Technology Affiliate

STORAGE DEVELOPER CONFERENCE
SDC 22

DNA Data Storage Alliance - At a Glance

History

- Formed on October 12th, 2020 by Ilumina, Microsoft, Twist and Western Digital
- Climbed to 60+ members by 2Q-2022
- Joined SNIA as a Technology Affiliate group as of Jun-2022

Mission

- Create and promote an interoperable storage ecosystem based on DNA as a data storage medium

Scope

- Educate the market to create awareness and adoption of DNA data storage
- Develop a DNA data storage industry technology roadmap to drive R&D and funding
- Develop standards and/or specifications as needed by ecosystem

DNA Data Storage Alliance

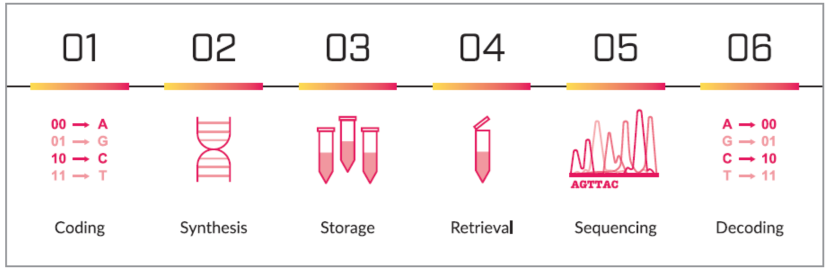
2022 Activities

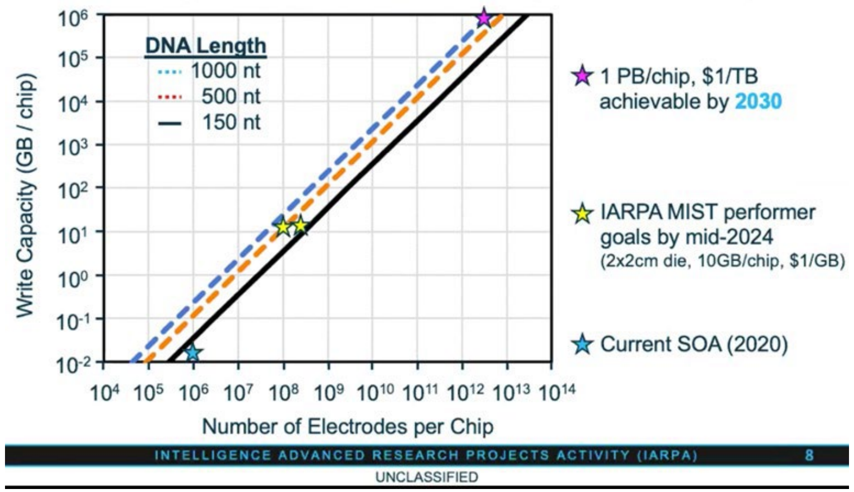
- Industry technology roadmap
- Start workgroups for potential standardization
- White Paper #2 - Market segments/use cases
- Newsletter
- Events

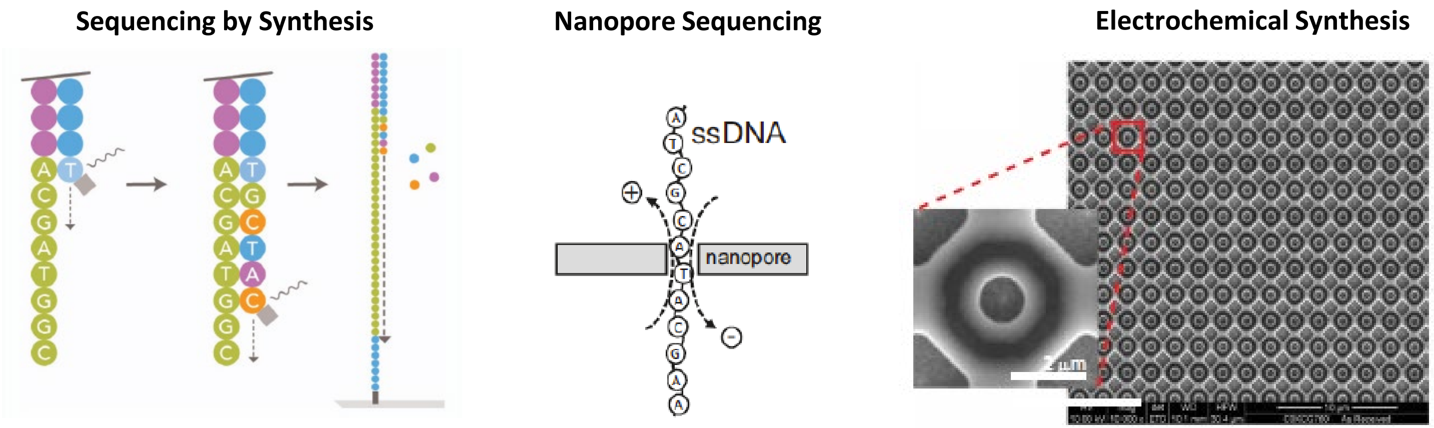
Industry Technology Roadmap

Roadmap for how DNA data storage can scale to commercial viability

- Key technologies and challenges in the pipeline
- Success metrics: capacity, transfer rates, cost, …







Guide for academic/industry research and investment

DNA Data Storage Alliance: TWG sub-groups

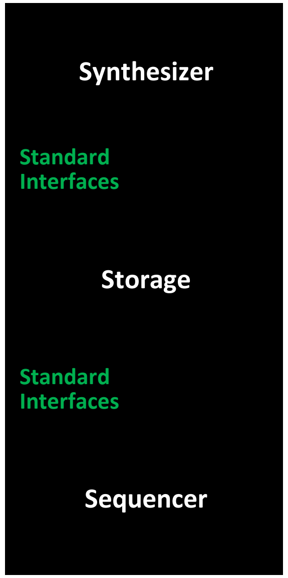
DNA Archive Rosetta Stone (DARS)

Create universal identifier describing how to decode/read rest of archive

Interoperable Interfaces

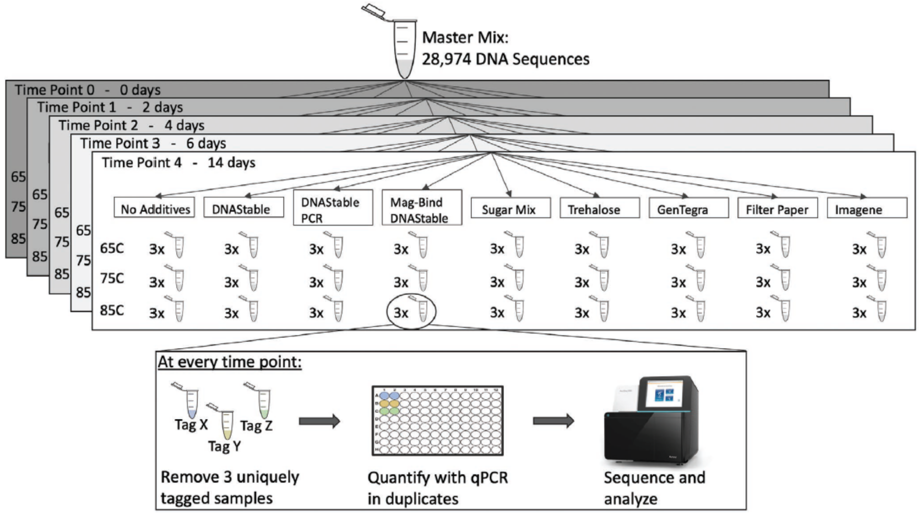
Ensure physical compatibility of synthesis, storage, and sequencing modules to ensure "plug-and-play" integration and avoid vendor lock in:

- plug and play swaps of instruments
- recovery of molecules for read, irrespective of supplier being existent at read time
- issues of fluidics in data centers



DNA Digital Data Retention

Define standard metrics and verification methods to enable comparison of data retention in DNA-based data storage solutions (retention properties, complexity of retrieval, etc.)



An Empirical Comparison of Preservation Methods for Synthetic DNA Data Storage
L. Organick, B. H. Nguyen, R. McAmis, W. D. Chen, A. X. Kohll, S. D. Ang, R. N. Grass, L. Ceze, K. Strauss, *Small Methods* 2021, 5, 2001094.

A disclaimer before closing

- DNA is not like other media; it needs some explaining
  - "Does this mean I'm going to be storing my music collection in my dog?"
- It is important to clarify

Storing digital data in synthetic DNA molecules in no way requires the use or creation of any cells, organisms, or life!

We are 'simply' using chemical mechanisms to store digital bits in DNA molecules instead of using electromagnetic or optical mechanisms to store bits in silicon, magnetic, or other materials.

"I ain't nothin' but a hound dog…."

THANK YOU

Enjoy rest of the track and come join the efforts to create a DNA data storage ecosystem!!

DNA DATA STORAGE ALLIANCE
A SNIA Technology Affiliate

Subscribe to our newsletter on our website
https://dnastoragealliance.org

And make sure to check out:
Preserving our Digital Legacy: An Introduction to DNA Data Storage

Follow us on Twitter
@DnaDataStorage

Follow us on LinkedIn
@dna-data-storage-alliance