SNIA

FMS

the Future of Memory and Storage

DNA Data Storage

DNA DATA STORAGE ALLIANCE

Dave Landsman (dave.landsman@wdc.com)
Distinguished Engineer - Western Digital
Board Co-Chair – DNA Data Storage Alliance
Board Member - SNIA

Context of DNA data storage

- Healthcare, astronomy, climate science, sports, smart cities and vehicles, governments, municipalities, etc. seeking to save ever larger data sets
- Increasingly expensive and impractical to save all the data with existing storage technology
- Opportunity cost of throwing away valuable information getting higher, and this effect being accelerated by the emergence of AI/ML

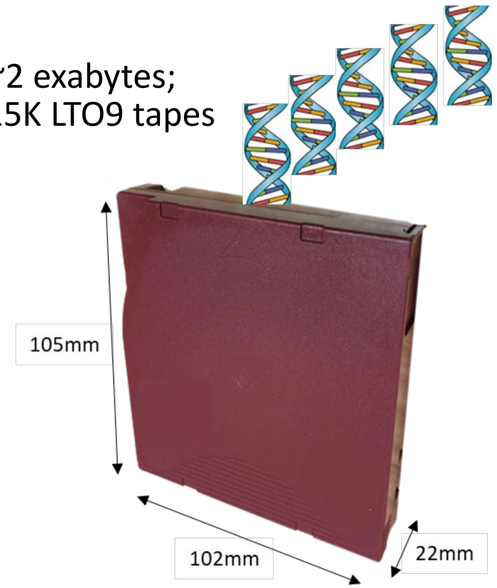
Source: Karl G. Jansky Very Large Array - NRAO/AUI/NSF

The “save/discard data” choice is becoming acute

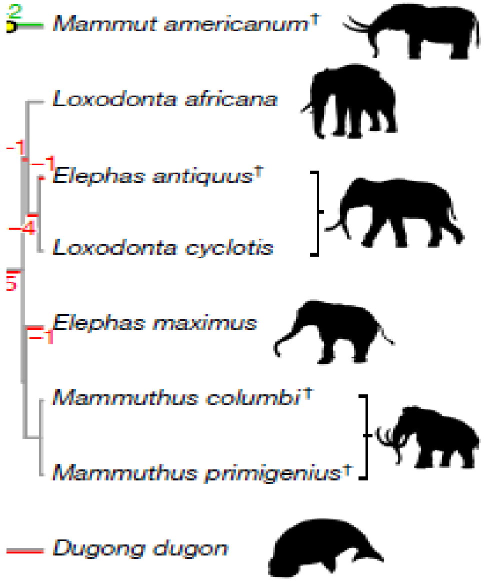
Why consider DNA? – Ultimately, TCO

1. DNA bits are <u>very</u> small, ~1nm^3

~2 exabytes;
~115K LTO9 tapes



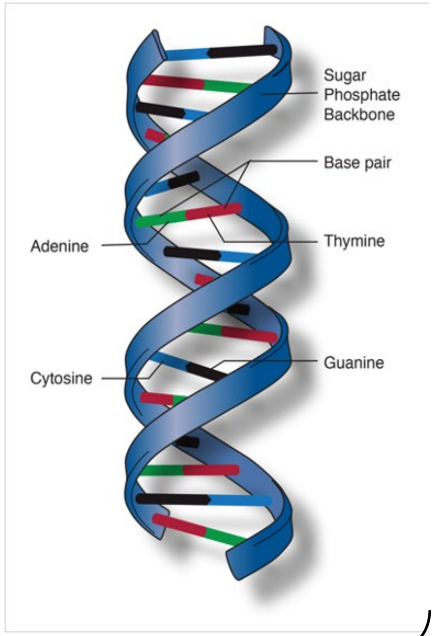
2. DNA bits are <u>extremely</u> durable – record 2M+ years



Kjær, K.H., Winther Pedersen, M., De Sanctis, B. et al. A 2-million-year-old ecosystem in Greenland uncovered by environmental DNA.

Nature 612, 283–291 (2022). https://doi.org/10.1038/s41586-022-05453-y

3. DNA is format immutable; i.e., it can always be read



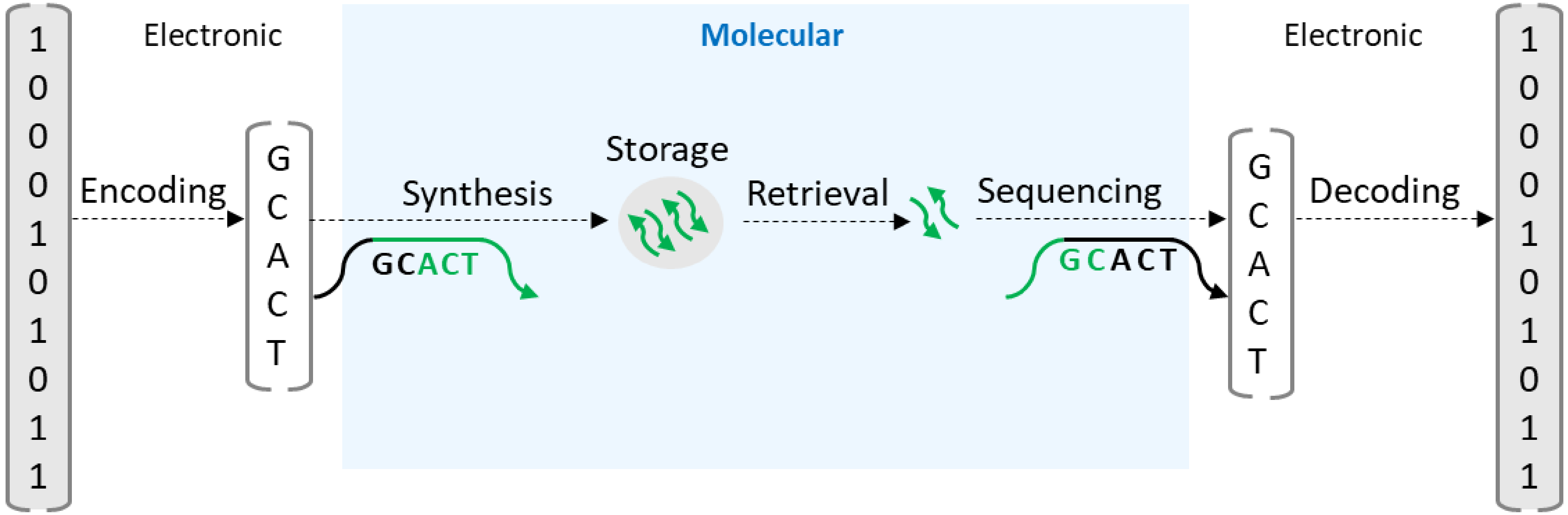
- Incredibly space efficient
- Room temp storage
- Few/No technology migrations

And in addition:
- **Virtually free copies**
- **Green**

DNA DATA STORAGE ALLIANCE
A SNIA Technology Affiliate

SNIA

The DNA Data Storage Pipeline



Example Coding

A = 00
C = 01
T = 10
G = 11

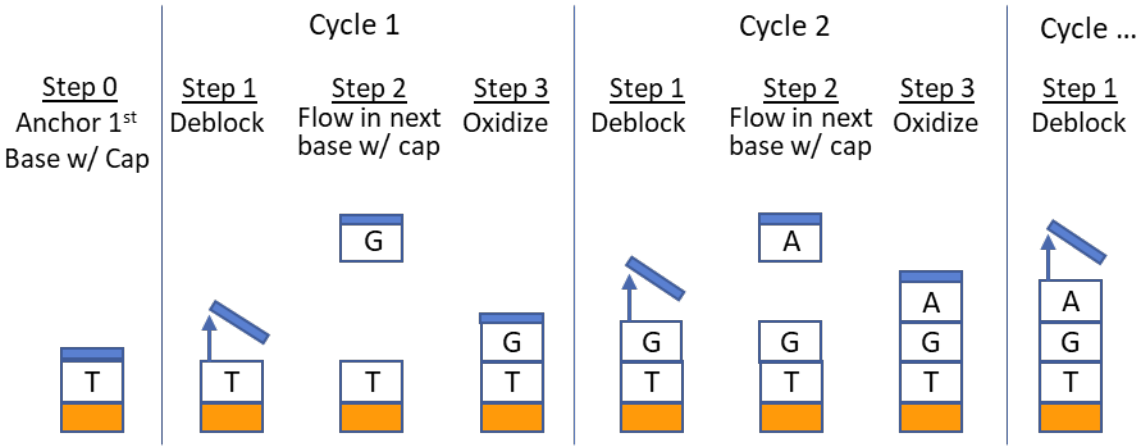
Basis of what makes DNA data storage feasible

- Information in DNA molecules is encoded in four bases (AGCT)
- We can (today) construct DNA molecules w/ defined sequences
- Hence, we can now encode digital bits in DNA molecules

Synthesis – Two general approaches today

Base by Base (chemical and enzymatic)

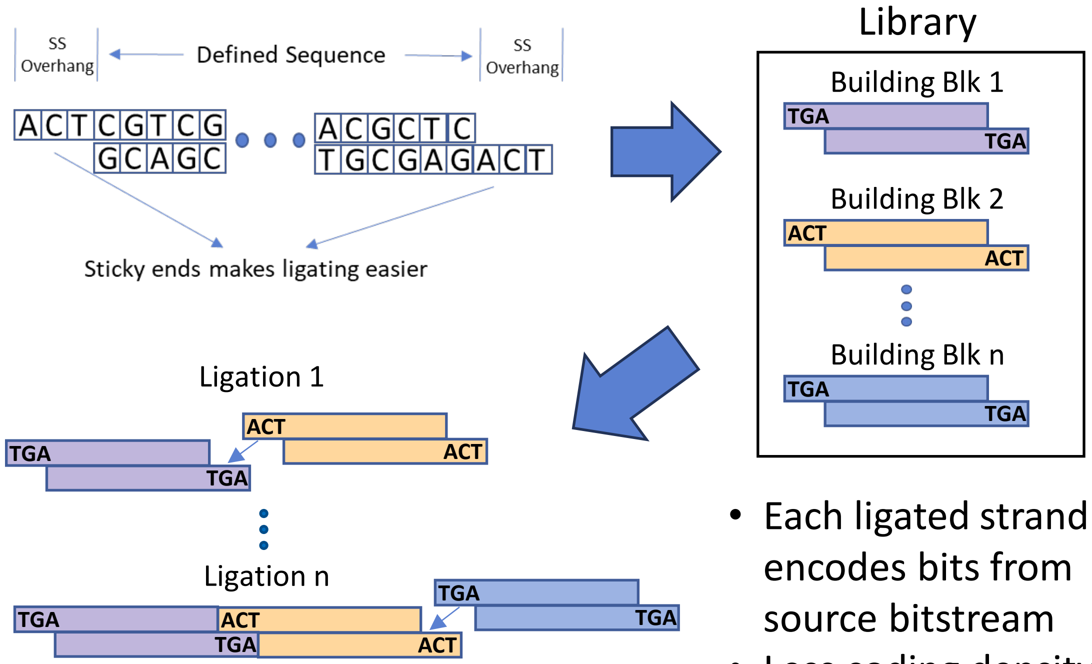
Short oligo model



- Direct bit-to-base mapping from bitstream
- Practical limit of 200-300 bases

DNA Assembly

Long oligo model

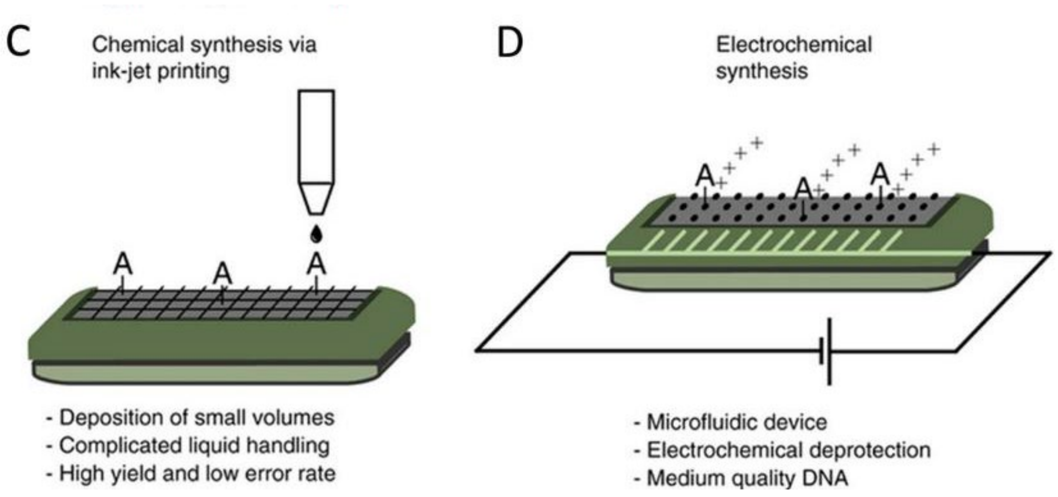


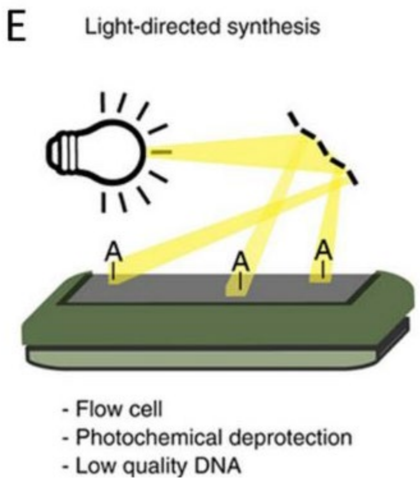
- Each ligated strand encodes bits from source bitstream
- Less coding density but other attributes; e.g., library created only once

- Different economies, performance, security, etc.
- Active innovation and competition ongoing

Synthesis – Implementation Modalities

Parellism is the trick, # synthesis sites per chip

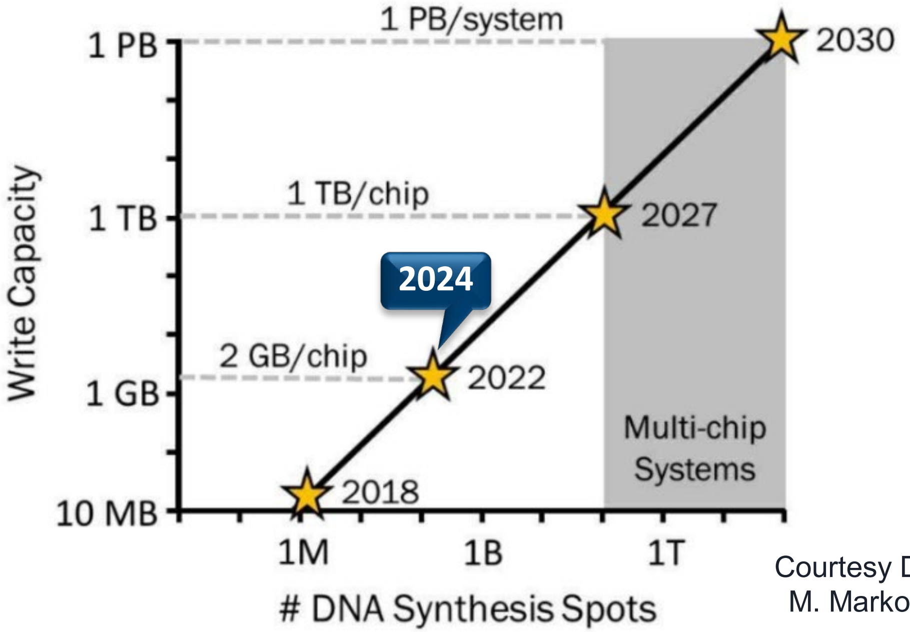




Andrea Doricchi, Casey M. Platnich, Andreas Gimpel, Friederikee Horn, Max Earle, German Lanzavecchia, Aitziber L. Cortajarena, Luis M. Liz-Marzán, Na Liu, Reinhard Heckel, Robert N. Grass, Roman Krahne, Ulrich F. Keyser, and Denis Garoli. ACS Nano 2022 16 (11), 17552-17571
DOI: 10.1021/acsnano.2c06748

2022 IARPA Roadmap for DNA Synthesis

Assumes ssDNA, 150 nt in length (20 nt flanking primers), encoded at 1 bit/nt.

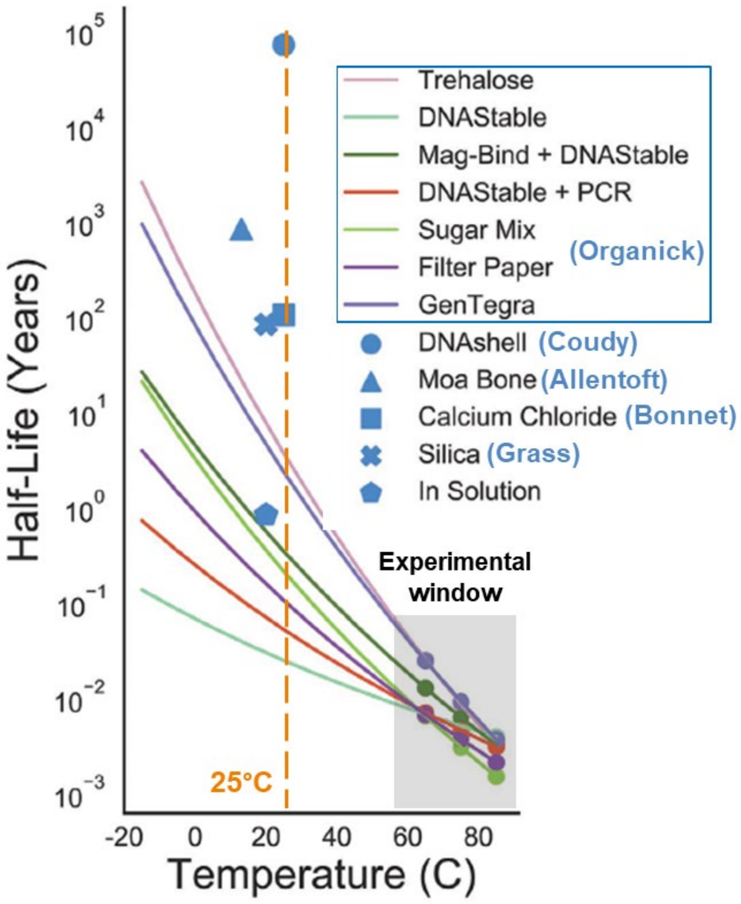


Courtesy David M. Markowitz

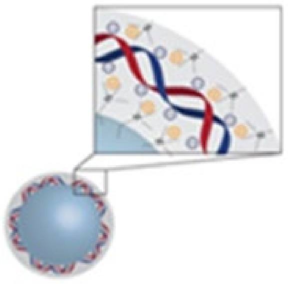
- Solutions nearing 256M spots, 2GB / run (typically 24 hrs)
- Needs more performance/cost scaling but approaching needs of early adopters

DNA Storage/Preservation

- DNA stays intact, even at room temp, if protected from water, light, oxygen
- We can predictably extend durability of DNA media by applying additives and/or containment methods
- Commercial solutions emerging with different durability properties, cost, and complexity of use



Edited from figure 2b, Organick et al , An Empirical Comparison of Preservation Methods for Synthetic DNA Data Storage. Small Methods. 2021; 5(5): e2001094



Silica nanoparticles (x years)

Vials (y years)

Sealed stainless steel (z years)

DNA Storage/Preservation Standard

Data Stability Evaluation Method for DNA Containment Systems

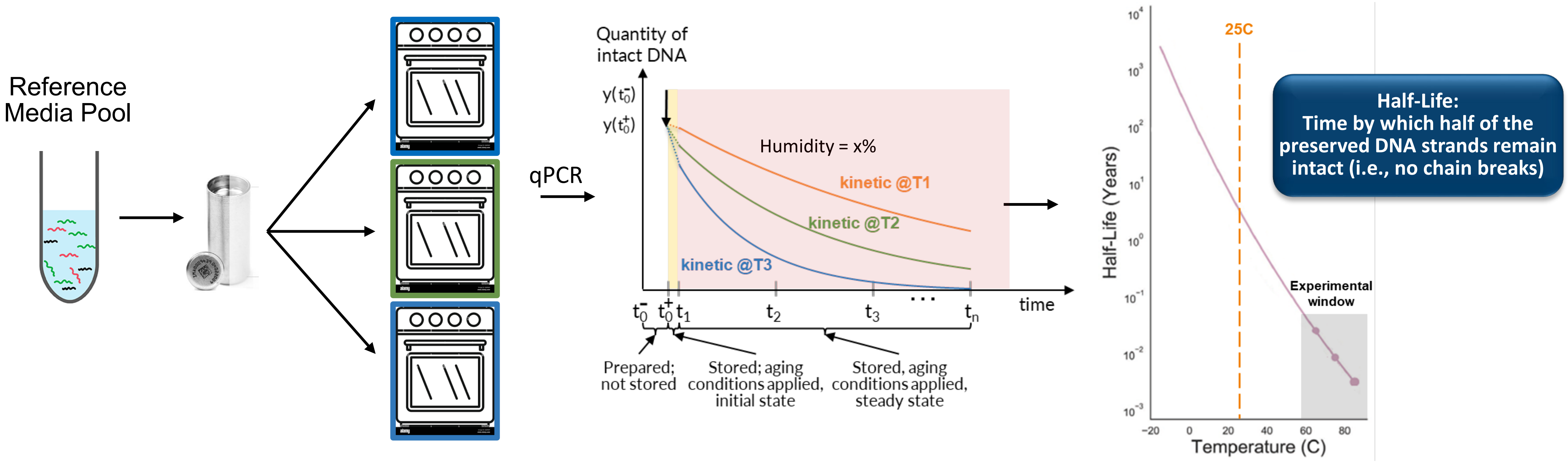
"Apples-to-apples" comparison of durability claims

DNA Containment System (DCS)

Accelerated aging at various temperature, humidity conditions

Measured remaining material (i.e., uncut strands)

Arrhenius curve fit of degradation data set



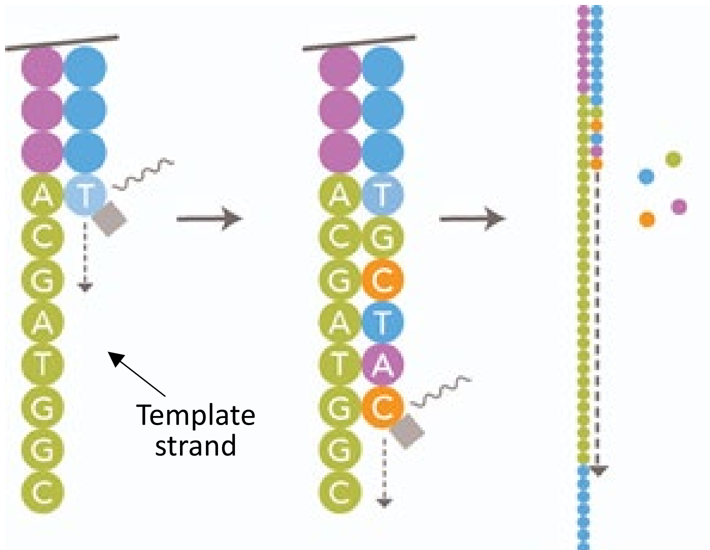
https://www.snia.org/tech_activities/publicreview

DNA DATA STORAGE ALLIANCE

A SNIA Technology Affiliate

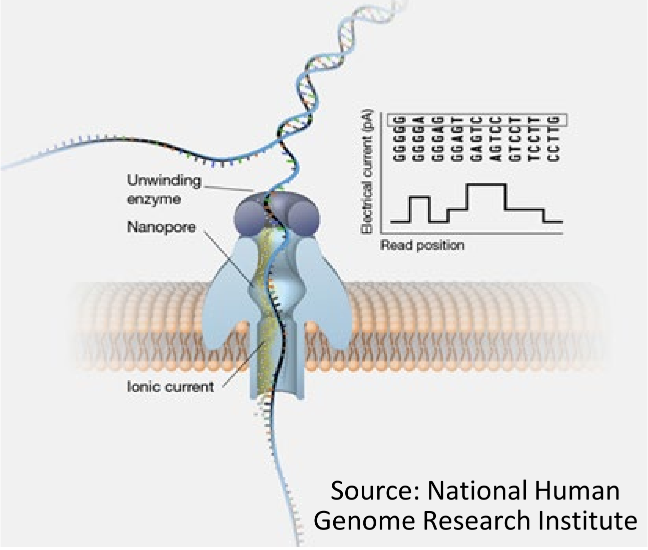
Sequencing – Two methods predominate today

- Sequencing by Synthesis (SBS)
  - From template DNA strand, build a complementary strand (hence "synthesis")
- Nanopore
  - Guide DNA strand through very small channel
  - Bases directly read as strand transits pore



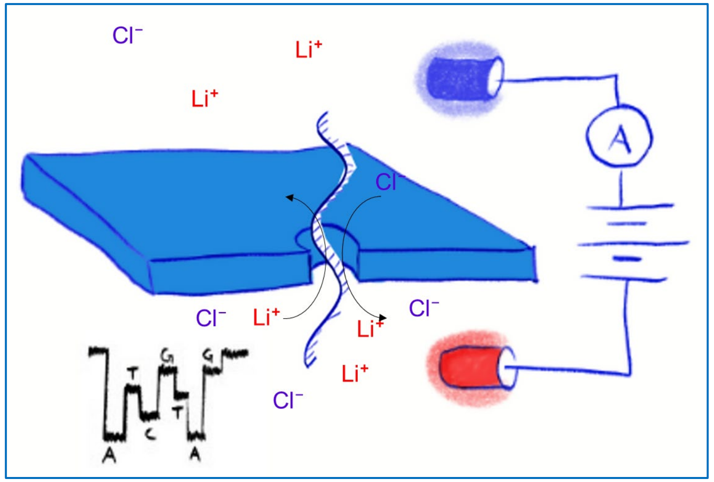
Source Illumina

Biological Nanopore



Source: National Human Genome Research Institute

Solid State Nanopore

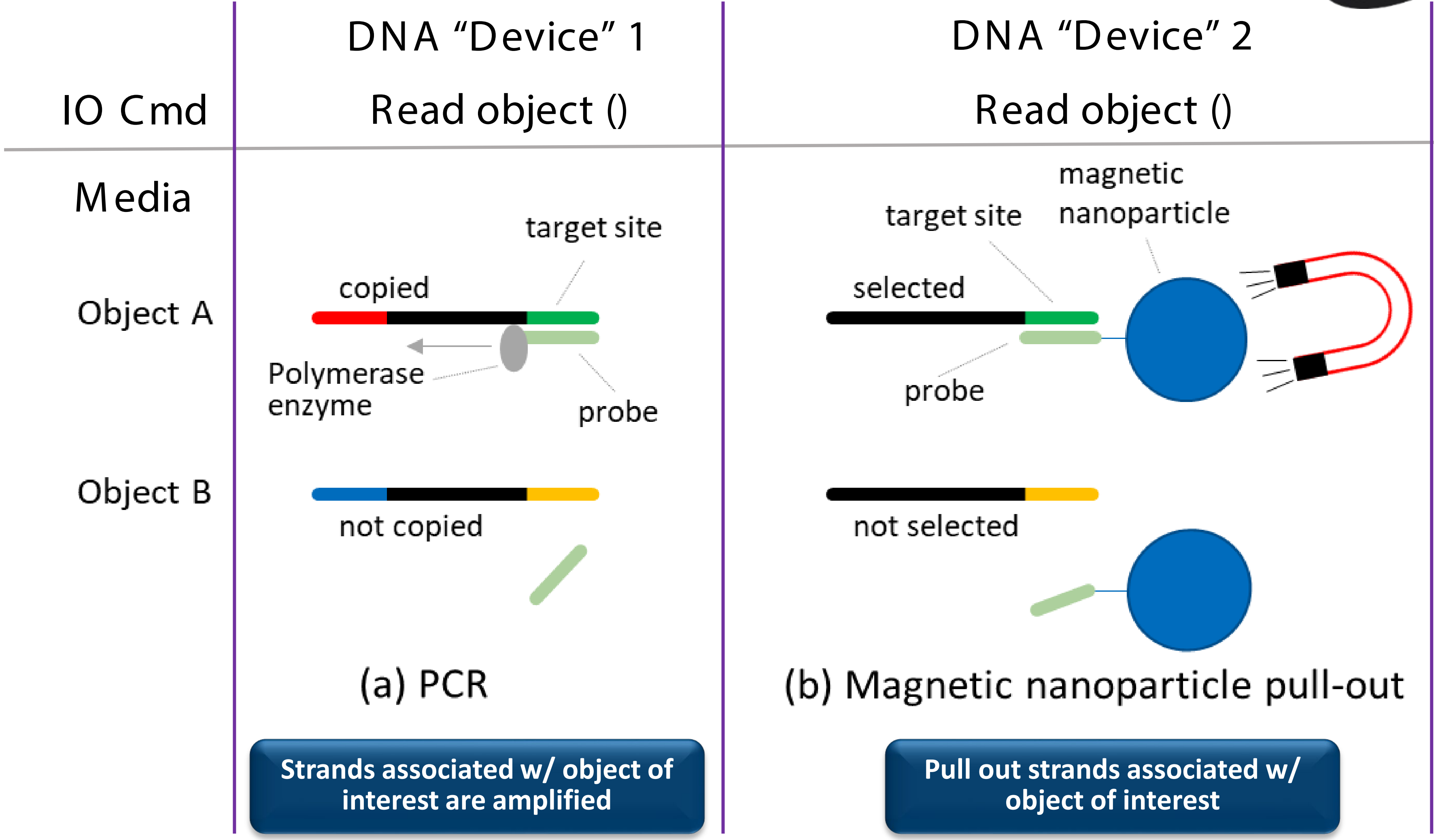


www.solidstatenanopore.com/post/nanopore-basics

- Approaching 7 terabases/day; ~1 raw TB/day (assuming 1 bit/base); <=~100 MB/day w/ protocol overhead
- Similar to case with synthesis, needs more scaling, but approaching requirements of some customers

Random Access

- Same IO operation; different chemical mechanisms in the device
- Analogous to two conventional storage devices with different designs servicing the same read command

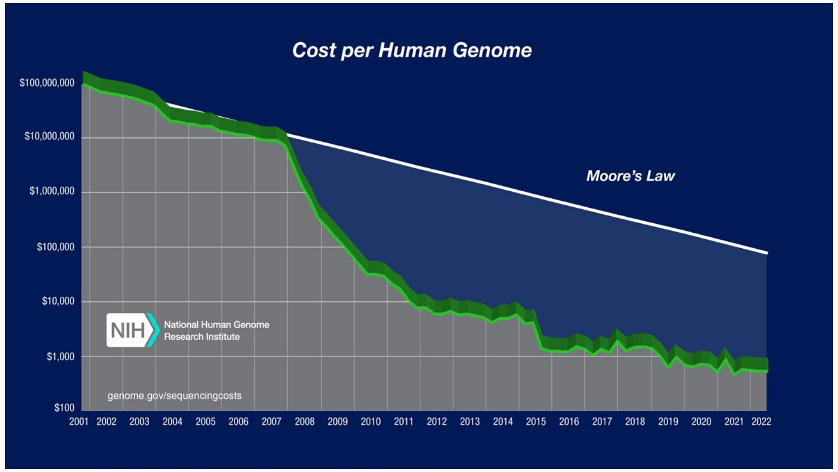


Source: Figure 7, D. Landsman, K. Strauss, "The DNA Data Storage Model", Computer, vol. 56, no. 07, pp. 78-85, 2023.doi: 10.1109/MC.2023.3272188 url: https://doi.ieeecomputersociety.org/10.1109/MC.2023.3272188

When is DNA data storage “commercially viable”?

- “Commercially viable” is in eye, and checkbook, of the beholder
  - DNA does not replace, but complements existing storage
  - Today’s DNA data storage price/perf numbers and quality metrics assume medical, scientific markets and use cases

Example:
- Cost of sequencing typically quoted in terms of price of human genome sequencing
- The # times a sample must be read to get accurate reading (coverage factor) for this use case is typically 30X; for data storage, probably 10X
- Lower coverage factor enables lower costs and sequencing performance points for data storage



Source: National Human Genome Research Institute

- Bottom line
  - DNA data storage has been shown to work on scalable technology platforms
  - Expect initial solutions for early adopters over next five years

DNA Data Storage Alliance

40+ member organizations today

Mission

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

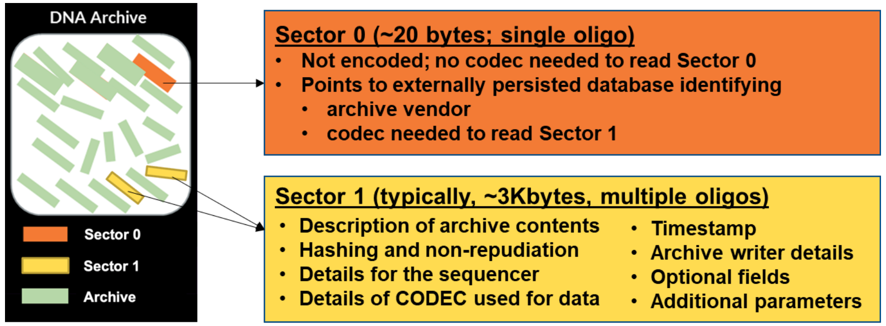
Scope

- Educate the market to create awareness and adoption of DNA data storage
- Influence and drive R&D and funding
- Develop standards and specifications to encourage evolution of the ecosystem

Twist Bioscience, Western Digital, Catalog, Quantum, Entegris, imagene, Biomemory, MISL University of WA, Dell Technologies, Lenovo, Microsoft, TU/e Eindhoven University of Technology, cacheDNA, EY, eureKARE, IMEC, Los Alamos National Laboratory, IBM, ISM Technion, DNA Script, Seagate, DNALI, DNAalgo, iol University of New Hampshire InterOperability Laboratory, Université Côte d'Azur, Monash University, UT Dallas, IRIDIA, Oligo Archive, Digital Bedrock, The University of Arizona Center for Applied NanoBioscience & Medicine, Philipps Universität Marburg, ICOS Newcastle University, BioSistemika, Reichman University, bioisac, Genomika, ipt Instituto de Pesquisas Tecnológicas, cinémathèque suisse, 21e8, Molecular Assemblies, Digital Preservation Coalition, Leibniz Universität Hannover

DNA Data Storage Alliance Standards Work

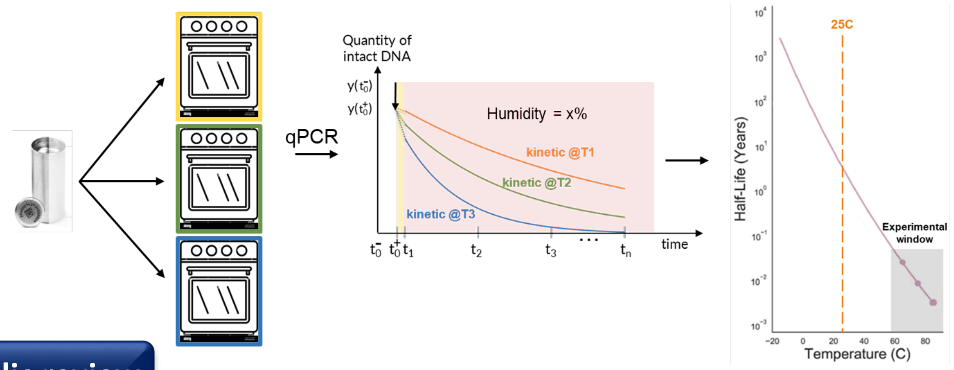
1) DNA Archive Rosetta Stone - MBR



https://www.snia.org/tech_activities/standards/curr_standards

Published

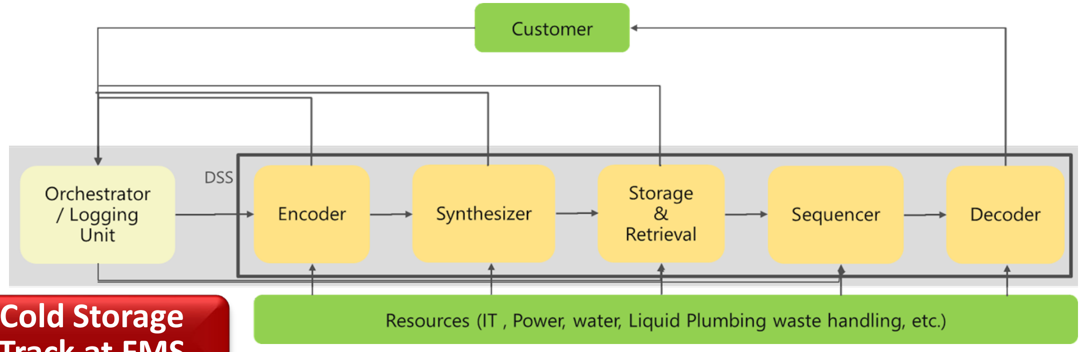
2) Data Stability Evaluation – Compare Half-Life



Public review

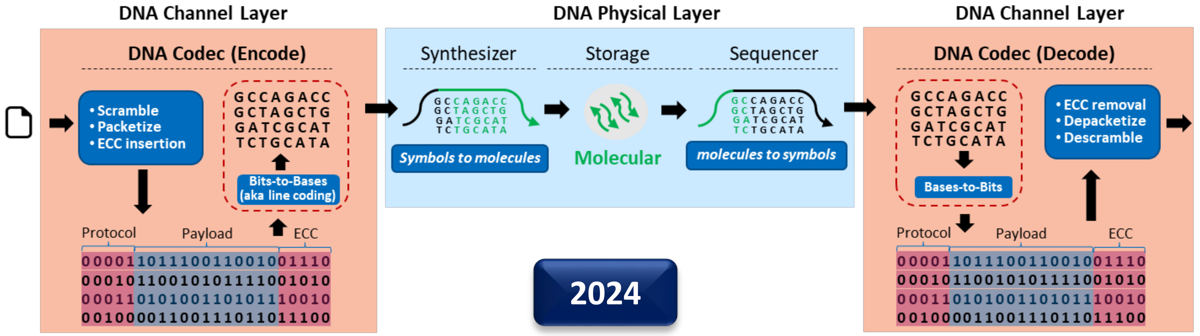
https://www.snia.org/tech_activities/publicreview

3) Interoperable Interfaces – DNA in the Datacenter



Cold Storage Track at FMS

4) Open-Source Codecs - DNA Channel



2024

And more ongoing: Biosecurity, Nanopore channel characterization, …

DNA DATA STORAGE ALLIANCE
A SNIA Technology Affiliate

Come join the fun

SNIA® DNA DATA STORAGE ALLIANCE

www.snia.org/groups/snia-dna-technology-affiliate
Twitter: @DnaDataStorage
LinkedIn: @dna-data-storage-alliance
Email: info@dnastoragealliance.org

DNA data storage fits OSI model

- App, Presentation, Session layers essentially the same as traditional storage
- Transport-Data Link layer functions (DNA Channel) implemented by DNA Codec
- PHY is chemistry vs. electrical

OSI Model

Layer	Functions
Application	• Network resource access Protocols/Services • HTTP, FTP, SMTP, …
Presentation	• Translate, encrypt, compress, serialize, … • Ensure data in usable format for transmission
Session	• Establish, manage and terminate sessions • Application Programming Interfaces
Transport	• Reliable end-to-end packet delivery • Error recovery
Network	• Move packets from source to destination • Internetworking/Routing
Data Link	• Define format on the wire; bits to frames, … • Hop-to-hop delivery
Physical	• Transmit bits over a physical medium

DNA Data Storage Model

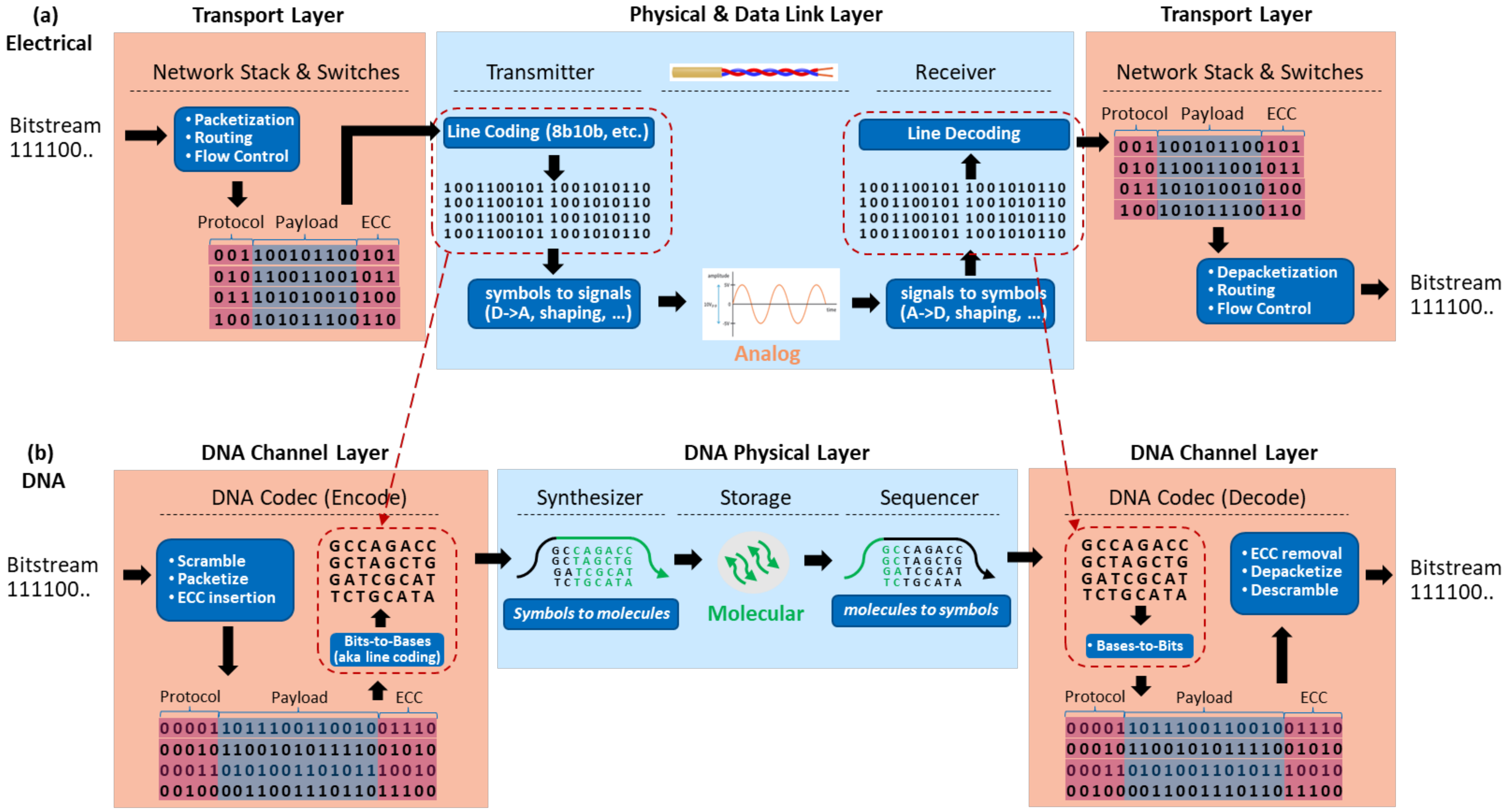
Layer	Functions	Domain
Application	• Cataloguing, Archive Management, …	Digital (1100…)
Presentation	• Encryption/Compression	Digital (1100…)
Session	• Interface for DNA archive access	Digital (1100…)
DNA Channel (Codec SW)	• Transformations • Segmentation/Packetization • Error Detection/Correction	Digital (1100…)
DNA Channel (Codec SW)	• Bits-to-Bases • Optional DNA space protocol	DNA (AGCT…)
DNA Physical (Chemistry)	• Synthesis, Storage, Sequencing • Retrieval: Primers & Probes	DNA (AGCT…)

Source: Figure 2, D. Landsman, K. Strauss, "The DNA Data Storage Model", Computer, vol. 56, no. 07, pp. 78-85, 2023.doi: 10.1109/MC.2023.3272188 url: https://doi.ieeecomputersociety.org/10.1109/MC.2023.3272188

OSI mapping: DNA and network model analogies

Network

DNA



Source: Figure 5, D. Landsman, K. Strauss, "The DNA Data Storage Model", Computer, vol. 56, no. 07, pp. 78-85, 2023.doi: 10.1109/MC.2023.3272188 url: https://doi.ieeecomputersociety.org/10.1109/MC.2023.3272188

DNA DATA STORAGE ALLIANCE
A SNIA Technology Affiliate

DNA Storage/Preservation Standard

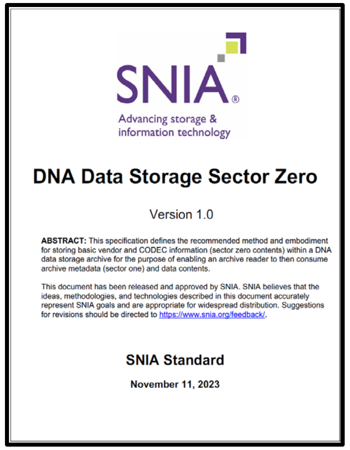
DNA Archive Rosetta Stone - “MBR” for a DNA archive

Sector 0 (~20 bytes; single oligo)

- Not encoded; no codec needed to read Sector 0
- Points to externally persisted database identifying
  - archive vendor
  - codec needed to read Sector 1

Sector 1 (typically, ~3Kbytes, multiple oligos)

- Description of archive contents
- Hashing and non-repudiation
- Details for the sequencer
- Details of CODEC used for data
- Timestamp
- Archive writer details
- Optional fields
- Additional parameters



SNIA
Advancing storage & information technology

DNA Data Storage Sector Zero

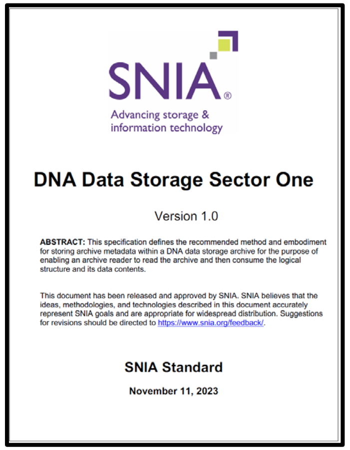
Version 1.0

ABSTRACT: This specification defines the recommended method and embodiment for storing basic vendor and CODEC information (sector zero contents) within a DNA data storage archive for the purpose of enabling an archive reader to then consume archive metadata (sector one) and data contents.

This document has been released and approved by SNIA. SNIA believes that the ideas, methodologies, and technologies described in this document accurately represent SNIA goals and are appropriate for widespread distribution. Suggestions for revisions should be directed to https://www.snia.org/feedback/.

SNIA Standard

November 11, 2023



SNIA
Advancing storage & information technology

DNA Data Storage Sector One

Version 1.0

ABSTRACT: This specification defines the recommended method and embodiment for storing archive metadata within a DNA data storage archive for the purpose of enabling an archive reader to read the archive and then consume the logical structure and its data contents.

This document has been released and approved by SNIA. SNIA believes that the ideas, methodologies, and technologies described in this document accurately represent SNIA goals and are appropriate for widespread distribution. Suggestions for revisions should be directed to https://www.snia.org/feedback/.

SNIA Standard

November 11, 2023

https://www.snia.org/tech_activities/standards/curr_standards

DNA DATA STORAGE ALLIANCE
A SNIA Technology Affiliate