DNA DATA STORAGE
ALLIANCE

# PRESERVING OUR DIGITAL LEGACY:

# AN INTRODUCTION TO DNA DATA STORAGE

JUNE 2021

A PUBLICATION OF THE DNA DATA STORAGE ALLIANCE
www.dnastoragealliance.org

# ACKNOWLEDGEMENTS

**This paper is endorsed by the following DNA Storage Alliance members:**

- Ansa Biotechnologies
- Battelle
- CATALOG
- Center for Applied Nanobioscience & Medicine (University of Arizona)
- Claude Nobs Foundation
- Digital Preservation Coalition
- DNA Script
- EPFL (École Polytechnique Fédérale de Lausanne)
- Functional Materials Lab (Swiss Federal Institute of Technology, ETH Zurich, Switzerland)
- George Church Lab (Harvard University)
- I3S lab (Université Côte d'Azur and CNRS)
- Illumina
- Imagene
- Imec
- Iridia
- Kioxia
- Los Alamos National Labs
- Microsoft
- Molecular Assemblies
- Molecular Information System Lab (University of Washington)
- OligoArchive (Imperial College)
- PFU America, Inc., a Fujitsu Company
- Quantitative Scientific Solutions (QS-2)
- Quantum
- Seagate Technology
- Semiconductor Research Corporation (SRC)
- Spectra Logic
- Twist Bioscience
- Western Digital

This document is made available by the Authors and Contributors listed above. The technology embodied in this document may be subject to intellectual property rights, including patents owned by such companies. No intellectual property license, either implied or express, is granted to you by this document. This document is provided on an as-is basis without any express or implied warranty.

# TABLE OF CONTENTS

## CONTENTS

## FIGURES

# FOREWORD

## About the authors

As members of the **DNA Data Storage Alliance,** we are committed to the future use of synthetic DNA to solve global data storage problems. We contribute our knowledge and varied perspectives to this document to inform you about this exciting new way of preserving and accessing the zettabytes of valuable information needed to preserve our unfolding human narrative.

## About the alliance

The DNA Data Storage Alliance was formed in October 2020 by Illumina, Microsoft, Twist Bioscience and Western Digital. Our mission is to create and promote an **interoperable storage ecosystem** based on manufactured DNA as a data storage medium. Our initial aim is to educate the public and raise awareness about this emerging technology. In addition, as the methods and tools for commercially viable DNA data storage become better understood and more widely available, the Alliance will consider the creation of **specifications and standards** (e.g., encoding, physical interfaces, retention, file systems) to promote the emergence of interoperable DNA data storage based solutions that complement existing storage hierarchies. The Alliance neither certifies nor endorses specific products or applications.

## About this document

This paper presents DNA data storage fundamentals in an accessible way for both technically curious readers not directly involved in computing or biological domains and for IT business, computer science or electrical engineering readers interested in the benefits, technical issues, and cost of ownership of this storage medium. It describes:

- The unprecedented demand to store and mine data and the emerging need for a new archival storage tier to meet this demand. (Sections 1 and 2)
- How DNA data storage can meet the needs of this emerging tier. (Section 3)
- The DNA data storage pipeline (i.e., the process of converting digital data into DNA, storing it, and reading it back) and the economics of DNA data storage. (Sections 4 to 6)
- The underlying technologies used for DNA data storage. (Sections 7-10)

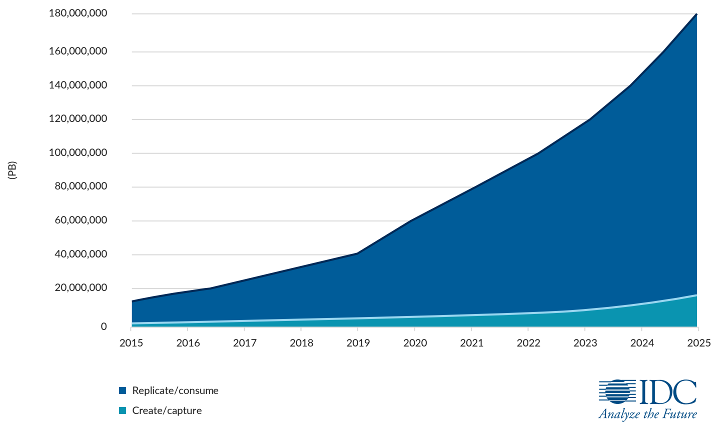
The Alliance is currently working on a second document that will describe the markets and use-cases for DNA-based storage. It will also examine, how to bridge the gap from where DNA data storage technology is now to the point of its commercial viability and widespread use.

We heartily encourage companies and institutions intrigued by the challenges of storage scaling to join our work to advance this transformational approach. If you wish to get involved or just get more information, please visit www.dnastoragealliance.org, or contact us at info@dnastoragealliance.org.

# 1 STATE OF DIGITAL DATA GROWTH: THE DATA OVERWHELM

The global Information Age – characterized by the creating, buying, selling, and hoarding of data – is overwhelming our ability to analyze, store, power, and secure this prized commodity. Big data has become a big problem. In this section, we briefly survey the data explosion and factors affecting it.

According to IDC's Global DataSphere[1], data generated globally (new and replicated copies) is expected to grow at a 23% compound annual growth rate (CAGR) from 2020-2025, reaching 180 zettabytes in 2025. Compare this staggering growth to a total of 3 zettabytes of data generated in 2010.[2] In only three years, newly created data more than doubled from 30 zettabytes (2017) to 64 zettabytes (2020)!



*Source: IDC, Worldwide Global DataSphere Forecast, 2021–2025: The World Keeps Creating More Data — Now, What Do We Do with It All?, #US46410421*

**Figure 1 - IDC, Worldwide Global DataSphere Forecast, 2021-2025**

In addition to the raw quantity of new data, IDC notes that the **ratio of replicated to initially captured data is also increasing (Figure 1).** Many cloud-based storage architectures maintain multiple copies of data to enhance the likelihood that it will be recovered intact and to avoid wait times as storage devices conduct error recovery; it is often quicker to simply retrieve the data from another device or database that holds a replica of the data.

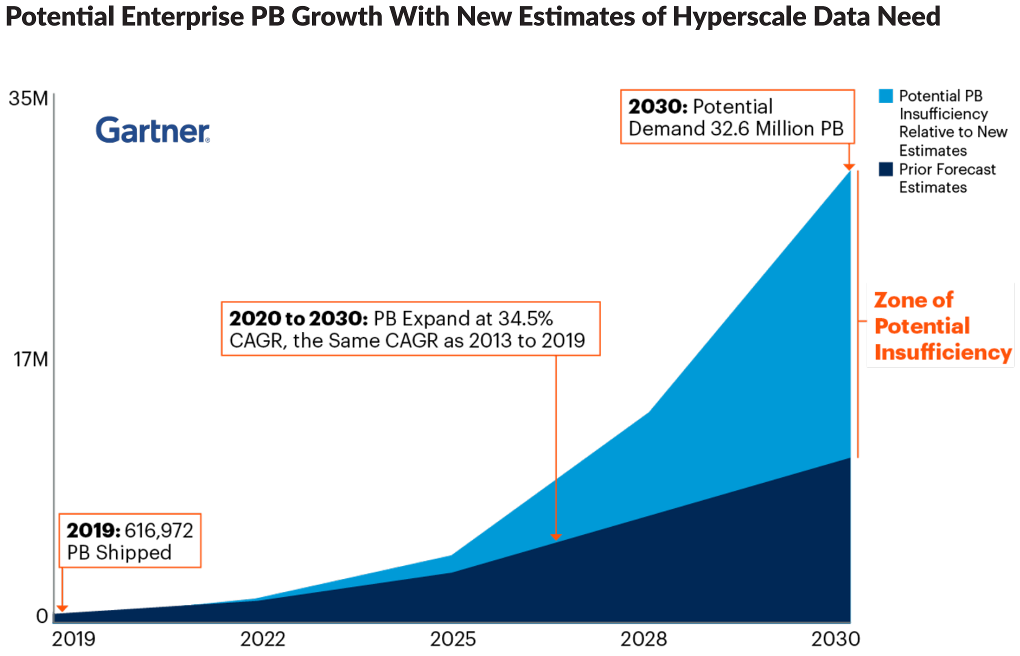
**Karl G. Jansky Very Large Array - Source: NRAO**

**Data retention and data mining** are also driving the "data overwhelm". Users in fields as diverse as robotics, smart cities, autonomous vehicles, healthcare, astronomy, climate science, and others, seek to save ever larger data sets for future data mining in efforts to remain competitive and/or drive scientific discovery. For example, radio astronomers are collecting huge volumes of data which, if stitched together, can literally yield new knowledge about the observable universe.  We cannot know today what data will become relevant to a new discovery tomorrow.  Similarly, sensors from intelligent/autonomous vehicles generate very large data streams. Discarding too much risks losing data vital for, say, accident reconstruction, traffic/municipal planning, etc. 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 weighted in the direction of saving more of this raw data for potential future mining.

Further, governments are enacting legislation that affects the retention of data (e.g., the Health Insurance Portability and Accountability (HIPAA) Act and the Sarbanes-Oxley (SOX) Act for all publicly and some privately traded companies). The retention period requirement for many regulatory frameworks may span decades.

Thus, general data growth rates, the commercial/scientific potential of data mining, and regulatory requirements are all driving the trend to store more data for longer time periods.

While the factors above are driving unprecedented storage demand growth, **the supply of storage is not keeping pace with the burgeoning demand.** IDC StorageSphere estimates that the total installed base of storage is expected to grow at 19% CAGR (2020-2025).[3]  On the other hand, Gartner refers to a coming **Zone of Potential Insufficiency (Figure 2)** and estimates that hyperscale vendor storage demand has already outstripped this rate, growing at nearly 35% CAGR from 2013 to 2019, and expected to potentially soar to as much as 50% CAGR from 2020 through 2030. [4]  We next explore some reasons for this.



*Source: Gartner Market Trends: Evolving Enterprise Data Requirements—How much is Not Enough? Published 14-Jul-2020 —ID G000724101 by analysts John Monroe, Robert Preston*

**Figure 2 – How much is not enough? (Gartner)**

# 2 STATE OF DIGITAL STORAGE

## 2.1 Historical storage technology scaling

Innovation in the storage industry has realized staggering improvements in density, size and total capacity. The first hard disk drive (HDD) was introduced in 1956 and was a refrigerator-size device with a capacity of 5 megabytes and a price of $10K/megabyte. With gradual improvements in magnetic recording technology, we have reached the 18-20 terabyte scale in a 3.5-inch form factor. The average selling price of nearline HDDs in 2019 was ~$20/terabyte, nine orders of magnitude less than the 1956 HDD. [5] The industry's next-generation HDDs will further scale density using energy-assisted magnetic recording (i.e., bits written using heat or microwave energy applied at the write head), achieving further cost scaling.

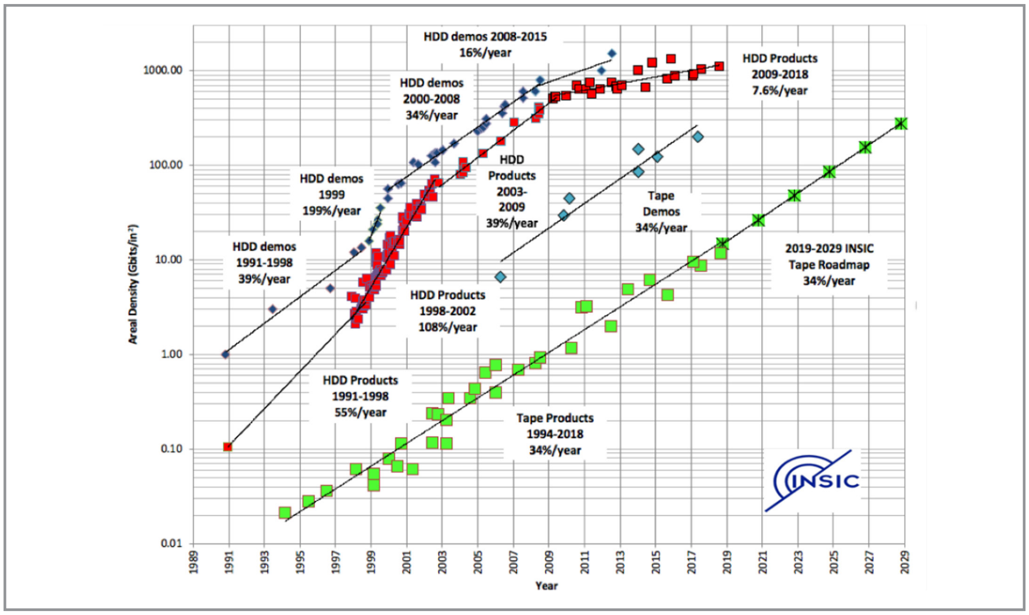


**Figure 3 – Areal Density Trends - HDD & Tape**

Tape and HDDs are magnetic recording siblings invented within a year of each other. [6] In use since 1951, tape remains a viable medium to backup or archive digital data. What began mid-century as storing data at 128 characters/inch on 8 tracks in magnetic tape now stores 10 gigabits per square inch. The areal density (i.e., the number of bits a medium can store per unit area) of tape is expected to grow at a 34% CAGR from 2019 – 2029. Figure 3 shows areal density trends for magnetic media over the past 30 years.[7]

Solid state drives (SSDs) are based on NAND flash memory and were introduced in 1991. SSDs are accessed using the same basic block storage model as HDDs but are many times faster, and more expensive, than HDDs. The first SSD could store 20 megabytes at $50/megabyte. The current state of the art enables the storage of multiple bits per memory cell and the stacking of memory cells on top of each other in 3D structures (100 layers is now common), enabling development of dense SSDs with capacities up to 100 terabytes. [8]

## 2.2 Challenges for today's archival storage technologies

Despite ongoing improvements in media scaling, key challenges remain for today's storage technologies when considered for zettabyte scale and long storage duration.

### 2.2.1 Storage maintenance and replacement costs

Today's storage media (magnetic, semiconductor, etc.) can, with proper care, retain data for decades. However, like any physical assets, they wear out over time and degrade. As a result, their status must be periodically checked and monitored to ensure data integrity (i.e., fixity checks).

Further, the intrinsic format of the media is tightly coupled with the techniques used to read and write it. Computing history is littered with storage devices that can no longer be read because their readers or physical media formats have become obsolete for technological or commercial reasons. As a result of this, data stored on any of today's storage devices are periodically re-written onto new generations of devices to ensure continued access. While some deep archives store the actual devices with which the data was written, this also poses impracticalities.

### 2.2.2 Density limitations

In 1975, Gordon Moore revised his original 1965 prediction and formulated what became known as Moore's Law; namely, that the number of transistors that can be packaged in an integrated circuit would double every two years. This prediction has held since it was stated, representing a CAGR of about 40%. For storage, media density growth rates have varied. For example, the CAGR of HDD areal density went from 108% for 1998-2002, to a 39% CAGR for 2003-2009, and to a 7.9% CAGR for 2009-2018 (see Figure 3). Though advances such as energy-assisted recording are emerging to further push HDD areal densities, the overall trend in areal density for magnetic media is slowing. NAND flash has hit periodic scaling limits; 2D NAND, in which memory cell size was reduced in the planar (x-y) dimension, began hitting scaling limits around 2012, but 3D NAND (building cells upward in the z-dimension) enabled a resumption of bit growth. 3D will itself eventually hit limits.

These trends pose capital expenditure and operational cost challenges to today's archival storage solutions at zettabyte scale. This does not imply that today's storage solutions will become obsolete. Rather, it suggests the need for a new level in the storage hierarchy that can cost effectively scale to meet the explosive data growth in the evolving storage ecosystem.

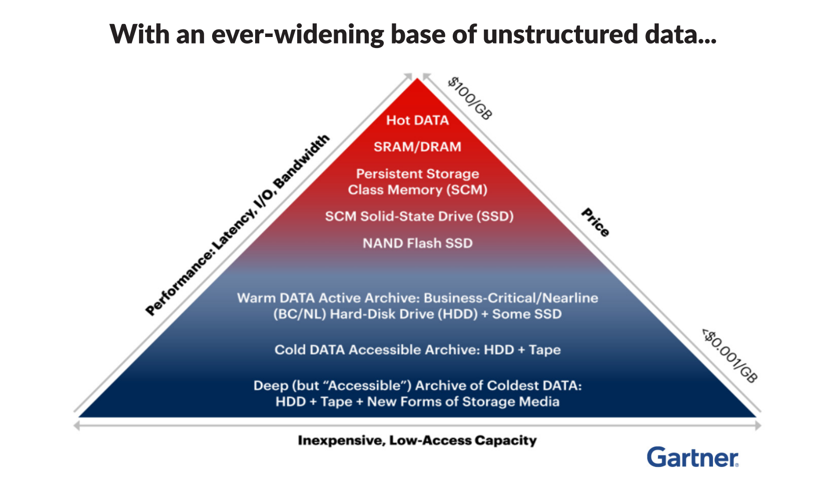
### 2.2.3 Energy and sustainability concerns

Some estimates indicate that data centers consumed around 1% of total global electricity in 2018, which could triple or quadruple within the next decade; if energy efficiency improvements do not continue, data center electricity usage could grow to 3-13% of **total global electricity consumption** by 2030. [9] Further, traditional storage devices are made from materials that must be mined. HDD and tape solutions, in particular, rely on rare earth metals with complicated supply chains, posing sustainability concerns.

## 2.3 The total cost of ownership for storage media

It is important to view the storage hierarchy in terms of total cost of ownership (TCO).

Storage can be divided into tiers based on the frequency of data access **(Figure 4).** Data accessed frequently ("hot") is stored on high-performance devices (i.e., SSDs). Data accessed somewhat frequently ("warm") is generally stored on HDDs. Data accessed infrequently ("cold") is generally stored on tape.

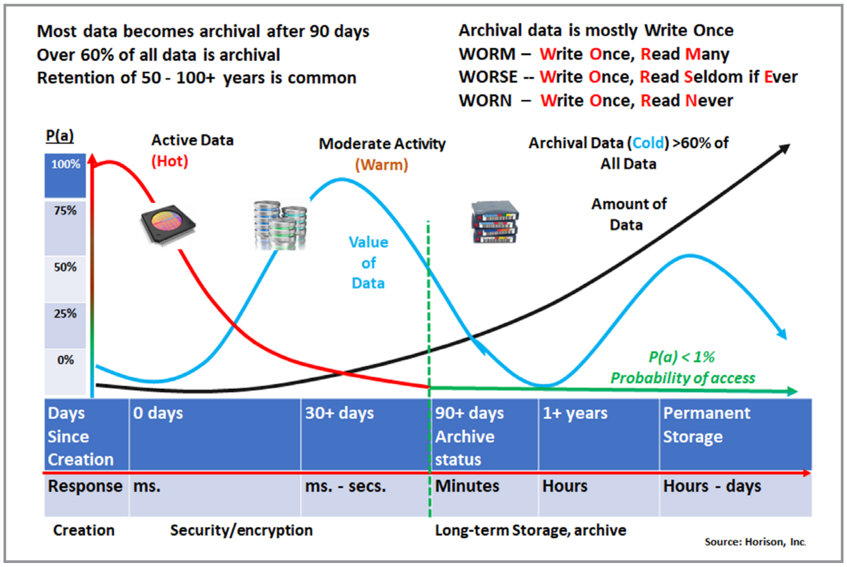

With an ever-widening base of unstructured data...
$100/GB
Hot DATA
SRAM/DRAM
Persistent Storage Class Memory (SCM)
SCM Solid-State Drive (SSD)
NAND Flash SSD
Warm DATA Active Archive: Business-Critical/Nearline (BC/NL) Hard-Disk Drive (HDD) + Some SSD
Cold DATA Accessible Archive: HDD + Tape
Deep (but "Accessible") Archive of Coldest DATA: HDD + Tape + New Forms of Storage Media
Performance: Latency, I/O, Bandwidth
Price
<$0.001/GB
Inexpensive, Low-Access Capacity
Gartner

*Source: Gartner Product Manager Insight: Tape to the Future?*
*Published 23 October 2020 – ID G000724101 by analysts John Monroe, Robert Preston*

**Figure 4 - The Evolving Storage Pyramid - Gartner**

As we move up the pyramid, storage media acquisition and replacement cost drive up the TCO. Further, storage devices higher up in the pyramid consume more power compared to devices in lower tiers, again driving TCO up. The total bits in each storage tier is inversely proportional to the underlying cost of the tier. In datacenters, SSDs and HDDs are typically refreshed every three to five years, and tape every seven to ten years. The need for such frequent replacement adds to maintenance cost, increasing TCO.

Another trend driving up TCO is that **the amount of cold data is growing faster than data in the other tiers;** that is, we are storing more data for longer. It is also increasingly common that less than 1% of this data is accessed more than 90-120 days after creation (e.g., historical archives of all kinds, photos, videos, climate data, astronomical data, and sports).[10] **Figure 5** shows how the access frequency (red-to-green line), commercial value (blue line), and quantity (black line) of data trend over time. Note in particular that while value drops substantially by around the one year mark, there is a subsequent rise in value as we move into the "permanent storage" range, and this is the part of the lifecycle where the amount of data continues growing rapidly. This "data cooling vs. value" trend



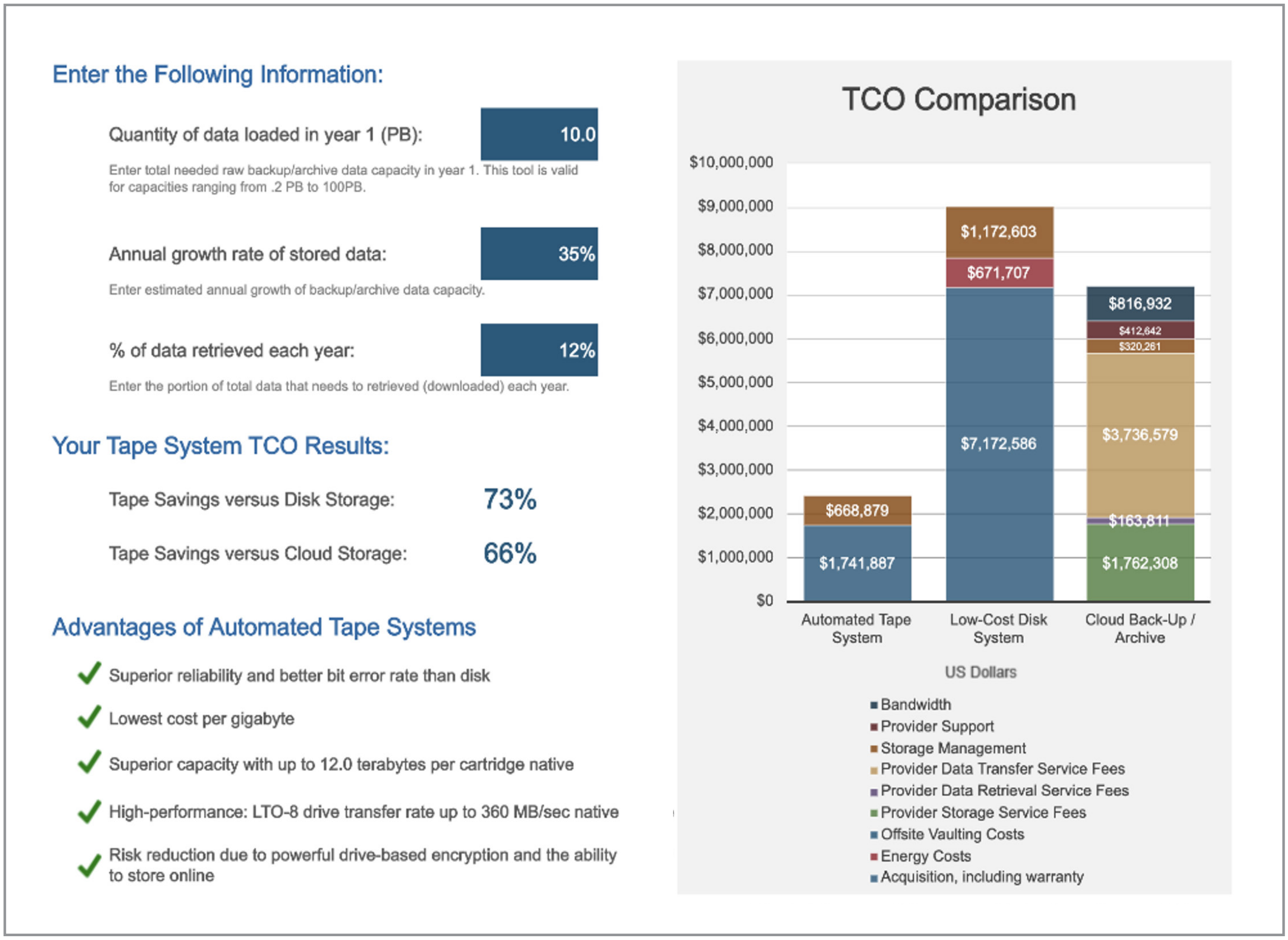
Source: Horison, Inc.

**Figure 5 - Data Temp/Value/Quantity vs. Time - Horison, Inc.**

reinforces the value of data mining (Section 1), and emphasizes the need for a cost-effective storage tier with greater scale than HDD or tape.

### 2.3.1 Calculating total cost of ownership

A thorough TCO calculation for storage includes factors such as: acquisition cost of hardware and media, the timeframe over which the data will be stored, the cost of writing the data, the annual growth rate of data stored, the amount and frequency of data retrieved, the number of copies stored, the number of years between migrations, the cost of electricity and facilities, the cost of migrations, the cost of employees, and more. The Fujifilm TCO calculator **(Figure 6),** is an example of a tool that helps analyze an organization's storage TCO. [11] We discuss the economics of DNA data storage in Section 6.



Enter the Following Information:

Quantity of data loaded in year 1 (PB): 10.0

Enter total needed raw backup/archive data capacity in year 1. This tool is valid for capacities ranging from .2 PB to 100PB.

Annual growth rate of stored data: 35%

Enter estimated annual growth of backup/archive data capacity.

% of data retrieved each year: 12%

Enter the portion of total data that needs to retrieved (downloaded) each year.

Your Tape System TCO Results:

Tape Savings versus Disk Storage: 73%

Tape Savings versus Cloud Storage: 66%

Advantages of Automated Tape Systems

- ✓ Superior reliability and better bit error rate than disk
- ✓ Lowest cost per gigabyte
- ✓ Superior capacity with up to 12.0 terabytes per cartridge native
- ✓ High-performance: LTO-8 drive transfer rate up to 360 MB/sec native
- ✓ Risk reduction due to powerful drive-based encryption and the ability to store online

**Figure 6 - Fujifilm Storage TCO Calculator**

# 3 DNA AS A STORAGE MEDIUM

One solution to reduce both the physical and carbon footprints of traditional storage, while significantly reducing TCO in the archival tier, is DNA-based data storage.

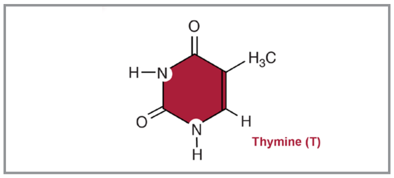
When stored properly, DNA data can last reliably for thousands of years with little to no power consumption or need for maintenance/refreshing.  Storage density, durability and low power consumption of DNA-based data storage radically reduce its TCO, making it a strong contender for long-term archival data storage (see Section 6).

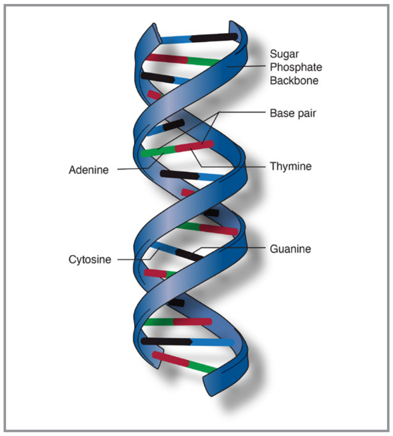
Unlike today's storage, where media are pre-manufactured and delivered empty, molecules that represent DNA-stored data are created on demand, and the information is directly encoded in the way the synthetic DNA molecule is assembled.

## 3.1 Biological v synthetic (manufactured) DNA

DNA, or deoxyribonucleic acid, is nature's system for reliable, long-term genetic information storage.  It is a molecule composed of two polymer chains that form a double helix. Each chain contains a sequential string of nucleotide monomers (also referred to as bases). There are four naturally occurring DNA bases: adenine (A), thymine (T), cytosine (C), and guanine (G). The two chains are kept together in the double helix conformation by hydrogen bonds between bases in the opposite chains, where "complementary" bases are matched in A-T and C-G pairs across from each other in the chain. In nature, DNA usually exists in a **d**ouble **s**tranded helix (dsDNA), but it also occurs in some organisms as a **s**ingle **s**tranded polymer chain (ssDNA). Either dsDNA or ssDNA may be useful for DNA data storage.

In the context of digital data storage, however, DNA is manufactured: **the creation of the DNA data storage medium does not require or use --- nor does the resulting stored data result in the creation or modification of --- any cells, organisms, or life.**





**Courtesy: National Human Genome Research Institute**

## 3.2 Properties of DNA for archival storage

DNA has unique properties that make it an ideal medium for storing archival data for many years, decades, centuries or even millennia.

### 3.2.1 Media durability

DNA is the molecule of choice for information storage in biological systems. It can remain intact for thousands of years at room temperature in a dry atmosphere.[12] An international team led by researchers at the Centre for Palaeogenetics in Stockholm discovered and successfully sequenced DNA from Columbian mammoth remains that are up to 1.2 million years old.  This chemical stability ensures that data encoded in DNA can be safely preserved over very long time periods.

### 3.2.2 Maintenance simplicity

Today's storage media must undergo periodic fixity checks to ensure that their data remains readable. Due to the durable nature and other properties of DNA, we expect its maintenance at rest to be much simpler than legacy storage solutions and as such will not incur significant post-creation data retention costs.

### 3.2.3 Format immutability

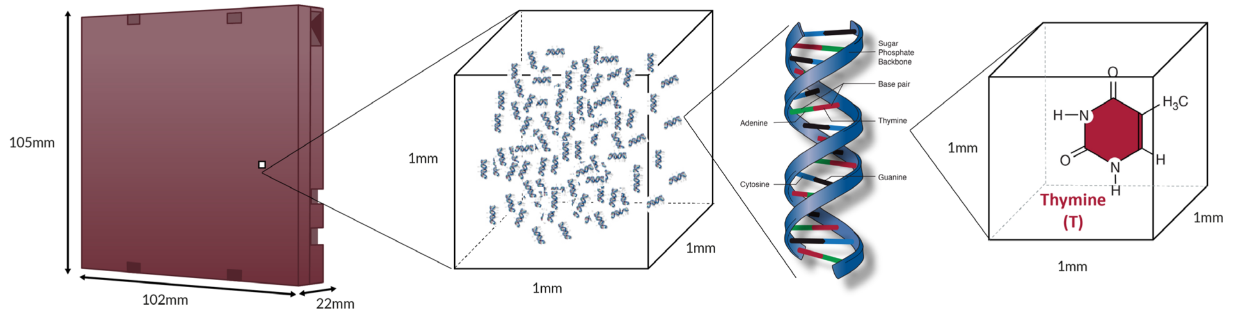
A fundamental factor distinguishing DNA as a storage medium is its molecular structure, which is universal. Digital data archived in DNA today will be chemically readable in thousands of years. This property offers a significant advantage for DNA-based storage vs. legacy storage.

As noted in section 2.2, with existing storage technologies, the physical structure and format of the media, and the methods used to read and write the media, are technically coupled.  This creates the risk that the devices needed to read archival data are not available, which in turn causes the need for the periodic migration of data to new generations of media/devices. In contrast, the immutable format of DNA ensures that the DNA in which digital data is stored will always be able to be read, and can be decoded as long as the encoding with which it was written (a logical construct vs. a physical device) is available. Data migration is minimal or unnecessary.

### 3.2.4 Density

Storage density for magnetic media has historically been defined as **areal density,** i.e., the number of bytes a medium can store per unit area, since bits in magnetic storage media are generally laid out on a two-dimensional plane. However, DNA makes possible a wide diversity of form factors, including three-dimensional storage. To compare data density between magnetic storage and DNA, it is more appropriate to use volumetric densities.

DNA bases, on the order of tens of atoms in size, occupy a volume of about one cubic nanometer; thus, even when accounting for significant practical system overheads, the number of DNA bits storable in a 1mm$^3$ volume is estimated at 9 terabytes, about half the capacity of an 18 terabyte LTO-9 magnetic tape. If the space inside an LTO cassette (approximately 235,000 mm$^3$) were filled with DNA bits, the cassette would hold about 2,000,000 terabytes of data, or about 115,000 times the capacity of an LTO-9 tape.



**If the space inside an LTO cassette, ~235,000 mm$^3$, were filled with DNA bits, the cassette would hold about 2,000,000 terabytes of data, or about 115,000 times the capacity of an LTO-9 tape**

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

**Single DNA Base fits in ~1nm$^3$**

DNA will enable storage capacity on a scale never imagined, let alone considered economically viable, compared to today's storage technologies.

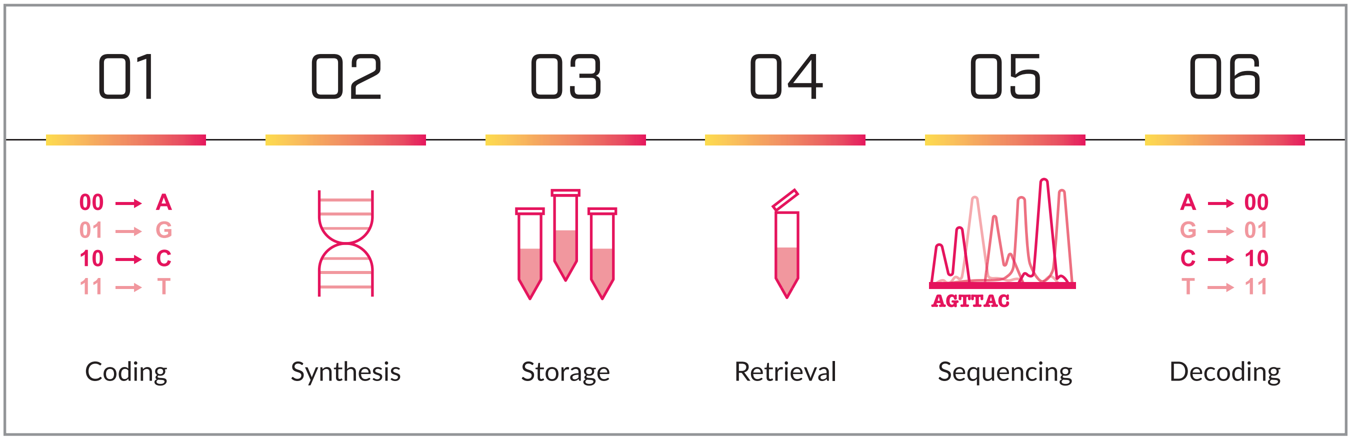
### 3.2.5 Energy efficiency and sustainability

Compared to today's datacenters with today's storage technologies, data stored in DNA consumes minimal to no resources while at rest.  While current datacenters use a significant amount of power and land, with DNA data storage, these requirements will be negligible. Finally, due to its durability and density, disposal of DNA should have much less environmental impact than the disposal of obsolete tape drives or HDDs.

### 3.2.6 Cost

When dealing with archival data that needs to last decades or longer, a storage medium that doesn't incur additional cost over time is very attractive. See Section 6 for details.

# 4 THE DIGITAL DATA TO DNA PIPELINE



To store data in DNA, the original digital data is **encoded** (mapped from 1's and 0's to sequences of DNA bases), then **synthesized** (written), and stored. When the stored data is needed again, the DNA molecules are **sequenced** (read) and **decoded** (re-mapped from DNA bases back to 1's and 0's). The next paragraphs address each step in turn.

## 4.1 Encoding ("converting bits to bases")

The basic concept of encoding for DNA data storage is the process of converting the 1's and 0's of the original digital data into sequences of the bases (ACGT) that comprise DNA molecules. Encoding methods are tightly coupled to the synthesis and sequencing methods being used, enabling acceptable bit density, compensating for error rates, and enabling the segmentation of the original binary data to DNA strands and the reassembly of those DNA strands back to binary data. See Section 7 for details.

## 4.2 Synthesis ("writing")

Synthesis is where DNA manufacturing occurs. Based on a series of chemical steps, the DNA molecules, as determined by the encoding step, are assembled in various ways that mirror the "bits-to-bases" or other encoding methods. See Section 8 for details.

## 4.3 Physical storage of DNA

After synthesis, the DNA is encapsulated for long-term preservation and deposited in a library where pools of DNA are stored. There are multiple types of encapsulation, including sealing DNA in capsules with inert gas or mixing it with chemicals that help to preserve it. See Section 9 for details.

## 4.4 Retrieval (from libraries)

After storage, and once the data is needed, the encoded DNA is retrieved from its library and prepared for sequencing. Often, this process also includes making copies of the molecules for sequencing methods that are molecule intensive and for cases where more copies serve distribution or further storage needs.

## 4.5 Sequencing ("reading")

Sequencing is the process of determining the identity and order of the DNA bases (ACGT) in a segment of DNA. Various sequencing methods are in use today (e.g., sequencing by synthesis (SBS), nanopore sequencing). These employ various methods (e.g., optical, pH-based, electrical) of sensing the actual bases in the strands of DNA being read.  See Section 10 for details.

## 4.6 Decoding ("converting bases to bits")

Decoding involves mapping the bases in a string of sequenced DNA back to digital data. Importantly, it involves performing error correction to recover from any potential errors that may have occurred during synthesis, preservation, and sequencing, e.g., errors in individual base sequences or from strings of bases being lost. Once decoding is completed, the data is reassembled in digital form and returned to the user.

# 5 DNA TOOLS

Over the past few decades, the field of biotechnology has made remarkable progress in developing tools to read, write, and manipulate DNA. Existing and emerging applications in life sciences, agriculture, and energy production/storage are driving this demand.

A particularly useful process is **PCR, or polymerase chain reaction,** which selectively copies DNA. PCR is used to create multiple replicas of data stored in DNA and for random access, i.e., selecting a pre-tagged subset of molecules to copy from a larger pool.  Other technologies have emerged in applications such as gene editing; these could be applicable in the future to DNA data storage, as well.

As DNA data storage continues to emerge as a viable application, it will further drive the creation of new tools that will continue to lower the TCO for DNA data storage, as well as enable new functionality.

# 6 ECONOMICS OF DNA DATA STORAGE

We now highlight some economic aspects of DNA data storage. Today, writing (synthesis) and reading (sequencing) DNA for data storage are not practical at scale; however, the trends are promising and progress in both is critical for the adoption of DNA data storage solutions. Synthesis costs are fundamental for all use cases, while sequencing costs are particularly important to enable archival use cases where reading the data frequently is required.

**Figure 7** shows trends since 1990, measured as "price per base". These trends have been driven mainly by scientific/medical applications. We'll next explore how these trends relate to cost per bit in the context of DNA data storage.

**Price Per Base of DNA Sequencing and Synthesis (circa 2017)**

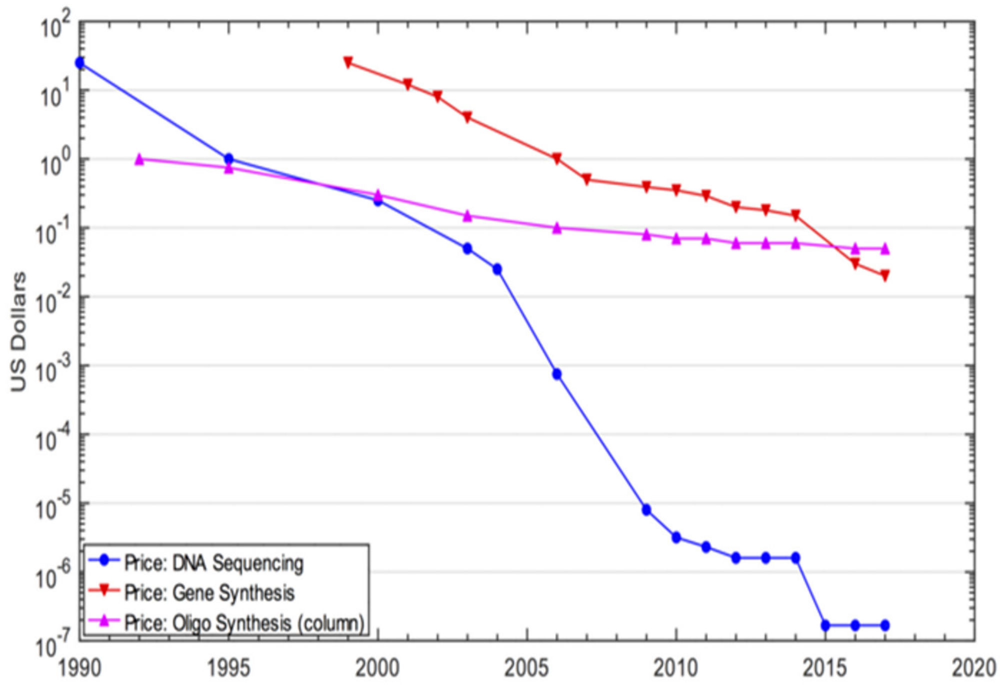


**Figure 7 - DNA Synthesis/Sequencing Costs Dr. Robert Carlson, Bioeconomy Capital**

## 6.1 Synthesis

Synthesis costs for DNA data storage are dependent on both how bits are encoded into DNA bases, and also the specific methods of synthesizing the DNA. Since today's commercial applications do not include DNA data storage, synthesis pricing estimates directly related to DNA data storage are somewhat hard to characterize. Nonetheless, IARPA, which is funding work in this area with their Molecular Information Storage (MIST) program, has laid out a target roadmap towards a synthesis cost of $1/gigabyte by 2024 and $1/terabyte by 2030. **(Figure 8)**

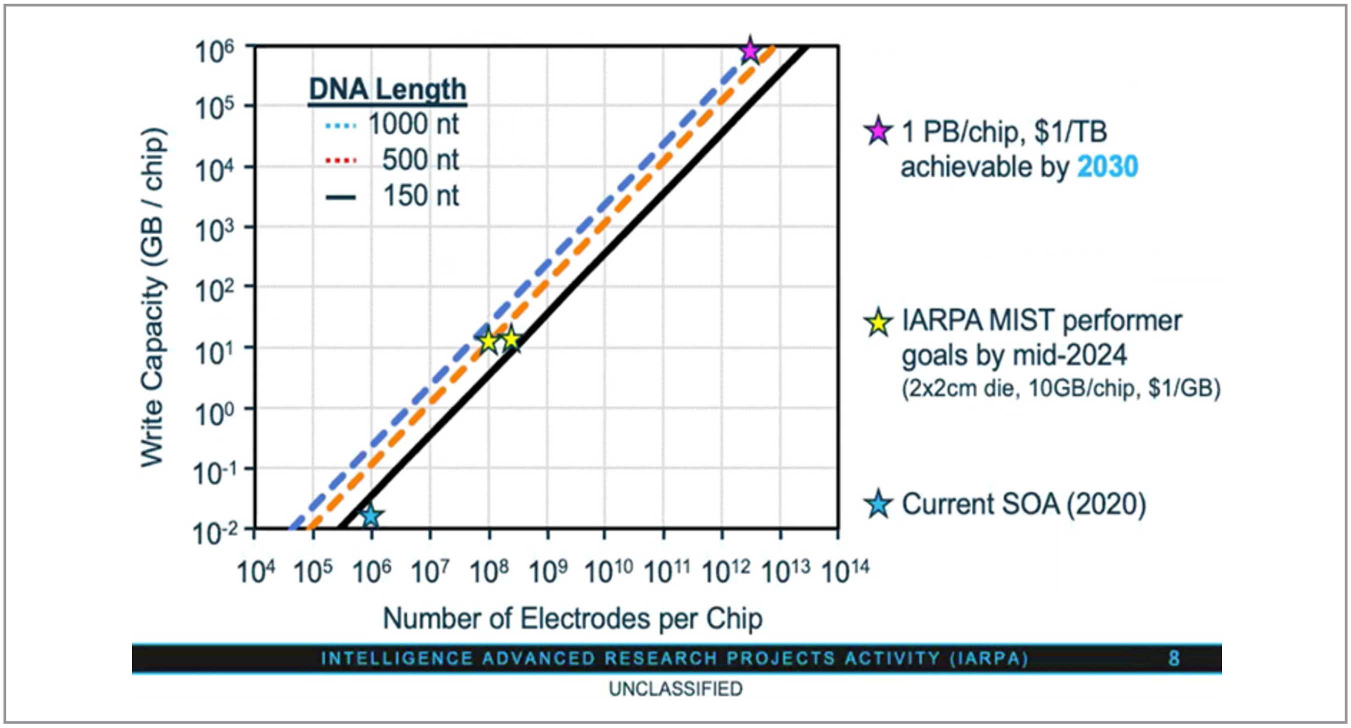


**Figure 8 - Scaling Potential for DNA Synthesis Chips (IARPA)**

With many companies and organizations working on DNA synthesis for DNA data storage applications, as well as continued industry investment in health sciences applications, there is good reason to expect that synthesis prices will continue their downward trends towards levels which can make DNA storage commercially viable.

In addition to the cost trends, DNA synthesis for data storage has an interesting property related to legacy storage. For legacy storage, the cost of the first or any subsequent copy of a dataset is the same as the cost of writing the original copy; a cost of media capacity in each case. In contrast, for DNA data storage, creating the first copy of a dataset has a cost associated with the synthesis, but the cost of creating subsequent copies is essentially zero due to the properties of tools like PCR, in which replicas are a natural outgrowth of the process. This “free replica” attribute of DNA-based storage coincides usefully with the trend of today’s hyperscale storage systems, which increasingly retain multiple copies of data.

## 6.2 Sequencing

When we talk about how much digital data can be read during DNA sequencing, what does this really mean?

We can get a perspective on this by considering a common use case, sequencing the human genome. The National Human Genome Research Institute (NHGRI) estimates that the human genome sequencing cost declined from $100 million in 2001 to $1000 in 2020 (Figure 9). Also per NHGRI, a whole human genome contains about 6 billion DNA bases, so if we assume that we can encode one bit of digital data per DNA base, a human genome can encode about 6Gbits (0.75 gigabytes) of data, which equates to about $1300/gigabyte at the $1000/human genome price point.

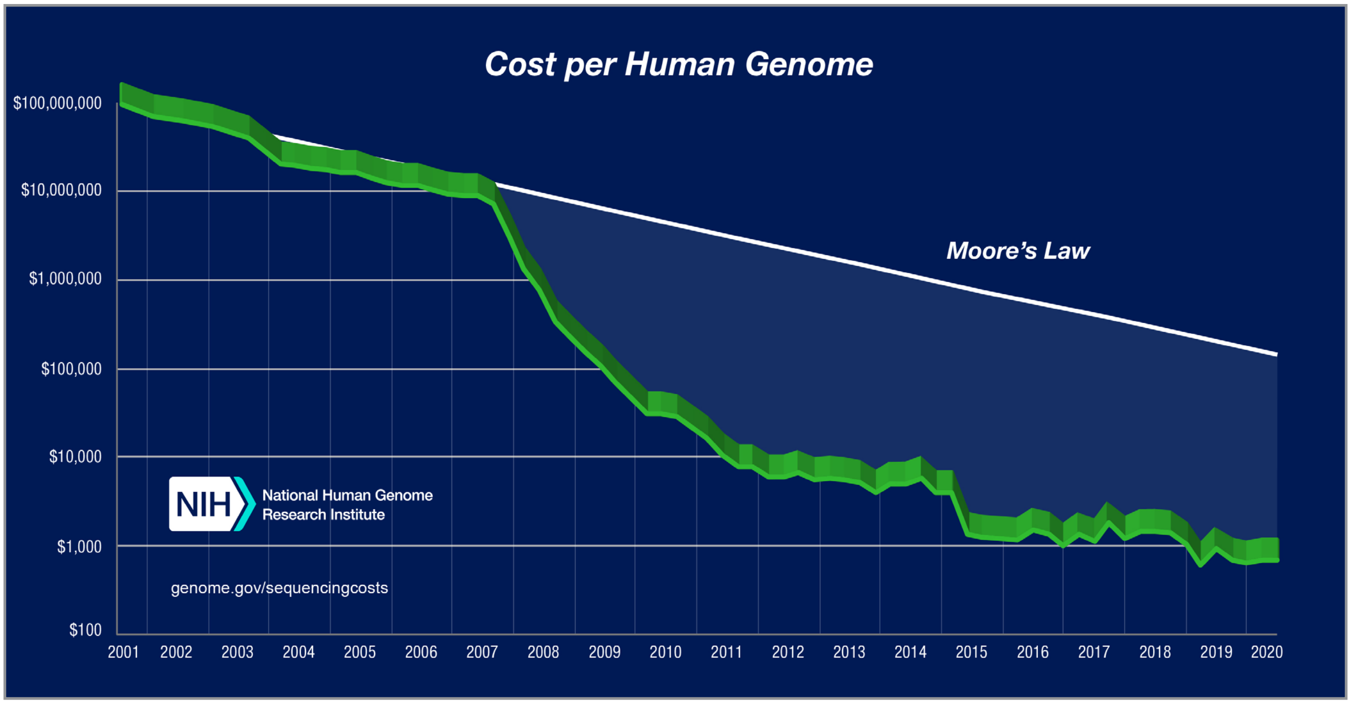


**Figure 9 – Cost per Human Genome – Nat'l Human Genome Research Institute**

The $1300/gigabyte cost derived from the NHGRI example is consistent with the listed throughput of today's high end commercial DNA sequencing platforms which, when converted to digital data carrying capacity, ranges from $800/gigabyte to $1500/gigabyte, again assuming 1 bit per DNA base (i.e., 8 gigabases = 8 gigabits = 1 gigabyte).[1]

Further, Illumina and others have stated that sequencing the human genome for only $100 will become possible on the highest throughput sequencing platforms within a few years. [13] This represents another 10x reduction, which equates to roughly $130/gigabyte. Given the IARPA synthesis target of $1/terabyte by 2030, and the fact that with today's mainstream storage technology there is no difference in read and write cost, it seems reasonable to assume that the DNA data storage ecosystem will strive for a similar cost structure for sequencing.

[1] **Assumptions:** 1) sequencing cost only; 2) 1 bit data per base (assuming 50% for error correcting, indexing and other overheads); 3) 10x read coverage (read 10 times to ensure no fragments missed) required for DNA storage use case.

## 6.3 Storage and maintenance

When examining archival storage costs, one must look not simply at reading and writing costs, but at the total cost over time. **Figure 10** shows the cost of **writing** and **storing** data (but not retrieving it, which is use-case dependent). It compares projected costs for cloud, tape, and DNA data storage over time for a petabyte (Pbyte) of data, assuming various price points for DNA, selected for comparison purposes only. The analysis assumes that, for DNA, there is no periodic data migration and only nominal fixity checks and energy required for storage.

We observe in the figure that, over time, as DNA writing costs decline, legacy storage and maintenance costs begin to greatly exceed DNA-based storage costs.

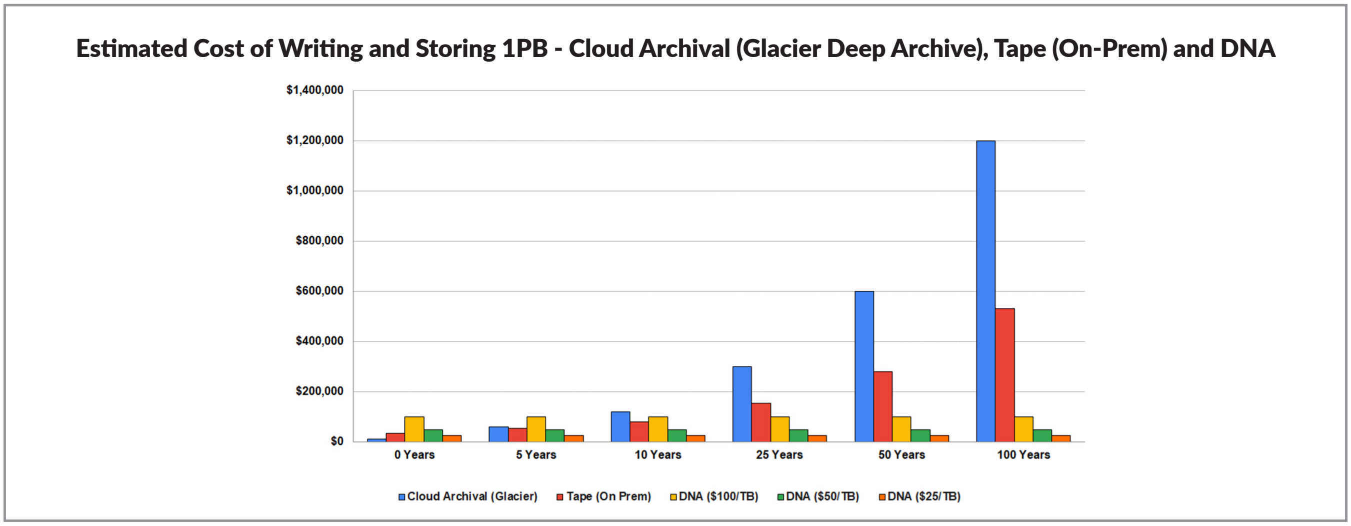


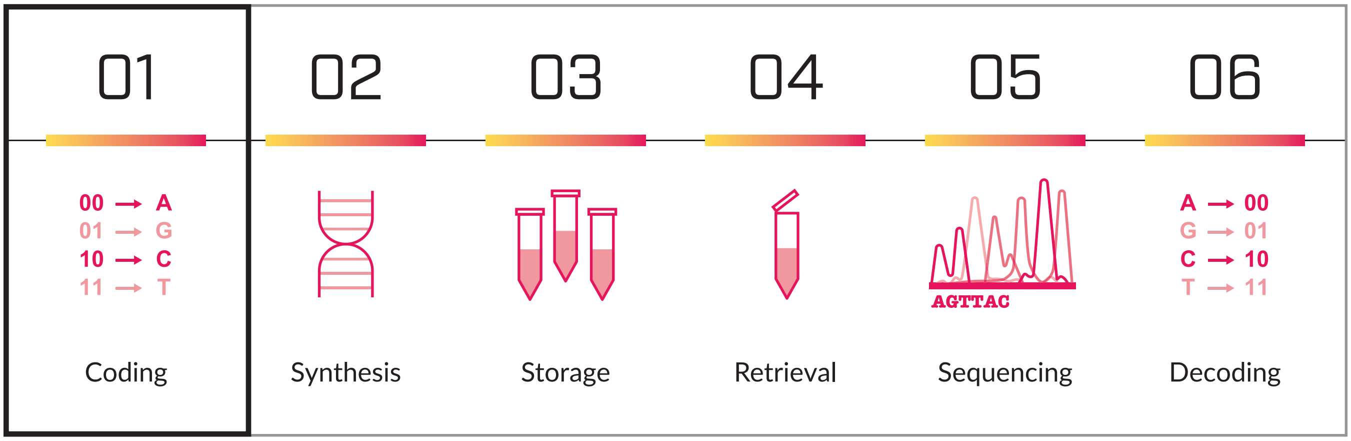
**Figure 10 - Estimated Total Cost of Writing and Storing - Legacy vs. DNA**

- 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

## 6.4 Summary – Economics of DNA data storage

Though the field of DNA data storage is nascent, the fundamental cost trends of manufactured DNA synthesis and sequencing for data storage continue to fall. For synthesis, the IARPA MIST project sets a goal of $1000/terabyte by 2024 and $1/terabyte by 2030 **(Figure 8).** Sequencing costs have already dropped dramatically and are approaching the $100/human genome milestone [14], which equates to a data storage cost of about $130/gigabyte, in the next few years. The DNA data storage ecosystem will strive to reach a cost structure for sequencing that is similar to that projected for synthesis. Reductions in cost for both synthesis and sequencing are critical for the adoption of DNA data storage and, assuming we achieve such reductions, then this, combined with the vastly lower physical storage and maintenance costs of DNA as a storage medium compared to today's storage technologies, will make a compelling argument for including DNA as a new layer in the archival data storage hierarchy.

# 7 THE CURRENT STATE OF DNA ENCODING



Encoding DNA for data storage is the process of converting original digital 1's and 0's into a sequence of bases (ACGT) that comprise DNA molecules. The specific encoding algorithms are technically intertwined with the underlying chemistry of synthesis and sequencing methods, and so the encoding method affects, and is affected by, the overall process complexity, scalability, data density, data reliability, and thus the cost, of any proposed DNA data storage system. For example, we could assign the value 00 to "A", 01 to "C", 10 to "G" and 11 to "T". With this encoding scheme, the digital string 0111011000 would be encoded by the bases ATCGA, and synthesized using a base-by-base synthesis method (see Section 8.1). Another encoding scheme is to create a library of small (e.g., 20-30 bases) oligonucleotides, which represent "letters" in an "alphabet" (or "symbols" in a "font"). In this method, the original digital 1's and 0's are encoded to the letters in the alphabet. This approach is compatible with the ligation method of synthesis (see Section 8.2).

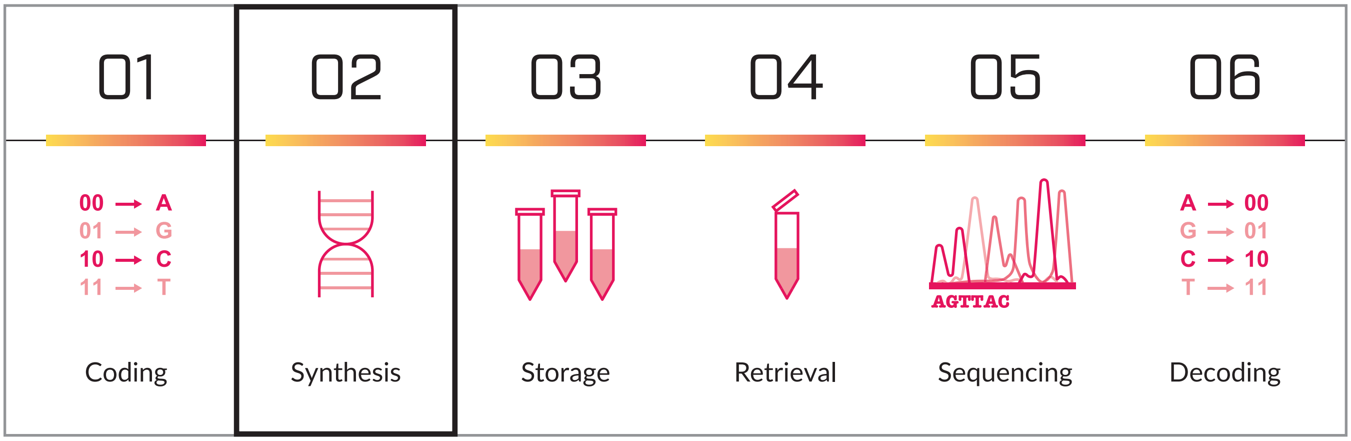
In the context of DNA data storage, the process of **synthesizing to storing to sequencing** can be considered in some ways analogous to "transmission" of digital data over an electrical interface. In the electrical case, 1's and 0's are converted to analog wave forms, at various amplitudes and frequencies, at the transmitter, and the wave forms are converted back to 1's and 0's, at the receiver. Error correction code bits (ECC) are added to the digital bit stream before transmission, and stripped out again once converted back to digital on the far side of the receiver, to detect/correct transmission errors. Bits are altered (scrambling patterns) at the receiver to avoid the transmission of certain patterns of 1's and 0's on the wire, which can create electrical interference during transmission, and thus transmission errors. In general, the way that bits are encoded for transmission on an electrical interface is optimized for the electrical properties of that particular interface (parallel, serial, transmission speed, etc.).

When encoding for "transmission" over DNA, the way the 1's and 0's are mapped to DNA bases prior to synthesis, and how DNA bases are mapped back to 1's and 0's during sequencing, is roughly analogous to the digital to analog to digital conversion during electrical transmission. ECC bits and scrambling pattern bits are added to the data stream before synthesis (transmitter), and removed during sequencing (receiver), to detect/correct and minimize errors, such as the mistranslation of bits to/from DNA bases, or the loss of DNA strands.

Another important aspect of DNA encoding for DNA data storage is **segmentation and addressing.** Since there are practical limits on the length of synthetic DNA strands, all encoding schemes today encode address information to enable the segmentation of long digital bit streams into DNA sub-segments, which are subsequently re-assembled during sequencing and decoding. Various addressing schemes are used to enable segmentation, for example the use of **fields** (sequences of bases which encode addresses are embedded at specific locations in the DNA sub-segments), **implicit mapping** (the encoding of bits such that the addressing for a sub-segment is embedded in the encoded data in each DNA sub-segment), or **external tags** (chemical or other types of physical identifiers attached to the DNA molecules). Other methods are being developed.

Encoding for DNA is a very active field, as dynamic a part of the evolution of DNA data storage as the underlying synthesis and sequencing technologies themselves.

# 8 THE CURRENT STATE OF DNA SYNTHESIS



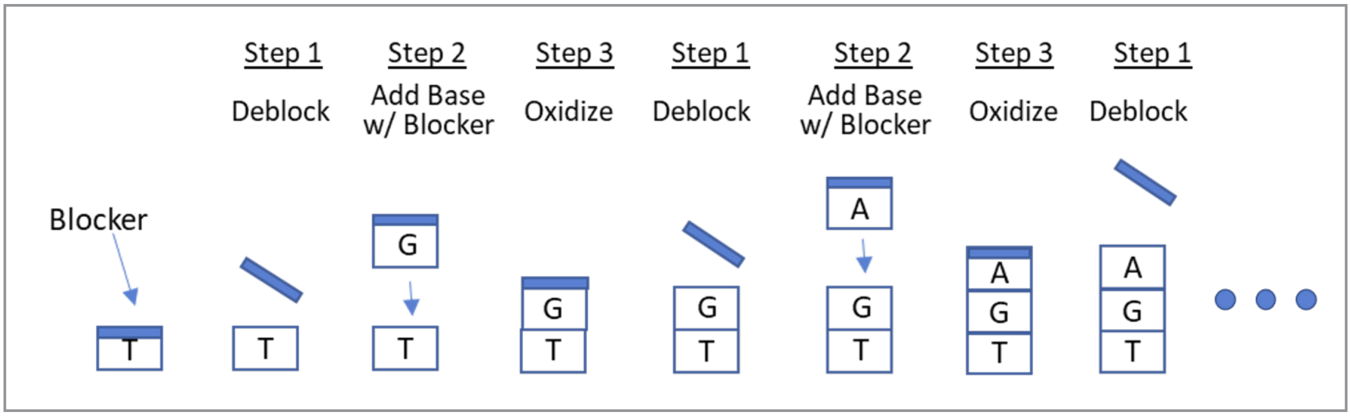
DNA synthesis is the chemical fabrication of sequences of nucleic acids. Most biological research and bioengineering involve synthetic DNA, which can include short DNA sequences up to much longer sequences, assembled from shorter ones.

Today, when considered in the context of digital data storage, all methods of DNA synthesis are orders of magnitude slower in total throughput than any existing storage technology; they will need to be massively parallelized to enable DNA data storage to become cost competitive with traditional data storage technologies. The methods of synthesis described below are undergoing continuous innovation and attempts at performance scaling.

## 8.1 Base-by-base synthesis – chemical & enzymatic

The mainstream technique for synthesis today is chemical synthesis. An emerging technique, called enzymatic synthesis, is maturing in the last few years and will start to enter the market soon. Both techniques build molecules base-by-base, using the process shown roughly below.

The process starts with a base, bound with a "blocker," i.e., a chemical element that can be attached to a base at the end of the DNA strand being synthesized and thus protect it during the process.  A loop then begins: (1) the strand is de-blocked, (2) a new base is added with a blocker on top of the strand, and (3) the new base is bound to the strand.  Then the process is repeated. In general, with the processes used today, the longest strand of synthetic DNA that can be constructed using base-by-base synthesis, while maintaining acceptable error rates, is 200-300 bases.



### 8.1.1 Chemical synthesis (phosphoramidite)

At present, all commercial synthetic DNA is custom-built using the phosphoramidite synthesis method, developed by the biochemist Marvin H. Caruthers. In this approach, oligonucleotides are synthesized from building blocks that replicate natural bases. The process has been automated since the late 1980s and is used to form desired genetic sequences for applications in medicine and molecular biology as well as for data storage. This method is currently the most robust, best tested and highest quality way to construct synthetic DNA.

Price is the main challenge with the phosphoramidite approach. Though declining every year, it remains far higher than that for mainstream storage applications. Another challenge is the speed of writing DNA using this method. Most technologies still rely on a sequential one-by-one addition of nucleotides to the growing strand, and the speed of liquid handling in microfluidic devices limits production speed. New methods and technologies, already being tested, show great promise for significantly increasing speed and reducing cost through parallelism.

### 8.1.2 Enzymatic synthesis

Starting in the mid-2010s, several research groups began work on an alternative to phosphoramidite chemistry that uses enzymes to synthesize DNA, in a cyclic process similar to chemical synthesis.

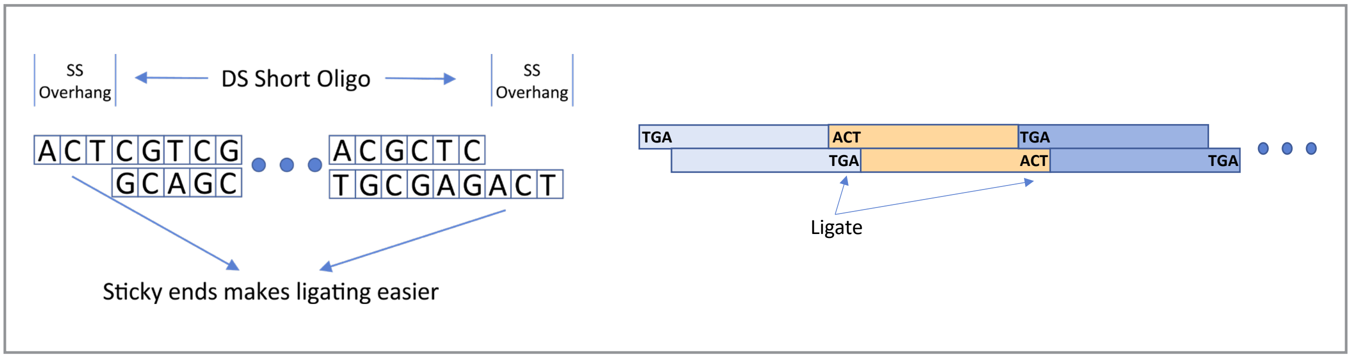
Enzymatic synthesis technology promises the use of only aqueous reagents, which produce fewer waste byproducts and are thus more sustainable. It can also speed up synthesis to achieve higher throughput and increase polymer length, and thereby data density, to reduce storage costs.

Enzymatic synthesis has not yet reached the commercial market; however, it is progressing rapidly, and proofs of concept for the enzymatic synthesis of 150 base long oligonucleotides, with an error rate low enough for data storage, were achieved in 2018. The first products are planned to ship before the end of 2021.

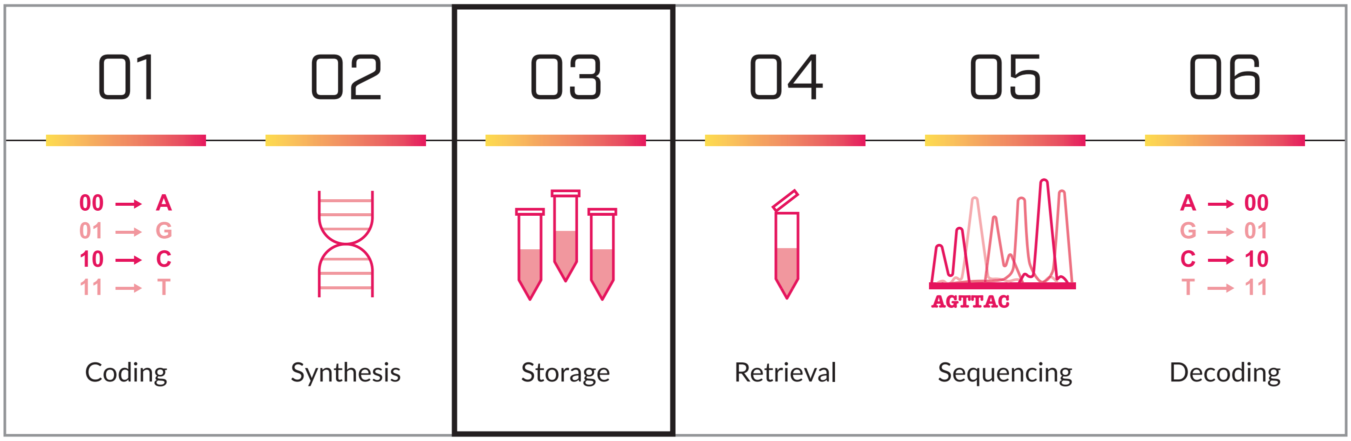
## 8.2 Synthesis by ligation

Another synthesis technique is being used to synthesize long strands of DNA for storage-based applications. The basic notion is to create a library of predefined short oligonucleotides using base-by-base synthesis techniques such as those described in the previous section, and then "stitching" the short oligonucleotides together (i.e., **ligation**) to produce long oligonucleotides (tens to hundreds of times longer than those constructed using base-by-base methods) with acceptable error rates. A longer oligonucleotide building block means it may be possible, depending on encoding methods, to amortize the cost for error correction, sub-segment re-assembly, etc., over a larger data payload or, in other words, enable lower protocol overhead.

One example of creating long oligonucleotides from short oligonucleotides using ligation is shown below. We start with a set of double-stranded short oligonucleotides that have single-stranded overhangs (left), where each short oligonucleotide represents an encoded data unit or symbol. The ligation step (right) combines these short oligonucleotides into longer ones. The basic idea is that the complementary DNA bases in the overhang segments pair naturally; an additional enzymatic process then creates permanent bonds between them, enabling the construction of the longer sequences.



# 9 PRESERVING DNA FOR DATA STORAGE



Once DNA is synthesized and encoded with digital data, the actual physical storage of the medium involves several factors.

Any DNA protection technology must use some packaging materials. Thus, practical aspects --- e.g., container cost, amount of data per container, time, plus cost to package/un-package --- must be considered in the full context of DNA data storage applications.  Also very important is automation for physical storage and retrieval, including collecting the outputs of synthesis, preparation of the DNA for physical storage, recovery of the material to service read requests, and its preparation for the reading process. We do not cover this in detail here but instead focus on media decay and general protection strategies.

## 9.1 Mechanisms of DNA decay

DNA is known to be degraded by interactions with some small organic molecules (xenobiotics), UV irradiation, water, enzymes, microorganisms, oxygen, ozone and other atmospheric pollutants.

Of the various risks, water is by far the primary degradation factor, in particular because it is necessary for the action of oxidants or enzymes. While the half-life of DNA embedded in buried ancient bone fossils has been estimated to be 512 years at 25 °C and, given optimal protection conditions, can reach more than 100,000 years, the half-life of DNA exposed to moisture degrades significantly.  Thus, storage strategies for DNA must carefully address moisture-related problems.

## 9.2 DNA media protection technologies

Currently, the two general classes of protection strategies are molecular-scale and macroscopic protection. A DNA data storage system may combine both.

1. **Molecular approach.** Individual DNA molecules are embedded into a matrix material designed to prevent diffusion of water and oxygen to the individual DNA molecules (aka **chemical encapsulation**). Due to the relatively high diffusion rates of water in polymers, organic molecules and water soluble salts, most suitable matrices consist of inorganic materials such as glasses.

2. **Macroscopic approach.** A dry DNA sample is stored in the presence of an inert gas in a hermetic container, e.g., a metallic capsule (aka **physical encapsulation**). As long as the integrity of the container can be guaranteed and oxygen and water diffusion can be controlled, chemical interactions of the data carrying DNA molecules can be avoided.

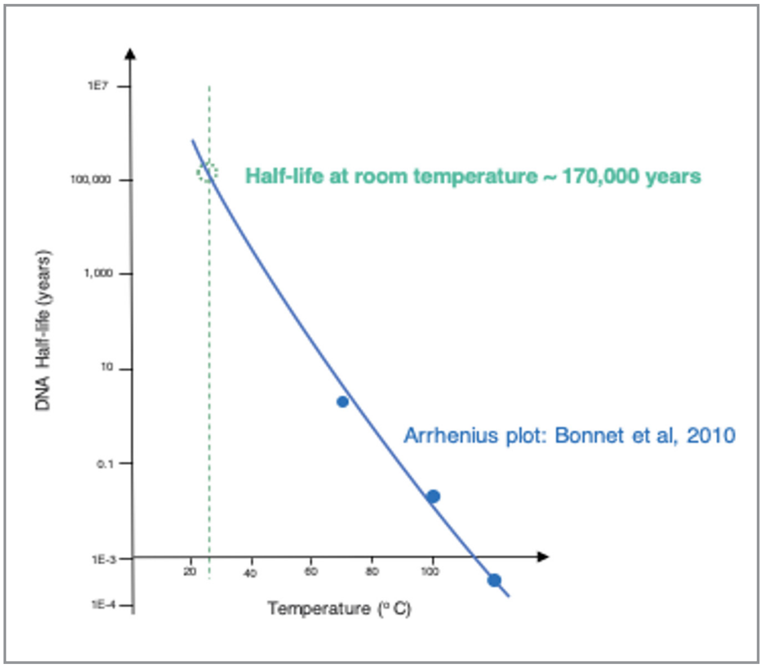
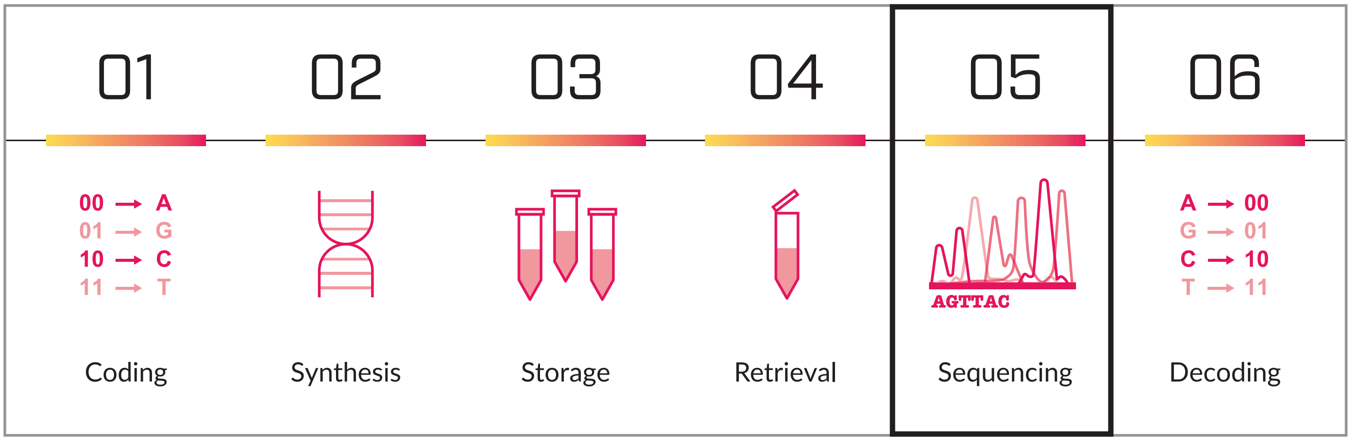


**Figure 11 - Half-life of DNA without Moisture Courtesy: Imagene**

**Figure 11** shows the effects of temperature on the storage of DNA media provided that moisture is controlled. We see from the curve that the temperature requirements of DNA retention are not very rigorous, with the half-life being extremely long even at room temperature.

# 10 THE CURRENT STATE OF DNA SEQUENCING



DNA sequencing is a general term describing a variety of techniques that detect and report the order of the bases in a strand of DNA. There is a long history of DNA sequencing technologies. The 1st generation of breakthrough DNA sequencing technology, Sanger "chain termination" sequencing [15], which was used to sequence the human genome for the first time, enabled many biological applications. Since the mid 1990's, a new sequencing technology, called **Next Generation Sequencing (NGS)**, has been developed and has broadened the usage of DNA sequencing. NGS uses massive parallelization to achieve breakthroughs in throughput, scalability, and speed. While NGS itself represents a diverse set of methods, there are two broad categories of NGS sequencing in commercial use today: **sequencing-by-synthesis (SBS)** and **nanopore sequencing**; some examples of these solutions are presented below.

## 10.1 Sequencing by synthesis (SBS)

Sequencing-by-synthesis gets its name because the approaches all rely on taking a single-stranded template (sample) piece of DNA and then leveraging the fact that the bases in a DNA molecule (ACGT) pair to one another in a complementary way (A-to-T and C-to-G) to synthesize the complementary strand to the template. Conceptually, SBS methods all do the following:

1. Begin with an original double-stranded DNA (dsDNA) sample and break it into segments of single-stranded DNA (ssDNA) of a desired length; this step is generally referred to as **library preparation**.

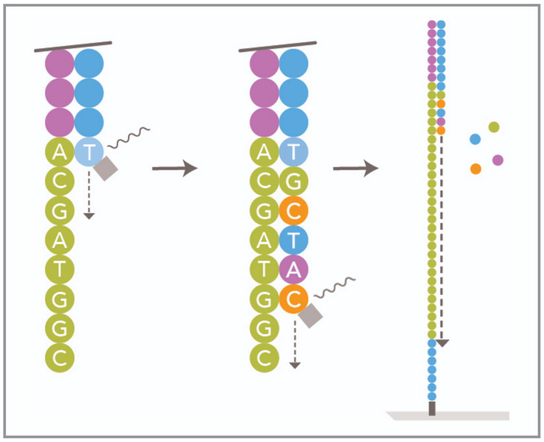


**Figure 12 - SBS Sequencing**
**Courtesy: Illumina, Inc.**

2. Take the prepared ssDNA library and add it to a **flow cell** in the sensing instrument. Then, for each ssDNA segment in the library, create a dsDNA segment (hence **sequencing-by-synthesis**) using a **polymerase** to "incorporate" the new strand of complementary bases into the original ssDNA segment.

3. Each incorporation event generates a signal (optical, electrical, release of an ion, etc.) that can be detected and, due the complementarity of how bases bond in DNA, used to interpret the composition of the original ssDNA segment from the library.

Illumina (then Solexa) originally pioneered SBS in 2006; since then, various methods have been deployed, with the main ones noted below:

- **Illumina SBS** is based on the imaging of fluorescence-tagged nucleotides. DNA library is added to a flow cell and then amplified into clusters, after which the synthesis step begins. Four types of fluorescence-tagged, reversible terminator bases are added, and non-incorporated nucleotides are washed away. A camera takes images of the fluorescently labeled nucleotides. Then, the dye, along with the terminal 3' blocker, is chemically removed from the DNA, allowing the next cycle to begin.

- **Pacific Biosciences Single Molecule Real time (SMRT) Sequencing Technology** is an SBS technology based on real-time imaging of fluorescently labeled bases as the polymerase synthesizes DNA along ssDNA template molecules. The technology generates long, continuous reads, with an average length of 15kb (kilo-bases) at single molecule resolution. [16]

- **Ion Torrent Semiconductor Sequencing Technology**, by ThermoFisher Scientific, is another SBS technology, which directly translates information encoded in DNA bases into digital information (0, 1) on a semiconductor chip. No modified nucleotides or optics are used. In this method, when a nucleotide is incorporated into the strand of original DNA being sequenced by a polymerase, a hydrogen ion is released as a byproduct, which changes the pH of the solution in which the incorporations are happening. Ion Torrent sequencers directly measure these pH changes and thus identify the base to rapidly produce the sequencing results. [17]

## 10.2 Nanopore sequencing

**Nanopore sequencing** uses a different underlying mechanism than SBS. In nanopore sequencing, a strand of DNA is passed through a pore in some kind of membrane surrounded by an electrolyte solution. With an electrical bias applied across the membrane, strands of DNA move through the pore, and a detection event (voltage, chemical, etc.) is registered, enabling the direct detection of the bases in the original DNA strand. The term **nanopore**

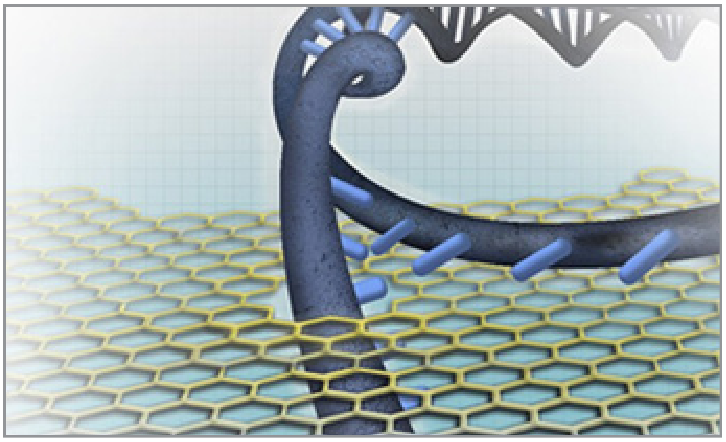


**Figure 13 - Nanopore Sequencing**
**Source: Jonathan Bailey, NHGRI, genome.gov**

comes from the requirement that the pore needs an opening of a few nanometers wide to permit sensing with base-level resolution.  Nanopore DNA sequencing detections can be streamed in real time and thus can produce immediate access to results. It is also possible to sense other molecules, not just DNA, using nanopore techniques.

The most widely deployed nanopore DNA sequencing solution today is from Oxford Nanopore Technology (ONT), which uses a biological pore embedded in a lipid membrane to make (electronic) sensing more precise. Semiconductor-based nanopore solutions are also being developed.

# 11 SUMMARY

There are phases in the lifecycle of technologies in which huge transformations lurk.  We believe we are on the cusp of one such transformation today regarding archival storage and DNA.

Information is being digitized on a massive scale, by servers in datacenters, by mobile devices, and by networks of sensors everywhere around us.  Artificial intelligence techniques and ubiquitous processing power are making it possible to mine this massive ocean of data; however, integral to harnessing this data as knowledge is the ability to store it for long periods of time.

Legacy storage solutions have scaled extensively over the years, but the areal density of magnetic media (HDD and tape), which enables today's mainstream archival storage solutions, is slowing, and the size of libraries is becoming unwieldy. In short, data growth is outpacing the scalability of today's storage solutions. The industry needs a new storage medium that is more dense, durable, sustainable, and cost effective in order to cope with the expected future growth of archival data.

DNA, nature's data storage medium, enters this picture at a time when synthesis and sequencing technologies for advanced medical and scientific applications are enabling the manipulation of synthetic DNA in ways previously unimagined. There are credible predictions that, by 2030, DNA synthesis could reach a cost of $1/terabyte and that DNA sequencing may reach similar levels. The scale of DNA data storage is unprecedented; the volume of space inside an LTO tape cartridge is estimated to hold 100,000 times the number of DNA-bits as an LTO-9 tape in that same cartridge. The durability of DNA and uniformity of the DNA molecular structure are ideally suited to long-term archival storage.  Finally, DNA is an inherently environmentally friendly medium in terms of power, space, and sustainability, which will place significantly lower burdens than legacy storage technologies on our fragile ecosystem.

The intersection of the "data overwhelm" (Section 1) and our ability to manipulate synthetic DNA offers a vision of a new layer in the storage hierarchy that could radically change the scale of what we store and how long we store it. Preserving our digital legacy in turn opens possibilities to extract, and even create or discover, new knowledge.

The DNA Data Storage Alliance is committed to helping preserve the knowledge, culture, art, and other precious resources that our past generations have created for the future enjoyment and enrichment of all. Through innovation and collaboration, we will continue advocating for and advancing DNA data storage as a new tier in the archival storage domain.  If you wish to get involved or just get more information, please visit our web site,  www.dnastoragealliance.org,  or contact us at  info@dnastoragealliance.org.

# 12 GLOSSARY

The glossary below, which defines technical terms used in this paper, has been excerpted from the DNA Storage Glossary[18] and modified as needed to fit the terms as used in this document.

## Molecular biology terminology

| | |
|---|---|
| **(sequencing) adapters** | DNA sequences designed to facilitate DNA sequencing, e.g. by allowing interaction with the flow cell. Located on the 3' and 5' ends of DNA samples prepared for sequencing |
| **amplification** | Replication or copying of DNA using **in vitro** methods such as PCR. |
| **base** | A nitrogen-containing compound that forms a crucial component of a nucleotide. There are four types of nitrogen bases: adenine (A), thymine (T), guanine (G) and cytosine (C). Commonly (but incorrectly) used interchangeably with nucleotide. |
| **bp (base pair)** | A pair of hydrogen-bonded nitrogen bases within a nucleotide. Also used as a unit of measure of the length of dsDNA.  Typically, the 'p' is omitted. See below for examples:<br>kb (or kbp): kilo base pairs (1,000 base pairs)<br>Mb (or Mbp): mega base pairs (1,000,000 base pairs)<br>Gb (or Gbp): giga base pairs (1,000,000,000 base pairs)<br>Use of lowercase 'b' and consideration of context is important to avoid confusion with **bytes** and **bits**. |
| **coverage (or depth of coverage)** | Refers to the number of times a specific nucleotide is sequenced. 15x coverage indicates that each nucleotide was read on average 15 times during a sequencing run (or it was present in 15 reads). A '30x genome' is one that has been sequenced so that each nucleotide of the genome was read 30 times on average. |
| **DNA** | Deoxyribonucleic acid (DNA) is a biological molecule composed of repeating units, called nucleotides, predominantly arranged within a double helix formation. The order of nucleotides which can contain one of four possible bases – adenine (A), thymine (T), guanine (G) and cytosine (C) – forms the genetic code present within nearly all lifeforms. |

| | |
|---|---|
| **DNA polymerase** | An enzyme that replicates DNA using a template DNA strand in a 5′ to 3′ direction according to complementary base pairing. Some polymerases are classified as high-fidelity DNA polymerases due to a proof-reading ability. |
| **DNA sequence** | The order of nucleotides in DNA. |
| **error** | Processes such as DNA synthesis, sequencing and DNA polymerases have an associated error rate (or fidelity for DNA polymerase). Errors can range from:<br>• insertion or deletion: additional nucleotides are added or deleted. Commonly abbreviated as indels.<br>• substitutions: replacement of one nucleotide with another. |
| **flow cell** | Typically a small microfluidic device onto which DNA samples are applied prior to loading in a sequencing device. The flow cell is where sequencing chemistry occurs.<br>(Entry within a DNA sequencing context) |
| **library** | In a next generation sequencing (NGS) context, DNA samples prepared and ready for DNA sequencing. Typically prepared by the following simplified workflow of DNA fragmentation, adapter ligation and quantification. |
| **next generation sequencing (NGS)** | A high-throughput approach to DNA sequencing allowing numerous DNA fragments to be sequenced (or read) far more rapidly and cheaply than was possible with previous technology such as Sanger sequencing. |
| **oligonucleotide (oligo)** | Refers to the repeating unit of a DNA molecule. Each nucleotide consists of a phosphate group, a sugar group and a nitrogen base.  Short single strand (ss) DNA sequences of a few to hundreds of nucleotides in length. Primers used in PCR are oligonucleotides. |
| **polymerase chain reaction (PCR)** | An in vitro technique that enables DNA to be copied. This is achieved through repeated thermal cycling and the use of primers, DNA polymerases, and nucleotides. A single thermal cycle consists of three steps: denaturation, annealing and extension. |
| **primer** | A short ssDNA sequence typically 18 to 22 nucleotides long. Designed to anneal to specific target DNA sequences. Serves as a starting point for DNA synthesis by DNA polymerases. |
| **read** | A DNA fragment that has been, or is being, sequenced. |

| | |
|---|---|
| **sequencing** | Determining the order of nucleotides in a DNA molecule. Essentially, the reading of the DNA sequence. |

## Computer technology terminology

| | |
|---|---|
| **Areal density** | The number of bits a medium can store per unit area |
| **binary** | A language that forms the basis of all computer information, in which there can only be two values: either 0 or 1. Each binary digit is known as a bit. |
| **bit** | A single binary digit: either 0 or 1. Bits can be combined to represent different values e.g. two bits can represent four values: 00, 01, 10, 11. The values these bits represent is dependent upon the coding notion used. |
| **byte** | A basic unit of memory that is usually 8 binary digits, or bits, in length. |
| **string** | A sequence of characters or symbols. |

## Information theory terminology

| | |
|---|---|
| **channel** | A communications or storage channel is any medium over which information is transmitted and received, or where information is stored and retrieved. Examples are wireless transmission links, twisted copper wires in the telephone infrastructure, magnetic hard disks, or DNA when used for storage. |
| **decoding** | Process of translating coded data back into information; the reverse of encoding. |
| **encoding** | Process of converting information such as text characters into **codewords** based on the specifications of a specific **code**; the reverse of **decoding**. |

| | |
|---|---|
| **error** | An incorrectly read symbol, for example a symbol transmitted as 1 and flipped to a 0 by the channel, or a DNA sequenced where the 14th symbol was synthesized as a G but sequenced as an A. Note that errors in information theory are different from insertions, deletions or erasures. They correspond to a substitution in biological sequence processing. |
| **rate** | The information content in every symbol transmitted over a channel. Supposing a perfectly compressed source, when using a block code (linear or not), the rate is the length of the information sequence divided by the length of the codeword assigned to it. |

## DNA data storage controlled vocabulary

| | |
|---|---|
| **bit** | Note word has differing meanings in Computer Science and Information Theory. Use of the full word 'bit', as recommended by ISO/IEC 80000-13:2008 reduces possible confusion with **bases (bp)**. See below for examples:<br>kbit: kilobit (1,000 bits)<br>Mbit: megabit (1,000,000 bits)<br>Gbit: gigabit (1,000,000,000 bits) |
| **byte** | A basic unit of data storage. Use uppercase 'B', as recommended by ISO/IEC 80000-13:2008, combined with SI prefixes of measure to refer to varying numbers of bytes. See below for examples:<br>kB: kilobyte (1000 bytes)<br>MB: megabyte (1,000,000 bytes)<br>GB: gigabyte (1,000,000,000 bytes)<br>TB: terabyte (1,000,000,000,000 bytes)<br>PB: petabyte (1,000,000,000,000,000 bytes)<br>EB: exabyte (1,000,000,000,000,000,000 bytes)<br>ZB: zettabyte (1,000,000,000,000,000,000,000 bytes)<br>Use of uppercase 'B' and consideration of context is important to avoid confusion with **bp** abbreviations and **bits**. |

| | |
|---|---|
| **SSD** | Solid State Drive. |
| **HDD** | Hard Disk Drive. |
| **TCO** | A measure of the total cost of owning something (equipment, factory, etc.), which includes not only the initial purchase price but operating expenses over the life of the asset. |
| **WORM** | Write Once Read Many. Read-only data for which the intention is to access it frequently. |
| **WORN** | Write Once Read Never. Read-only data for which the intention is to never access it unless there is an unexpected event. Backups are usually WORN as we don't need to access them on a daily/monthly basis. |
| **WORS** | Write Once Read Seldom. Read-only data for which the intention is to access it very infrequently (once every few years) unless there is an unexpected event. Backups or archival copies of cold data are usually WORS as we don't need to access them on a daily/monthly basis. |

[1] IDC, Worldwide Global DataSphere Forecast, 2021–2025: The World Keeps Creating More Data — Now, What Do We Do with It All?, IDC Doc #US46410421

[2] https://datareportal.com/reports/digital-2021-global-overview-report

[3] IDC,  Worldwide Global StorageSphere Forecast, 2021–2025: To Save or Not to Save Data, That Is the Question, IDC Doc #US47509621, March 2021

[4] Gartner Market Trends: Evolving Enterprise Data Requirements —How Much Is Not Enough? (ID G00727554)

[5] Worldwide Hard Disk Drive Forecast Update, 2019–2023 (IDC #US45667319, December 2019).

[6] http://www.insic.org/wp-content/uploads/2019/07/INSIC-Technology-Roadmap-2019.pdf

[7] https://www.insic.org/areal-density-chart/

[8] https://www.businesswire.com/news/home/20180319005742/en/Nimbus-Data-Launches-the-World%E2%80%99s-Largest-Solid-State-Drive-%E2%80%93-100-Terabytes-%E2%80%93-to-Power-Data-Driven-Innovation

[9] https://davidmytton.blog/how-much-energy-do-data-centers-use/

[10] https://horison.com/publications/tiered-storage-2020

[11] https://datastorage-na.fujifilm.com/tco-tool/

[12] Bonnet et al., Nucleic Acids Research, 2010 (https://doi.org/10.1093/nar/gkp1060)

[13] Next Generation Sequencing (NGS) market Market report Size, Growth and Trends Market Report.  DeciBio, Dec. 2020

[14] https://www.bloomberg.com/news/articles/2019-02-27/a-100-genome-within-reach-illumina-ceo-asks-if-world-is-ready

[15] Sanger, Nicklen, Coulson, DNA sequencing with chain-terminating inhibitors (https://doi.org/10.1073/pnas.74.12.5463);

[16] Roberts et al, 2013 (https://doi.org/10.1186/gb-2013-14-7-405)

[17] https://www.thermofisher.com/us/en/home/life-science/sequencing/next-generation-sequencing/ion-torrent-next-generation-sequencing-technology.html

[18] Hesketh EE, Sayir J and Goldman N. Improving communication for interdisciplinary teams working on storage of digital information in DNA [version 1; peer review: 2 approved]. F1000Research 2018, 7:39 (https://doi.org/10.12688/f1000research.13482.1)

DNA DATA STORAGE ALLIANCE

A PUBLICATION OF THE DNA DATA STORAGE ALLIANCE

www.dnastoragealliance.org