PERSPECTIVE

The Future of DNA Storage in Revolutionizing Biological Data Management

Rongrong Chen[1], **Xue Li**[1,*] and **Ben Cao**[2]

[1] The Key Laboratory of Advanced Design and Intelligent Computing, Ministry of Education, School of Software Engineering, Dalian University, Dalian 116622, China
[2] School of Computer Science and Technology, Dalian University of Technology, Dalian 116024, China

Abstract

Compared with traditional storage media, biological data storage has advanced more rapidly in capacity, diversity, and lifespan, and it also enables continuous data retention. The DNA molecule, as Nature's own archival medium, provides unparalleled density, longevity, and passive durability, making it a compelling foundation for the next generation of "cold" and "deeply cold" archives. Since 2019, progress across the stack—coding for insertion–deletion–missing channels, large-scale random access, enzyme writing, nanopore-native retrieval, and chemically robust expansion—has transformed DNA storage from provocative demonstrations into mature technologies with early end-to-end prototypes. From this perspective, we believe biological data management (BDM) is the first area where DNA storage can have practical impact, including raw sequencing archives, microscopic images, clinical omics data, and compliance-driven retention of de-identified records. We synthesize the latest technology levels, highlight what is actually working in today's labs, distinguish challenges from bottlenecks (particularly write costs/latencies and standardization), and propose a 2025–2030 roadmap with specific milestones in coding, writing, access, and media preservation. Finally, this paper proposes guidelines for integrating DNA archives into biological institutes, biobanks, and hospital systems as a supplementary layer to tape/object storage, and outlines a research agenda covering privacy, chain of custody, and sustainability.

Keywords: DNA storage, biological data management.

Submitted: 06 September 2025
Accepted: 17 September 2025
Published: 26 September 2025

Vol. 1, **No.** 2, 2025.
10.62762/JAIB.2025.924847

***Corresponding author:**
✉ Xue Li
xueleecs@gmail.com

1 Introduction

With the rapid development of high-throughput sequencing technology and omics research, the field of biology is generating massive data at an unprecedented rate. According to the cumulative data volume since the Human Genome Project, the growth rate of global biological data has exceeded Moore's Law, and it is estimated that by 2030, the amount of data in the life science field will exceed the sum of astronomical observation data and social media data [1]. How to store this data efficiently, at low cost and in the long run has become a core challenge.

Citation
Chen, R., Li, X., & Cao, B. (2025). The Future of DNA Storage in Revolutionizing Biological Data Management. *Journal of Artificial Intelligence in Bioinformatics*, 1(2), 51–57.

Traditional storage media (such as magnetic tapes, hard drives, and solid-state drives) face bottlenecks in storage density, lifespan, and energy consumption. DNA, as the natural carrier of life's genetic information, is considered an ideal material for next-generation data storage due to its extremely high information density, stability over millions of years, and ability to replicate itself [6]. The DNA storage process is illustrated in Figure 1.

In recent years, DNA storage technology has gradually moved from proof of concept to laboratory applications and is expected to become an important pillar of biological big data management in the future. Against this background, DNA storage is gradually attracting attention as an emerging data storage paradigm. DNA is considered a potential core infrastructure for next-generation biological data management for its high-density storage, long-term stability and biocompatibility [7, 12]. Recently, DNA data storage technology has become a research hotspot due to its ultra-high density ($\sim$1 EB/mm^3) and long-term stability (millions of years). The development of this field can be summarized into three aspects: encoding and writing, reading technology and random access strategy. All directions promote each other and promote the technology to be practical. A comprehensive analysis of these works is shown in Table 1, which outlines the innovative contributions, coding robustness, and demonstrated scales of current DNA storage methods.

In terms of coding optimization, researchers have gradually shifted from error correction to fault tolerance. Early studies [7] adopted a fountain code architecture that optimized storage capacity and resisted base replacement errors, but could not cope with insertions/deletions (indels). The HEDGES proposed in 2020 achieves direct correction of indel for the first time, while de Bruijn assembly [16] enhances the robustness of DNA breakage/recombination through graph structure. In addition, the Extended Alphabet [4] introduces chemically modified bases to enhance the encoded symbol set and indirectly increase the information density. The composite HEDGES-nanopore system in 2024 further combines the characteristics of sequencing equipment to achieve portable and fast decoding [13], reflecting the coordinated optimization of encoding and hardware. The rapid development of DNA synthesis technology has promoted the rapid implementation of DNA storage. Traditional chemical synthesis [2, 3] has high cost and limited throughput, while enzymatic synthesis has greatly improved the writing efficiency through single-base resolution synthesis. DNA-DISK [5] integrates microfluidic control and enzymatic synthesis, realizing the full-process for the first time, laying the foundation for the external deployment of the laboratory technologies. Meanwhile, advances in nanopore sequencing [18] reduce read complexity and make random access more efficient.

Random access can overcome the problem of delay in sequential access and improve data reading efficiency. Large-scale data storage requires efficient retrieval, and primer addressing [19] amplify target files by PCR, but rely on pre-designed large primer libraries. Primer-Disk [11] innovatively adopts solid-phase PCR, supporting multiple readings of a single disk, while Microsoft Similarity Search [10] explores molecular-level content addressing, directly comparing DNA sequences to achieve the retrieval of 1.6M image library, providing a paradigm for future database applications. Despite significant progress, DNA storage still faces bottlenecks such as cost (synthesis/sequencing), speed (write/read latency), and system integration. In the future, it is necessary to combine molecular biology, nanotechnology and automation platforms to promote the technology from laboratories to commercialization. With the deep integration of synthetic biology and information technology, DNA storage may be the first to be implemented in cold data archiving, biocompatible storage and other fields. This Perspective will systematically review the development context of DNA storage, explore its application prospects in biological data management, and propose future-oriented research directions.

2 Why DNA, and why biological storage medium for data first?

Density & longevity. The information density of DNA (theoretically a few exabytes per cubic millimeter) and its stability over millennia under proper packaging are incomparable with disks or optical media, and its refresh cycle (3-10 years) can avoid the repeated processing and energy costs associated with traditional storage media. Early silica encapsulation demonstrated zero error recovery after simulated centuries of accelerated aging, establishing a reliable path to the "time capsule" archive.

Retrieval model fits "cold" workloads. Biological archives (FASTQ/BAM/CRAM formats, original microscope images, spectral cubes) are rarely accessed, but must be retained to ensure compliance and the

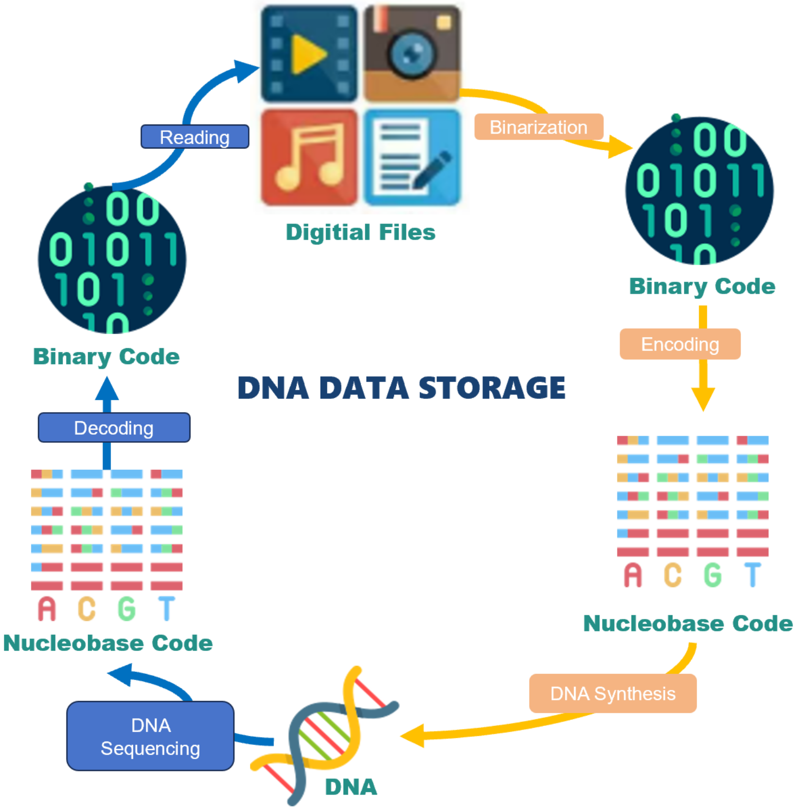


Figure 1. The overall process of DNA storage.

scientific reproducibility of research. The "write once, rarely read" configuration file of DNA storage is very consistent with such workloads and is a clear goal for programs such as IARPA's Molecular Information Storage (MIST).

Sustainability. Lifecycle analysis shows that once written to the chemical scale, DNA archives can provide lower greenhouse gas emissions and energy consumption than hard disks or tapes.

3 The technical stack: where progress is real

3.1 Encoding & error control in an indel-dominated channel

The DNA storage channel is prone to unique error types, most notably insertions and indels, which pose a greater challenge than simple base substitutions. Natively correcting the missing code for insertion - for example, HEDGES by Press et al. [13] and its optimization of nanopores is a standard choice for robust decoding of a large number of sequences.

On the other hand, the rate less fountain code (DNA Fountain) pushed channel capacity while controlling GC content, homopolymer lengths and primer limits. Together with de Bruijn-graph assembly and learning enhancement decoders, these codec pipelines improve density and retrieval speed.

In addition to A/C/G/T, the extended base alphabet also provides higher symbol entropy, and in principle the synthesis cost is lower for each memory bit [17], although standardization and read errors of artificial bases are still open questions.

3.2 Writing: from phosphonamidite to enzymatic and combinatorial

Phosphonamidite chemical synthesis technology is mature, but it is expensive and wasteful when synthesizing long oligomers on a large scale. Enzymatic DNA synthesis (EDS) is typically used in dynamic control mechanisms, with significant advances over the past 2-3 years, including multiple single base fidelity and benchtop systems that democratize iterations. Two complementary "speed" strategies are emerging: One is starting directly from scratch with precision-benchmarked process control – once stabilized, this strategy is expected to reduce waste and potential costs [14].

Combination Assembly (e.g., Catalog): Assemble information from pre-made fragments to achieve very efficient write throughput at the expense of a certain density. Recent peer reviews and community reports outline prototypes that connect storage and

Table 1. Comprehensive analysis of current DNA storage methods.

Work (Year)	Core Contribution	Write Medium / Method	Indel-Robust Coding	Random Access	Demonstrated Scale
DNA Fountain [7]	Capacity-approaching, robust fountain-code architecture	Chemical synthesis	Substitution-robust via fountain + constraints	PCR primer addressing	2.14 MB stored & perfectly retrieved
Organick et al. [12]	Large-scale random access (35 files, >200 MB)	Chemical synthesis	Code/constraint mix	Yes (primer library)	>13 M oligos, file-level retrieval
HEDGES [13]	Code that directly corrects indels	— (coding layer)	Yes (indels)	Compatible	Reference datasets
Extended alphabet [15]	Adds chemically modified bases to boost symbol set	Chem. synthesis + modified bases	Code-compatible	Potential	Prototype system
De Bruijn assembly [16]	Robust recovery from breaks/rearrangements via assembly	— (decoding strategy)	Indel/break tolerant	Pool-level	Multi-file demos
Similarity search [10]	Molecular-level similarity search over 1.6 M images	Chemical synthesis	N/A	Content-addressable	1.6 M-item DB
DNA-DISK [5]	Automated end-to-end write-store-read on DMF	Enzymatic + microfluidics	N/A	On-chip workflows	End-to-end platform
Primer-Disk [11]	Indexed solid-phase PCR enabling reusable random reads	Chem. synthesis + solid-phase PCR	N/A	Yes (≥20 reads/file)	Multi-file demos
Silica encapsulation [8]	Long-term, error-free recovery after aging	N/A	N/A	N/A	83 kB exemplar
DNA-of-things [9]	Embedding DNA in objects via silica beads	N/A	N/A	Object-level access	3D-printed materials

molecular computing. End-to-end automation is coming: DNA-DISK integrates write, store and read on the digital microfluidic platform, improving data storage reliability and reducing user operation time, which is an important system milestone.

3.3 Storage & preservation media

Chemical encapsulation is key for biological data that must last decades to hundreds of years. Silica encapsulation remains the most proven method, and "DNA of-things" demonstrates embedding into materials without affecting recovery. Alternative fillers, such as including amber-like solid polymers, are broadening room temperature storage and logistics options. Industrial products are now undergoing room temperature DNA/RNA protection assessment and being integrated into commercial workflows, marking the readiness for biobank pilot deployments.

3.4 Reading & access

Illumina remains the leading technology in terms of accuracy; nanopore sequencing is highly attractive for on-demand access and on-site filtering. A key operational requirement is that PCR-free random access methods using nanopore "adaptive sampling" and indexed PCR demonstrate a practical retrieval model that can avoid the need to centralize the entire archive. At the system level, large-scale random accesses for file-level searches from tens of millions of oligomers have been demonstrated, and similarity searches learned on DNA-encoded databases suggest a mix of molecular-level computing and storage.

4 Bottlenecks and how to break them?

Despite significant progress in DNA storage in recent years, there are still several systemic challenges in moving from laboratory prototypes to institutional archiving-challenges that not only affect technical economics but also operability, reliability, and governance compliance. Write costs and delays remain the most critical limiting factors, and even as enzyme-driven synthesis (EDS) continues to mature, the cost per base and write processing time per file still dominate the total cost of ownership. At the same time, institutional archiving requires the ability to enable file or record-level access without relying on resequencing of the entire DNA pool. In addition, the complex noise characteristics (synthesis and sequencing bias, context-related insertion and storage-induced breakage) of the DNA channel put higher demands on the encoding scheme. Regarding their liability of claims about media lifespan, although DNA storage performs well in accelerated aging experiments, institutions still need to monitor stored data regularly and maintain off-site copies when actually deploying room temperature storage. Finally, governance, privacy and ethical issues cannot be ignored. Even synthetic DNA must avoid being mixed with biological samples, while ensuring the integrity of barcode tracking, audit logs and encrypted signatures.

In order to break through these bottlenecks, it is necessary to coordinate the promotion from both technology and management. Improving EDS fidelity and parallelism, or adopting combined assembly to balance the relationship between throughput and cost, can significantly improve write economy. At the same time, high-grained access can be achieved and operational complexity can be reduced through standardized metadata architecture, index allocation and adaptive sequencing strategies. An open community testing platform and cross-technical error tracking dataset will accelerate the improvement of comparability and the optimization of coding solutions, while ensuring reliability across different media and packaging conditions. In media management, periodic checksum verification and dual-point mirroring in the early stages of deployment can help verify long-term storage performance, while industrial capsules and polymer packaging offer a viable approach to enhancing durability and operational simplicity. At the governance level, through de-identification, barcode tracking, audit logs and encrypted metadata, DNA archives can be ensured that DNA archives meet institutional requirements in privacy and compliance.

5 Conclusion

DNA data storage is expected to move from laboratory concepts to commercial applications during 2025–2030, and gradually become the most durable, sustainable and compact archival layer in the Biological Data Management (BDM) systems. It will not replace existing thermal/temperature storage, but is positioned at the last layer of "forever data". If it can continue to advance in writing economy, standardized random access and media engineering, DNA will move towards normalized deployment in the next five years.

In the short term (2025–2026), institutions can initiate small-scale pilot archives (1–10 TB) and select non- raw data (e.g. microscopic imaging datasets, instrument logs) as the initial objects for archiving. The recommended architecture is to create a DNA capsule image alongside conventional object storage; to use a combination of HEDGES-based indel-tolerant encoding and fountain/RaptorQ outer coding, and to employ a conflict-free primer library for file addressing. Capsule media (such as silica or polymers) need to be combined with tamper-proof barcodes and checksums are recorded by LIMS/ELN for traceability and compliance.

In the medium term (2027–2028), DNA storage can be extended to institutional-level archives (50–200 TB), partially replacing tape-based deep cold storage, and is especially suitable for long-term data that requiring storage periods of ≥20 years. To alleviate writing bottlenecks, institutions can introduce combinatorial assembly or high-throughput enzyme-driven synthesis (EDS) and adopt the Primer-Disk random access architecture. In terms of data extraction, adaptive sampling technology based on nanopore sequencing is expected to support on-premises targeted retrieval and enhance the practical accessibility of archived data.

In the long term (2029–2030), DNA storage will enter the phase of interoperability and ecosystem development. It is expected to align with the specifications of the DNA Data Storage Alliance to establish a unified medium ID, metadata schema and exchange format, while ensuring reliability and comparability through public testing platforms and cross-organization "truth set" verification. At the same time, the properties of DNA molecules themselves will also support emerging applications, such as content retrieval based on molecular similarity, which can achieve analogous searches in large-scale microscopic or omics data corpus.

At the practical level, successful deployment of DNA archives requires attention to the following points:(1)Data category selection: (prioritizing immutable, long-term analyzable raw data, (e.g. FASTQ and reference images); (2)Media and containers: selecting options that withstand ambient temperature fluctuations and maintaining off-site copies; (3)Coding schemes: considering GC and homologous sequence constraints to ensure primer libraries are reusable; (4)Access strategies: using PCR/Primer-Disk for conventional audits, nanopore devices for temporary extraction; (5)Governance and compliance: treating DNA capsules as regulated records, which require comprehensive documentation, copying, and verification throughout their lifecycle; and(6)Costing: comparing $/GB write and search costs with those of LTO tape and cloud archiving solutions.

In summary, the future of DNA storage does not rely on one-time "hot" demonstrations, but requires a series of quiet but solid institutional-level pilots to verify link management, total cost of ownership and retrieval reliability. Once this transition is completed, by 2030, DNA storage will naturally be integrated into the institutional data lifecycle policies as an integral part of the typical three-layer architecture: "Compression → Cloud/Tape → DNA", In this architecture, each capsule is assigned a unique barcode, and all data can be traced and verified within the system. This evolution will transform DNA storage from "frontier exploration" to a standard component of a biological data management system.

Data Availability Statement

Not applicable.

Funding

This work was supported without any funding.

Conflicts of Interest

The authors declare no conflicts of interest.

Ethical Approval and Consent to Participate

Not applicable.

References

[1] Li, X., Wang, B., Lv, H., Yin, Q., Zhang, Q., & Wei, X. (2020). Constraining DNA sequences with a triplet-bases unpaired. *IEEE transactions on nanobioscience, 19*(2), 299-307. [Crossref]

[2] Cao, B., Li, X., Wang, B., He, T., Zheng, Y., Zhang, X., & Zhang, Q. (2025). Achieving handle-level random access in an encrypted DNA archival storage system via frequency dictionary mapping coding. *Patterns*. [Crossref]

[3] Cao, B., Zheng, Y., Shao, Q., Liu, Z., Xie, L., Zhao, Y., ... & Wei, X. (2024). Efficient data reconstruction: The bottleneck of large-scale application of DNA storage. *Cell Reports, 43*(4). [Crossref]

[4] Carmean, D., Ceze, L., Seelig, G., Stewart, K., Strauss, K., & Willsey, M. (2018). DNA data storage and hybrid molecular–electronic computing. *Proceedings of the IEEE, 107*(1), 63-72. [Crossref]

[5] Choi, Y., Bae, H. J., Lee, A. C., Choi, H., Lee, D., Ryu, T., ... & Kwon, S. (2020). DNA micro-disks for the management of DNA-based data storage with index and write-once–read-many (WORM) memory features. *Advanced Materials, 32*(37), 2001249. [Crossref]

[6] Church, G. M., Gao, Y., & Kosuri, S. (2012). Next-generation digital information storage in DNA. *Science, 337*(6102), 1628-1628. [Crossref]

[7] Erlich, Y., & Zielinski, D. (2017). DNA Fountain enables a robust and efficient storage architecture. *Science, 355*(6328), 950-954. [Crossref]

[8] Grass, R. N., Heckel, R., Puddu, M., Paunescu, D., & Stark, W. J. (2015). Robust chemical preservation of digital information on DNA in silica with error-correcting codes. *Angewandte Chemie International Edition, 54*(8), 2552-2555. [Crossref]

[9] Koch, J., Gantenbein, S., Masania, K., Stark, W. J., Erlich, Y., & Grass, R. N. (2020). A DNA-of-things storage architecture to create materials with embedded memory. *Nature biotechnology, 38*(1), 39-43. [Crossref]

[10] Bee, C., Chen, Y. J., Queen, M., Ward, D., Liu, X., Organick, L., ... & Ceze, L. (2021). Molecular-level similarity search brings computing to DNA data storage. *Nature communications, 12*(1), 4764. [Crossref]

[11] Ma, J., Yang, Y., Pei, B., Mi, S., Xiong, Z., & Ouyang, L. (2025). Primer-Disk-Enabled DNA Data Storage System with Index and Record-Many-Read-Many Features. *Advanced Science*, e02367. [Crossref]

[12] Organick, L., Ang, S. D., Chen, Y. J., Lopez, R., Yekhanin, S., Makarychev, K., ... & Strauss, K. (2018). Random access in large-scale DNA data storage. *Nature biotechnology, 36*(3), 242-248. [Crossref]

[13] Press, W. H., Hawkins, J. A., Jones Jr, S. K., Schaub, J. M., & Finkelstein, I. J. (2020). HEDGES error-correcting code for DNA storage corrects indels and allows sequence constraints. *Proceedings of the National Academy of Sciences, 117*(31), 18489-18496. [Crossref]

[14] Rasool, A., Hong, J., Hong, Z., Li, Y., Zou, C., Chen, H., ... & Dai, J. (2024). An Effective DNA-Based File Storage System for Practical Archiving and Retrieval of Medical MRI Data. *Small Methods, 8*(10), 2301585. [Crossref]

[15] Ren, Y., Zhang, Y., Liu, Y., Wu, Q., Su, J., Wang, F., ... & Zhang, H. (2022). DNA-Based Concatenated Encoding System for High-Reliability

and High-Density Data Storage. *Small Methods*, *6*(4), 2101335. [Crossref]

[16] Song, L., Geng, F., Gong, Z. Y., Chen, X., Tang, J., Gong, C., ... & Yuan, Y. J. (2022). Robust data storage in DNA by de Bruijn graph-based de novo strand assembly. *Nature communications*, *13*(1), 5361. [Crossref]

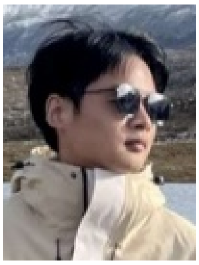
[17] Liu, D., Xu, D., Shi, L., Zhang, J., Bi, K., Luo, B., ... & Ping, Z. (2025). A practical DNA data storage using an expanded alphabet introducing 5-methylcytosine. *Gigabyte*, *2025*, gigabyte147-0. [Crossref]

[18] Winston, C., Organick, L., Ward, D., Ceze, L., Strauss, K., & Chen, Y. J. (2022). Combinatorial PCR method for efficient, selective oligo retrieval from complex oligo pools. *ACS Synthetic Biology*, *11*(5), 1727-1734. [Crossref]

[19] Zheng, Y., Cao, B., Zhang, X., Cui, S., Wang, B., & Zhang, Q. (2024). DNA-QLC: an efficient and reliable image encoding scheme for DNA storage. *BMC genomics*, *25*(1), 266. [Crossref]

Rongrong Chen currently is a researcher in Dalian University, and she graduate from Liaoning Normal University. Her main research directions are AI and DNA Storage. (Email: chenrongrong@dlu.edu.cn)

Xue Li currently a lecturer at the School of Software Engineering, Dalian University. She obtained her Ph.D. from China University of Petroleum (East China) and the nus-duke Medical School of the National University of Singapore. Her main research directions are bioinformatics and DNA computing, and she has published over ten papers in journals such as JBHI and BIB. (Email: xueleecs@gmail.com)



Ben Cao received his Ph.D. in Computer Science from Dalian University of Technology (DLUT), where he was jointly trained with A*STAR, Singapore. His research focuses on the intersection of computer science and biotechnology, with a commitment to solving biological problems using computational methods. His work spans areas such as DNA data storage and security, coding theory, gene expression, artificial intelligence, and large language models (LLM). Dr. Cao has published over 20 papers as the first author or co-first author in international journals such as Cell Reports, IEEE Transactions, and BIB, including one ESI hot paper and two ESI highly cited papers. He actively contributes to the academic community, serving as the Associate Editor of the Journal of Artificial Intelligence in Bioinformatics, and has been invited to review for approximately 20 SCI journals, including IEEE TNNLS, TCBB, TMBMC, TNB, Computational and Structural Biotechnology Journal, Future Generation Computer Systems, Scientific Reports, IEEE Communications Letters, PeerJ, BMC Bioinformatics, PeerJ Computer Science, and others. (Email: bencaocs@gmail.com)