Unraveling the Promise of Computing DNA Data Storage: An Investigative Analysis of Advancements, Challenges, Future Directions

Zarif Bin Akhtar*

MPhil Postgraduate Researcher, Master of Philosophy (MPhil) in Machine Learning and Machine Intelligence, Department of Engineering, University of Cambridge

* Corresponding author. Email: zarifbinakhtarg@gmail.com; zarifbinakhtar@ieee.org
Manuscript submitted April 10, 2024; revised April 30, 2024; accepted May 3, 2024.
DOI: 10.18178/JAAI.2024.2.1.122-137

Abstract: This research is dedicated to a comprehensive examination of DNA-based data storage systems, emphasizing the fundamental principles of DNA computing and its applications for long-term data archival. The primary objective is to critically assess the capabilities and limitations of DNA in meeting the exponential demand for robust and sustainable data storage solutions. Employing an in-depth iterative exploration, the research delves into recent breakthroughs in synthetic biology tools and unconventional computing functional methodologies in line with artificial intelligence (AI) technicality features, particularly emphasizing the transformative machine intelligence impact of the interdisciplinary collaborations within biomedical engineering and AI application domains. The findings illuminate the promise of DNA data storage as a viable solution for addressing the escalating data storage demands of the digital era, shedding light on its efficiency and integration of AI, scalability, and transformative potential. Ultimately, this research contributes to the development of efficient and scalable DNA data storage technologies, the role of AI and machine intelligence integration highlighting their immense significance in the ever-evolving landscape of data management towards the future and beyond.

Keywords: Artificial Intelligence (AI), Biomedical Engineering (BME), Deep Learning (DL), DNA Data Storage, Machine Learning (ML), synthetic biology tools, unconventional computing.

1. Introduction

In the digital age, the incessant surge in data generation has precipitated an urgent demand for innovative and sustainable data storage solutions that transcend the limitations of traditional electronic mediums. Amidst this burgeoning landscape of data management, the convergence of biological systems and computational science has sparked a renaissance in the exploration of many unconventional computing paradigms, notably the transformative potential of DNA data storage [1–3]. In harnessing the intrinsic information encoding capabilities of DNA molecules, contemporary research endeavors have embarked on a quest to unlock the promises and possibilities of leveraging biological substrates for long-term data archival and retrieval. The exponential growth of data-driven technologies, ranging from AI, cloud computing infrastructures to IoT ecosystems, has accentuated the need for data storage solutions that offer unparalleled durability, compactness, and longevity. In response to this pressing demand, the fusion of synthetic biology tools, unconventional computing methodologies, and DNA-based information storage systems has engendered a paradigm shift in the realm of data storage technology.

AI technology for EHR implementations has a great future scope of boom in terms of medical healthcare informatics and will be explored in the years to come. AI infusions with medical informatics in terms of technical computing pipelined with biomedical engineering (BME) has a wide array of domains and areas where AI will play an impacting factor. Both perspectives in terms of bioinformatics and biocomputing accelerations will be developed with machine intelligence integrations in the near years to come. In that retrospective, data computing, data management and especially storage functionality and digital perseverations of information will become a pivotal factor to tackle and where this research investigative exploration sheds light on for the future directions on that pathway. By imbuing DNA molecules with the ability to encode and store vast amounts of digital information, researchers and practitioners have embarked on a pioneering journey to explore the frontiers of DNA data storage, with the vision of revolutionizing the landscape of digital information preservation and management [4–6].

In this context, this investigative analysis delves into the multifaceted dimensions of DNA data storage, unraveling the intricacies of its advancements, challenges, and future directions. By amalgamating insights from interdisciplinary research domains, this exploration endeavors to illuminate the transformative impact of DNA data storage on the evolution of data storage technologies, thereby offering a comprehensive understanding of its potential as a sustainable and robust archival solution in the digital era. Through an in-depth examination of the current state-of-the-art techniques, theoretical frameworks, and technological innovations, this research aims to elucidate the opportunities and constraints associated with DNA data storage, paving the way for a comprehensive roadmap towards its practical integration within the broader landscape of data management infrastructures. DNA, as a data storage medium, demonstrates superior efficiency and longevity compared to traditional digital methods. Its benefits include better compression, higher physical density, increased stability, and lower energetic costs [7–9]. The integration of metadata within the DNA sequence is essential to ensure efficient data retrieval and management, thereby facilitating long-term accessibility and utilization. Moreover, the imperative of prioritizing bio cybersecurity measures cannot be overstated, as the potential risks associated with synthetic DNA-encoded malware pose a significant threat to data integrity and security during DNA sequence analysis. Implementing robust security protocols is crucial to safeguard the integrity and privacy of stored data and prevent unauthorized access or malicious tampering. Standardizing the coding and decoding processes is imperative to facilitate seamless data retrieval and interpretation, minimizing the risk of errors that could lead to data loss. Ensuring the reliability of protocols for handling DNA samples is paramount to maintaining the integrity and longevity of stored information [10–12]. Furthermore, the inclusion of specific identifying features within the DNA structure serves as a crucial mechanism for future recognition and interpretation of DNA data storage. Integrating markers or sequences that distinguish synthetic DNA from biological entities aids in preserving the accessibility and interpretability of stored data for future generations, even in the absence of explicit knowledge about DNA data storage techniques. While DNA presents immense potential as an intelligent data storage medium, several challenges related towards scalability, standardization, metadata integration, bio cybersecurity, and the requirement for specialized tools must be addressed to facilitate its widespread implementation. Collaboration among interdisciplinary experts, coupled with a forward-thinking approach, is essential to unlock the full potential of DNA data storage, promising substantial benefits such as low energy costs, high-density storage, and long-term stability.

2. Methodology

For this investigative analysis, a meticulous investigation and synthesis of a broad array of background research information were conducted, encompassing influential research studies, scholarly articles, and technical reports relevant to DNA data storage, synthetic biology, and unconventional computing.

This extensive analysis process not only established a comprehensive understanding of the current landscape but also facilitated the synthesis of a nuanced knowledge base crucial for the investigative analysis. In conjunction with the available knowledge review, a systematic collection and analysis of primary data were undertaken. This involved the meticulous aggregation of empirical data and case studies focusing on the implementation of DNA data storage systems across various experimental contexts.

Employing both qualitative and quantitative methodologies, the analysis aimed to discern prominent trends, patterns, and challenges inherent in DNA data storage technologies, thus contributing to a comprehensive and insightful assessment. Moreover, insights data from various types of interviews and consultations with leading experts in molecular biology, computational genomics, and synthetic biology were gathered and retrieved. These discussions provided invaluable perspectives on the feasibility, scalability, and ethical implications of integrating DNA-based storage solutions within the broader spectrum of data storage technology.

The input from these esteemed professionals significantly enriched the analytical framework and facilitated a more holistic understanding of the subject matter. In order to provide a comprehensive evaluation, a comparative analysis framework was employed, compared to the efficacy and efficiency of DNA data storage systems with conventional electronic storage methods. Multiple case studies investigations were examined to assess the performance, data retrieval speed, and long-term stability of DNA-based storage architectures across diverse environmental conditions.

This comparative approach enabled a thorough assessment of the advantages and limitations of DNA data storage, thus contributing to a complexify understanding of its potential as a viable data archival solution. Lastly, drawing upon the insights gleaned from the information retrieval, data analysis, expert consultations, and comparative studies, a comprehensive roadmap was devised. This roadmap outlined key research priorities, technological advancements, and interdisciplinary collaborations necessary to address existing challenges and expedite the practical integration of DNA data storage within mainstream data management infrastructures. This forward-looking approach aimed to chart a clear path for the future development and implementation of DNA data storage technologies.

3. Background Research and Available Knowledge

The concept of DNA digital data storage involves encoding and decoding binary data into synthesized DNA strands. While DNA demonstrates high storage density, its practical application is limited due to cost constraints and slow read and write times [1–5]. Various methods for encoding data in DNA are actually possible, including encoding text through translation of letters into corresponding ***"codons"*** and encoding arbitrary data by converting it into ternary data [6–8]. The incorporation of error correction mechanisms and its associated protective measures is vital for ensuring the accuracy and integrity of the stored data.

The use of DNA for digital data storage has a rich history, dating back to early conceptualization by scientists in the timeline years of the 1950s and 1960s. Notable milestones include the encoding of images and entire books into DNA, as well as the development of error-correcting coding schemes to enhance data retention accuracy [9–11]. Recent advancements in the wide array of diverse domains have showcased the development of more sophisticated and especially complex level of the encoding methods and automated systems for data retrieval and decoding. Additionally, the integration of DNA storage in live organisms through synthetic biology and CRISPR gene editing presents a novel avenue for further exploration [12, 13].

The exploration of in vivo direct DNA data recording systems, leveraging optogenetically regulated recombinases and engineered ***"molecular recorders,"*** represents a groundbreaking approach in terms of DNA data storage [14–16].

The methodology's ability to encode light-based stimuli into DNA and perform data processing operations within biological systems underscores the versatility and potential applications of DNA digital data storage [17, 18]. The concept of the ***"DNA of Things"*** (DoT) introduces a paradigm shift by embedding digital data into physical objects, enabling the independent storage of information off the grid [19, 20].

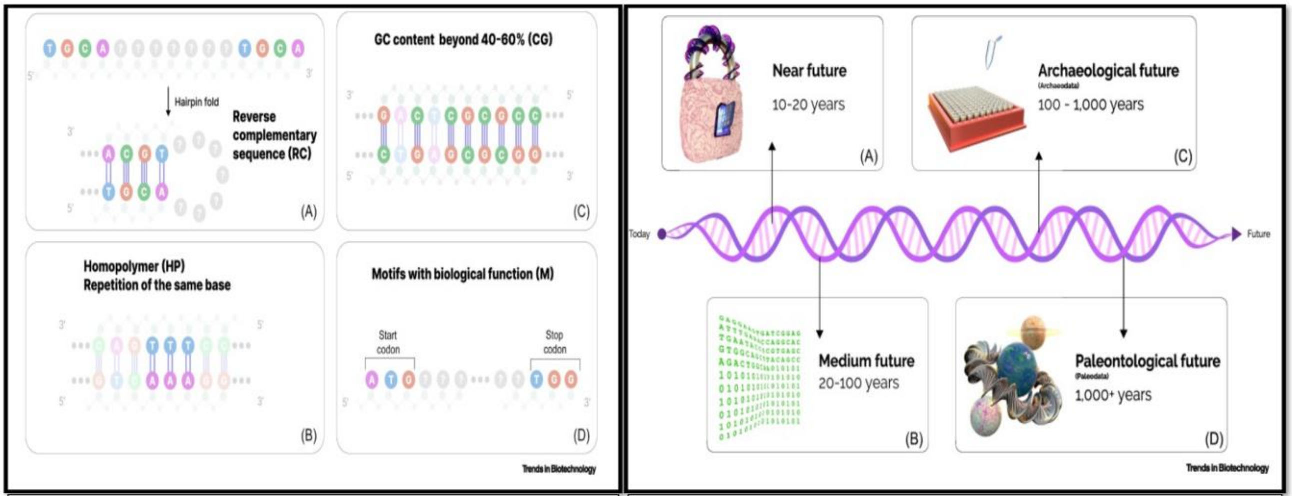
These advancements collectively highlight the evolving landscape of DNA digital data storage and its potential for revolutionizing data storage and archiving methodologies across various fields, paving the way for the development of more efficient and sustainable data storage solutions in the near future.

4. Biomimetic Systems and Genetic Code as an Intelligent Data Storage

The potential of biomimetic systems and the genetic code as a means of intelligent data storage is becoming increasingly apparent. Inspired by the robustness and efficiency of DNA, researchers have been exploring ways to leverage its unique properties for digital information storage. Unlike traditional binary systems, DNA offers greater capacity and density due to its four-state representation and 3D storage potential. With 3.2 billion base pairs in the human genome alone, the data density of DNA far exceeds that of contemporary hard disks. Furthermore, the stability and longevity of DNA make it an attractive option for long-term data preservation, with the capability to survive for hundreds of thousands of years under suitable storage conditions. This characteristic, along with the relatively low energy costs associated with DNA storage, presents a compelling case for its integration into data storage technologies. Current advancements in DNA data storage involve a comprehensive five-step process, including encoding, writing, storing, reading, and decoding. Each step is carefully designed to address the challenges posed by the unique properties of DNA, such as avoiding specific sequences, maintaining GC content, and preventing errors that can occur during synthesis and sequencing. Various techniques for encoding and decoding data onto DNA have been proposed, ranging from direct mapping to more complex algorithms that maximize the use of DNA's storage potential. Additionally, the storage of DNA can be achieved in vitro, in devices, or even within living organisms. While in vitro storage methods require controlled environmental conditions which are required to prevent DNA degradation, in vivo storage within microorganisms offers scalability and self-repair capabilities.

Despite significant progress, challenges remain, such as the need for improved synthesis throughput and sequencing speeds to make DNA data storage more feasible for practical applications. Nevertheless, ongoing research and advancements in DNA synthesis and sequencing technologies continue to drive the development of DNA data storage towards becoming a viable solution for long-term data preservation. To provide an idea Fig. 1 illustrates a representation of the recent trends in biotechnology. For the widespread use of DNA data storage in the future, several critical aspects must be standardized. It is vital to ensure that metadata associated with the stored data is contained within the DNA itself, allowing for autonomous retrieval of the information. By incorporating necessary information, such as the encoding method and access protocols, directly into the DNA sequence, the risk of data loss due to the unavailability of external resources can be minimized. However, such an approach raises concerns regarding bio cybersecurity breaches, necessitating the implementation of robust security measures within the DNA data storage pipeline. In the medium term, as various companies introduce their unique coding and decoding algorithms, establishing a standardized approach to retrieve data stored over decades is imperative. While competition among tech giants may foster innovation, a unified retrieval mechanism will ensure continued access to data, even as companies evolve and technologies change. To safeguard the accessibility of data stored over centuries, it is essential to standardize every step of the DNA data storage process, including coding, decoding, sample handling, library preparation, and sequencing. Creating a consistent methodology for data retrieval can prevent potential loss or misinterpretation of information due to technological discrepancies.

Looking into the distant future, when our contemporary society's knowledge might be lost or distant from the current era, it becomes crucial to establish methods for identifying non-biological data stored in DNA. Integrating specific sequences or rare isotopes into the DNA can serve as a distinguishing watermark, clearly indicating the synthetic nature of the stored data. Implementing such measures will ensure that future societies can recognize and access the valuable information encoded within the DNA, even if they lack the knowledge of the contemporary technological processes.



(A) Self-reverse complementary DNA (RC) sequences form secondary 3D structures due to regions that are complementary to each other. These structures can hinder sequencing and synthesis processes, leading to random errors. (B) High-throughput sequencing methods are prone to errors when they encounter long consecutive repetitions of the same base, known as homopolymers (HPs). (C) GC content (GC), representing the percentage of G and C bases in the sequence, should be maintained at around 50% to ensure molecular stability and to better control the melting temperature during synthesis. (D) Caution must be exercised to prevent the synthesis of DNA sequences (M) that might trigger protein production if they interact with the necessary biological machinery. Avoidance of open reading frames (ORFs) and potentially pathogenic sequences is crucial.

(A) For use in the next 10–20 years, whoever stores data into DNA should be sure that the metadata needed to access the stored information must be available at the same place if not part of the DNA. Also, bio cybersecurity protocols need to be part of the DNA data storage pipeline. (B) For use in the next 20–100 years, there is the need for a single way of retrieving data that has been stored for decades; that is, a standard coding and decoding algorithm. (C) For the use of stored information 100–1000 years from now, protocols for reading the DNA data need to use a consistent methodology, from sample handling to sequencing method and primer usage. (D) When someone needs to read data stored for more than 1000 years, it is imperative to identify the DNA sample as being not natural so it is entirely distinct from genetic material in living organisms. It is critical to signalize that the pattern on the DNA is somehow related to a binary file.

Fig. 1. A visual representation of recent trends in biotechnology.

5. Unlocking the Power of DNA, Data Storage Revolutionization Advantages and Challenges

The concept of DNA data storage was initially proposed by scientists in the 1990s, and it saw its first practical demonstration in 2012 by a team of researchers at the European Bioinformatics Institute led by Nick Goldman. This groundbreaking experiment involved encoding a large text file into DNA sequences, synthesizing the DNA, and successfully retrieving the original data with high accuracy. This pivotal achievement has since catalyzed rapid growth in the field of DNA data storage, with numerous groups and companies now actively pursuing more efficient and cost-effective methods for utilizing DNA as a storage medium.

DNA data storage has emerged as a response to the growing demand for secure, dense, and long-term data storage solutions. DNA, known for its role in carrying genetic information, possesses key characteristics that make it an ideal candidate for data storage, including its high density, long-term stability, and built-in error correction mechanisms. The process of storing data in DNA involves converting digital data into DNA sequences comprised of the four chemical building blocks of DNA, and these sequences can then be synthesized and stored for later retrieval. One of the primary advantages of DNA data storage is its exceptionally high storage density, with a theoretical capacity of up to 1 exabyte per gram, surpassing that of traditional data storage media. DNA's long-term stability has been demonstrated through its ability to persist for thousands of years if properly preserved, making it an ideal option for archival storage purposes, especially for scientific data and historical records. Despite these advantages, there remain several technical challenges associated with DNA data storage.

The cost and complexity of DNA synthesis and sequencing have traditionally been significant obstacles, but recent advancements in these technologies have rendered DNA data storage increasingly feasible and cost-effective. Furthermore, error correction poses a notable challenge in DNA data storage, prompting the development of specialized algorithms and techniques to ensure data accuracy, including redundancy strategies and error correction algorithms that mitigate the risks associated with sequencing errors.

6. Artificial Intelligence (AI) Integration within DNA Data Storage: A Groundbreaking Development under progress

The integration of artificial intelligence (AI) in the field of DNA data storage holds significant promise for advancing the capabilities and efficiency of this technology. AI's contributions encompass various key areas, including optimizing DNA sequences for data storage, enhancing error correction mechanisms during DNA sequencing and data retrieval, streamlining the DNA synthesis process, and facilitating data compression prior to encoding into DNA sequences. AI algorithms can help optimize the encoding of binary data into DNA sequences, taking into account critical factors like data density, error correction, and sequencing efficiency. Moreover, they can improve the accuracy of error correction methods employed during DNA sequencing, thereby enhancing the reliability and precision of the data retrieval process. Additionally, AI can streamline the DNA synthesis process by optimizing synthesis conditions, ultimately reducing the associated costs and time required for DNA synthesis. Furthermore, AI-driven data compression techniques can effectively minimize the amount of DNA needed to store data, thereby making DNA data storage more practical and accessible for a broader range of applications. As AI technology continues to evolve, its increasing role in the advancement of DNA data storage is expected to significantly contribute to the development of more efficient and cost-effective solutions in this rapidly evolving field. In a groundbreaking development, a research team led by Associate Professor Poh Chueh Loo from the National University of Singapore (NUS) has pioneered a revolutionary ***'biological camera'*** that utilizes living cells and their inherent biological mechanisms to encode and store data. This innovative approach, recently published in Nature Communications, introduces a novel model for data storage, resembling a digital camera, and has the potential to significantly impact the data storage industry. By leveraging DNA's exceptional storage capacity, stability, and durability, the team's ***'BacCam'*** system offers a promising solution to the ever-growing data overload problem, marking a significant milestone in the integration of biological systems with digital technologies.

The world's escalating data generation has sparked a quest for an alternative storage solution that transcends traditional methods and addresses the environmental impact of resource-intensive data centers. Acknowledging DNA's unparalleled data storage capabilities, the team highlights its ability to store vast amounts of information, underlining that a single gram of DNA can hold over 215,000 terabytes of data. Overcoming the limitations of existing DNA storage research, which primarily involves synthesizing DNA strands outside of cells, the team's ***'BacCam'*** system utilizes live cells as a ***'data bank,'*** eliminating the need for external DNA synthesis and offering a more accessible and scalable approach. The ***'BacCam'*** system operates by employing optogenetics to imprint light signals onto DNA within cells, akin to capturing images on photographic film. The researchers utilize barcoding techniques and machine-learning algorithms to organize and reconstruct the stored images, mimicking the functionality of a digital camera's data capture, storage, and retrieval processes. Notably, the system demonstrates the ability to capture and store multiple images simultaneously using different light colors, representing a significant advancement in the field of DNA data storage. Associate Prof. Poh's team envisions their pioneering work as a stepping stone toward further innovation in the realms of recording and storing information, emphasizing the integration of biological and digital systems as a key avenue for future exploration and development.

7. DNA Tagging Technology, Available Market Solutions with Horizon Expansion and Possibilities

The development of DNA-tagging technology presents a promising approach to addressing the limitations of conventional tagging systems. While various existing technologies like UPC barcoding and RFID have their advantages, they are often unsuitable for small, flexible, or hidden objects. Leveraging DNA's high storage capacity and stability, DNA tagging offers a potentially massive storage medium that can withstand the test of time.

The key to successful DNA tagging lies in the design process, encompassing crucial decisions such as the type and size of information, coding methods, storage techniques, and data extraction processes. Although the concept of DNA tagging has been in development since the 1980s, the technology has encountered challenges, particularly regarding the instability of nucleic acids due to environmental factors. Protecting DNA from degradation agents remains a primary concern, prompting the use of protective materials like silica, polymers, and gels.

Decoding the DNA tag is another obstacle that researchers are addressing through various methods, including PCR, smartphone-based assays, and CRISPR-based readout techniques such as SHERLOCK. The use of portable DNA sequencing technologies, like SmidgION developed by Oxford Nanopore Technologies, shows promise in overcoming some of the existing obstacles. The procedure for DNA-based tagging involves applying DNA to physical objects, which can be integrated into the object's structure or labeled for identification. Extracting the DNA and reading the tag involves processes such as DNA amplification and sequencing, culminating in the decoding and transformation of the sequence into valuable information. Despite the challenges, ongoing advancements in this field highlight the potential of DNA tagging technology in revolutionizing the tagging and tracking of various physical items. The development of DNA-tagging systems has expanded significantly within the recent years, offering innovative solutions for the secure identification and authentication of various tangible objects. One such system, Porcupine, developed in collaboration between the University of Washington and Microsoft, utilizes portable end-to-end molecular tagging technology, incorporating highly separable nanopore signals for efficient readout processes.

Various other research and commercial enterprises have also contributed to this field, introducing a range of DNA-tagging technologies for diverse applications. These technologies encompass an array of methodologies, including DNA ink, DNA-enclosed silica capsules, and adhesive DNA markings, each designed to provide unique advantages for specific industries and market segments.

Companies such as SigNature DNA, Haelixa, and SelectaDNA have implemented DNA tagging into various marking systems, including RFID devices, labels, and serial numbers, with applications in product authentication and anti-counterfeiting measures. Some companies, like Holoptica and DNA Technologies, have integrated synthetic DNA tags into inks for brand protection and art authentication, utilizing unique photoluminescent properties for enhanced security measures.

Despite the potential benefits of DNA-tagging technologies, challenges persist, such as limitations in scalability due to the requirement for laboratory analysis and the potential for false positives or contamination. Additionally, concerns over the vulnerability of these tags to next-generation sequencing methods underscore the need for ongoing research and development in this rapidly evolving field. Nonetheless, the expansion of these technologies demonstrates their potential to revolutionize security measures across a wide range of industries and applications. To provide a better understanding concerning the perspective Fig. 2 provides a visual representation relating to the matter. The future of DNA-tagging technology appears promising, with its potential applications extending across various industries and fields.

While this technology is still relatively underutilized compared to more conventional tagging methods, advancements in DNA synthesis and encoding technologies are anticipated to make it more practical and accessible for a broader range of applications.

Table 1. An illustrative visual representation of all the data materials

Company name, country of origin and launch date	Main features of the technology	Selected markets	References
Applied DNA Sciences, United States, 1983	Botanical DNA fragments, detection by PCR and CE, an encapsulation system	Product authentication, supply chain traceability, brand protection, anti-counterfeiting, textiles, pharmaceuticals, etc.	[21–26]
Haelixa, Switzerland, 2016	Synthetic DNA tags, detection by PCR, DNA enclosed in silica	Product authentication, supply chain traceability, intellectual property protection, etc.	[26–30]
Selectamark Security Systems (SelectaDNA), United Kingdom, 1986	Laboratory analysis of DNA for owner identification if microdots are absent (DNA serves as an alternative authentication solution)	Asset protection and recovery, securing high-value items, art and jewelry authentication, IT equipment and vehicle security, forensic applications, theft prevention and deterrence, etc.	[31–33]
TraceTag (CypherMark), United Kingdom / Norway, 2001	Synthetic DNA with unique primers, authorized access to primer sequences, detection using qPCR	Brand safeguarding, industrial applications, cash security, security of documentation, oil and fuel tracking, anti-counterfeiting measures, etc.	[34–36]
Holoptica, United States, 2012	Synthetics DNA tags (100 nucleotides), integration with inkjet cartridges	Artwork, documents and assets protection, verifying product authenticity, food tracking, etc.	[37–39]
DNA Technology, United States, 1993	DNA-laced ink, combination of DNA synthetic segments and optical taggants	Memorabilia and collectibles, limited edition artwork, pharmaceuticals, apparel and luxury goods, health and beauty industry, etc.	[40, 41]
Tagsmart, United Kingdom, 2015	Synthetic DNA tags, secure Certificate of Authenticity	Artwork, securing collectibles, verifying paper documents, book manufacturing, etc.	[42, 43]
DNA Guardian, Australia, 2007	UV-detectable stain, detection using pyrosequencing	Asset marking, crime prevention, artwork protection, theft deterrence, etc.	[44, 45]
Aanika Biosciences, United States, 2018	Genetically modified *Bacillus subtilis* as an encapsulation system for DNA tag	Agriculture and food production, textiles, etc.	[46–48]

Notably, improvements in sequencing technologies, such as the development of portable sequencers, are expected to facilitate the integration of DNA tagging into everyday practices without the need for specialized laboratory environments. The potential applications of DNA tagging are diverse, ranging from the authentication of high-value art and pharmaceutical products to the monitoring of weaponry and the tracing of valuable minerals like diamonds. The technology shows promise in addressing challenges related to data security and encryption, potentially offering solutions for secure data storage and communication in sensitive sectors like defense and healthcare [49, 50]. By combining the characteristics of DNA with existing cryptographic techniques, DNA tagging has the potential to provide enhanced security in various applications [51, 52].

The concept of DNA tagging is also proposed to extend into the realm of the metaverse, linking physical objects to their digital representations and enabling the establishment of unique identifiers for tangible objects, similar to nonfungible tokens (NFTs). Moreover, the technology may find applications in forensic science, animal tracking, and industrial process management, showcasing its versatility and potential across different domains. The interdisciplinary nature of DNA tagging necessitates collaboration among experts from various fields, emphasizing the need for continued research and development to address existing challenges and further expand the technology's potential applications. With ongoing advancements and increased adoption, DNA tagging has the potential to revolutionize security measures and authentication processes across industries and sectors. The integration of DNA into physical objects through tagging processes introduces a plethora of potential applications, encapsulated within the concept of the ***"DNA-of-things."*** This novel approach holds the promise of providing unique identification features akin to living organisms, potentially linking the DNA polymer utilized with the individual or organization possessing the object, such as artworks and inventions.

Moreover, the use of DNA tagging presents an opportunity to bridge the gap between the virtual and physical realms, facilitating the validation of connections between tangible items and their digital counterparts, thereby potentially enhancing the verification process for nonfungible tokens (NFTs) associated with physical objects. The versatility of DNA tagging extends to the realm of forensic science, where it can be employed to combat criminal activities by enabling the tagging of items with DNA sprays, offering an advantageous tool to track and identify criminals. Additionally, the technology can be leveraged for animal tagging without the need for intrusive tracking devices, ensuring minimal disruption to natural animal behavior. The combination of DNA tagging capabilities with data storage potential and robust cryptographic techniques opens the door to its application in military and intelligence operations, serving as a formidable deterrent to emerging threats posed by quantum computers, which could render current encryption methods obsolete.

In the context of cybersecurity, the use of DNA labeling can serve as a potent defense mechanism against potential threats, with the integration of encryption methodologies enabling the linkage of user genetic material, thereby creating a highly challenging obstacle for unauthorized access.

DNA tagging holds significant promise in enhancing the safety and security of food and pharmaceutical production, making it exceedingly difficult or nearly impossible to counterfeit products, especially in light of the alarming rise in counterfeit medication distribution globally. As underscored by the World Health Organization (WHO), the counterfeit drug market's expansion has had severe consequences, leading to a notable increase in counterfeit-related deaths and exacerbating public health challenges, especially the cases during the ongoing of the COVID-19 pandemic timeline.

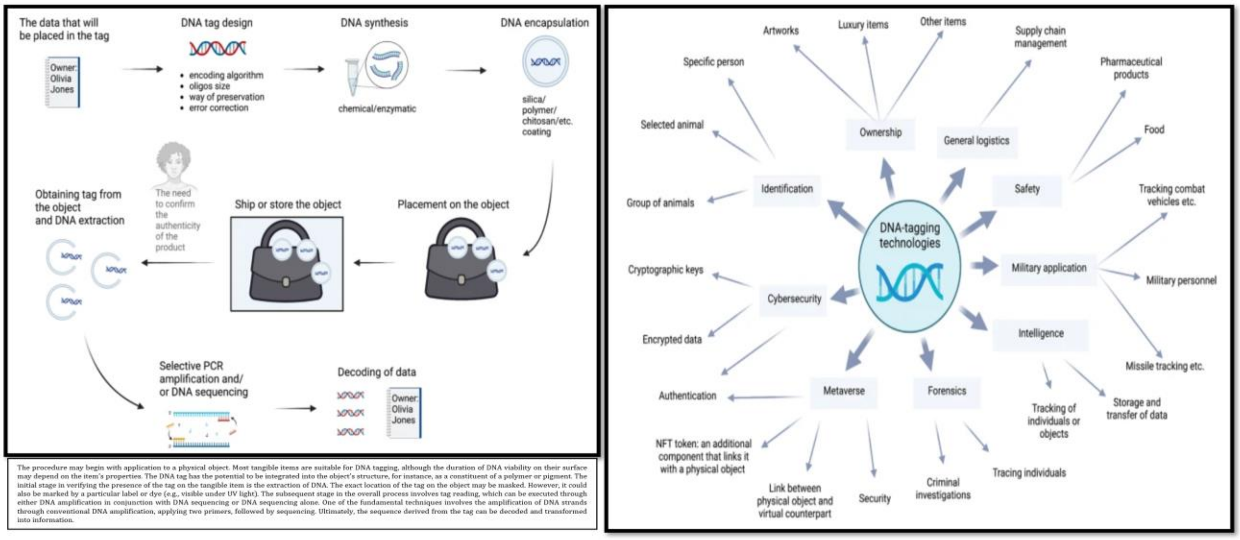
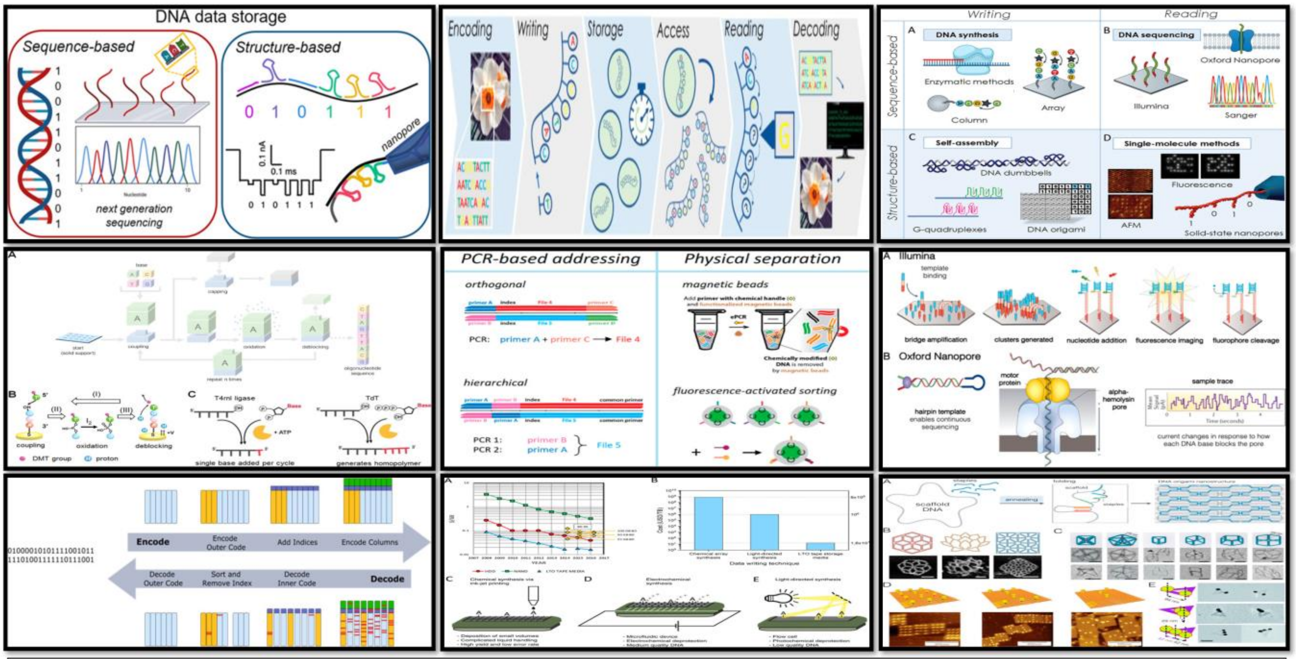


Fig. 2. The DNA tagging technology retrospective.

8. Results and Findings

The field of DNA data storage holds the promise of achieving high-density digital information storage, but it currently faces challenges that hinder its competitiveness against conventional storage technologies.

Notably, the limitations in writing speed and cost remain significant obstacles, with the current record for DNA digital data storage at approximately 200 MB, and single synthesis runs taking around 24 hours. Despite these challenges, efforts are underway to address the constraints through advancements in encoding schemes, writing and reading processes, and storage procedures. Emerging chemical and enzymatic processes are driving down the costs and time associated with writing DNA-based information, thereby reducing barriers for both sequence-based and structure-based storage approaches. To improve data readout, advancements in DNA sequencing techniques are crucial, with the potential for new chemistries, such as unnatural nucleotides and molecules modulating DNA structure, expanding the parameter space for both sequence- and structure-based DNA data storage. Additionally, the integration of solid-state nanopores and optical techniques in readout processes is expected to enhance accuracy and speed without relying on enzymes. The utilization of DNA nanotechnology in assembly procedures and the application of computational methods offer avenues for further development and deployment of DNA data storage systems. Archival storage is predicted to be a significant application for DNA data storage, leveraging the long-term stability of DNA under appropriate storage conditions. Despite the established stability of DNA, further research is required to evaluate the longevity of noncovalently assembled DNA nanostructures and their readability after prolonged periods. The implementation of dynamic properties within DNA databases, allowing for operations such as data erasure, rewriting, and updates, could enhance the practical viability of DNA-based storage systems, presenting an opportunity for improved efficiency and cost-effectiveness. A concerted multidisciplinary effort is essential to propel the field of DNA data storage forward, with collaboration among researchers from diverse areas necessary to develop advanced chemical techniques, instrumentation, characterization methods, and automated analysis tools. A holistic bottom-up approach is crucial for designing the entire process, from data encoding to decoding, underscoring the need for collaborative efforts across various scientific disciplines, including mathematics and polymer chemistry. To provide a visual representation of the matter Fig. 3 and Fig. 4 provides an overview concerning the perspective.



Overview of next-generation sequencing technologies presently used in DNA data storage. (A) Illumina sequencing generates clusters of identical single-stranded oligonucleotides. As the complement is synthesized using spectrally distinct, fluorescently tagged nucleotides, the identity of each base along the strand can be determined through the color of emission. (B) Oxford Nanopore measurements do not require fluorescent dye molecules. As the oligonucleotide passes through the protein pore, the three-dimensional shape of each base will modulate the ionic current, which results in a current-time trace that corresponds to the specific sequence. (C–E) Comparison of different DNA synthesis platforms and their characteristic traits. (C) Printing technology is primarily used by Twist and Agilent. (D) Electrochemical synthesis is employed by Custom Array. (E) Antkowiak et al. used light-directed synthesis. ***Images adapted with permission from Springer Nature.***

Fig. 3. An Overview of the findings concerning the research investigation.

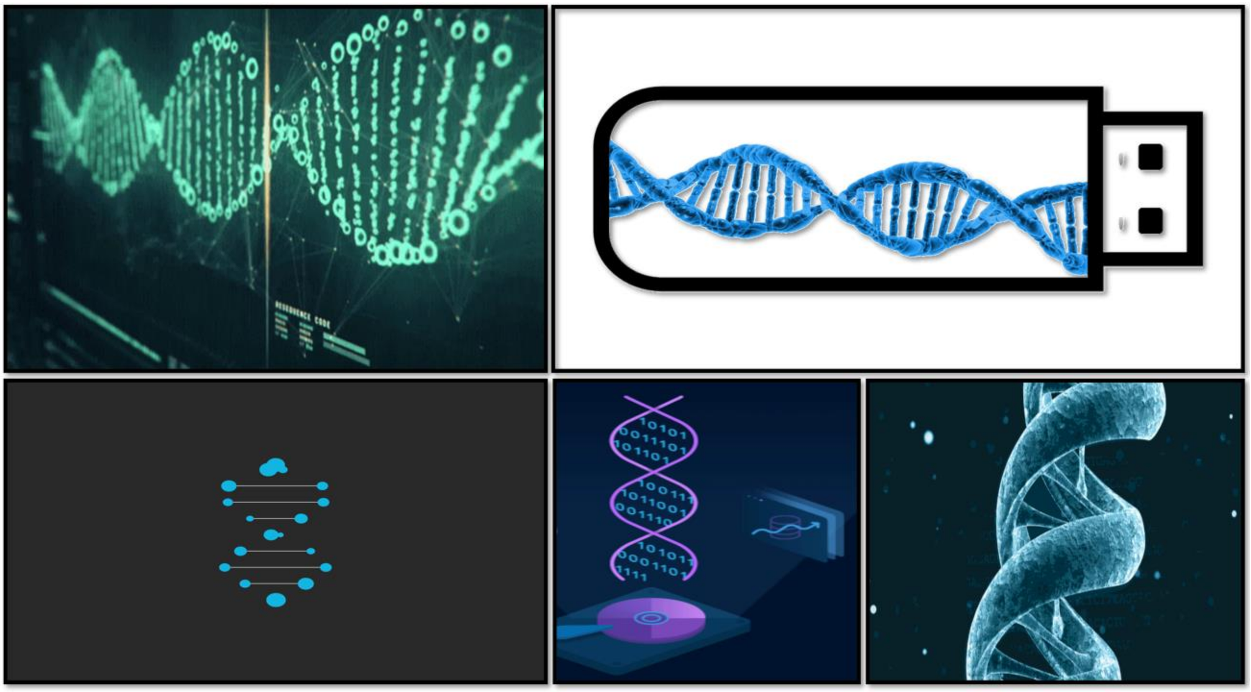


Fig. 4. An overview of the possible integrations of DNA Data Storage with AI and machine learning.

There are various company and organizations who over the years have been working and developing and are currently providing various types of features and technical computing functionalities for the selected typesets within terms of many specific markets. This research has been resourced and have been retrieved from various types of domains, distributions, packages and platform of data and their availability of the materials, some of which are not all publicly available due to containing many types of private information within them but the data which can be designed and developed, publicly available are mentioned and referenced accordingly within Table 1 for further clarification.

To be more precise on the matter, the original imaging and the various types of data some of which are not all publicly available, because they contain private information. The types of provided data that support the research results with its associated findings and information of the research investigations explored are also referenced where appropriate.

9. Discussions and Future Directions

The potential of DNA data storage is gaining traction, evident in recent market initiatives like the collaboration between Netflix and Twist Bioscience that encoded the Biohackers series into DNA. The technology's applicability is currently focused on cold and glacial storage, with its use in warm or hot storage facing challenges in writing, reading, and decoding data. However, advancements in writing platforms are anticipated, with expectations of achieving massive parallel DNA synthesis of gigabytes in a single chip in the coming years and terabytes within the next decade. The remarkable capacity of DNA data storage to hold vast amounts of information in a minuscule space and its exceptional stability, facilitating retrieval at significantly faster rates than traditional storage mediums, underscore its potential.

Despite these promising prospects, substantial breakthroughs are required, particularly in enhancing the speed and length of synthetic DNA writing. Furthermore, for the technology to be more widely adopted in hot storage applications, efficient means of locating specific files and the reusability of reagents and nucleotides in the writing process must be addressed. While the current hurdle lies primarily in the writing speed, enzymatic synthesis holds promise for addressing this challenge by enabling faster and longer data strand writing. Perhaps in the years to come hopefully this will be accelerated into a much better scope of both from a funding standpoint and a research integration for various types of the device peripherals in terms of medical equipment's and technical computing.

At the same time, while cost remains a concern, the increasing adoption of the technology is expected to render it more affordable, especially considering the benefits of energy-efficient storage systems compared to conventional data servers. Notably, various companies are actively involved in different aspects of the DNA data storage pipeline, from CODEC development to synthesis and sequencing. This concerted effort is expected to pave the way for widespread access to the technology within the next decade, enabling the general public and businesses to store various types of binary data into DNA, potentially ensuring its preservation indefinitely.

In the dynamic landscape of data storage, the emergence of DNA data storage stands out as a promising solution to address the persistent challenges faced by traditional methods. These regulatory directions need to be mapped towards into the technical underpinnings, advantages, potential applications, challenges, and future prospects of DNA data storage.

Researchers and investors are presented with a visionary solution in DNA technology for the future of data storage, spanning applications in archiving, research, finance, and technology sectors. While enthusiasm for DNA data storage is warranted, a pragmatic approach is essential, considering both its potential and ongoing research. Strategic investment in this technology offers future-proof capabilities, surpassing current electronic and magnetic storage limitations.

The stability and longevity demonstrated by DNA make it a prudent choice for organizations seeking lasting data storage solutions. Incorporating DNA data storage diversifies an organization's data management strategy, mitigating risks associated with conventional technologies. This forward-thinking approach positions organizations as leaders in technological innovation, at the forefront of the convergence between biotechnology and information technology. Early adopters gain a competitive edge, setting industry standards in efficiency and technological prowess. The wide-ranging applications of DNA data storage make it invaluable across sectors.

From preserving historical records to managing vast research data, it offers unparalleled security and stability for finance and technology sectors. Investing in this technology acts as a catalyst for collaborative innovations, fostering partnerships between biotechnologists, data scientists, and IT professionals. Beyond its technological benefits, DNA data storage addresses environmental concerns, aligning with the increasing focus on sustainability and corporate responsibility. As organizations seek green technologies to reduce their carbon footprint, DNA data storage provides a sustainable solution.

However, the implementations must balance enthusiasm with pragmatic considerations. Current challenges, including high costs, slower access speeds, and technical complexities, require careful planning. A phased investment approach, starting with pilot projects and research initiatives, allows organizations to understand and leverage DNA data storage effectively. DNA data storage technology represents a strategic move towards innovative, sustainable data management. Positioned at the cutting edge of data-driven advancements, organizations adopting this technology are prepared to leverage its immense potential as it evolves into a pivotal solution for managing the ever-growing digital landscape.

10. Conclusions

DNA data storage represents a promising technology offering secure, dense, and long-term data storage solutions. Despite existing technical and cost-related challenges, the potential advantages of DNA data storage have led to intensive research and development efforts in this field. As the technology continues to advance, it is expected to become increasingly feasible and cost-effective, potentially addressing the growing demand for secure and long-term data storage in the future. The amount of data that can be stored in DNA is determined by various factors, including the size of the DNA sequences, the encoding method used, and the accuracy of the sequencing and data retrieval processes.

While the theoretical storage capacity of DNA is exceptionally high, reaching up to 1 exabyte per gram, practical limitations exist due to the current constraints of DNA synthesis and sequencing technologies, as well as the associated costs and complexities. However, ongoing research and advancements in DNA synthesis and sequencing technologies are likely to increase the amount of data that can be stored in DNA, aided by the integration of artificial intelligence (AI) in this domain. It is important to recognize that while DNA data storage holds the potential for very high storage densities, it may not be suitable for all data storage requirements.

Specifically, DNA data storage is best suited for large-scale data archives that necessitate long-term storage, such as scientific data and historical records. However, it may not be well-suited for applications demanding high-speed data retrieval or frequent data access and updates. These considerations highlight the need for continued research and development to address the challenges and maximize the potential of DNA data storage in diverse data storage contexts.

Supplementary information

The original imaging and the various types of data some of which are not publicly available, because they contain private information. The provided data that support the findings and information of the research investigations are referenced where appropriate.

Conflict of Interest

There are no Conflict of Interest or any type of Competing Interests for this research.

Author Contributions

Described in details within the Acknowledgements section.

Funding

No Funding was provided for the conduction of this research.

Acknowledgments

The main prospect and the scope for this research was conducted and idea perspective investigations with the manuscript writing was done by the author himself. All the datasets, data tools, data models, data sources which have been retrieved and used for the conduction of this research are mentioned and referenced where appropriate.

References

[1] Ceze, L., Nivala, J., & Strauss, K. (August 2019). Molecular digital data storage using DNA. *Nature Reviews. Genetics*. *20(8)*, 456–466. doi:10.1038/s41576-019-0125-3. PMID 31068682. S2CID 148570002.

[2] Akram, F., Haq, I. U., Ali, H., & Laghari, A. T. (October 2018). Trends to store digital data in DNA: An overview. *Molecular Biology Reports. 45(5)*, 1479–1490. doi:10.1007/s11033-018-4280-y. PMID 30073589. S2CID 51905843.

[3] Panda, D., Molla, K. A., Baig, M. J., Swain, A., Behera, D., & Dash, M. (May 2018). DNA as a digital information storage device: hope or hype? *3 Biotech. 8(5), 239*. doi:10.1007/s13205-018-1246-7.

PMC 5935598. PMID 29744271.
[4] Shankland, S. (29 June, 2019). Startup packs all 16GB of Wikipedia onto DNA strands to demonstrate new storage tech - Biological molecules will last a lot longer than the latest computer storage technology, Catalog believes. *CNET*.
[5] Roquet, N., Bhatia, S. P., Flickinger, S. A., Mihm, S., Norsworthy, M. W., Leake, D., & Park, H. (2021-04-20). DNA-based data storage via combinatorial assembly. bioRxiv: 2021.04.20.440194. doi:10.1101/2021.04.20.440194. S2CID 233415483.
[6] Smith, G. C., Fiddes, C. C., Hawkins, J. P., & Cox, J. P. (July 2003). Some possible codes for encrypting data in DNA. *Biotechnology Letters, 25(14), 1125–1130*. doi:10.1023/a:1024539608706. PMID 12966998. S2CID 20617098.
[7] Goldman, N., Bertone, P., Chen, S., Dessimoz, C., LeProust, E. M., Sipos, B., & Birney, E. (February 2013). Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. *Nature, 494(7435), 77–80*. Bibcode:2013Natur.494...77G. doi:10.1038/nature11875. PMC 3672958. PMID 23354052.
[8] Lee, H. H., Kalhor, R., Goela, N., Bolot, J., & Church, G. M. (June 2019). Terminator-free template-independent enzymatic DNA synthesis for digital information storage. *Nature Communications, 10(1), 2383*. Bibcode:2019NatCo..10.2383L. doi:10.1038/s41467-019-10258-1. PMC 6546792. PMID 31160595.
[9] Kalhor, R., Kalhor, K., Mejia, L., Leeper, K., Graveline, A., Mali, P., & Church, G. M. (August 2018). Developmental barcoding of whole mouse via homing CRISPR. *Science, 361(6405)*. doi:10.1126/science.aat9804. PMC 6139672. PMID 30093604.
[10] Lim, C. K., Yeoh, J. W., Kunartama, A. A., Yew, W. S., Poh, C. L. (July 2023). A biological camera that captures and stores images directly into DNA. *Nature Communications, 14(1), 3921*. doi:10.1038/s41467-023-38876-w. PMC 10318082. PMID 37400476.
[11] Rebrova, I. M., Rebrova, O. Y. (2020). Storage devices based on artificial DNA: the birth of an idea and the first publications. *Voprosy Istorii Estestvoznaniia i Tekhniki (in Russian). 41(4)*, 666–76. doi:10.31857/S020596060013006-8. S2CID 234420446.
[12] Extance A(September 2016). How DNA could store all the world's data. *Nature. 537(7618)*, 22–24. Bibcode:2016Natur.537...22E. doi:10.1038/537022a. PMID 27582204.
[13] Limbachiya, D., Dhameliya, V., Khakhar,. M, Gupta, M. K. (25 April 2016). On optimal family of codes for archival DNA storage. *Proceedings of 2015 Seventh International Workshop on Signal Design and Its Applications in Communications (IWSDA).* (pp. 123–127). arXiv:1501.07133. doi:10.1109/IWSDA.2015.7458386. ISBN 978-1-4673-8308-0. S2CID 7062541.
[14] Organick, L., Ang, S. D., Chen, Y. J., Lopez, R., Yekhanin, S., *et al.* (March 2018). Random access in large-scale DNA data storage. *Nature Biotechnology. 36(3), 242–248*. doi:10.1038/nbt.4079. PMID 29457795. S2CID 205285821.
[15] Patel, P. (2018-02-20). DNA data storage gets random Accesscf. *IEEE Spectrum: Technology, Engineering, and Science News*.
[16] Langston, J. (2019-03-21). With a "hello," Microsoft and UW demonstrate first fully automated DNA data storage. *Innovation Stories*.
[17] Appuswamy, R., Le, Brigand, K., Barbry, P, Antonini, M., Madderson, O., Freemont, P., McDonald, J., Heinis, T. (2019). OligoArchive: Using DNA in the DBMS storage hierarchy (PDF). *Proceedings of Conference on Innovative Data Systems Research (CIDR)*.
[18] Tabatabaei, S. K., Wang B., Athreya, N. B., Enghiad, B., *et al.* (April 2020). DNA punch cards for storing data on native DNA sequences via enzymatic nicking. *Nature Communications, 11(1)*, 1742.

Bibcode:2020NatCo..11.1742T. doi:10.1038/s41467-020-15588-z. PMC 7142088. PMID 32269230.

[19] Molteni, M, (2019-12-09). *These Plastic Bunnies Got a DNA Upgrade. Next up, the World? Wired.*

[20] Hotz, R. L. (2019-12-09). Scientists store data in synthetic DNA embedded in a plastic bunny. *Wall Street Journal*. Retrieved 2019-12-09.

[21] Han, S., Guo, X., Jin, X., & Shen, X. (1998). Methode a Technique Analytique de Code D'acide Nucleique Utilisee Dans des Etiquettes Infalsifiables. WO patent WO1998006084A1.

[22] Pierre, A., Bonnet, J., Vekris, A. & Portier, J. (2001). Encapsulation of deoxyribonucleic acid molecules in silica and hybrid organic-silica gels. *J. Mater. Sci. Mater. Med., 12*, 51–55.

[23] Gootenberg, J. S. *et al.* (2017). Nucleic acid detection with CRISPR-Cas13a/C2c2. *Science, 356*, 438–442.

[24] Qian, J. *et al.* (2020). Barcoded microbial system for high-resolution object provenance. *Science, 368*, 1135–1140.

[25] SmidgION. Oxford nanopore technologies. Retrieved 2023, from https://nanoporetech.com/products/smidgion

[26] Hashiyada, M. (2004). Development of biometric DNA ink for authentication security. *Tohoku J. Exp. Med., 204*, 109–117.

[27] Sharief, S. A., Chahal, P., & Alocilja, E. (2021). Application of DNA sequences in anti-counterfeiting: current progress and challenges. *Int. J. Pharm., 602*, 120580.

[28] Altamimi, M. J. *et al.* (2019). Anti-counterfeiting DNA molecular tagging of pharmaceutical excipients: An evaluation of lactose containing tablets. *Int. J. Pharm., 571*, 118656.

[29] Puddu, M., Paunescu, D., Stark, W. J. & Grass, R. N. (2014). Magnetically recoverable, thermostable, hydrophobic DNA/silica encapsulates and their application as invisible oil tags. *ACS Nano, 8*, 2677–2685.

[30] Hogan, M. E., Jung, L., Richter, N., Sun, Y. & Szczepanik, M. (2018). Nucleic acid coated submicron particles for authentication. WO patent WO2018156352A1.

[31] Ponnuswamy, N. *et al.* (2017). Oligolysine-based coating protects DNA nanostructures from low-salt denaturation and nuclease degradation. *Nat. Commun., 8*, 15654.

[32] Brown, J. & Reichert, B. (2019). Compositions for use in security marking. US patent US20120135413A1.

[33] Sleat, R. & Van Lint, G. (2014). Marking apparatus for nucleic acid marking of items. US patent US8783194B2

[34] Dollinger, G. D. (1990). Methods for tagging and tracing materials with nucleic acids. WO patent WO1990014441A1.

[35] Berk, K. L. *et al.* (2021). Rapid visual authentication based on DNA strand displacement. *ACS Appl. Mater. Interfaces, 13*, 19476–19486.

[36] Jung, L., Hayward, J. A., Liang, M. B. & Berrada, A. (2018). DNA marking of previously undistinguished items for traceability. US patent US9919512B2

[37] Crowley, J., Rademan, K. & Tan, E. (2022). The textile tracer assessment. Fashion for good reports. From: https://reports.fashionforgood.com/report/textile-tracer/chapterdetail?reportid=712&chapter=8

[38] Securing Industry. (2018). Inkjet codes with sigNature® Molecular tags for supply chain traceability and brand protection. From: https://www.securingindustry.com/free-papers-and-media/inkjet-codes-with-signature-molecular-tags-for-supply-chain-traceability-and-brand-protection/s60/a7378/

[39] Mead, B. K. *et al.* (2014). The role of DNA-stained currency in gang robberies within the United Kingdom. *J. Forensic Sci. 59*, 264–267.

[40] Sniderman, D. (2017). Are synthetic chemical markers needed in lubricants? *Tribol. Lubr. Technol., 73*,

38–46.

[41] Ullrich, C., Luescher, A. M., Koch, J., Grass, R. N. & Sax, H. (2022). Silica nanoparticles with encapsulated DNA (SPED) to trace the spread of pathogens in healthcare. *Antimicrob. Resist. Infect. Control, 11*, 4.

[42] DENIMSANDJEANS. How soorty is leveraging haelixa DNA to get brands into traceability—Denimandjeans | Global trends, news and reports | Worldwide. From: https://www.denimsandjeans.com/news/how-soorty-is-leveraging-haelixa-dna-to-get-brands-into-traceability/58183 (2022).

[43] Paunescu, D., Puddu, M., Soellner, J. O. B., Stoessel, P. R. & Grass, R. N. (2013). Reversible DNA encapsulation in silica to produce ROS-resistant and heat-resistant synthetic DNA 'fossils'. *Nat. Protoc., 8*, 2440–2448.

[44] SelectaDNA. (2023). SelectaDNA forensic coding. From: https://www.selectadna.co.uk

[45] Pakendorf, M. (2019). Combinatorial DNA barcoding for security applications, Doctoral Dissertation, University of Leicester, Leicester.

[46] Bonnell, A. (2010). *Covert Tagging*, Pira International Ltd, United Kingdom,

[47] Brand Integrity International. (2023). DNA smartMarks—BrandProtection. From: https://brandintegrity.com.au/dna-smartmarks

[48] Holoptica. (2023). DNA technology. From: https://holoptica.com/dna-technology/

[49] Akhtar, Z. B. (2024). The design approach of an artificial intelligent (AI) medical system based on electronical health records (EHR) and priority segmentations. *J. Eng.*, 1–10 https://doi.org/10.1049/tje2.12381

[50] Akhtar, Z. B., & Gupta, A. D. (2024). Advancements within molecular engineering for regenerative medicine and biomedical applications an investigation analysis towards a computing retrospective. *Journal of Electronics, Electromedical Engineering, and Medical Informatics, 6(1)*, 54-72. https://doi.org/10.35882/jeeemi.v6i1.351

[51] Akhtar, Z. B., & Gupta, A. D. (2024). Integrative approaches for advancing organoid engineering: From mechanobiology to personalized therapeutics. *Journal of Applied Artificial Intelligence, 5(1)*, 1–27. https://doi.org/10.48185/jaai.v5i1.974

[52] Akhtar, Z. B. (2023). Accelerated computing a biomedical engineering and medical science perspective. *Annals of the Academy of Romanian Scientists Series on Biological Sciences*, *12(2)*, pp. 138–164. Academia Oamenilor de Stiinta din Romania. https://doi.org/10.56082/annalsarscibio.2023.2.138