UNIVERSITAT ROVIRA i VIRGILI

DNA-Based Data Storage: A Bibliographic Review of Cryptographic, Steganographic, and Cellular Recording Applications

Aleksandra Anna Chaczko Korzeniowska

Final project for the bachelor's degree in Biotechnology

Academic mentor:
Dr. Albert Bordons de Porrata-Doria, Emeritus Professor, Dpt. Biochemistry and Biotechnology, albert.bordons@urv.cat

Rovira i Virgili University, Tarragona, Spain
June 2025
Double Degree in Biochemistry and Molecular Biology and Biotechnology

Table of contents

1.- Abstract....3
2.-Introduction....4
2.1-The Evolution of Data Storage: Challenges in the Digital Era....4
2.2.- DNA as a Data Storage Medium....5
2.3.- Key Milestones in DNA Data Storage....5
2.4.- Encoding, Storage, and Retrieval of Information in DNA....6
2.4.1.-*In Vivo* Storage....7
2.4.2.-*In Vitro* Storage....8
2.4.3.- Encapsulated and Protective Preservation....9
2.5.-Cryptography and Steganography in DNA Storage....9
2.6.-*In Vivo* DNA Storage and Cellular Applications....10
3.- Hypothesis and Objectives....12
4.-Methodology....12
5.-Results and Discussion....13
5.1.-Cryptography and Steganography Approaches....13
5.1.1.- Recombinant DNA Technique....12
5.1.2.-CRISPR-Cas9 Technique....14
5.2.- Molecular Recording Devices: Strategies for Cellular Information Storage....16
5.2.1.-Recombination-based Cellular Recording....16
5.2.2.- Implementation of Genome Editing for Molecular Recording....19
5.2.3.- DNA Methylation for Biological Recording....24
6.-Conclusions....25
7.-Self-Assessment....25
8.-Bibliography....26

1.- Abstract

DNA has specific characteristics and properties that make it a powerful medium for information storage, with the potential to replace current data storage systems. DNA-based data storage has many applications, not only for preserving digital information, but also for transmitting it, such as in cryptography and steganography. Other applications are related to cellular processes, like recording the cellular environment or studying cell lineages and their evolution. This project reviews the properties of DNA that enable it to work as an information storage system, how data can be encoded and later retrieved, and presents some of the most common current applications. Finally, an example of each application is presented, along with the main limitations of each technique and possible future developments.

Keywords: DNA-based data storage, *in vivo*, cryptography, steganography, cellular tracking, cell lineages.

2.-Introduction

2.1-The Evolution of Data Storage: Challenges in the Digital Era

For centuries, various kinds of materials have been used as media for recording information, ranging from wood, bone, and stone to the revolutionary advancements of computer science and digital information storage. The digital storage of information in magnetic, optical, and electronic media, as well as its transmission through the internet, has promoted the expansion of next-generation science, technology, and the arts (Dong et al., 2020). However, according to the International Data Corporation, global data storage demand was expected to reach 175 ZB (1 Zettabyte = 10^{21} bytes = 10^{12} GB, which is 1.75×10^{14} GB) by 2025, thereby putting existing storage technologies under considerable strain (Reinsel et al., 2018). The storage density required to accommodate such an amount of information far exceeds the capacity of any current system. Current storage techniques are increasingly prone to obsolescence and decay, while their non-biodegradable materials contribute to environmental pollution (Shrivastava & Badlani, 2014). Additionally, the high costs associated with data maintenance and transfer, combined with limited lifespans and frequent data losses, further emphasize the urgent need for novel information storage solutions (Dong et al., 2020).

Key aspects of storage media include density (bits per unit of physical volume), retention (the period during which data remains recoverable), access speed (latency and bandwidth for data retrieval), and the energy cost associated with storing and accessing data (Ceze et al., 2019). Scientists and researchers have explored various methods for storing data across different media, with molecular data storage emerging as a particularly attractive approach. Inspired by the solution that nature has employed for over three billion years, they have turned to deoxyribonucleic acid (DNA) as a medium for information storage (Dong et al., 2020).

2.2.-DNA as a Data Storage Medium

The concept of DNA-based data storage emerged in the 1960s, introduced by computer scientists and engineers after the seminal work of Watson and Crick on DNA as the carrier of genetic information (Davis, 1996). Nevertheless, the technological limitations in DNA synthesis and sequencing at that time prevented the practical implementation of this idea.

The main reasons why DNA is considered a promising medium for the next generation of digital information storage are as follows (Ping et al., 2019). DNA exhibits remarkable stability compared to conventional storage media, owing to its double-helix structure and base-stacking interactions. It endures significantly longer than silicon-based devices and, when stored under controlled conditions—protected from light, humidity, and extreme temperatures—it can remain intact for centuries or even millennia, as demonstrated by the successful retrieval of DNA from fossils thousands of years old (Ceze et al., 2019). Moreover, DNA possesses a high storage density. 1 gram of single-stranded DNA can store up to 455 exabytes (EB = 10^{18} bytes) of data, offering a density of up to 10^{18} bytes per mm^3, which is approximately six orders of magnitude denser than the densest media available today (Ceze et al., 2019). Regarding

accessibility, DNA can be easily and rapidly replicated through PCR, which represents a significant advantage given that maintaining and organizing digital data consumes a substantial amount of time (Ceze et al., 2019). Last but not least, currently, sequencing and chemical synthesis technologies enable us to read and write the information stored in DNA (Ping et al., 2019). Thus, the use of DNA for data storage is up to a million times more efficient than that of a modern personal computer.

2.3.- Key Milestones in DNA Data Storage

The idea of using DNA for digital storage was first indirectly implemented in 1988, when Joe Davis converted the image *Microvenus* into a binary sequence and encoded it into a 28-base-pair-long DNA molecule, which was then inserted into *Escherichia coli* (Davis, 1996). Following Davis's project, most subsequent efforts up until 2012 focused on *in vivo* data storage, meaning the information was stored within living cells. This approach appears to have been a strategic choice, as it facilitated DNA sequencing at the time (Ceze et al., 2019). Other *in vivo* approaches have been proposed, along with some *in vitro* methodologies that focus not only on data storage but also on information encryption.

However, these early attempts at DNA as data storage only had a capacity of less than tens of Bytes—a small amount of data with little scalability for practical usages (Dong et al., 2020). It was not until the work of Church and Goldman (Church et al., 2012; Goldman et al., 2013) that larger amounts of data were stored successfully in DNA molecules, reaching 739 KB. Moreover, the data stored in both studies contained not only texts but also images, sounds, PDFs, etc., which confirmed that DNA can store a wide variety of data types. In **Figure 1**, a timeline of experiments with data volumes is shown. There is a clear exponential rate of progress in capacity when jumping from *in vivo* experiments to *in vitro* ones (Ceze et al., 2019). Many other complex methods were developed, each of them following a different procedure to encode the message in a DNA sequence. By the end of 2018, the amount of data stored in DNA exceeded 200 MB. Along with the development of DNA synthesis and sequencing techniques, new DNA storage methods kept emerging (Dong et al., 2020).

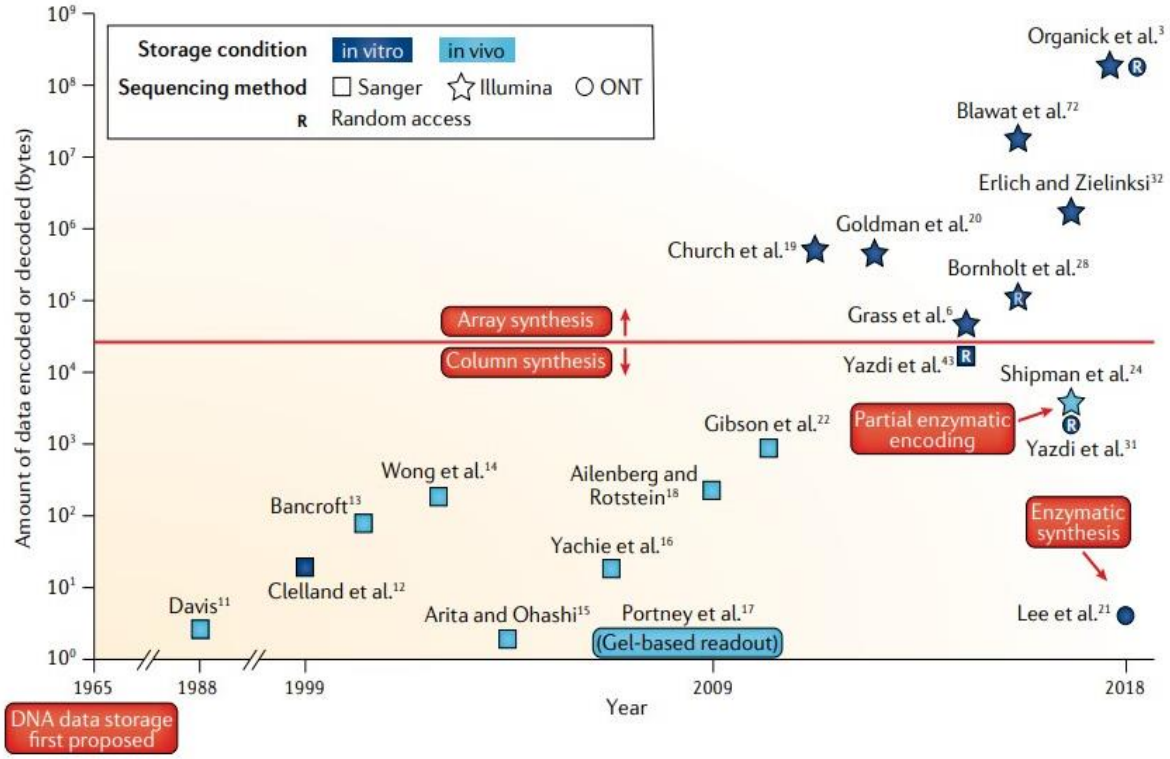

Figure 1. Timeline of major published works on digital data storage with DNA. A timeline of DNA-based storage studies, including how the DNA data were synthesized, stored, and read, and whether retrieval supported random access. Extracted from (Ceze et al., 2019).

2.4.- Encoding, Storage, and Retrieval of Information in DNA

DNA data storage, regardless of whether it is *in vivo* or *in vitro*, consists of four major steps, as illustrated in **Figure 2**. First, digital data in binary format is encoded into a DNA sequence using a specific algorithm. Arithmetic coding is also introduced to reduce the redundancy and compress the data, improving the coding efficiency (Zhang et al., 2022). The resulting DNA sequence is then synthesized, and multiple copies are produced. Solid-phase synthesis, based on phosphoramidite chemistry, can be performed either on a column (low-throughput) or an array-based solid support (high-throughput). Since DNA sequences have a limited length, the data must be divided into smaller fragments that are later reassembled. To facilitate this reassembly, an index can be included in each fragment, or overlapping segments can be distributed across different DNA sequences. Once synthesized, the DNA sequence is cloned and stored within a biological cell (*in vivo*) or, more commonly, stored *in vitro*, such as being frozen in solution or dried down. Robust preservation methods slow down DNA degradation and improve the recoverability of information. When the data stored in the DNA is requested, the corresponding DNA needs to be retrieved and sampled. The ability to choose specific data from a large set can be performed via magnetic bead extraction with probes mapped to data items or PCR using primers associated with the encoding process. Finally, automated sequencing instruments generate the reads that correspond to the detected molecules. The most common sequencing methods are Sanger (low-throughput), sequencing-by-synthesis

instruments (high-throughput, for example, by Illumina), and nanopore sequencing (for instance, Oxford Nanopore Technologies (ONT)). The reads are decoded back into the original format of the message (Ceze et al., 2019; Zhang et al., 2022).

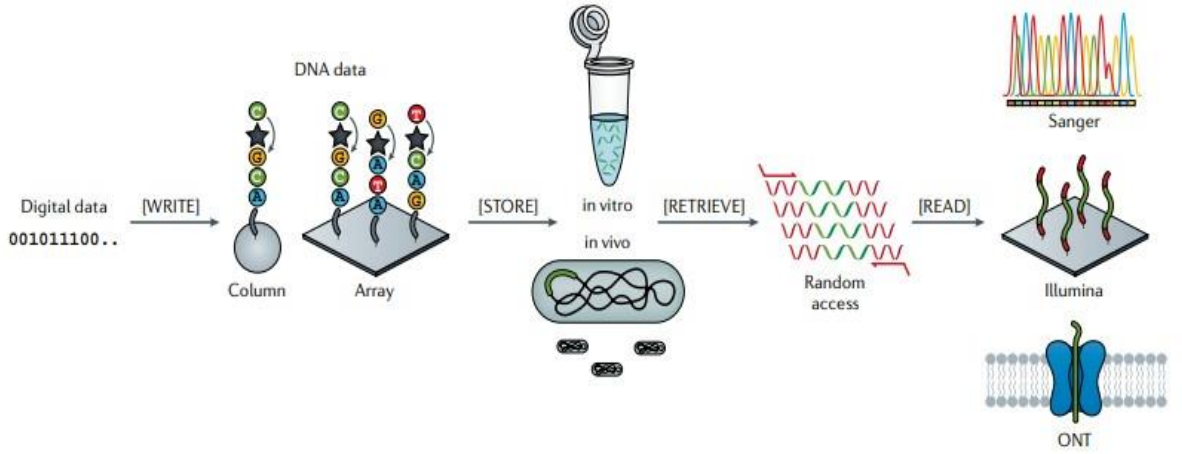

Figure 2. Overview of the major steps of digital data storage in DNA. Extracted from (Ceze et al., 2019).

However, synthesis and sequencing of DNA are error-prone processes. Studies have reported an error of approximately 1% per base position, i.e., for a given position in a DNA sequence, after synthesizing and sequencing, 1% of the total reads will have an error in that position. Moreover, DNA manipulation, PCR, and storage may cause data loss, resulting in an under-representation of certain sequences in the mix. Thus, it is common to apply error-correcting codes on raw storage media. A whole field of computer science is dedicated to developing coding schemes that deliver digital data reliably over noisy media and communication channels. As for DNA correcting codes, besides correction of substitution errors, there are also codes for insertions and deletions, which makes the process even more challenging (Ceze et al., 2019). Examples of such codes are Fountain code, Rapid Tornado Code (RaptorQcode), or Reed-Solomon code (RS code), which can recover possible chain loss, base substitution, deletion, and other errors (Zhang et al., 2022).

2.4.1.-*In Vivo* Storage

Synthetic DNA, as genetic information, can be stably preserved and replicated in cells and bacteria. As mentioned at the beginning, Joe Davis (Davis, 1996) initiated the preservation of information *in vivo*. Later on, to address the issue of storage capacity in a living system, different approaches have been introduced. For instance, a mixed culture method of bacterial cells carrying large oligonucleotide pools of 2304 Kbps synthetic DNA, which corresponds to 445 KB, was developed (**Figure 3A**) (Hao et al., 2020). One year later, a 254,886 bp yeast artificial chromosome (YAC), encoding 37,782 bytes of digital data (**Figure 3B**), was designed (W. Chen et al., 2021). Apart from importing exogenous information libraries, other information writing methods based on recombinase and clustered regularly interspaced short palindromic repeats/CRISPR-associated (CRISPR-Cas) have also been developed (Zhang et al., 2022), which will be discussed in more detail later.

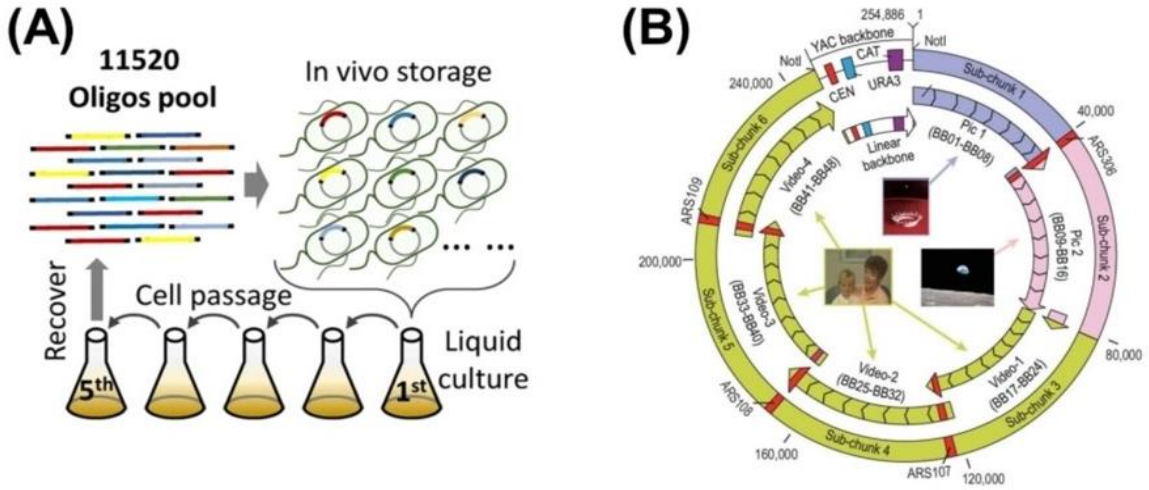

Figure 3. *In vivo* storage approaches. (A) The mixed culture of bacterial cells carrying oligo information pools and (B) the ultralong artificial yeast artificial chromosome (YAC) information platform. Extracted from (Zhang et al., 2022).

However, it is unlikely that *in vivo* DNA data storage will become a viable alternative for exogenous digital data storage. This is primarily due to its low storage density, limited by the biological constraints of living cells, and the complexity involved in modifying or integrating artificial DNA with natural genomic DNA. Nevertheless, scientists have found ways to leverage this *in vivo* DNA's capacity for data storage. *In vivo* DNA data recording enables applications such as logging cellular history and environmental changes, and even synthesizing "molecular tapes" that record dynamic, time-ordered molecular events within cells, which can later be read and analyzed (Ceze et al., 2019). Further details on this topic will be provided in a later section.

2.4.2.-*In Vitro* Storage

Since, as mentioned earlier, *in vivo* approaches have limitations, DNA data is usually chemically synthesized and stored in a single physical DNA pool, mainly in liquid or solid form. The liquid state allows for fast access and reading of the DNA; however, it is prone to damage and degradation. Cryopreservation can prolong storage time, but it increases energy consumption and equipment requirements. Moreover, repeated freezing and thawing cycles may cause DNA strand breakage and data loss. Since water and oxygen affect the integrity of the DNA, dehydrated storage on filter paper was a common technique for preservation and transportation. Nowadays, commercial preservation plates that inactivate bacteria, molds, etc., are available on the market. Most of the provided strategies include systems to read the stored information directly. For instance, indium tin oxide-coated glass plates with electrodes or Quick Response (QR)-encoded micro-disks with primer sequences and file names of the DNA information, facilitating the index and access (Zhang et al., 2022).

2.4.3.- Encapsulated and Protective Preservation

Inspired by nature and fossils, strategies of encapsulating DNA information molecules in inorganic components have been developed. An example is the silica shell encapsulation strategy, which enables DNA molecules to withstand temperatures up to 200 ºC and reactive oxygen species. Its storage lifetime at -18 ºC was predicted to be more than 2 million years by extrapolation of the Arrhenius equation. A less time-consuming method with unlimited capacity involved encapsulating DNA in fossil-mimicking materials. Three types of alkaline earth metal salts (CaP, CaCl2, and MgCl2) were co-precipitated with DNA. Moreover, the recovery of the information was achieved by simply dissolving the coprecipitated salt. These strategies, based on inorganic components, effectively enhance the storage lifetime of DNA data. Current research focuses on improving DNA stability and enabling long-term storage at room temperature (Zhang et al., 2022).

2.5.-Cryptography and Steganography in DNA Storage

The field of cybersecurity protects the stored and transmitted information from different types of threats and attacks. The attacks on shared information may cause it to be disrupted, corrupted, or stolen. Cryptography and steganography are techniques that securely convey information from source to destination. Cryptography aims to alter the message in such a way that it becomes unreadable to a third party (Alhabeeb et al., 2021). However, the complicated form of the encrypted message indicates the existence of valuable information and may arouse the suspicions of hackers and reduce security. By contrast, steganography is the science of hiding information in a regular carrier, the security of which is achieved by concealing the existence of the message (Huang et al., 2023).

One of the many applications of using DNA as data storage is DNA-based cryptography and steganography. DNA is proposed as the ultimate concealing medium that avoids or mitigates the drawbacks of other media. As shown in **Figure 4**, three components are combined to provide security for sensitive information transmitted over an untrusted channel. To embed the hidden message in the cover medium, three main techniques can be used, which are insertion, substitution, and complementary rules-based algorithms (Alhabeeb et al., 2021).

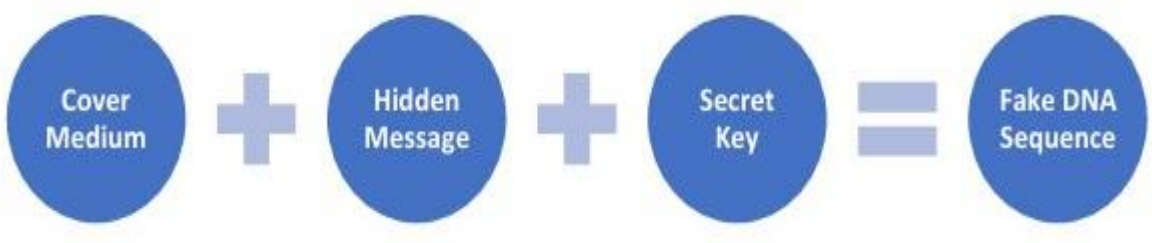

Figure 4. DNA steganography process. Extracted from (Alhabeeb et al., 2021).

2.6.-*In Vivo* DNA Storage and Cellular Applications

Although the storage capacity is limited, in comparison to storage *in vitro*, *in vivo* storage takes advantage of the efficient cellular machinery of DNA replication, proofreading, and long-chain DNA maintenance. The compatibility between DNA data storage and living systems has prompted the development of DNA-based cellular recording techniques (**Figure 5A**), which support live recording of biochemical events *in situ* in living organisms as a generalized concept of information storage (Dong et al., 2020). By applying novel synthetic biology techniques, engineered organisms can record cellular events using DNA as a recording medium (David et al., 2023). Numerous applications of DNA-based cellular recording have been demonstrated to date, such as the diagnosis of disease biomarkers (Loveless et al., 2021), capturing horizontal gene transfer events (Munck et al., 2020), tracking cellular lineages throughout embryonic development (Chan et al., 2019), and constructing genetic circuits for therapeutic applications (Kempton et al., 2022) (**Figure 5B**).

Changes in the genetic information of organisms are possible due to the development of gene editing technologies, such as natural and engineered DNA targeting and modifying enzymes, which can serve as writing modules in DNA storage systems. Other approaches, such as temporal RNA-seq and biosensors, have been proposed to perform biological tracking. However, DNA-based storage stands out due to its advantageous characteristics (Jang & Yim, 2024). Data stored in DNA has the advantage of being permanent within a cell, while transcriptome circuits that store information at the RNA or protein level need constant input of energy, meaning that cell death results in information loss (Ceze et al., 2019).

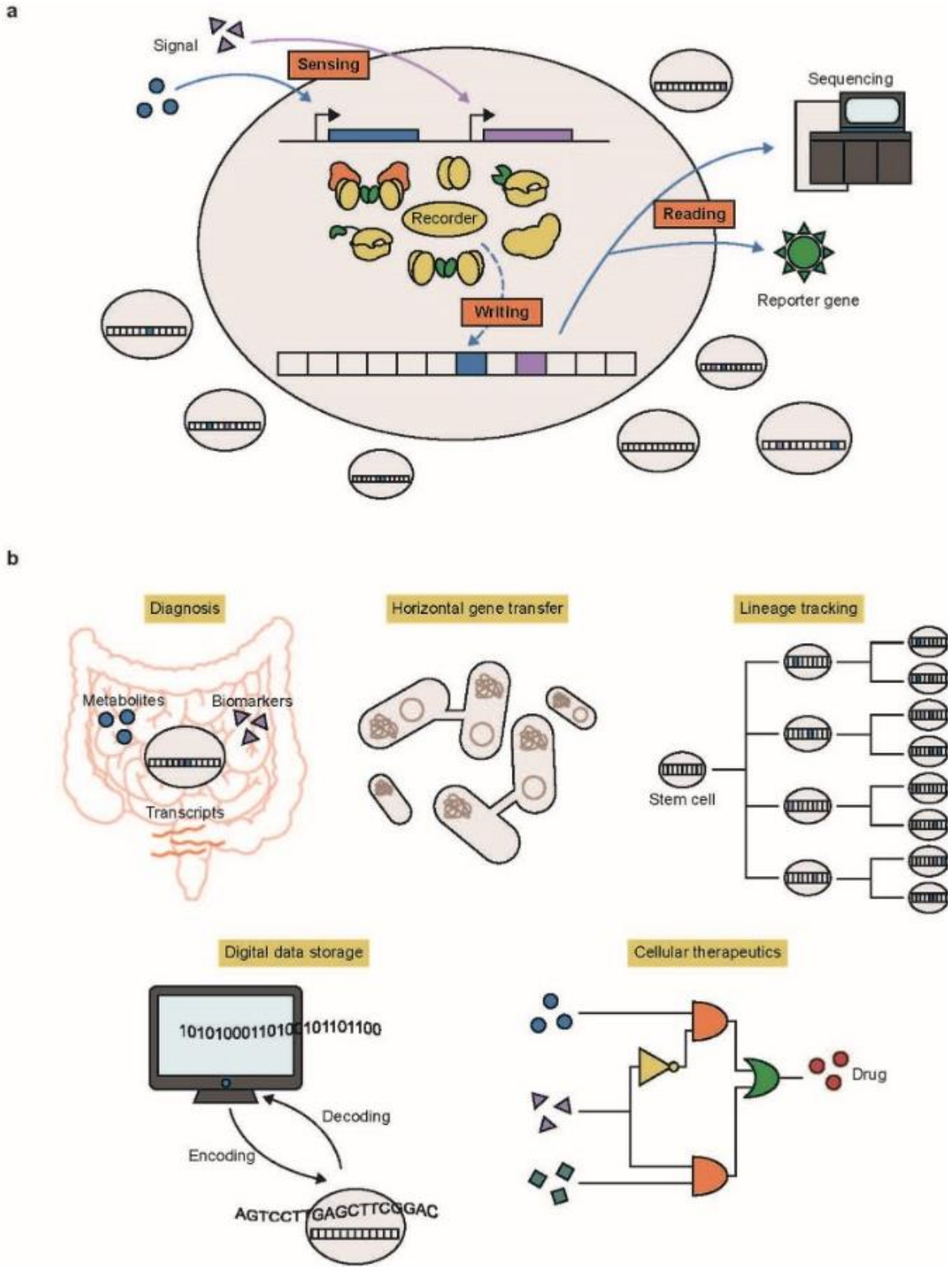

Figure 5. In vivo DNA-based storage and its cellular applications. (A) In DNA-based cellular recording, various environmental or cellular signals activate molecular recorders. Once activated, these recorders alter the DNA sequence or epigenetic states to store the data. The recorded data can be retrieved through sequencing or reporter gene expression. (B) Examples of DNA-based cellular recording applications. Extracted from (Jang & Yim, 2024).

3.- Hypothesis and Objectives

The hypotheses of this project are:

- DNA can function as both a storage and concealment medium for cryptographic and steganographic purposes.

- DNA-based storage applied to cellular tracking enables the recording and study of biochemical events.

The main objectives of this study are:

- To review and analyze *in vivo* methodologies for DNA-based cryptography and steganography, highlighting their respective advantages and limitations.

- To describe and compare current genetic engineering strategies used for DNA-based cellular memory and lineage tracking.

4.-Methodology

Three online academic platforms were used to gather relevant literature: PubMed, Web of Science, and Google Scholar. The initial search strategy involved two keywords and their related terms, which were DNA data storage (“DNA-based data storage”, “DNA digital storage”) and encryption techniques (“encoding”, “decoding”, “cryptography”, “steganography”). The selected date range was from January 1, 2019, to April 1, 2025. However, due to limitations encountered after performing the first search, the search scope was subsequently expanded. A second set of keywords was used, covering both DNA data storage (“DNA-based data storage”, “DNA digital storage”) and cellular recording (“in vivo data storage”, “molecular event tracking”, “biological data recording”, “cellular tracking”). This extended search covered the period from January 1, 2019, to May 1, 2025, which was the last date of information retrieval. Some earlier literature was included for introductory purposes.

5.-Results and Discussion

Initially, the main topic to be discussed was the methods of cryptography and steganography. However, since their use is more common *in vitro* than *in vivo*, not many studies were found on this subject. Therefore, the focus was broadened to include various methods for cellular tracking, aiming to cover a wider range of techniques and relevant applications within the fields of cellular biology and biotechnology related to DNA-based storage.

5.1.-Cryptography and Steganography Approaches

5.1.1.-Recombinant DNA Technique

The Double Sequence Recombinant and DNA Technology Method (DSRDTM) approach was developed for encryption and hiding of messages (Wang et al., 2019). First, to hide the message within the DNA sequence, the information was encoded following a coding rule. Then, applying an injective mapping rule table, which is a transformation rule, the DNA-encoded message was transformed into a "fake" DNA sequence. Thus, the secret message was transformed twice. Next, the DNA sequence, additionally with selectable markers, was embedded into a plasmid through recombination. Finally, the recombinant DNA was inserted into *Escherichia coli*. The cells were amplified to reach a detectable level and were concealed with dummy cells. The cells of interest could be screened by selective cultures (Wang et al., 2019). A diagram and procedure for this data hiding approach are shown in **Figure 6**.

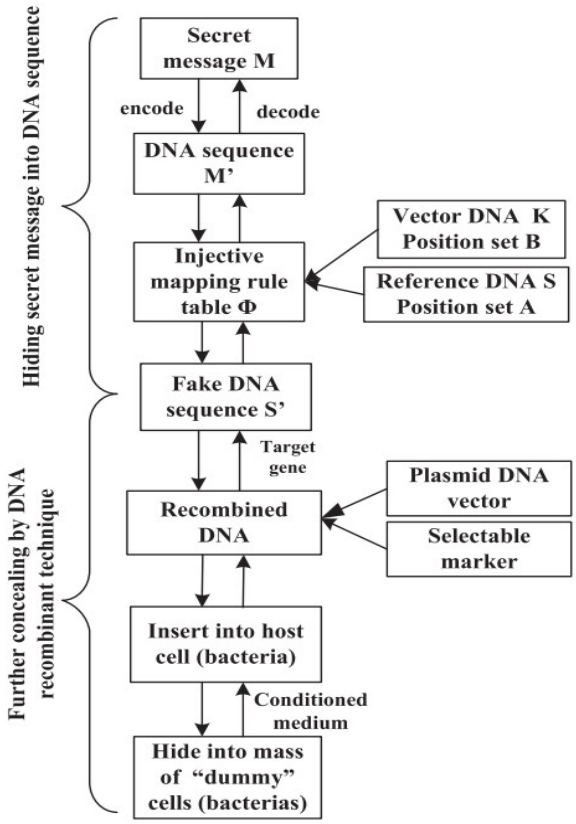

Figure 6. Procedure for hiding information in DNA. Extracted from (Wang et al., 2019).

An experimental procedure was carried out to demonstrate that a message can be successfully embedded within bacterial cells and later decoded through DNA sequencing and the application of specific algorithms. In this case, the encoded message was “CHINA”. The security of the system was evaluated by calculating the probability of breaching each step, revealing that the method possesses high robustness, stability, and security. These findings underscore its potential for cryptographic and steganographic applications in DNA-based storage systems (Wang et al., 2019).

The main limitation of this approach, often found in the *in vivo* studies, lies in the biological capacity of the host organism to embed and maintain the encoded message. In the referenced experiment, only a short message was encoded. However, practical applications would likely require the storage of longer sequences, making it essential to consider the volume of information that can be reliably stored. Importantly, the study did not address the maximum storage capacity achievable with this method. Morever, message length is not the only constraint. For efficient encoding and proper functioning of the designed sequences, the synthetic DNA must meet several criteria, including CG content, melting temperature, Hamming distance, a distance to avoid hybridization reactions, and free energy requirements (Wang et al., 2019). These parameters significantly restrict the pool of viable sequences.

Among the disadvantages of the approach, a fundamental concern is the need to transmit additional information, beyond the host cells themselves, to the recipient. Specifically, elements such as the selectable marker, restriction enzymes, and the injectable mapping rule table are required for successful message retrieval (Wang et al., 2019). Thus, even though the method is robust and secure, interception of these elements could lead to unauthorized decoding. Another major limitation is the potential loss of the message due to mutations in the host genome or the death of the host cells. To minimize this risk, strict control over experimental conditions is necessary. Moreover, the inserted sequence must be carefully designed to avoid disrupting essential genes or interfering with host metabolism. Assessing the biological impact of the embedded DNA is critical to ensuring cell viability and stability of the stored information.

5.1.2.-CRISPR-Cas9 Technique

As in the previous approach, most DNA-based hiding methods in living cells rely on the insertion of the message into a specific position within the plasmid. However, with the advancement of next-generation sequencing (NGS) techniques, it is now possible to perform DNA *de novo* sequencing or metagenomic sequencing without the need for any prior sequence information. This implies that an attacker could potentially identify the location of the hidden message through NGS and alignment with reference genomes from natural cells (M. Chen, 2019). To avoid this outcome, in this study, they proposed to hide the message in different sites of the host cells using CRISPR-Cas9. The process followed by the sender, the receiver, and a possible attacker is visualized in **Figure 7**. Briefly, the information is encoded into DNA code, and then, using a substitution algorithm, transformed into the secret message (SM), which is inserted along with other random sequences into the DNA of the host. The receiver obtains the

cells and retrieves the SM by PCR with the key to decode and recover the initial information (M. Chen, 2019).

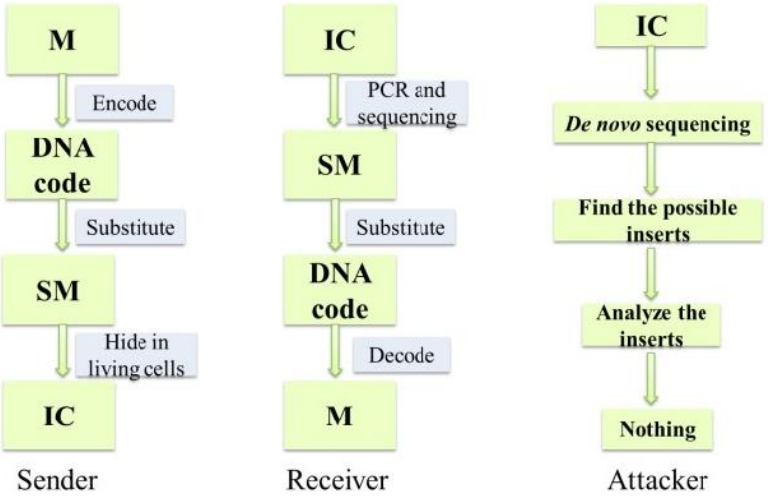

Figure 7. Flowchart of the sender, the receiver, and the attacker. Extracted from (M. Chen, 2019).

The theoretical example and experimental procedure were explained. Specifically, two strains were produced: one containing the secret message, which was divided into two parts (SM_M and SM_H) and ligated with random sequences in between (PM1, PM2, and PM3), which were inserted into the plasmid using restriction enzymes; and the other containing a random sequence (PM4), constructed using the CRISPR-Cas9 system (M. Chen, 2019). A schematic representation is shown in **Figure 8**.

.

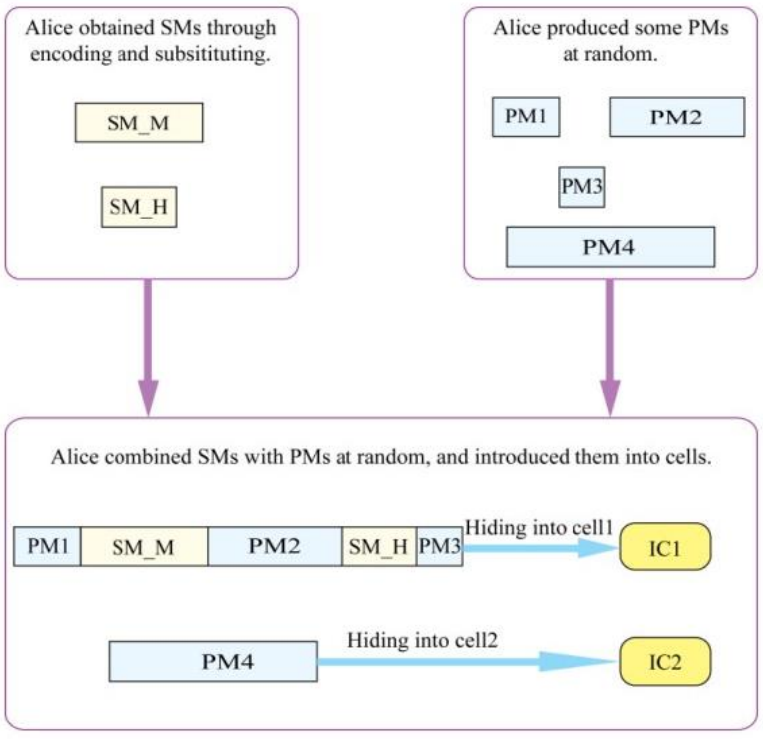

Figure 8. Production of information cells (IC) by the sender ("Alice"). Extracted from (M. Chen, 2019).

In comparison to the previously described method, the main advantage of this approach is the proposed solution for the capacity storage in the host cell. If the DNA code is longer than a certain length, it can be divided into several groups, which are substituted individually, and an individual key index is created. Following this method, messages of any length can be encoded. However, two drawbacks were defined for this approach. The first one is the expenses associated with the DNA synthesis and sequencing, which can reach $2,000. The second one is the time required for the process to be completed, which can take several weeks (M. Chen, 2019).

Although the cost and time required for DNA synthesis and sequencing have been significantly reduced thanks to advances in biotechnology, DNA-based cryptography and steganography are still more commonly implemented *in vitro* rather than *in vivo*. The main advantages of *in vitro* DNA storage include the ability to perform high-throughput oligonucleotide synthesis, a lower error rate due to the absence of mutation-inducing cellular processes, and greater ease of handling and manipulation in laboratory conditions. However, this approach also presents notable disadvantages, such as higher maintenance costs and a faster degradation rate of DNA molecules when stored outside of living systems (Ping et al., 2019).

Due to the limited amount of information available regarding DNA-based cryptography and steganography procedures carried out *in vivo*, the scope of this project has been broadened. Consequently, the following section will explore cellular tracking approaches that utilize *in vivo* DNA-based storage.

5.2.- Molecular Recording Devices: Strategies for Cellular Information Storage

5.2.1.-Recombination-based Cellular Recording

Recombination-memory biosensors represent one of the earliest innovations of synthetic cellular recording (Kalvapalle et al., 2024). The site-specific recombinases recognize flanking sequence motifs and enzymatically cleave, invert, and integrate the intervening region of DNA (Ceze et al., 2019). Thus, the presence, absence, or orientation of the DNA segment is used to interpret the cellular environment (David et al., 2023). Generally, the readout process is performed by sequencing, PCR, or reporter protein, whereas the writing of the molecular information is controlled chemically with inducers that activate the expression of the recombinase (Ceze et al., 2019).

Two types of recombinases can be distinguished: serine integrases and tyrosine integrases. Serine integrases invert or excise site-specific DNA in an irreversible manner depending on the orientation of the flanking sites. The sites, known as attB and attP, when found in the same orientation, result in the excision of the in between DNA; if found in the opposite orientation, the DNA region is inverted. This recombinase system has been used to design complex genetic circuits, including history-dependent logic in bacteria and lineage tracing in animal models. In contrast, tyrosine recombinases, such as FLP (flippase) and Cre, act on two identical sites, catalyzing reversible excision, integration, and inversion (Guiziou et al., 2023). Specifically,

Cre recombinase recognizes and acts on DNA sequences called *lox* sites. When these *lox* sites are oriented in the same 5' to 3' direction, Cre excises the DNA segment located between them (Garabello et al., 2024).

A serine recombinase-based approach was developed to assess the endurance of recombinase-based designs (Kalvapalle et al., 2024). In the study, memory sensors were created to record and retain binary information upon exposure to specific analytes in *Escherichia coli* (**Figure 9**). Concretely, the used analytes were sugar arabinose and the signal 3-oxo-C12-homoserine lactone (C12 AHL), a quorum-sensing molecule used in microbial communication. For this purpose, two plasmids were designed: the first plasmid encodes the recombinase under the control of a promoter inducible by the presence of the analyte; the second plasmid serves as the memory unit, where recombinase activity is recorded through the inversion of a DNA sequence, which results in the activation or deactivation of the expression of a reporter protein (concretely, the green fluorescent protein, GFP). Additionally, to the GFP signal, a more direct method for reading out the memory, qPCR was used to quantify the amount of DNA in each state (Kalvapalle et al., 2024).

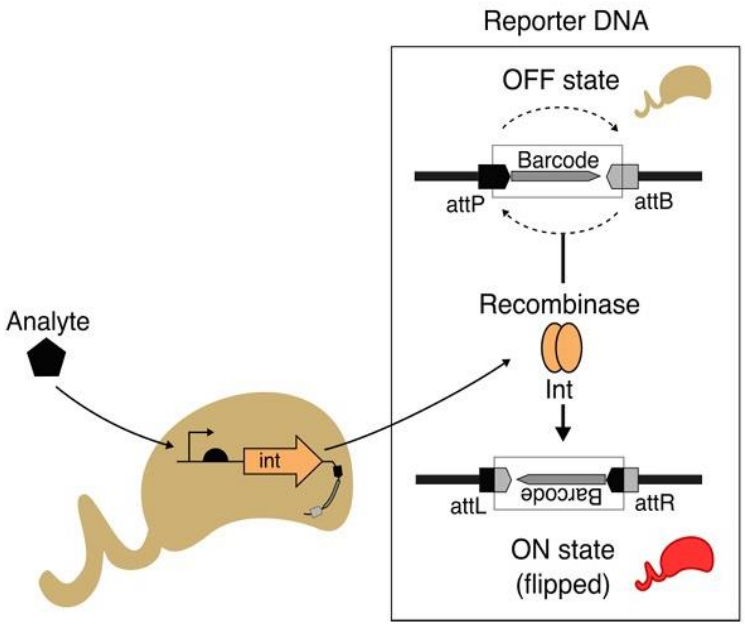

Figure 9. A recombinase-memory biosensor works by conditionally expressing a recombinase enzyme when it senses an analyte. The recombinase binds the DNA attachment sites attP and attB and reverses the DNA segment flanked by the sites, thereby encoding information as a genetic memory in the DNA. Extracted from (Kalvapalle et al., 2024).

The results showed that although increasing analyte concentrations initially led to a higher fraction of the population in the ON state, a downward trend was observed at the highest concentration, due to system saturation and stress. Moreover, after performing serial batch cultures, the researchers discovered that both ON and OFF states were unstable over time. A background ON state signal was detected in absence of analyte due to sufficient baseline recombinase expression, and, when the analyte was added, the ON state signal decreased

notably after the second day in the case of arabinose memory and around the fourth day for the C12-AHL memory due to the burden arised from fluorescent protein expression (Kalvapalle et al., 2024). A diagram explaining the instability of the OFF and ON states is shown in **Figure 10**.

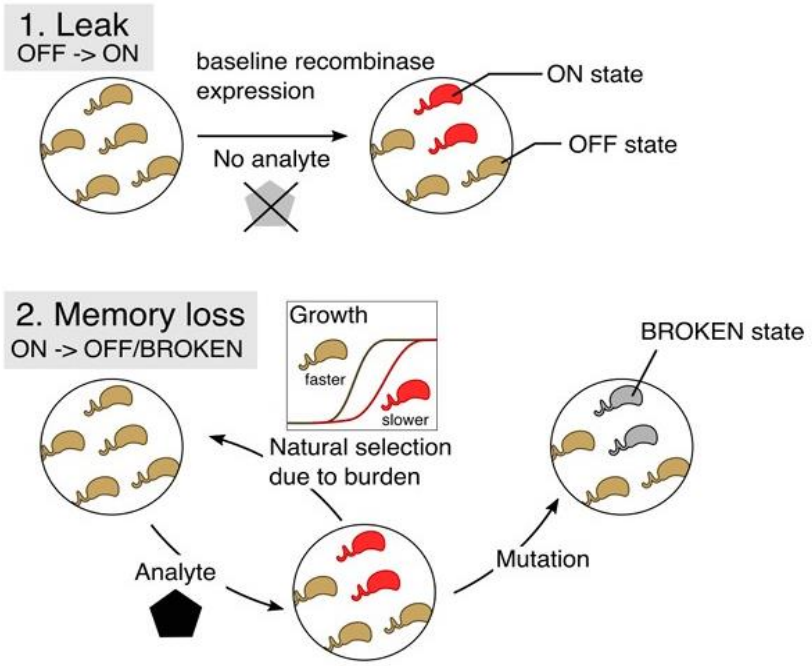

Figure 10. Mechanisms predicted to underlie memory instability. (1) Leaky recombinase expression in the absence of the analyte is predicted to yield a signal in the OFF state (brown). With time, the recombinase accumulates to sufficient levels to flip the memory to the ON state (red). (2) The metabolic burden of fluorescent protein expression is predicted to make the ON state unstable. Cells in the ON state are either outcompeted by cells in the OFF state or they accumulate mutations that abolish fluorescent protein expression, resulting in a BROKEN state (gray). Extracted from (Kalvapalle et al., 2024).

This disadvantage was addressed through three combined strategies: decrease of recombinase translation to lower the background signal; removal of non-essential intergenic regions to minimize mutations; and using sequence-based changes once the analyte is detected, such as flipping a non-functional DNA region, instead of protein expression. As a result of these implementations, cellular burden and background signal were reduced, leading to more stable responses to the analyte (Kalvapalle et al., 2024).

Although a long-term environmental biosensor was achieved, the studies were conducted under conditions of constant or increasing analyte concentrations. Since, after the modification, the ON state is stable and irreversible, signal accumulation is expected when the analyte concentration increases. However, it was not assessed whether a reduction in analyte concentration would lead to a decrease in the detected signal. Moreover, following this approach, it may not be possible to switch the ON signal back to the initial OFF state, and signal reduction might result from cell death or mutations. A proposed solution for this

limitation is the usage of recombinases capable of reverting the inversion, enabling the switching between states in response to fluctuating inputs.

Additionally, the memory biosensors require further optimization for effective use in their original environment. To achieve this, the most abundant microbes from the environment of interest should be isolated and programmed by conjugating the plasmid to design an optimized memory biosensor, adapted to the cells. Subsequently, mixed microbial communities should be studied in chemostat bioreactors to evaluate the effects of the biosensor. Finally, the transformed microbes should be released into the environmental samples to perform one last test (Kalvapalle et al., 2024).

Tyrosine recombinases have also been used to develop environmental sensing approaches. A novel technology was proposed in a preprint to record intracellular signals (Kato et al., 2025). The system was composed of two DNA domains: a Response domain, which expresses an enzyme with DNA editing activity under cellular signals, such as stress; and the Memory domain, where the DNA editing enzyme alters the DNA sequence. To assess the approach's potential, they designed a Response domain that detects WNT (Wingless-related Integration site) signaling pathway. They selected the TCF-LEF response element (RE), a transcriptional enhancer activated by WNT, followed by a Cre recombinase open reading frame. Then, in the Memory domain, an inverted mKate2 sequence flanked by two different LoxP variants, causing its flipping and correct orientation upon Cre-mediated recombination. This unit was placed downstream of an EF1α promoter, enabling red fluorescent protein expression. After transfecting HEK293 cells treated with activators for WNT signaling, the responses were observed using fluorescence microscopy and quantitative Polymerase Chain Reaction (qPCR), which confirmed successful intracellular event recording in the Memory domain. The results suggested that the system successfully recorded the intracellular signaling events (Kato et al., 2025). Once again, as a limitation, the study only performed analysis under increasing concentrations of WNT activators; it was not demonstrated whether the approach would be fit for recording continuous variation of signal. Moreover, as the study is currently available as a preprint and has not undergone peer review, the results and conclusions should be interpreted with caution.

5.2.2.-Implementation of Genome Editing for Molecular Recording

Genome editing involves targeted alterations of genomic sequences, such as insertions, deletions, and substitutions. Among various techniques, such as zinc-finger nucleases (ZFNs) and transcription activator-like effector nucleases (TALENs), it is important to highlight that CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) technology has significantly facilitated programmable genome engineering, ultimately enabling the development of diverse DNA-based recording systems. CRISPR-Cas is a microbial adaptive immune system that protects the cell from viruses. The system enables the bacteria to capture and store a sequence from the viral genome, so if the same virus attacks again, the stored CRISPR information guides the Cas enzyme to cut and disable the virus (Jang & Yim, 2024).

In the field of genome editing, the most commonly used nuclease is Cas9, a programmable DNA-cutting enzyme (Ceze et al., 2019). CRISPR-Cas9 is a class of ribonucleoprotein complexes formed by a Cas9 protein and a single guide RNA (sgRNA). Cas9 can create DNA cleavage at desired genomic positions that are guided by precise base pairing between sgRNA and DNA, adjacent to a protospacer-adjacent motif (PAM) (Guo et al., 2023). The CRISPR-Cas9 nuclease causes DNA double-stranded breaks (DSBs), leading to irreversible insertions or deletions (indels) that are inherited in the next cell generation. The accumulation of these mutations could be used as barcodes for cellular events (Jang & Yim, 2024). An improvement of this approach is to combine CRISPR-Cas9 barcoding with single-cell RNA sequencing (scRNA-seq), allowing a more robust lineage tracing (Wagner & Klein, 2020) and cell type heterogeneity exploration.

In the following study, a Cas9-based method was developed to report cellular states and lineage history of embryogenesis in mice by including heritable indels (Chan et al., 2019). For this purpose, two types of cassettes have been synthesized: the target site cassette, containing three "cut sites" (also referred to as target sites) preceded by an "integration barcode" (intBC), embedded into the 3'UTR of a constitutively transcribed fluorescent protein (FP) for transcriptome profiling; and the guide cassette, containing three independently transcribed sgRNAs (corresponding to the three "cut sites") to record multiple and distinct signals (**Figure 11A-B**). To demonstrate the information recovery, K562 cells were transduced with a complex library of unique intBCs and propagated for 18 days. Then, the target sites were amplified from a single cell barcoded cDNA, and the intBCs in each cell were identified as present or absent, detecting the cell populations (Chan et al., 2019).

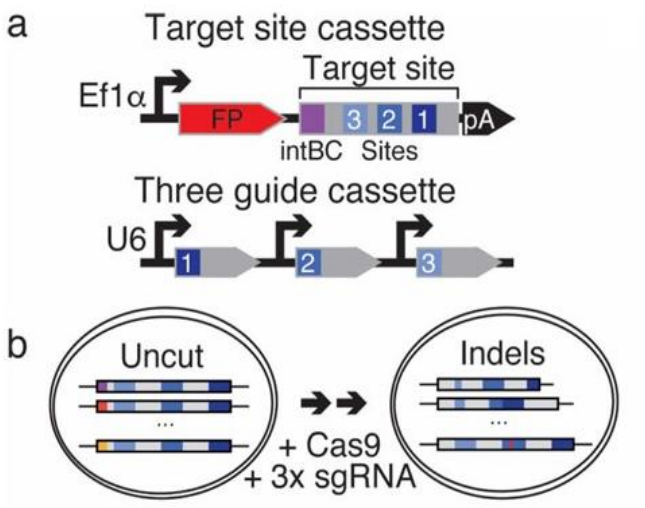

Figure 11. Optimization of a multi-purpose molecular recorder. (A) Target site (top) and three guide (bottom) cassettes. The target site consists of an integration barcode (intBC) and three cut sites for Cas9-based recording. Three different single guide RNAs (sgRNAs) are each controlled by independent promoters (in this study, mU6, hU6, and bU6). (B) Molecular recording principle. Each cell contains multiple genomic, intBC distinguishable target site integrations. sgRNAs direct Cas9 to cognate cut sites to generate insertion (red) or deletion mutations. Here, Cas9 is either ectopically delivered or induced by doxycycline. Extracted from (Chan et al., 2019).

Then, the system was applied during mouse early development from totipotency onwards (Chan et al., 2019). The target site and the three guide cassettes were encoded into a single piggyBac transposon vector, a movable genetic element that recognizes specific DNA sequences (ITRs, inverted terminal repeats) and transposes the content between vectors and chromosomes (Zhao, 2016). The vector, transposase mRNA, and Rosa26::Cas9:EGFP sperm were injected into oocytes for tracing in all subsequent cells. Rosa 26 is a mouse locus commonly used for transgene insertion. It is a secure localization since it enables stable expression without alteration of the mouse's genes. Transferred embryos were then recovered after gastrulation (**Figure 12**), and the target sites from bulk placenta, yolk sac, and three embryonic fractions were amplified to study the lineage tracing by indel proportion similarity. ScRNA-seq data was collected from seven embryos, and the indel likelihood distribution was estimated. By comparing the expression signature in the tissue proportions and age of the embryo, the cells from lineage-traced embryos were compared to wild-type embryos. Phylogenetic reconstruction strategies were developed to trace cell lineages by summing the log-likelihoods for all indels that appear using likelihoods estimated from embryo data (Chan et al., 2019).

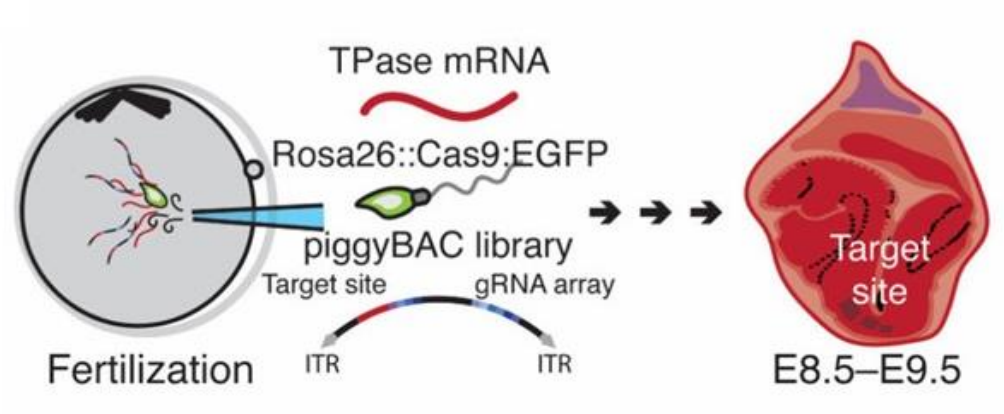

Figure 12. Lineage tracing in mouse from fertilization through gastrulation. Lineage tracing in mouse experiments. The target site (within mCherry's 3'UTR) and the three guide cassettes are encoded into a single piggyBac transposon vector (ITRs, inverted terminal repeats). The vector, transposase mRNA, and Rosa26::Cas9:EGFP sperm are injected into oocytes to ensure early integration and tracing in all subsequent cells after zygotic genome activation. Transferred embryos are then recovered after gastrulation. Extracted from (Chan et al., 2019).

Future modifications of this system could include environmentally responsive promoters to record stress, neuronal action potentials, or cell-to-cell communication. However, the main limitations of the CRISPR-Cas9 approach are the off-target activity of the Cas9 nuclease (Guo et al., 2023) and limited scalability, which is constrained by the number of cassettes that can be introduced. Moreover, the indels generated during DSB repair may lead to cellular toxicity by saturating the DNA repair machinery or activating pathways such as p53, potentially resulting in apoptosis or additional mutations. Additionally, when Cas9 expression is stable, background signals and possible off-target activity can lead to an altered development of the

organism. A more controlled method is the usage of nickase Cas 9 (nCas9), a nuclease without the ability to cause DSB, reducing cellular toxicity but retaining the ability to bind target sequences guided by sgRNAs (Jang & Yim, 2024).

A variation of the CRISPR-Cas9 system is the self-targeting CRISPR, also referred to as homing CRISPR. Self-targeting CRISPR is a modified system where the Cas9-gRNA complex targets the gRNA locus itself. Thereby, the self-targeting guide RNA (stgRNA), or homing guide RNA (hgRNA), contains a PAM region, serving both as a guide and a target site (**Figure 13A-B)** (Leeper et al., 2021). This approach has shown potential for mapping cell development and recording biological events. hgRNAs offer advantages over sgRNAs as a barcoding approach, such as targeting of their loci, without the requirement of a separate target locus, and production of a larger amount of allelic diversity, which is due to the non-randomness of non-homologous end joining (NHEJ) outcomes. Moreover, several studies have demonstrated that NHEJ outcomes of Cas9-induced DSB lead to a predictable set of mutations, limiting both the diversity of mutant alleles and their application for barcoding. Additionally, once the target site is mutated, sgRNA can't recognize it again. HgRNAs can proceed even after mutagenesis of the target site before losing the PAM or another essential part, allowing it to still operate beyond the initial limited set of mutant alleles and create a larger diversity (**Figure 13C**) (Leeper et al., 2021).

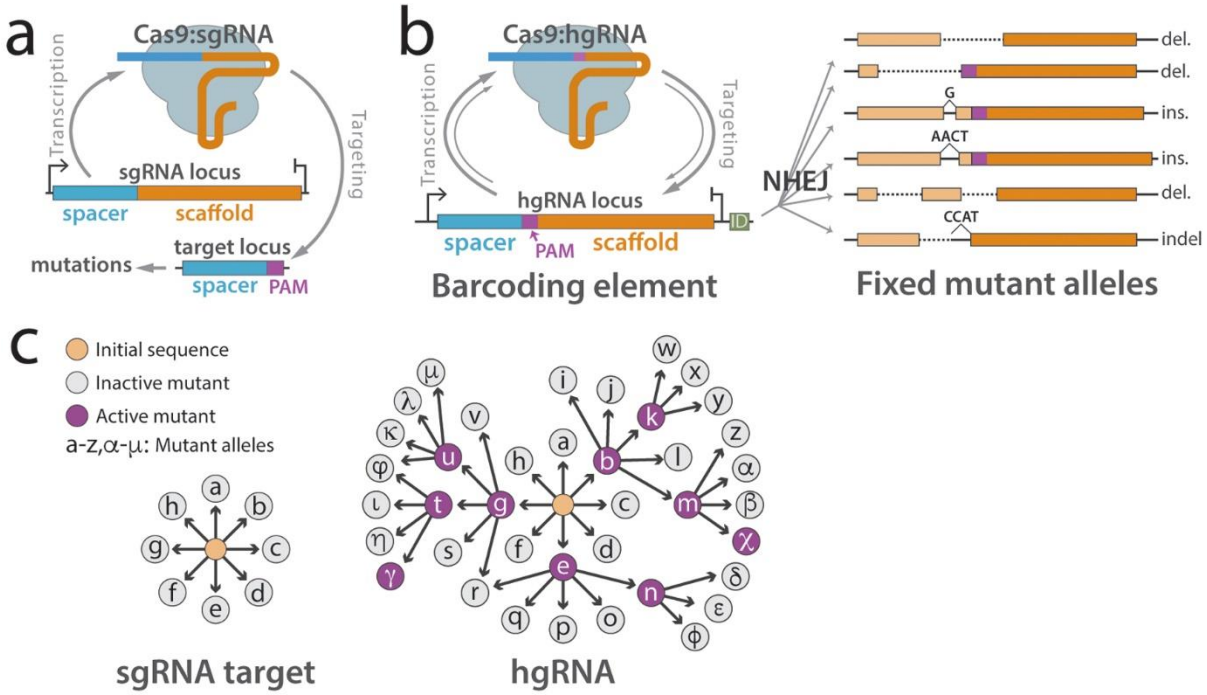

Figure 13. Principles of the homing CRISPR system. (A) Canonical CRISPR sgRNAs and Cas9 protein target the locus containing a spacer and PAM, leading to mutations in the target locus. (B) Homing CRISPR system barcoding element. The hgRNA transcript expressed by the locus forms a complex with Cas9 protein and targets the locus for double-strand break. As the NHEJ repair system repairs the cut, it introduces mutations in the hgRNA locus. These random mutations can effectively act as barcodes. (C) sgRNAs (left) are limited to a single round of targeting and mutagenesis. hgRNAs (right) can undergo multiple rounds of targeting and mutagenesis, expanding their allelic diversity. Extracted from (Leeper et al., 2021).

This homing CRISPR strategy was used for lineage barcoding in mouse. The "Mouse for Actively Recording Cells 1" line, or MARC1 line, is a mouse line with hgRNA barcoding elements, currently available at the Mutant Mouse Resource & Research Centers (MMRRC), supported by the National Institutes of Health (NIH). A library of hgRNAs was created and transfected into mouse embryonic stem (mES) cells, which were injected into blastocysts to generate chimeric male mice, the transgenic founders of the MARC1 line (**Figure 14**).

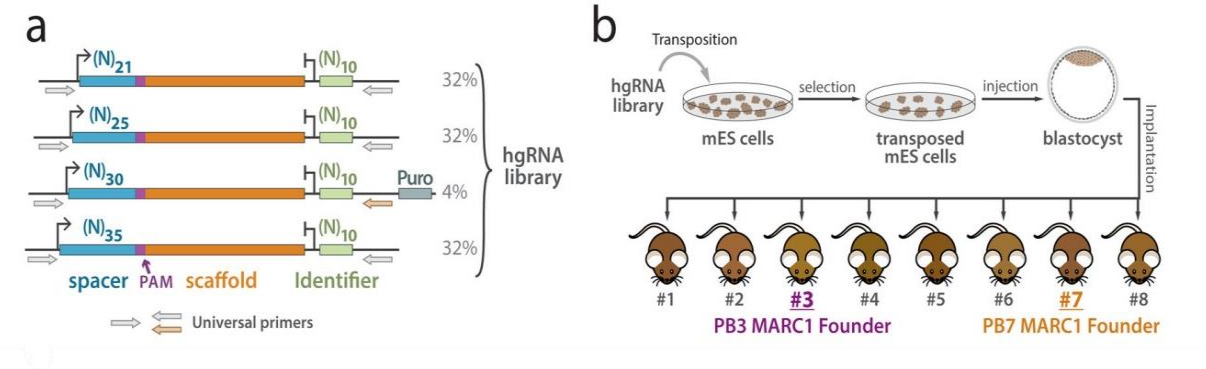

Figure 14. Strategy to generate mice with multiple hgRNA integrations. (A) Design of the PiggyBac hgRNA library for creating a transgenic mouse. Four hgRNA sublibraries with 21, 25, 30, and 35 bases of distance between transcription start site (TSS) and scaffold PAM were constructed and combined. The library contained a 10-base identifier downstream of the hgRNA scaffold to allow for hgRNA identification in the case of a large deletion. It also contained common primer-binding sites to allow for easy amplification of all hgRNAs simultaneously (B). Blastocyst injection strategy for producing hgRNA mice. The hgRNA library was transposed into mES cells. Cells with a high number of transpositions were enriched using puromycin selection and injected into E3.5 mouse blastocysts to obtain chimeras. Chimeras 3 and 7 were chosen as PB3 and PB7 MARC1 founders, respectively. Extracted from (Leeper et al., 2021).

The MARC1 founders were fertile and presented no morphological abnormalities. Their hgRNAs were inherited by Mendelian inheritance and were not positioned in coding exons. Since the hgRNAs were dormant, to study their activity, MARC1 males were crossed with Rosa26-Cas9 knock-in females with constitutive expression of Cas9. Samples from the progeny were collected at different embryonic stages, and the hgRNA loci were amplified and sequenced. The identifier sequence was used to group the reads corresponding to the same hgRNA and analyze the fraction of mutated spacers, resulting in a consistent average behavior for each hgRNA between different embryos (**Figure 15**) (Leeper et al., 2021).

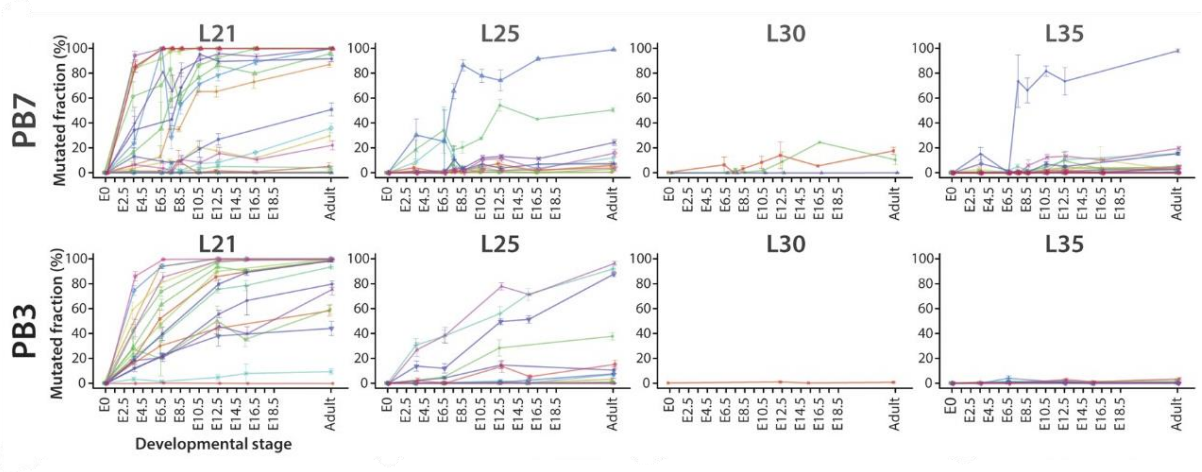

Figure 15. Activity of MARC1 hgRNAs. Activity profiles of all hgRNAs in embryonic and adult progenies of the MARC1 founder crossed with Cas9 knock-in females, broken down by hgRNA length. The fraction of mutated spacers in each hgRNA was measured. Lines connect the observed average mutation rates of the same hgRNA at different timepoints. Means ± SEMs are shown. Adult mice were sampled via ear punch at 21 days old. Extracted from (Leeper et al., 2021).

An advantage of the MARC1 system is the direct genetic integration of all necessary barcoding constructs into stable mouse lines, without the need for injecting constructs into individual embryos. Strategies that rely on injection may suffer from dilution of barcoding components due to degradation and cell division, and are technically challenging. Moreover, the MARC1 system, due to the distributed barcoding elements in the genome, is less prone to suffer from loss of mutant alleles as a result of large deletions, something very common in CRISPR targets in tandem. However, more loci are needed for the genotyping. The main limitation of the MARC1 system is its non-availability for scRNA-seq due to the lack of polyadenylation resulting from the usage of a RNA polymerase III promoter for transcription. Thus, single-cell lineage and transcriptome analysis need sorting of single cells into single wells, lowering the throughput and increasing the technical difficulty (Leeper et al., 2021).

5.2.3.- DNA Methylation for Biological Recording

The previously presented studies about CRISPR-Cas9 machinery rely on the endogenous DNA repair system after the nuclease creates the DSB (traditional Cas9) or a single nick (nCas9). However, the complexity of these pathways can make it difficult to limit the outcome to a single desired change. A more secure modality is to rewrite the epigenetic landscape without changing the underlying DNA sequence (Nuñez et al., 2021). Synthetic epigenetic circuits, especially those involving targeted DNA methylation, have regulated specific gene expression levels and durably retained cellular epigenetic memory. Moreover, by a combination of a methyltransferase and a demethylase, reversible epigenetic modifications can be achieved. However, methylation-based techniques for recording the temporal order of various signals have not been demonstrated. Moreover, the need for specific recognition sites for each methyltransferase may limit its application (Jang & Yim, 2024).

6.-Conclusions

Several factors must be considered for *in vivo* DNA storage in cryptography and steganography approaches, the most important being the maximum amount of information that a single cell can carry, as well as the potential incompatibilities and interactions between the information-carrying DNA and the host DNA. Although artificially encoded DNA is not prone to forming open reading frames, misexpression may emerge, and the biological consequences should be closely studied. Additionally, it has not been demonstrated that insertion of encoded DNA alters the host cell's expression, so methods must be developed to prevent the potential biological impact associated with the insertion of DNA fragments carrying non-biological information (Dong et al., 2020).

Regarding DNA-based cellular environment recording and lineage tracing, numerous enzyme-driven and genome editing strategies have been developed. Several of the most widely used systems have been reviewed, their limitations critically analyzed, and potential improvements proposed. However, the novelty of these technologies combined with the complexity of *in vivo* systems arouses questions about the best approach and their intrinsic limits, necessitating broader studies of these technologies.

7.-Self-Assessment

The process of writing this Final Degree Project in Biotechnology has been a valuable learning experience. Although my initial expectations were focused on the cryptographic and steganographic topics, due to the limited amount of content, I decided to broaden the scope of the project. In my opinion, this gave it greater depth and richness.

I have improved my ability to search for scientific information efficiently. At the beginning, it was difficult to understand some of the procedures and experimental approaches. However, after multiple readings, these concepts became clearer. This experience has helped me to adapt and learn independently.

Moreover, I gained a deeper understanding of emerging techniques and learned how to critically analyze results and suggest improvements. Overall, I am satisfied with the outcome of the project, although I recognize that I could have proposed additional solutions to the current challenges in the field.

8.-Bibliography

Alhabeeb, O. H., Fauzi, F., & Sulaiman, R. (2021). A Review of Modern DNA-based Steganography Approaches. *International Journal of Advanced Computer Science and Applications*, *12*(10). https://doi.org/10.14569/IJACSA.2021.0121021

Ceze, L., Nivala, J., & Strauss, K. (2019). Molecular digital data storage using DNA. *Nature Reviews. Genetics*, *20*(8), 456–466.

Chan, M. M., Smith, Z. D., Grosswendt, S., Kretzmer, H., Norman, T. M., Adamson, B., Jost, M., Quinn, J. J., Yang, D., Jones, M. G., Khodaverdian, A., Yosef, N., Meissner, A., & Weissman, J. S. (2019). Molecular recording of mammalian embryogenesis. *Nature*, *570*(7759), 77–82. https://doi.org/10.1038/s41586-019-1184-5

Chen, M. (2019). Hiding secret messages in living cells in the age of next-generation sequencing. *IEEE Access*, *7*, 184734–184742. https://doi.org/10.1109/ACCESS.2019.2961105

Chen, W., Han, M., Zhou, J., Ge, Q., Wang, P., Zhang, X., Zhu, S., Song, L., & Yuan, Y. (2021). An artificial chromosome for data storage. *National Science Review*, *8*(5). https://doi.org/10.1093/nsr/nwab028

Church, G. M., Gao, Y., & Kosuri, S. (2012). Next-Generation Digital Information Storage in DNA. *Science*, *337*(6102), 1628–1628. https://doi.org/10.1126/science.1226355

David, G., O'Keefe, R. T., & Cai, Y. (2023). Cellular Surveillance: DNA-Based Recording to Monitor and Memorize Biological Events. *GEN Biotechnology*, *2*(3), 197–210. https://doi.org/10.1089/genbio.2023.0018

Davis, J. (1996). Microvenus. *Art Journal*, *55*(1), 70–74. https://doi.org/10.1080/00043249.1996.10791743

Dong, Y., Sun, F., Ping, Z., Ouyang, Q., & Qian, L. (2020). DNA storage: Research landscape and future prospects. In *National Science Review* (Vol. 7, Issue 6, pp. 1092–1107). Oxford University Press. https://doi.org/10.1093/nsr/nwaa007

Garabello, E., Yoon, H., Reid, M. C., & Giometto, A. (2024). *Tunable Low-Rate Genomic Recombination with Cre-lox in* E. coli *: A Versatile Tool for Synthetic Biology and Environmental Sensing*. https://doi.org/10.1101/2024.10.02.616356

Goldman, N., Bertone, P., Chen, S., Dessimoz, C., LeProust, E. M., Sipos, B., & Birney, E. (2013). Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. *Nature*, *494*(7435), 77–80. https://doi.org/10.1038/nature11875

Guiziou, S., Maranas, C. J., Chu, J. C., & Nemhauser, J. L. (2023). An integrase toolbox to record gene-expression during plant development. *Nature Communications*, *14*(1). https://doi.org/10.1038/s41467-023-37607-5

Guo, C., Ma, X., Gao, F., & Guo, Y. (2023). Off-target effects in CRISPR/Cas9 gene editing. *Frontiers in Bioengineering and Biotechnology*, *11*. https://doi.org/10.3389/fbioe.2023.1143157

Hao, M., Qiao, H., Gao, Y., Wang, Z., Qiao, X., Chen, X., & Qi, H. (2020). A mixed culture of bacterial cells enables an economic DNA storage on a large scale. *Communications Biology*, *3*(1), 416. https://doi.org/10.1038/s42003-020-01141-7

Huang, C., Yang, Z., Hu, Z., Yang, J., Qi, H., Zhang, J., & Zheng, L. (2023). DNA Synthetic Steganography Based on Conditional Probability Adaptive Coding. *IEEE Transactions*

on Information Forensics and Security, *18*, 4747–4759. https://doi.org/10.1109/TIFS.2023.3285045

Jang, H., & Yim, S. S. (2024). Toward DNA-Based Recording of Biological Processes. In *International Journal of Molecular Sciences* (Vol. 25, Issue 17). Multidisciplinary Digital Publishing Institute (MDPI). https://doi.org/10.3390/ijms25179233

Kalvapalle, P. B., Sridhar, S., Silberg, J. J., & Stadler, L. B. (2024). Long-duration environmental biosensing by recording analyte detection in DNA using recombinase memory. *Applied and Environmental Microbiology*, *90*(4). https://doi.org/10.1128/aem.02363-23

Kato, S., Ikemoto, A., Isayama, J., Takimoto, T., Saya, H., & Hamada, K. (2025). *Proof-of-Concept of A DNA-Based Recording System for High-Throughput Functional Gene Screening*. https://doi.org/10.1101/2025.04.28.650920

Kempton, H. R., Love, K. S., Guo, L. Y., & Qi, L. S. (2022). Scalable biological signal recording in mammalian cells using Cas12a base editors. *Nature Chemical Biology*, *18*(7), 742–750. https://doi.org/10.1038/s41589-022-01034-2

Leeper, K., Kalhor, K., Vernet, A., Graveline, A., Church, G. M., Mali, P., & Kalhor, R. (2021). Lineage barcoding in mice with homing CRISPR. *Nature Protocols*, *16*(4), 2088–2108. https://doi.org/10.1038/s41596-020-00485-y

Loveless, T. B., Grotts, J. H., Schechter, M. W., Forouzmand, E., Carlson, C. K., Agahi, B. S., Liang, G., Ficht, M., Liu, B., Xie, X., & Liu, C. C. (2021). Lineage tracing and analog recording in mammalian cells by single-site DNA writing. *Nature Chemical Biology*, *17*(6), 739–747. https://doi.org/10.1038/s41589-021-00769-8

Munck, C., Sheth, R. U., Freedberg, D. E., & Wang, H. H. (2020). Recording mobile DNA in the gut microbiota using an Escherichia coli CRISPR-Cas spacer acquisition platform. *Nature Communications*, *11*(1), 95. https://doi.org/10.1038/s41467-019-14012-5

Nuñez, J. K., Chen, J., Pommier, G. C., Cogan, J. Z., Replogle, J. M., Adriaens, C., Ramadoss, G. N., Shi, Q., Hung, K. L., Samelson, A. J., Pogson, A. N., Kim, J. Y. S., Chung, A., Leonetti, M. D., Chang, H. Y., Kampmann, M., Bernstein, B. E., Hovestadt, V., Gilbert, L. A., & Weissman, J. S. (2021). Genome-wide programmable transcriptional memory by CRISPR-based epigenome editing. *Cell*, *184*(9), 2503-2519.e17. https://doi.org/10.1016/j.cell.2021.03.025

Ping, Z., Ma, D., Huang, X., Chen, S., Liu, L., Guo, F., Zhu, S. J., & Shen, Y. (2019). Carbon-based archiving: Current progress and future prospects of DNA-based data storage. In *GigaScience* (Vol. 8, Issue 6). Oxford University Press. https://doi.org/10.1093/gigascience/giz075

Reinsel, D., Gantz, J., & Rydning, J. (2018). *The Digitization of the World From Edge to Core*.

Shrivastava, S., & Badlani, R. (2014). Data Storage in DNA. *International Journal of Electrical Energy*, 119–124. https://doi.org/10.12720/ijoee.2.2.119-124

Wagner, D. E., & Klein, A. M. (2020). Lineage tracing meets single-cell omics: opportunities and challenges. In *Nature Reviews Genetics* (Vol. 21, Issue 7, pp. 410–427). Nature Research. https://doi.org/10.1038/s41576-020-0223-2

Wang, Y., Han, Q., Cui, G., & Sun, J. (2019). Hiding Messages Based on DNA Sequence and Recombinant DNA Technique. *IEEE Transactions on Nanotechnology*, *18*, 299–307. https://doi.org/10.1109/TNANO.2019.2904842

Zhang, Y., Ren, Y., Liu, Y., Wang, F., Zhang, H., & Liu, K. (2022). Preservation and Encryption in DNA Digital Data Storage. *ChemPlusChem*, *87*(9). https://doi.org/10.1002/cplu.202200183

Zhao, S. , J. E. , C. S. , G. Y. , S. A. , L. T. , L. L. , & Y. Z. (2016). PiggyBac transposon vectors: the tools of the human gene encoding. *Translational Lung Cancer Research*, *5*(1), 120–125.