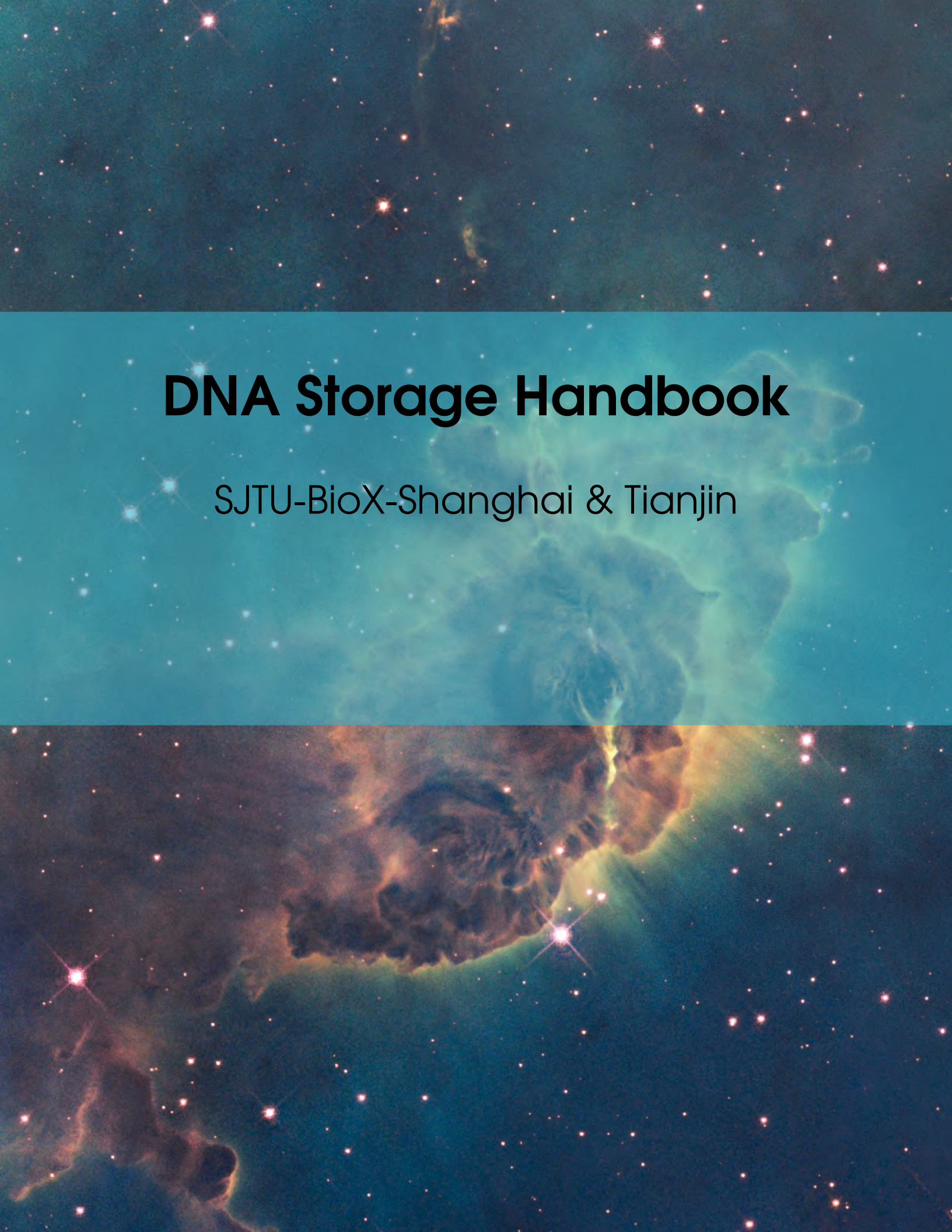


DNA Storage Handbook

SJTU-BioX-Shanghai & Tianjin

Dear readers,

This year, members from SJTU-BioX-Shanghai and Tianjin enjoy a lovable journey of iGEM. The two teams, both work in the track of fundamental advancement, are enamored in DNA data storage and try to solve different questions and create new application scenarios in this field.

This little brochure is one of the fruits of their close partnership and they believe it is the right time now to popularize this technique, which is due to change and shape the future of mankind.

Hope this little brochure gives you a general concept of DNA data storage. For more professional information, please refer to academic papers.

If you figure out any problem or bug when reading, please feel free to contact us.

Enjoy yourself!
Best wishes,
SJTU-BioX-Shanghai & Tianjin
2022.9

Contents

1. Background on DNA storage

1.1 The structure of DNA

DNA is a deoxyribonucleic acid composed of four different deoxyribonucleotides whose backbone is a phosphate ester. A deoxynucleotide contains one pentose, one base, and one phosphate.

The pentose in DNA is deoxyribose. The bases are divided into two groups: purines and pyrimidines, adenine (A), guanine (G), cytosine (C) and thymine (T). These four types of bases make up the A:T, G:C base pairs, which are the core of DNA for replication as well as transmission of genetic information. Base and deoxyribose together form a deoxynucleotide through glycosidic bond. The hydroxyl group on the fifth carbon of the deoxyribose molecule is condensed by dehydration with the phosphate group of another molecule to form deoxyribose monophosphate.

The structure of DNA can be divided into primary, secondary, tertiary and quaternary structures. The primary structure is a long-chain polymer formed by four kinds of deoxynucleotides connected with phosphodiester bonds, which indicates the arrangement order and connection mode of the bases. The secondary structure is the double helical structure formed by the reverse parallel winding of two deoxynucleotide chains. The tertiary structure is a special geometric conformation formed by further twisting and winding of DNA on the basis of the double helix structure, also known as superhelix structure. Quaternary structure is the highly ordered and dense structure of eukaryotic DNA.

DNA is a biological macromolecule with a defined sequence, the genetic material of most living organisms and carries a large amount of genetic information. The genetic information is embodied in two aspects: one is the different types of bases, the other is the sequence of the bases. If the bases are assigned a value, such as a&t = 0 or G&c = 1, the chemical signals can be converted into digital ones. It is a binary data storage material.

From the DNA composition and structure listed above, the followings are determined: DNA is stable, the reliability of using DNA to store data can be guaranteed, and the storage lifespan is also very considerable; The arrangement and combination of the four bases in DNA sequences can store a large amount of data, which can be used to process today's massive information data. DNA's data storage density is very high, it also requires relatively less energy for the same amount of time.

1.2 Background of DNA storage

The development of industrialization has brought prosperity in the current era of big data. However, the supply of big data storage cannot keep up with its production. According to Seagate, the world's largest maker of hard drives, 98.29% of data is expected to be unsalvageable in 2025 due to shortage of storage technology and size.

The mainstream storage media on the market can be divided into three categories based on storage principles, which are optical disk, magnetic storage (hard disk, magnetic tape, etc.) and semiconductor storage (flash memory, etc.).

The storage density of existing storage media is low. Take tape storage as an example, the current storage density can reach 1014 bits/cubic centimeter, while hard disk and flash memory are not as good as tape storage, the lack of storage density will lead to high operating costs and construction costs.

In terms of energy consumption, 1 zeabyte of data volume requires about 1000 exabyte data centers, and each exabyte data center requires about 70,000 square meters of floor space, 200 MW/year power consumption, and the total storage cost is up to $1 billion. Tape storage consumes less energy than storage and flash memory and is capable of storing data offline, but tape relies mainly on rare-earth metals. Although there are abundant rare earths around the world, they are rapidly declining due to over-exploitation and cross-border import and export trade, and the tape storage solution is not sustainable.

Service life is also a key factor limiting the development of data storage. Among the existing storage media, the lifespan of optical disc is 10-15 years, the lifespan of HDD storage and flash memory is 5-10 years, and the lifespan of tape storage is 15-30 years. Data storage systems need to periodically clean up corrupted data and replace faulty units. Low service life will lead to high maintenance costs for storage of zebyte data volume, so the market needs more stable storage media to support the rapidly growing data volume.

To sum up, low density storage media aggravates the overuse of land resources and excessive energy consumption. Short service life increases the cost of data migration and maintenance. The insufficient migration of large storage devices increases data security risks. In order to meet the emerging demand of massive data storage, the market is in urgent need of innovative storage media. In particular, the storage density, service life, energy consumption, data security and other factors should be significantly optimized and improved. DNA storage is one of the solutions for future data storage media. On the one hand, DNA is the storage medium that has the highest storage density known, the storage density can be achieved theoretically at 455 exabytes/gram. On the other hand, due to the stability of DNA, ancient DNA from 700,000 years ago can still be sequenced, which shows the timeliness of storage.

1.3 History of DNA storage

1.3.1 The Birth of Micro Venus

Back in 1959, Nobel Prize winner Feynman first proposed the concept of a molecular scale computer. DNA storage dates back as far as 1964, when Wiener and Neiman proposed molecular "genetic memory". In an interview in 1964, the concept of storing data with genetic information was first proposed, but it was only envisaged that information would be done in a computer, similar to genetic memory. At the same time, Neiman discussed possible directions for preliminary research to develop electromagnetic, electronic, and ionic methods for recording and reading information

in long-chain polymer molecules, and the concept of mapping digital information on biological genetics was briefly mentioned in Dawkins' book The Blind Watch-maker. But these are only theoretical descriptions of the concept of DNA data storage. A related landmark was Joe Davis's 1988 bioartwork Micro Venus, which encoded a 35-bit image of an ancient Germanic rune for Female Earth, proving experimentally for the first time that information could be stored in DNA.

1.3.2 A New Era of Biocomputing

In 1994, Adleman, the Turing Prize winner, solved a Hamilton shortest path problem for the first time using DNA molecules as the carrier of "data" and biological enzymes and biochemical operations as "operators", which opened a new era of biological computing. Baum of Princeton university, published in Science "Building an associative memory vastly larger than the brain", formally put forward the conception based on large capacity of DNA database storage system.

1.3.3 DNA information encryption and the establishment of the database

The concept of DNA storage has been mentioned and practiced ever since it was first experimented in the field of bio-art. A DNA-based dual steganography technique developed in 1999, in which researchers created a coded strand of DNA and hid it in a printed document, which was then sealed and sent through the United States Postal Service. Finally, the embedded message (JUNE 6 INVASION: NORMANDY) is successfully recovered in a laboratory environment. In 2000, Leier et al. implemented encryption and decryption methods that could be applied to DNA "barcoding". In 2001, Reif et al. of Duke University first constructed a small DNA database that could be accessed randomly.

1.3.4 Using DNA to store large amounts of data

In 2012, Church et al. developed a strategy for encoding arbitrary digital information in DNA, using DNA synthesis to store a book containing 53,426 words, 11 pictures, and a JavaScript program, totaling 5.27 MB. Its storage cost is about 100,000 times lower than that of the first generation of coding. In 2017, in the paper "DNA Fountain Enables a Robust and Efficient Storage Architecture," Erlich et al. introduced the Fountain code into the DNA coding system, called the "DNA Fountain," to achieve high data storage density.They used DNA fountain coding technology to pack an average of 1.6 binary bits into each base.This allowed them to store at least 60% more data than previously published methods. In 2018, Organick et al. stored up to 200 MB of data in DNA molecules, implemented random access in large-scale systems, and experimented with single-molecule sequencing for data reading and recovery. In 2021, Yingjin Yuan's team demonstrated a new approach to DNA data storage. The team designed and synthesized a 254,886 bp yeast artificial chromosome dedicated to data storage from de novo coding, and stored two classic images and a video in the efficiently assembled artificial chromosome with the help of the cutting-edge error correction coding in wireless communication.

2. Technology for DNA storage

2.1 DNA synthesis

DNA synthesis is one of the core technologies in DNA storage, and its efficiency and cost greatly affect the large-scale application of DNA storage technology. The DNA sequence for storing information is generated by an encoding system, which is not available in natural organisms. Therefore, in vitro DNA synthesis has become an essential part of DNA storage.

Natural DNA molecules consist of deoxynucleotides with different bases A, T, G, C. Synthetic DNA is based on chemical or biological methods (also known as enzymatic synthesis) in which these deoxynucleotide monomers are connected in a defined sequence, that is, DNA synthesis. Conventional DNA synthesis methods are based on the synthesis of single-stranded oligonucleotides.

According to the principles, DNA synthesis is generally divided into chemical synthesis and biological synthesis. Chemical synthesis is mainly based on the phospholamide method, in which the controlling methods include photochemistry, electrochemical method, inkjet printing method, integrated circuit control and so on. Biosynthesis was mediated by TdT, TDT-DNTP crosslink and mixed enzymes.

2.1.1 Chemical synthesis

Phospholipamide triester synthesis is the most widely used oligonucleotide synthesis method, which is also the synthesis method used by mainstream DNA synthesizers at home and abroad, including deprotection, coupling, capping and oxidation four steps cycle. (as shown in figure)

First, deoxyribonucleotides are compounds that contain multiple functional groups, and to ensure the formation of a 3'-5' phosphodiester bond under specific order, the active group that is not involved in the reaction needs to be temporarily protected. The 5' -OH of deoxyribose is protected by dimethoxy-triphenylmethyl (DMT), and the -OH of the 3' -terminal phosphate group is protected by β-cyanethyl (Ce). In addition, bases with primary amino groups (such as adenine A, guanine G, and cytosine C) must also be protected with benzoyl (BZ) or isobutanyl (Ib) groups.

DNA synthesis proceeds from the 3' end to the 5' end of the target product.The 3'-OH of the initial deoxyribonucleotide is covalently connected to the solid-phase support porous glass beads

(CPG) by a long alkyl arm. The synthesis process consists of four cycles: deprotection, coupling, capping and oxidation.

1. De-protection. The terminal nucleotide with the guard group pre-attached to the solid phase support is treated with dichloroacetic acid or trichloroacetic acid solution to remove the protective group DMT of the DNA deoxyribose loop 5' -OH and dissociate the 5'-OH.
2. The coupling reaction. Raw material for DNA synthesis (phospholipamide protected nucleotide monomers) is mixed with the activator tetrazolium to give a highly reactive nucleoside phosphite activation intermediate.The activation intermediate is coupled to the free 5' -OH on the terminal nucleotide, and the triester phosphite bond is rapidly formed.
3. Capping reaction. A very small amount (less than 2%) of the 5'-OH in the coupling reaction may not participate in the reaction. Acetylation of the unreacted 5' -Oh prevents further extension of the oligonucleotide chain in which it is located. This short fragment can be separated and removed during purification, which is beneficial to reduce synthetic errors and synthesize full-length DNA fragments.
4. The oxidation. The new 3', 5' -triphosphite bond formed in the reaction is very unstable and prone to fracture under acidic or alkaline conditions, so iodine solution is needed to oxidize the trivalent phosphite bond to a stable pentavalent phosphodiester bond in the reaction.

In each of the four steps, the nucleotide chain is extended one nucleotide in the 5'-OH direction until the DNA of a specific sequence length is fully synthesized. Each nucleotide extension takes about 10 minutes.The extended nucleotide chains are always fixed to the solid phase support, and excess reaction reagents or intermediates are removed by solvent flushing.

When the synthetic chain reaches the predetermined length, it is cut off from the solid phase support, the protective group is removed, and the desired product is obtained after separation and purification.

However, due to the incompleteness of chemical reactions at each step and the occurrence of side reactions (such as deadenosine in the process of deprotection), the longer the oligonucleotide synthesis chain, the lower the synthesis efficiency and the higher the synthesis error rate, which greatly limits the length and quality of oligonucleotide synthesis.

2.1.2 Biosynthesis method

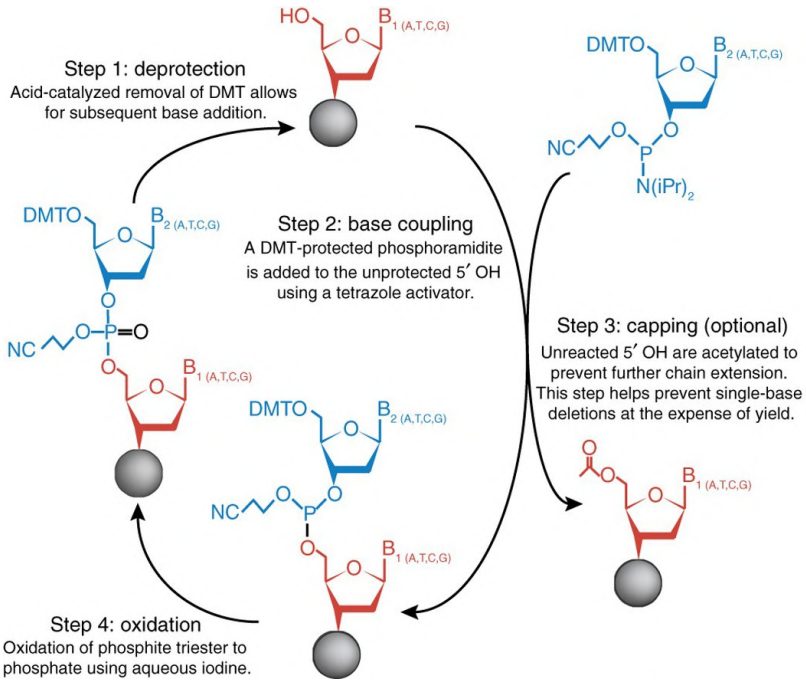
The traditional chemical synthesis of phosphamide is limited by the efficiency of chemical reactions, and the length of DNA synthesis products can only reach about 200-250 nt, which greatly limits the downstream application. Strong acids and oxidants are involved in the synthesis process, resulting in more chemical waste liquid harmful to the environment, which leads to the high cost of subsequent treatment. In recent years, the bioenzymatic DNA synthesis technology is usually carried out in the aqueous environment, which can effectively avoid the above problems and is expected to synthesize longer DNA molecules at a lower cost.

In nature, the in vivo synthesis of DNA molecules is mainly catalyzed by various DNA polymerases and relies on DNA templates for synthesis. DNA terminal transferases and some kinds of DNA polymerases can directly catalyze the synthesis of DNA strands without relying on existing DNA template molecules. Bioenzymatic DNA synthesis technology also borrows from different DNA synthesis methods in nature.

1. TdT enzyme mediated enzymatic synthesis reaction
   TdT mediated enzymatic synthesis is a hot topic in the research of enzymatic DNA synthesis. TdT is a template-independent enzyme that usually extends the DNA strand in a random

manner and can add four natural bases to the 3 ’end of the DNA strand. The enzymatic synthesis reaction mediated by TdT enzyme is to chemically synthesize nucleotide monomers with reversible termination groups by modifying nucleotide molecules, and then use TdT enzyme to continuously add the bases to the ends of the synthesized sequences. The basic principle is to control the chemically modified group on the nucleotide molecule through chemical reaction, so that the enzyme can only extend the target single base at a time, then remove the termination group, start the synthesis of the next target base, and finally achieve the control of the orderly connection of the target base in the DNA chain.

2. Tdt-dntp cross-link mediated enzymatic synthesis reaction
   The basic principle is that the TdT enzyme is first combined with a single deoxynucleoside triphosphate with a reversible adaptor at the 3 ’end to form the TDT-DNTP cross-link. Whenever the 3’ end of the new synthetic strand is exposed during DNA synthesis, the TDT-DNTP cross-link is connected to the 3 ’end of the strand, and the new target base is also introduced. At the same time, the TdT enzyme also remains at the 3 ’end to prevent additional monomers from being added.
3. Mixed enzyme mediated enzymatic reaction
   The basic principle is that the 3 ’end of each primer is composed of reversible termination nucleotides (rtNTP), containing a specific combination of enzymes or ribozymes with terminal transferase activity, which are extended by constant repetition to synthesize 300 nt length primers. Compared with the synthetic fragments with the same length, the accuracy was significantly improved compared with the chemical synthesis method because the de novo synthesis stepwise ligation steps were reduced.



2.2 DNA modification

Target-specific targeting and processing of digital information in living cells remains a great challenge. Recent developments in gene editing provide tools that can repair damaged sequences and modify specific genes within living cells. These advances provide technical means to rewrite foreign information in vivo.

In this study, a total of three plasmids were designed. The first plasmid was a DNA information plasmid containing the original information of the site to be modified, the second plasmid was a gRNA array targeting the site to be edited, and the third plasmid was a Cytosine base editor (CBE), the tool to implement DNA modification. It is made up of four components, namely dCas9 protein, sgRNA, cytosine deaminase, and uracil DNA glycosylase inhibitor (UGI). The function of dCas9 is to recognize and bind with PAM sites, sgRNA is to accurately identify the target gene sequence, cytosine deaminase is to change cytosine C into uracil U, UGI is to inhibit the natural repairment of uracil-guanine intermediates in DNA, preventing cells from correcting the modified bases back.

These four components together constitute the mechanism of action of the study: After gRNA targets CBE to the site to be edited, cytosine C on the DNA distal to the PAM site is changed to uracil U by cytidyl deaminase. UGI then inhibits the repair of this mispairing by the cell body. Thus, DNA polymerase may recognize U as T during DNA replication. Thus, the conversion from C-G base pair to A-T base pair is achieved, which means the effect of single base editing is achieved.

In order to achieve multi-site editing, ultra-long gRNA arrays need to be synthesized. Due to the similar structure of each segment of gRNA, there will be many repetitive sequences in the ultra-long array, which increases the difficulty of synthesis. Therefore, we used the following methods: tRNA sequences were inserted between the gRNAs at both ends, and the tRNA was excised by post-transcriptional modification of the cell itself to release the gRNA. An ultra-long gRNA array was constructed by using a yeast type II syntactic promoter and terminator for transcription of every three gRNAs.

In addition to the DNA modification used in the experiments of our project, the team of Kai Liu and Hongjie Zhang from the Department of Chemistry, Tsinghua University constructed a living cell information storage and rewriting system based on gene editing technology. The recent work titled "In vivo processing of digital information molecularly with targeted specificity and robust Reliability ", published in Science Advances.

To rewrite complex information, they built a two-plasmid system based on CRISPR-Cas12A and a phage-derived recombinase. The INFO plasmid contained the target DNA sequence carrying the original digital information, while the Help plasmid contained a template expressing Cas12a and phage-derived recombinases. After the expression of phage-derived recombinase and Cas12a was activated by tetracycline and arabinose, respectively, the target DNA fragments present in the INFO plasmid were selectively excised by Cas12a under the guidance of crRNA. The donor DNA carries the rewritten information and two homologous arms of approximately 500 base pairs (bp) each. The target DNA fragment is then replaced and the information plasmid is reconstituted with the help of phage-derived recombinase. Thus, a two-plasmid system is used to selectively modify the original information encoded in the DNA sequence of the INFO plasmid by replacing the target DNA with donor DNA.

2.3 Encryption of DNA information

Combining DNA storage with gene editing technology can be applied to the field of information encryption. The information to be encrypted is encapsulated by traditional encryption methods, and

a biological encryption method is added on this basis. The original information is modified and hidden, the PAM site is manually designed on the DNA sequence of the encrypted information, and the unique key of the corresponding information is designed by base editing technology. The encrypted information can be obtained successfully if and only if both the encrypted DNA sequence and the gRNA sequence are possessed. The key is set in the form of plasmid, and the physical barrier is used to completely isolate the information leakage in the network environment. Only when you have the corresponding decryption plasmid can you edit the DNA sequence and decrypt the information successfully. This method can be combined with DNA multi-site editing technology to achieve large-scale biological information encryption.

2.4 DNA encapsulation

The preservation of DNA molecules is an important aspect of DNA storage. Despite DNA molecules have half-lives up to 521 years, DNA exposed to air is highly susceptible to degradation by DNA enzymes. During long-term storage, DNA molecules may also undergo damage such as chemical bond breaks and base mutations, so their half-life decreases significantly as the length of the molecular chain increases. In terms of the form of encapsulation, there are two types of encapsulation: physical and biological.

2.4.1 DNA physical encapsulation

The preservation stability of DNA solutions and dry powders differs slightly. Due to the risk of DNA hydrolysis, DNA solutions can be stored stably at room temperature for a maximum of six months, while their shelf life can be extended to two years with -20° C freezing.

In order to maintain the integrity of DNA structure and avoid large-scale DNA breakage and degradation, a better method of DNA preservation is dry powder at -80° C. However, cryopreservation not only takes up a lot of space and requires the acquisition of an oversized cryogenic unit, but also requires an accompanying management system for large sample volumes, which will increase the cost of preservation apportioned to individual samples. Some scientists have also proposed to take advantage of the low-temperature properties of the polar regions and build nature conservation repositories to reduce maintenance costs, while the construction of polar infrastructure and supporting equipment is also very costly. Most importantly, cryopreservation does not provide absolutely safe long-term preservation of DNA samples, but only delays the process of DNA oxidation and hydrolysis.

The lyophilized powder of DNA has been considered to be more durable, and scientists have developed a variety of preservation methods for DNA dry powder, including three types of pore plates made of Biomatrica, alginate and polyvinyl alcohol, which facilitate the stable preservation of DNA molecules. DNA dry powder can also be adsorbed and preserved in a special filter paper (US5807527) for up to 3 years. In recent years, scientists have also found that some alkaline earth metal salts can also enhance the stability of DNA dry powder state molecules. On the other hand, metal ions such as Fe2+, Cu2+, etc. may also generate free radicals through Fenton reaction, which may cause damage to DNA molecules.

Preservation of DNA in a special housing allows it to be protected from nucleases, oxygen, ultraviolet radiation, ionizing radiation or other chemical products that are toxic to DNA molecules. chemical products that are toxic to DNA molecules. By referring to the model of DNA preservation in fossil bones, researchers have mimicked the isolation of fossils from environmental, the researchers developed a series of methods to encapsulate DNA molecules by mimicking the isolation of fossils

from moisture and reactive oxygen species in the environment. Among them, DNA molecules can be encapsulated in silica particles. Using accelerated aging tests, it is assumed that the DNA molecules can be preserved for thousands of years.

2.4.2 DNA Biological encapsulation

The basic principle of encapsulating DNA information in the form of plasmids is to use in vitro gene splicing to link DNA inserts with plasmids to form recombinant plasmids. DNA is ligated in vitro to form a recombinant plasmid into E. coli, and the recombinant plasmid replicates as the E. coli increases in value, thereby encapsulating DNA information is encapsulated within the cell.

Biological encapsulation in the form of plasmids is divided into the following main steps.

1. DNA sequence synthesis: obtaining the DNA sequence used to carry the coding information, which can be synthesized artificially by chemical or enzymatic principles.
2. recombinant plasmid construction: Using in vitro splicing techniques, the DNA fragments obtained by synthesis are assembled together with the linearized plasmid to form a recombinant plasmid.
3. Transformation: The receptor cells are treated with some special methods, and the permeability of their cell membrane is changed so that they can take in foreign DNA. recombinant plasmids are mixed with receptor E. coli cells to achieve the transformation of recombinant DNA molecules.
4. Screening of transformed cells: The plasmids currently used as vectors mostly contain ampicillin or kana resistance genes. The recombinant plasmids can only grow in the corresponding resistant media, while bacteria without plasmids will die.
5. Storage of strains and plasmids: Plasmids can be stored at -20°C for a long time. The strains can be stored at -20°C or -80°C in a culture medium containing 20%-50% glycerol.

In addition to plasmids, there is also biological encapsulation in the form of artificial chromosomes. The use of artificial chromosomes to encapsulate DNA information has similarities to traditional optical disc storage, i.e., it can be written once and read multiple times. In this method, the long DNA fragment encoding the information is synthesized and assembled in vitro, and the writing is done by in vivo assembly of the cell. As long as the carrier cells are cultured, the written cells can be replicated quickly and inexpensively to produce uniform copies of the data. Although the current production cost, i.e., synthesis and assembly, is high, copying of information can be achieved by cell culture, which is more economical than the DNA molecular library storage model.

2.5 DNA sequencing

DNA sequencing refers to the use of gene sequencing techniques to obtain the order of bases in the target DNA fragment, adenine A, thymine T, cytosine C, and guanine G.

The principle of DNA storage is essentially the conversion of the binary code of a digital file into the quadratic code of DNA bases and write the information through DNA synthesis. Therefore, to read out the information stored in a DNA fragment, it is necessary to first determine Therefore, to read out the information stored in a DNA fragment, the base sequence of that DNA fragment needs to be determined first.

2.5.1 Solexa Sequencing

The Solexa Sequencing System uses Sequencing-By-Synthesis as its basic design concept, and Bridge PCR and Reversible Terminator as its core technologies.

Bridge amplification refers to the preparation of a single-stranded DNA library that is complementary to a single-stranded primer on the surface of a Flow Cell, with one end immobilized on the chip and the other end randomly complementary to another primer segment nearby, also immobilized, forming a bridge. Bridge type After 30 rounds of amplification-denaturation cycle, a monoclonal DNA cluster of about 1000 copies is formed, which achieves the required signal intensity for the sequencing reaction.

In each round of sequencing reactions, the Solexa sequencing system uses four different dNTPs specifically fluorescently labeled to polymerize with DNA clusters bearing DNA template signals. The other unbound dNTPs, DNA polymerase and fluorescent groups are removed and a new round of reaction is started. These dNTPs with the chemically cleavable portion at the 3' end are the reversible terminators of the polymerization reaction.

During the sequencing process, when a fluorescently labeled dNTP is involved in the polymerization reaction, the fluorescent signal carried by this dNTP can be identified by laser excitation and imaging to complete signal acquisition, followed by cleavage to facilitate polymerization of the next dNTP, and so on, until the final sequencing of the template DNA fragment, base by base, is achieved.

It should be noted that the principle of DNA template replication in bridge-based amplification technology is similar to the chain reaction in nuclear fission. The advantage of this exponential replication method is that the replication speed is fast. However, replication errors can be generated and accumulated during the next round of replication using the replicates as templates, which may lead to distortion of a small amount of DNA information.

2.5.2 DNBSEQ Sequencing

DNA nanoball technology includes the generation, preparation and loading of DNA nanoballs. Among them, the generation and preparation of DNBs mainly employ single-strand circular DNA and Rolling Circle Amplification (RCA): the long strand of DNA is randomly broken by ultrasound or enzyme to form a template DNA fragment, which is joined into a circle by the action of The DNA long strand is randomly interrupted by ultrasound or enzymes to form a template DNA fragment, which is joined into a circle, i.e. sscirDNA, by the action of connectors; then, the circle is replicated by rolling, and the replication product is spatially wound to form a nanosphere DNB containing 300-500 copies; these prepared DNBs are then uniformly loaded onto the sequencing carriers (Flow Cell) and attached and fixed to the prefabricated nanoscale activation sites to form a Patterned Array

The amplification principle is that the original single-stranded circular DNA is always used as the template for synthesizing new copies, and the chance of amplification errors in all copies at the same position is very small, which also effectively avoids the accumulation of PCR amplification errors. With Patterned Array, DNBs are arranged in a matrix grid of active sites on the sequencing mount, and all active sites are neatly spaced. The spacing of all active sites is kept neat and consistent, and only one DNB is bound to each site, which ensures that the optical signals between DNBs do not interfere with each other. This ensures the accuracy of sequencing and also improves the efficiency of the sequencing carriers, achieving excellent imaging efficiency and The sequencing slides can be used to achieve excellent imaging efficiency and optimal reagent usage. Such a single sequencing slide can be arranged with billions of active sites.

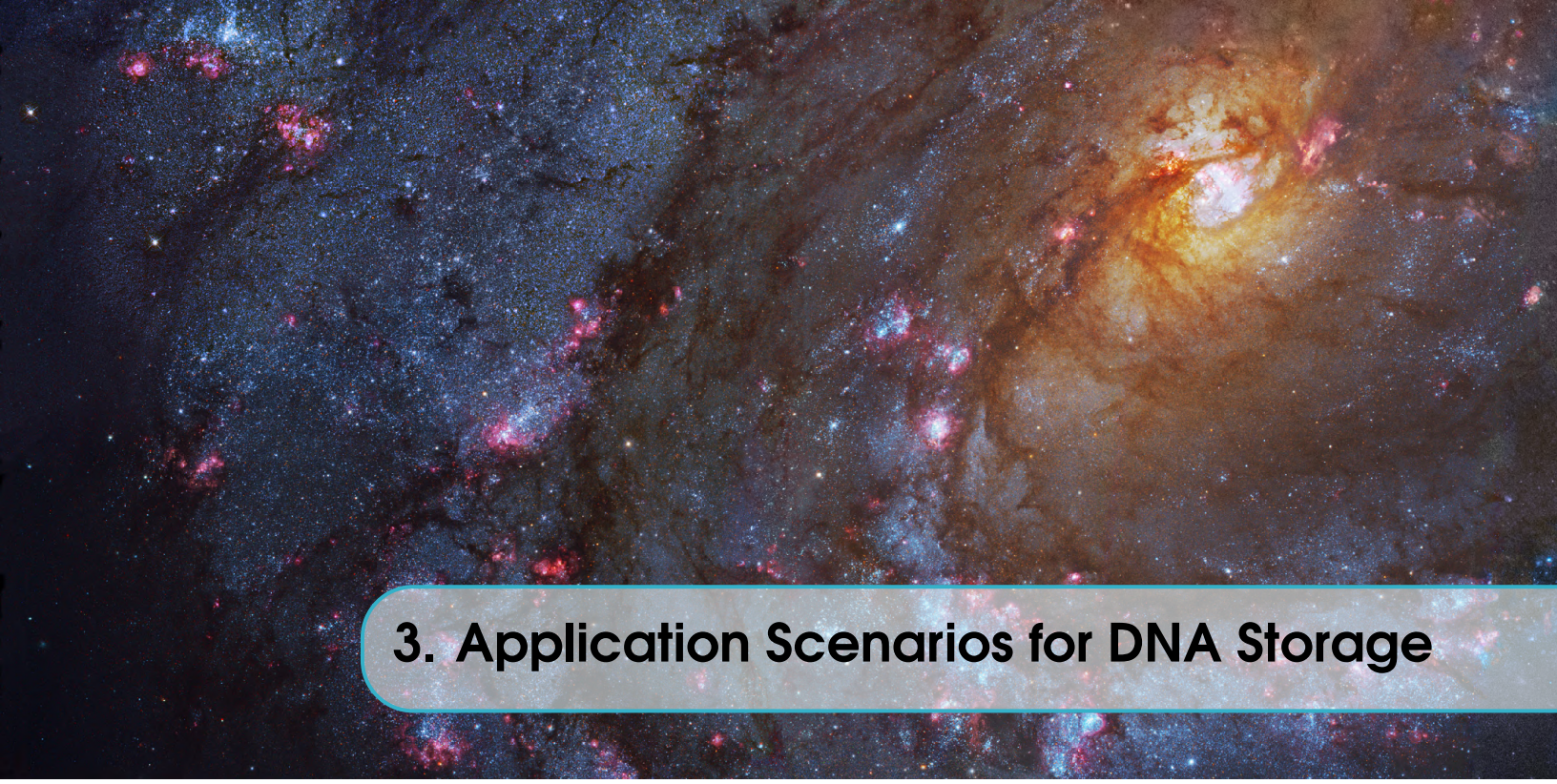
After the DNB is loaded onto the sequencing carrier, the DNBSEQ sequencing platform uses Combinatorial Probe-Anchor Synthesis to polymerize the sequencing primer-anchored molecules and fluorescent probes on the DNB, while using a high-resolution imaging system to capture, read and identify the optical signal to obtain the sequence information of a single base, and then adds The

sequence information of a single base is then obtained by adding regeneration reagents, eluting the fluorescent group, performing the next reaction, and obtaining the sequence information of the next base. After 50-150 cycles of single- and double-ended reactions, the base sequence information is combined into a complete DNA sequence by an algorithm.

2.5.3 Single-molecule sequencing technologies

The main technical routes for single-molecule sequencing include Zero Mode Waveguides, ZMWs and Nanopore sequencing technologies. This technology is characterized by the fact that no amplification of DNA templates is required and that real-time detection of DNA molecules can be achieved based on longer reads. The technology uses an optical module based on a zero-mode waveguide well that allows light to illuminate only the nanopore in which a single DNA polymerase/template molecule is immobilized. molecules. The zero-mode waveguide hole is a 10-50 nm diameter hole, and when the laser hits the bottom of the zero-mode waveguide hole, it illuminates only a small amount of light. When the laser hits the bottom of the zero-mode waveguide hole, it illuminates only a small area where the DNA polymerase is immobilized. It is only in this region that the base-carrying Only in this region is the base-carrying fluorophore activated and thus detected, dramatically reducing background fluorescence interference.

The basic principle of nanopore sequencing technology is that when a nanopore is filled with a conductive solution, a certain voltage is added at both ends and a molecular template passes through the nanopore to generate a measurable current. When the diameter of the nanopore is exactly one nucleotide, single-stranded DNA or RNA of up to 1000 bases will pass through the nanopore sequentially under the action of electric field, causing a change in the current intensity. Due to the different spatial conformations of the four bases, the intensity of the nanopore current change is different. Each of the four bases generates a specific current peak, which can be used to determine the different bases and realize high-speed real-time sequencing. Since the original current signal is very weak, noisy and random, the performance in terms of base identification accuracy is lower than the current mainstream high-throughput short-read sequencing technology.



3. Application Scenarios for DNA Storage

In today's world of explosive data growth, the ultra-high capacity of DNA storage can solve the problem of insufficient storage media capacity. Application scenarios based on the research and development of DNA storage include big data storage, novel data encryption, molecular diagnostics based on DNA computing and other applications.

3.1 Big Data Storage

The large data base, variety and fast growth rate are typical features of current big data, and how to store large amount of data centrally has become an urgent problem to be solved. Due to the characteristics of DNA storage such as high information density, long preservation life, energy saving and environmental protection, slow reading and writing speed, and difficulty of random reading and writing, DNA storage is currently mainly applicable to long-term archiving of cold data. Because DNA storage can greatly reduce carbon emissions, it is useful for building a new type of big data storage and achieving green and low-carbon data centers. It will play an important role, which is reflected in.

1. Data storage power consumption. DNA long-term offline storage does not consume power, compared to tape, hard disk and other traditional media, power consumption is smaller.
2. Infrastructure power consumption. Traditional storage media require the support of supporting equipment such as air conditioners, humidifiers and dehumidifiers, UPS, batteries, voltage regulators, etc., but these devices will bring additional power losses. In contrast, DNA storage requires less power for infrastructure.
3. Land resource occupation. DNA storage itself and supporting infrastructure have a small footprint and can realize the storage of large amounts of data while consuming limited land resources.
4. Manufacturing materials are environmentally friendly. Due to the low storage density, traditional storage media requires a large amount of material to produce. Compared with DNA storage, the A small amount of DNA can store a large amount of data, so it is more environmentally friendly.

3.2 New type of data encryption

In addition to private information such as property information, health and physiological information, biometric information, and identity information, DNA storage can also be used for new types of data encryption, including personal and collective private information, due to the high storage density and low energy consumption of DNA storage.

In addition to being used as large-capacity archival storage in data centers, DNA storage can also be used for personal encryption of private information, or anti-counterfeiting of important items.

Personal private information includes property information, health and physiological information, biometric information, identity information, network identification information, etc., which may endanger personal property and personal safety.

For sensitive and highly confidential information, it is generally recommended to keep it encrypted on offline devices; to prevent illegal access, two technologies, steganography and encryption, are usually used to take advantage of DNA's inherent high storage density, high parallelism, low energy consumption, and small size, which have certain incomparable and alternative superiority in the field of steganography and encryption.

Steganography is a technology that hides sensitive information such as text, images, audio, video or files in some boundary quality, authorized only to a specific knowledgeable person. The following properties of DNA sequences can be used in steganographic and encrypted carriers:

1. The high data density of DNA, even when combined with the consumption caused by encoding and data redundancy, DNA storage far exceeds traditional technologies in terms of data volume per unit area.
2. DNA is extremely small and highly steganographic. There is little difference between man-made DNA and natural DNA sequences, making it difficult to distinguish whether man-made DNA contains confidential information or not.
3. When DNA with confidential information is mixed with other DNA, there is little possibility of tampering or falsifying the information.
4. DNA can be used in combination with other technologies for a wider range of applications.

3.3 Molecular diagnostics based on DNA computing

DNA computing is an emerging computational technology. It contains disciplines such as informatics, mathematics, physics, and nano fields, and has great potential for application due to its DNA molecular The DNA molecules have great potential for application due to their high-density information storage capacity, powerful parallel computing capability and molecular recognition ability, and are currently under It is currently in the early stage of research.

According to the results of the study, DNA molecule computation is based on the principle of parallel computation. If the computational power of a pair of DNA molecules complementary bases of a pair of DNA molecules is understood as 1, a DNA sequence of 1uM can compute about 1017 data in parallel. With DNA DNA computation can perform more operations in one second than existing supercomputers, and when combined with DNA When combined with DNA's parallel computing capabilities and molecular recognition, it enables fine, intelligent and complex molecular computations.

Compared with traditional technologies, DNA computing-based molecular diagnostic technology has powerful parallel computing capability and molecular recognition ability. Once matured and applied, it will be a significant change to clinical molecular diagnosis. It is still in the stage of theoretical improvement and scientific research.

4. Reference

[1]Ceze, Luis., Nivala, Jeff., Strauss, Karin., & Strauss, Karin.. (2019). Molecular digital data storage using DNA. Nature reviews. Genetics, 20(8), 456-466.

[2]Organick, Lee., Ang, Siena Dumas., Chen, Yuan-Jyue., Lopez, R&olph., & Yekhanin, Sergey.. (2018). Random access in large-scale DNA data storage. Nature biotechnology, 36(3), 242-248.

[3]Chen, Kaikai., Kong, Jinglin., Zhu, Jinbo., Ermann, Niklas., & Predki, Paul.. (2019). Digital Data Storage Using DNA Nanostructures and Solid-State Nanopores. Nano letters, 19(2), 1210-1215.

[4]Newman, Sharon., Newman, Sharon., Stephenson, Ashley P., Willsey, Max., & Nguyen, Bichlien H.. (2019). High density DNA data storage library via dehydration with digital microfluidic retrieval. Nature communications, 10(1), 1706.

[5]Anavy, Leon., Vaknin, Inbal., Atar, Orna., Amit, Roee., & Yakhini, Zohar.. (2019). Data storage in DNA with fewer synthesis cycles using composite DNA letters. Nature biotechnology, 37(10), 1229-1236.

[6]Koch, Julian., Gantenbein, Silvan., Masania, Kunal., Stark, Wendelin J., & Erlich, Yaniv.. (2019). A DNA-of-things storage architecture to create materials with embedded memory. Nature biotechnology, 38(1), 39-43.

[7]Lopez, R&olph., Lopez, R&olph., Chen, Yuan-Jyue., Dumas Ang, Siena., & Yekhanin, Sergey.. (2019). DNA assembly for nanopore data storage readout. Nature communications, 10(1), 2933.

[8]Lee, Henry H., Lee, Henry H., Kalhor, Reza., Kalhor, Reza., & Goela, Naveen.. (2019). Terminator-free template-independent enzymatic DNA synthesis for digital information storage. Nature communications, 10(1), 2383.

[9]Erlich, Yaniv., Erlich, Yaniv., Erlich, Yaniv., & Zielinski, Dina.. (2017). DNA Fountain enables a robust and efficient storage architecture. Science (New York, N.Y.), 355(6328), 950-954.

[10]Mayer, Clemens., McInroy, Gordon R., Murat, Pierre., Van Delft, Pieter., & Balasubramanian, Shankar.. (2016). An Epigenetics-Inspired DNA-Based Data Storage System. Angewandte Chemie (International ed. in English), 55(37), 11144-8.

[11]Meiser, Linda C., Antkowiak, Philipp L., Koch, Julian., Chen, Weida D., & Kohll, A Xavier.. (2019). Reading and writing digital data in DNA. Nature protocols, 15(1), 86-101.

[12]O'Reilly, Rachel K., Turberfield, Andrew J., & Wilks, Thomas R.. (2017). The Evolution of DNA-Templated Synthesis as a Tool for Materials Discovery. Accounts of chemical research, 50(10), 2496-2509.

[13]Zhang, Yueyue., Zhang, Yueyue., Li, Fan., Li, Min., & Mao, Xiuhai.. (2019). Encoding Carbon Nanotubes with Tubular Nucleic Acids for Information Storage. Journal of the American Chemical Society, 141(44), 17861-17866.

[14]Zeraati, Mahdi., Zeraati, Mahdi., Langley, David B., Schofield, Peter., & Moye, Aaron L.. (2018). I-motif DNA structures are formed in the nuclei of human cells. Nature chemistry, 10(6), 631-637.

[15]Organick, Lee., Ang, Siena Dumas., Chen, Yuan-Jyue., Lopez, R&olph., & Yekhanin, Sergey.. (2018). Erratum: Random access in large-scale DNA data storage. Nature biotechnology, 36(7), 660.

[16]Kahn, Jason S., Hu, Yuwei., & Willner, Itamar.. (2017). Stimuli-Responsive DNA-Based Hydrogels: From Basic Principles to Applications. Accounts of chemical research, 50(4), 680-690.

[17]Lapique, Nicolas., & Benenson, Yaakov.. (2017). Genetic programs can be compressed and autonomously decompressed in live cells. Nature nanotechnology, 13(4), 309-315.

[18]Demirer, Gozde S., Zhang, Huan., Goh, Natalie S., González-Gr&ío, Eduardo., & L&ry, Markita P.. (2019). Carbon nanotube-mediated DNA delivery without transgene integration in intact plants. Nature protocols, 14(10), 2954-2971. [19]Houlihan, Gillian., Arangundy-Franklin, Sebastian., & Holliger, Philipp.. (2017). Exploring the Chemistry of Genetic Information Storage and Propagation through Polymerase Engineering. Accounts of chemical research, 50(4), 1079-1087. [20]Anavy, Leon., Vaknin, Inbal., Atar, Orna., Amit, Roee., & Yakhini, Zohar.. (2019). Author Correction: Data storage in DNA with fewer synthesis cycles using composite DNA letters. Nature biotechnology, 37(10), 1237.