Engineering 29 (2023) 130–136

Contents lists available at ScienceDirect

Engineering

journal homepage: www.elsevier.com/locate/eng

Research
Material Science and Engineering—Article

Engineering DNA Materials for Sustainable Data Storage Using a DNA Movable-Type System

Zi-Yi Gong [a,#], Li-Fu Song [a,#], Guang-Sheng Pei [b], Yu-Fei Dong [a], Bing-Zhi Li [a,*], Ying-Jin Yuan [a]

[a] Frontiers Science Center for Synthetic Biology & Key Laboratory of Systems Bioengineering (Ministry of Education), School of Chemical Engineering and Technology, Tianjin University, Tianjin 300072, China
[b] School of Biomedical Informatics, The University of Texas Health Science Center at Houston (UTHealth), Houston, TX 77030, USA

ARTICLE INFO

Article history:
Received 7 April 2022
Revised 24 May 2022
Accepted 31 May 2022
Available online 1 July 2023

Keywords:
Synthetic biology
DNA data storage
DNA movable types
Economical DNA data storage

ABSTRACT

DNA molecules are green materials with great potential for high-density and long-term data storage. However, the current data-writing process of DNA data storage via DNA synthesis suffers from high costs and the production of hazards, limiting its practical applications. Here, we developed a DNA movable-type storage system that can utilize DNA fragments pre-produced by cell factories for data writing. In this system, these pre-generated DNA fragments, referred to herein as "DNA movable types," are used as basic writing units in a repetitive way. The process of data writing is achieved by the rapid assembly of these DNA movable types, thereby avoiding the costly and environmentally hazardous process of *de novo* DNA synthesis. With this system, we successfully encoded 24 bytes of digital information in DNA and read it back accurately by means of high-throughput sequencing and decoding, thereby demonstrating the feasibility of this system. Through its repetitive usage and biological assembly of DNA movable-type fragments, this system exhibits excellent potential for writing cost reduction, opening up a novel route toward an economical and sustainable digital data-storage technology.

1. Introduction

Global digitization has led to exponential growth in digital information [1]. However, the density and long-term stability of silicon-based storage technologies are approaching their theoretical limits [2,3]. Moreover, the manufacturing process of silicon-based storage media generates large quantities of environmental hazards, including the media themselves [2–5], which is not consistent with the global vision of green and sustainable development [6]. A dense, long-term stable, and sustainable storage technology is highly desirable to meet the future needs of digital data storage [2,7–12]. DNA is a naturally formed polymer carrying the genetic information of all life on earth [13,14]. Studies have revealed that DNA can also be utilized for the storage of massive amounts of artificial information [7,15–28]. DNA data storage utilizes the complex processes of DNA synthesis and sequencing for the "writing" and "reading" of information. Uniquely, data-encoding DNA molecules can be mixed together, forming a natural three-dimensional (3D) storage system. The extra dimension allows DNA data storage with the outstanding feature of ultra-high density, outpacing traditional planner media [9,19,28–33]. While recent studies have made significant progress in terms of data scale, stability, random access, data replication, and so on, cost has emerged as the key barrier to a practical DNA data-storage system [7,15–21,33,34]. The reading cost of DNA data storage is relatively high compared with that of traditional mediums [14,20]. However, the cost of reading technology is not a significant limitation to the practical applications of DNA data storage, considering the low reading frequency in its main application scenario of long-term archive storage. The main obstacle at present is the expensive and environmentally unfriendly data-writing process of chemical-based DNA synthesis. It has been estimated that a decrease of five to six orders of magnitude is required in the writing cost in order to outpace magnetic-tape-based storage technology [2,9,14,20].

Movable-type printing is the crystallization of the wisdom of ancient China. It uses pre-made movable types rather than fixed-order printing plates, which makes it possible to print any type of documents by placing prepared movable types in a certain order during the printing process [35]. This strategy greatly reduces the cost of text printing through the repetitive and flexible usage of

* Corresponding author.
E-mail address: bzli@tju.edu.cn (B.-Z. Li).
# These authors contributed equally to this work.

https://doi.org/10.1016/j.eng.2022.05.023

movable types. In this study, inspired by movable-type printing technology, we designed and implemented a movable-type-like system for DNA data storage. This DNA movable-type storage system utilizes natural pre-synthesized DNA fragments as basic writing units—that is, as DNA movable types. Rather than depending on DNA synthesis, the data-writing process is achieved by the selection and assembly of specific DNA movable types into longer storage units. A fast and reliable enzymatic assembly process has been designed and optimized for these short fragments, ensuring reliable data writing in DNA. This system, through its repetitive usage of pre-synthesized DNA movable types, shows tremendous promise for low-cost data writing. With further advancements, it has the potential to address the high-cost issue of current DNA data-storage technologies. Significantly, as far as our knowledge extends, this system is a pioneering effort in the realm of data writing, as it is the first system reported that can utilize natural DNA molecules produced by cell factories. This opens up a novel and sustainable approach to digital data storage.

2. Material and methods

2.1. Construction of DNA movable types

All polymerase chain reaction (PCR) amplifications were performed with Phanta Max Super-Fidelity DNA Polymerase (CAT#: P505-d2; Vazyme Biotech Co., Ltd., China) in a 50 μL volume. For the initial construction of each DNA movable type, a 45-base pair (bp) single strand DNA (ssDNA) fragment was used as templates for the PCR amplification. The 6 bp data encoding region is located in the center of the ssDNA fragment. The PCR amplifications were performed with the following primes: P1: 5′-GTCGGTCTCGTCAGATGTAGATAACTACGACGCTCGTGATCATCCATTCTCGTCTAC-3′; P2: 5′-GTGAAGACTCCTGAGTCATTCTCAGAATGACGTAGTTGGTGGTGGAACCTGAGAAGC-3′. The two primers were used to introduce two restriction sites of BsaI and BbsI. The resulting DNA movable-type fragment is 120 bp in length. The PCR amplifications were performed using the following program: an initial denaturation step at 95 °C for 30 s, followed by ten cycles of 95 °C for 15 s, 48 °C for 15 s, and 72 °C for 4 s. This was followed by 20 cycles of 95 °C for 15 s, 68 °C for 15 s, and 72 °C for 4 s, and a final extension step of 72 °C for 5 min, with a final hold at 4 °C.

To prepare a large quantity of DNA movable-type fragments, a different pair of primers with a shorter length was used, as follows: P1: 5′-GTCGGTCTCGTCAGATGTAGATAACTACGAC-3′; P2: 5′-GTGAAGACTCCTGAGTCATTCTCAGAATGAC-3′. The applied PCR amplification program was as follows: 95 °C for 30 s, 95 °C for 15 s, 59 °C for 15 s, and 72 °C for 4 s for 30 cycles; followed by 72 °C for 5 min and hold at 4 °C.

The PCR products were purified by gel purifications before usage. All the gel purifications were performed using the SPARKeasy Gel/DNA Extraction Kit (CAT#: AE0301-B; Shandong Sparkjade Biotechnology Co., Ltd., China).

2.2. Assembly of DNA movable types

The DNA movable types were digested with fast-digesting enzymes before assembly. Thermo Fisher's FastDigest series enzymes (CAT#: FD0294, FD1014) were used for all the restriction enzyme digestion experiments. To perform the digestion, 27.5 μL of the DNA movable-type solution, 5 μL of 10× buffer, and 1 μL of specific restriction enzyme were mixed in a tube and incubated at 37 °C for a controlled time ranging from 5 to 30 min. The digested DNA movable-type fragments were purified before assembly.

For the ligase-mediated assembly, 11 μL of each DNA movable-type fragment, 3 μL of 10× buffer, and 1.5 μL of T_4 DNA ligase (CAT#: M0202L; New England Biolabs (NEB), USA) were mixed in a tube, double distilled water (ddH_2O) was also added to make a total volume of 30 μL. After ligation, the mixture was inactivated by incubating it at 65 °C for 10 min, according to the manual.

In the series of experiments to optimize the ligation process, the same ligation system was used as above, except for the dosages of the ligase and ligation time. Ligase dosages of 0.50, 0.75, 1.25, and 1.50 μL were tested with a ligation time of 20 min. The resulting ligation samples were analyzed using the Agilent Bioanalyzer 2100 system (Agilent Technologies, Inc., USA).

2.3. Cost-estimation settings

In the biological preparation of the DNA movable types, plasmid extraction was performed using the TIANprep Mini Plasmid Kit (CAT#: DP103; Tiangen Biotech (Beijing) Co., Ltd., China). *Eco*RI-HF was used to digest the extracted plasmid, and the buffer used was CutSmart Buffer from NEB (CAT#: R3101L).

2.4. Data and materials availability

All the original algorithms proposed in this study were implemented with Python; the source codes are provided in Appendix A. All sequencing data can be retrieved at https://doi.org/10.6084/m9.figshare.19425236.

3. Results

3.1. The design of the DNA movable-type storage system

As illustrated in Fig. 1(a), DNA movable types with a specific length of data-encoding region are used as basic writing units. Each specific DNA movable type contains a unique sequence combination in the encoding region. All the DNA movable types of all possible sequence combinations must be pre-produced before the actual writing operations can take place. This makes it possible to write arbitrary digital information via the automatic selection and ordered assembly of specific DNA movable types. Importantly, these DNA movable-type fragments are utilized repetitively in this system. The initial synthesis of the DNA movable-type fragments adopts mature oligonucleotide chemical synthesis methods, and subsequent large-scale production can be carried out through biological manufacturing and separation. For data encoding, specific binary information is first transcoded into DNA strings according to the mapping schema shown in Table 1. To avoid decoding failures, one error-checking base is inserted for every two bases, using a strategy reported in a previous study [29]. The DNA sequence generated is then split into massive storage units. Next, to realize the writing of data, a high-throughput device is employed to pick the corresponding DNA movable types and produce the required storage units by means of an optimized process of fragment assembly, as illustrated in Fig. 1(b).

As the core module of the DNA movable-type storage system, a unified structure of the DNA movable types is crucial. As shown in Fig. 1(c), to support the assembly of two arbitrary DNA movable types with a defined order, two Type IIS restriction sites are respectively placed on the two ends of the DNA movable types. BsaI and BbsI are Type IIS endonucleases with the key feature of detached recognition and cleavage sites. With this important feature, the recognition sites can be eliminated after the enzymatic digestion by designing appropriate DNA movable types, thereby avoiding mis-cleavage of the assembled DNA movable types in the next round of assembly. In addition, the issue of self-ligation of DNA

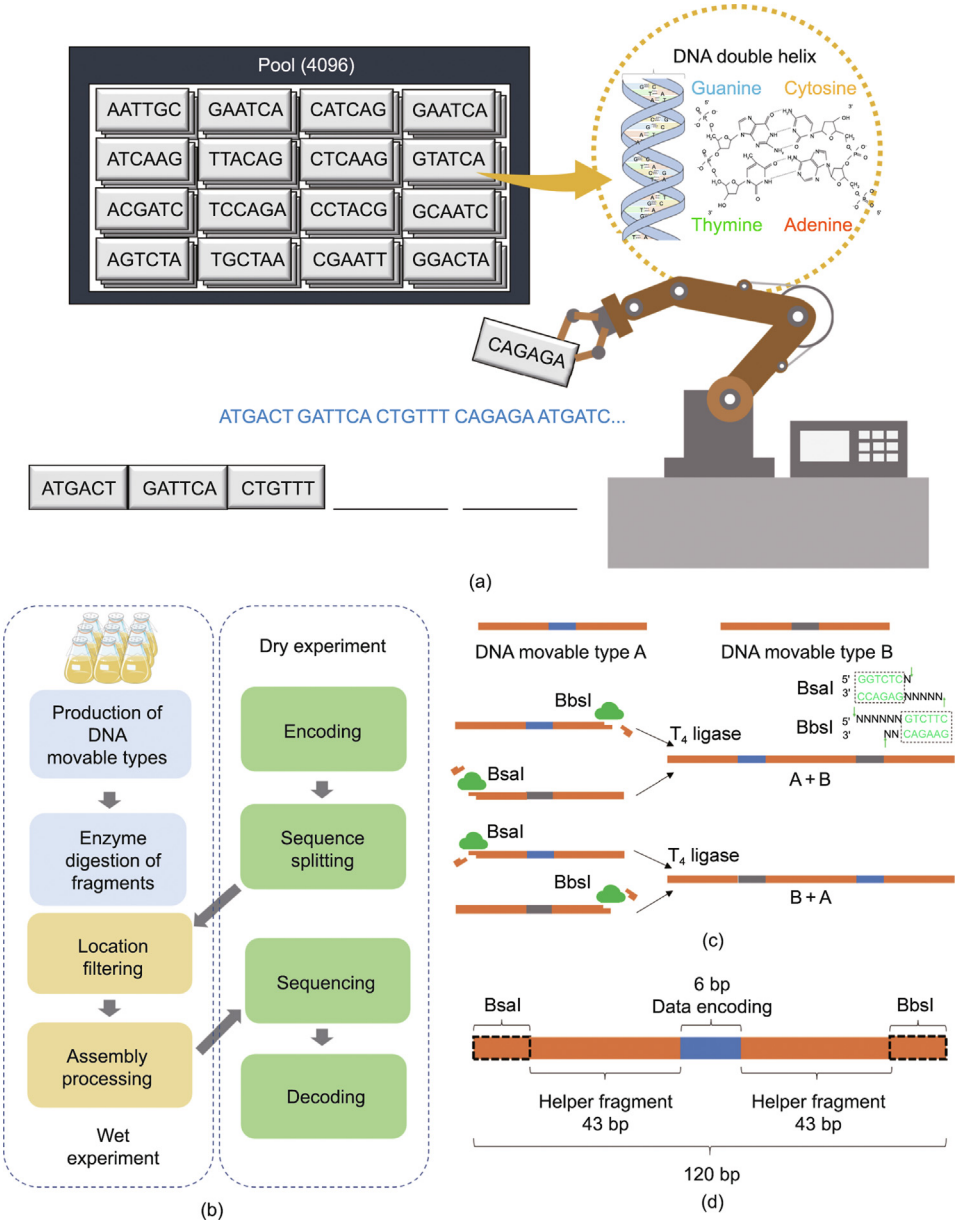

Fig. 1. The principle of the DNA movable-type storage system. (a) Illustration of the data-writing process of the DNA movable-type storage system. For the data-writing process, high-throughput automation equipment is employed to select the desired DNA movable types and assemble them into corresponding storage units with a length of 408 bp. (b) The overall workflow of the data-writing and -reading processes of the DNA movable-type storage system. (c) Diagram of the ordered assembly of DNA movable types. By selectively digesting either with BbsI or BsaI, the representative DNA movable types A and B can be assembled in a desired order (A + B or B + A) using T_4 ligase. Blue and grey areas indicate the data encoding regions. (d) Structure of the DNA movable types. The blue area in the middle stands for the data-encoding region; the two orange modules are helper fragments, which are two randomly generated sequences for improving the ligation efficiency. The two primer binding sets represented by black dotted boxes include two restriction enzyme sites of BbsI and BsaI, respectively. All the DNA movable types have a 6 bp data-encoding region and an overall length of 120 bp. There are 4096 (4^6 = 4096) possible sequence combinations for all 6 bp regions, yielding a total of 4096 unique pre-manufactured DNA movable types (a longer data-encoding region can also be applied, in which case the overall number of DNA movable types required to be pre-manufactured will be correspondingly enlarged).

movable types—a common issue with conventional Type II endonucleases that possess a palindromic sequence in the cleavage site—can be avoided. The two restriction sites have been designed to produce compatible sticky ends, enabling the assembly of two arbitrary DNA movable types via selective digestion of the two ends with the corresponding restriction enzyme. To avoid strand dissociation of the double strands of the DNA movable types, two helper fragments are added to increase the melting temperatures. This also increases the ligation efficiency of T_4 ligase, which is more efficient with double-stranded DNA (dsDNA). Finally, a basic fragment with a length of 120 bp is formed, as shown in Fig. 1(d) and Fig. S1 in Appendix A. The fragments—that is, the DNA movable types—are assembled into storage units after specific rounds of assembly to accomplish the data-writing process. All the storage units can be collected and stored at low temperature for long-term storage. To read the stored data, the storage units must be sequenced by an appropriate sequencing device, such as an Illumina sequencer.

The storage capacity of a system represents its storage potential, and whether it can flexibly cope with the data size is crucial for large-scale data storage. For storage units with a length of L in which the index length is x, the length of the data-encoding region is $L - x$. The information-storage capacity of a single base pair is 2 bits. Thus, the maximum total amount of information that can theoretically be

Table 1
Mapping schema for transcoding between binary information and the DNA string.

Four bits	Two bases
0000	AT
0001	AG
0010	AC
0011	AA
0100	TA
0101	TC
0110	TG
0111	TT
1000	GG
1001	GA
1010	GT
1011	GC
1100	CC
1101	CT
1110	CA
1111	CG

stored in storage units with a length of L is $N = 4^x \times (L - x) \times 2$ bits. It can be concluded that the storage capacity of the fragments increases with an increase in the index length; that is, when $x = L - 1$, N has the maximum value. Based on this formula, with storage units that are 24 bp long, the theoretical storage capacity of this system can reach about 17 terabytes (TB; 1 TB = 10^{12} bytes). In practice, the error correction (EC) codes should be excluded from the calculation. With the three-base block coding applied, the effective coding length of each storage unit is 16 bp ($24/3 \times 2 = 16$), and a minimal data-encoding region is two bases. Thus, we obtain a storage capacity of 134 megabytes (MB; 1 MB = 10^6 bytes). Importantly, the storage capacity of this system can be sustainably improved by using longer storage units, which can be achieved using either DNA movable types with a longer encoding region or more rounds of assembly, as shown in Table 2.

3.2. Design and optimization of the assembly process

The DNA movable-type fragments were digested and then the storage units were assembled via ligation-based assembly, using T_4 ligase. T_4 ligase is very sensitive to temperature. To optimize the assembly efficiency and shorten the assembly time, we tested the ligation efficiency of T_4 ligase at various reaction times (15 s, 30 s, 1 min, 2 min, 5 min, and 10 min) and temperatures (16, 26, and 37 °C). Figs. 2(a) and (b) show that ligation products can be observed after a 15 s reaction time, with the efficiency of ligation steadily improving with increased ligation time. In the experimental group, ligase at 16 °C showed good ligation efficiency within 2 min, making it suitable for cases with a tight assembly time. When the reaction is carried out at 26 °C, the ligation efficiency shows a more significant growth trend with the extension of time, particularly in the 5–10 min interval, which exceeds the effect at 16 °C for the same time. At this temperature, the concentration of assembled fragments is relatively high, allowing direct support for subsequent experiments without amplification. Moreover, there is no need for temperature adjustment facilities like a PCR instrument, making it easy to operate under normal environmental requirements. However, when the ambient temperature is changed to 37 °C, the ligation effects only slightly increase with prolonged ligation time. After incubation at 37 °C for more than 5 min, the effect is comparable to that of 1 min at 16 °C. When the temperature is increased to 45 °C, the ligation efficiency no longer shows a significant trend with time, but ligated fragments can still be observed, indicating that a certain ligation effect was achieved. Although the concentration of ligated fragment is low, they could be subsequently amplified using PCR before the next round of assembly.

These results indicate that reliable data writing can be achieved in the DNA movable-type storage system under normal operating temperatures ranging from 16 to 45 °C. The main factor contributing to the writing cost is the consumption of ligase. Therefore, reducing the amount of ligase used in the system could decrease the writing costs. In an additional experiment, the ligase added was gradually reduced from 1 unit to 5/6, 2/3, 1/2, and 1/3 of a unit at room temperature (26 °C). Although a slight decrease in ligation efficiency was observed as the ligase dosage was reduced, the subsequent assembly process remained reliable (Figs. 2(c) and (d)). This assembly-based writing process is highly adaptable and allows for balancing the cost and action time by adjusting the dosage of added enzymes and the reaction time.

3.3. Proof-of-concept verification using text information

A short text with the first names of three scientists who made great contributions to the discovery of the DNA double-helix structure—namely, "Watson, Crick, Franklin."—was used as the input information for the proof-of-concept verification. To deal with the base error of data in DNA synthesis, storage, amplification, and sequencing and to ensure the reliable recovery of information, block EC codes were introduced in the process of binary data conversion. Finally, a string of base sequences with a length of 144 bp was obtained. This sequence was split into a series of small fragments of 6 bp for a total of 24 DNA movable types (Table S1 in Appendix A). These DNA movable types were then found in the pre-generated fragment "pool" by means of location screening. Each DNA movable type is 120 bp in length, as shown in Fig. 1(d). Four DNA movable types were assembled into one storage unit after two rounds of assembly (Fig. 3(a)). For each assembly process, 24 nucleic acid residues were removed. Thus, the length of an assembled storage unit is 408 bp ($120 \times 4 - 24 \times 3 = 408$). As shown in Fig. 3(b), agarose gel electrophoresis analysis of the assembled products revealed a band corresponding to a length of 250–500 bp, as expected. To reduce the space occupied, the fragments can be mixed together. However, the storage units are out of order in this case. In order to decode the data correctly, the order information of the storage units must be encoded in a certain way. To achieve this, the first two DNA movable types of each storage unit were designed to be the same as the last two DNA movable types of the previous storage units, to determine the order of the storage units (Fig. 3(c)). A total of 11 storage units were assembled to encode the 24 bytes of text information. It should be noted that a small fraction of the storage unit fragments may be lost without

Table 2
Storage capacity of the designed DNA movable-type system for DNA movable types with varying lengths of data-encoding regions and for different rounds of assembly.

Encoding length of DNA movable types	Number of movable types	Rounds of assembly		
		1	2	3
6 bp	2 960	2 KB (12 bp)	134 MB (24 bp)	576 PB (48 bp)
9 bp	75 584	524 KB (18 bp)	9 TB (36 bp)	2 BB (72 bp)
12 bp	1 427 200	134 MB (24 bp)	576 PB (48 bp)	11 CB (96 bp)

The numbers shown in parenthesis represent the total length of data encoding regions after multiple rounds of assembly.
KB: kilobytes, 1 KB = 10^3 bytes; PB: petabytes, 1 PB = 10^{15} bytes; BB: brontobyte, 1 BB = 10^{27} bytes; CB: corydonbytes, 1 CB = 10^{36} bytes.

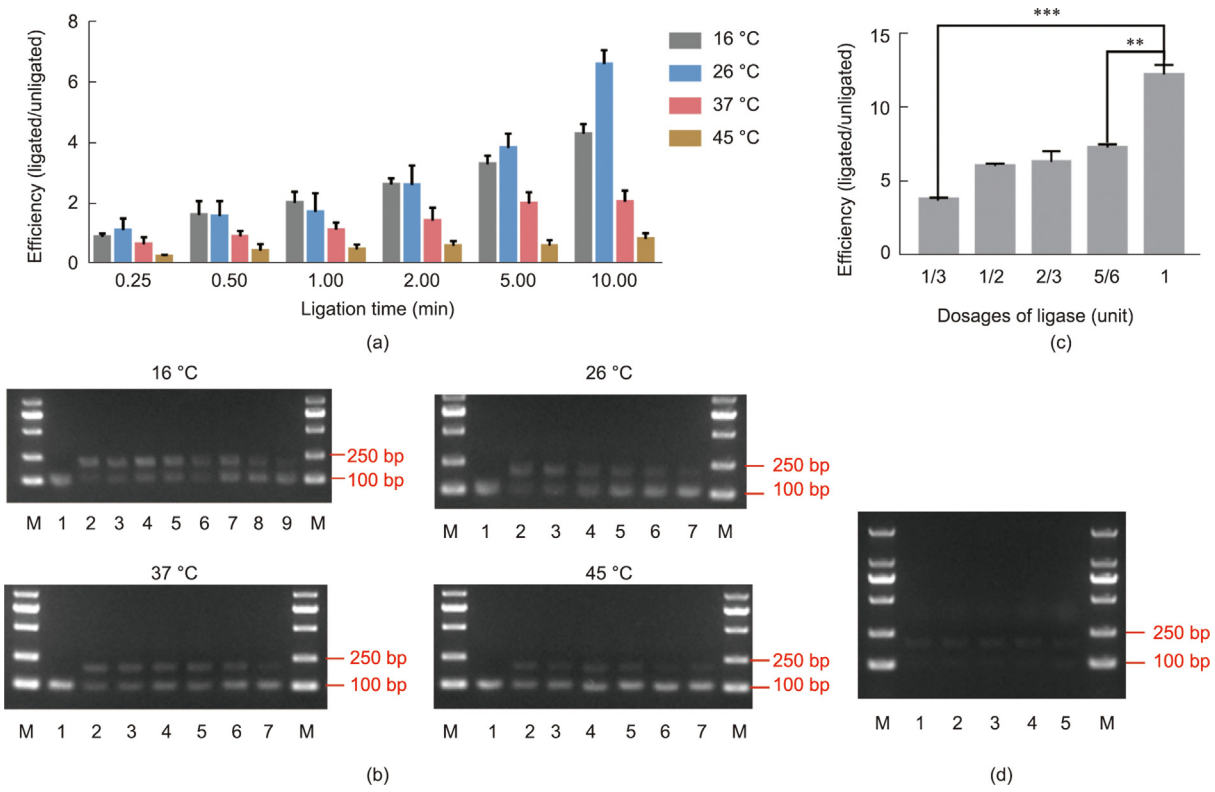

Fig. 2. Assembly process optimization. (a) Assembly efficiency under various temperatures and reaction times. (b) DNA agarose gel electropherogram at different temperatures and under different reaction times. Sample 1 is the control group without ligase. The reaction time of samples 4–9 at 16 °C and of samples 2–7 at 26, 37, and 45 °C is gradually reduced from 10 min to 15 s, respectively. For comparison, the connection results at 16 °C for 30 and 20 min are also tested, as shown in samples 2 and 3. (c) Assembly efficiency with various dosages of ligase. (d) DNA agarose gel electropherogram at various dosages of ligase. The dosage of enzyme per reaction unit in samples 1–5 was gradually reduced from 1 unit to 1/3 of a unit. M: marker.

the possibility of recovery when massive storage units are utilized for large data storage. Thus, additional erase codes are definitely required to handle this issue [7,15–19,36]. Data retrievals were performed independently with Sanger sequencing and Illumina sequencing methods. All storage units encoding tiny pieces of information were mixed in a tube for high-throughput sequencing via the Illumina sequencing platform. In both cases, perfect data recovery was achieved with the obtained sequencing reads (Figs. S2–S13 and Table S2 in Appendix A), demonstrating the feasibility of the proposed DNA movable-type storage system.

3.4. Estimation of the potential of low-cost data writing

The cost of this storage system is comprised of the production cost of the DNA movable types and the assembly cost for generating the storage units. The production cost of the DNA movable types has two parts: the initial construction cost and the reproduction cost. For the initial construction of all DNA movable types, the synthesis cost of each base in the DNA single strand is around 0.039 USD, based on previous reports [14,20], and the length of each single-strand template is 90 bp. For the 6 bp data-encoding region, there are 4096 ($4^6 = 4096$) different DNA movable types, considering all the sequence combinations. The two universal primers are both 31 bp in length. With an optimized production process, the required synthesized single-stranded template length was reduced to 45 bp and both universal primers are 57 bp in length, resulting in a synthesis cost of 7192.926 USD ($0.039 \times (45 \times 4096 + 57 \times 2) = 7192.926$).

It should be clarified that this initial manufacturing process only needs to be performed once. In other words, this initial construction cost is the research and development (R&D) cost, not the data-writing cost in practical usage. The reproduction of DNA movable types can be achieved via cheap PCR or even cell factories (Fig. S14 in Appendix A). Importantly, the DNA movable types can be used multiple times for the assembly of various storage units—that is, for the storage of various digital data. As shown in this study, even at a low dosage of 0.5 ng·μL^{-1}, efficient assembly can still be achieved. To achieve economical and accurate production of DNA movable-type fragments, large-scale production can be performed using cell factories, which can further reduce the writing costs. Considering that the commercial price of the kits used includes considerable profit and R&D recovery funds, and the cost of large-scale production would be much lower than this price. Therefore, 1% of the commercial price of the kits is used to calculate the cost more reasonably.

Assuming the volume of a single tube for cell cultivation is V_1, and take $V_1 = 5$ mL for the cost estimation. The main cost of the cell cultivation comes from the medium. Taking the price of the lysogeny broth (LB) medium for estimation, $a_1 = 0.63$ USD·L^{-1} ($63.16 \times 1\% \approx 0.63$), the cost of LB medium of 5 mL cell culture is $W_1 = a_1 V_1/1000 = 3.15 \times 10^{-3}$ USD. According to the plasmid extraction kit instructions, a maximum of 5 mL of plasmids can be extracted at one time. According to the kit specifications and price, the cost of a single plasmid extraction is $a_2 = 5.7 \times 10^{-3}$ USD ($0.57 \times 1\% = 5.7 \times 10^{-3}$), and the cost of a single test tube plasmid extraction is $W_2 = a_2 = 5.7 \times 10^{-3}$ USD. For the high-copy plasmid pUC19, a plasmid yield of 5 mL of bacterial liquid can reach $m_1 = 30$ μg, and the concentration of plasmid in the final extract is $c = 6$ μg·mL^{-1}. The main cost of enzyme digestion and product recovery comes from the reagents used in the enzyme digestion process. According to the price and dosage of the reagents used in digesting the extracted plasmid, the digestion cost is

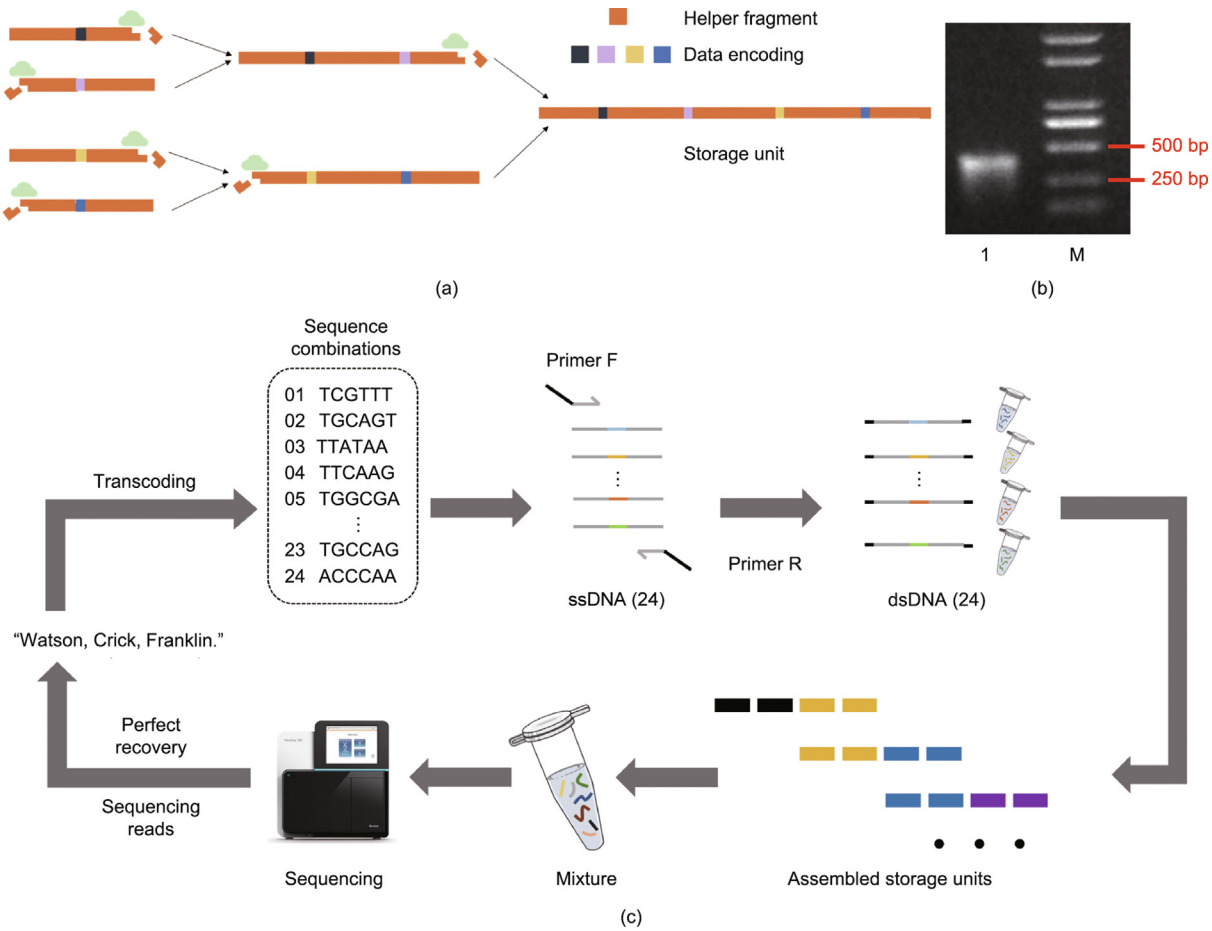

Fig. 3. Proof-of-concept experimental verification with short text information. (a) Illustration of the assembly process of storage units. (b) Gel electrophoresis analysis of the assembled storage units, which are designed to be 408 bp in length. (c) Overall workflow of the data-writing and -reading process. For the proof of concept, 24 bytes of text information were used as encoding data: "Watson, Crick, Franklin." Primers F and R mean forward primer and reverse primer, respectively.

approximately $a_3 = 0.79$ USD·mg^{-1} ($78.8 \times 1\% \approx 0.79$) of DNA, and the cost of the digestion and product-recovery steps is approximately $W_3 = a_3 m_1/1000 = 2.37 \times 10^{-2}$ USD. In this experiment, the OD_{600} (OD_{600} stands for the optical density at 600 nm) of the bacterial solution reached 2.4 after overnight cultivation, corresponding to a bacterial solution concentration of about $C = 2.4 \times 10^8$ mL^{-1}, while the copy number of the pUC19 plasmid was about $e = 600$. Assuming that the extraction efficiency of the short film segment $\eta = 50\%$, then the number of short fragments of information-storage DNA that can be extracted from 5 mL of bacterial solution is $n = C \cdot V_1 \cdot e \cdot \eta = 3.6 \times 10^{11}$. According to the results of the above calculation, the total cost of the production of short DNA fragments is approximately $W = W_1 + W_2 + W_3 \approx 3.26 \times 10^{-2}$ USD. The effective information-storage capacity of one storage unit in this system is $S = 6$ bytes; for each information-storage instance, the number of copies requiring the same DNA short fragments $n_1 \approx 1000$. As such, the biological preparation cost of DNA movable types is $W' = n_1 W/(nS \times 10^{-6}) \approx 1.5 \times 10^{-5}$ USD·MB^{-1}. This system has the potential to significantly reduce the writing cost of DNA data storage by several orders of magnitude.

4. Discussion and conclusions

DNA molecules are green materials that are abundant in nature. However, naturally produced DNA molecules cannot yet be used for the storage of artificial digital data. The reliance on expensive and hazardous DNA synthesis methods for data writing restricts the practical applications of current DNA data storage solutions [37]. In this study, inspired by movable-type printing technology, we implemented a DNA movable-type storage system. With this system, pre-produced DNA materials can be engineered as basic writing units for the storage of arbitrary digital information. Through the repetitive usage of these basic writing units, herein referred to as "DNA movable types," this system exhibits great potential for reducing the cost of data writing in DNA data storage. The data-writing process of this DNA movable-type storage system is achieved through rounds of assembly, producing the storage units with the desired length. As shown in Table 2, the storage capacity of this DNA movable-type storage system can be strikingly enhanced by adding more rounds of assembly. As shown here, in the case of DNA movable types with an encoding length of 6 bp, the storage capacity of this system is as high as a staggering 576 petabytes (PB; 1 PB = 10^{15} bytes) with merely three rounds of assembly. The storage capacity can be further enhanced by using DNA movable types with longer encoding regions. With encoding regions of 9 and 12 bp, the storage capacity could be increased to 2 brontobytes (BB; 1 BB = 10^{27} bytes) and 11 corydonbytes (CB; 1 CB = 10^{36} bytes), respectively, showcasing the system's potential for large-scale data storage. Alternatively, by incorporating additional rounds of assembly, the storage capacity could be greatly boosted without the need for an increase in the total number of DNA movable types. It should be noted that using the highest coding capacity scheme will result in lower coding efficiency. Coding efficiency is considered to be essential as it impacts the cost and physical density—two crucial parameters indicating the

advancement of a DNA storage system. Coding efficiency has gained great attention previously, especially in DNA synthesis-based methods where it directly affects the writing cost [16–18]. However, in DNA movable-type storage system, although coding efficiency is still important, the cost and density are less sensitive to it. Theoretically, high-density and low-cost data storage can be achieved by reducing the strand copies of the DNA movable-type fragments, even with low coding efficiency. In practice, as long as high density and low cost are achieved, relatively low coding efficiency is no longer an issue. The total number of DNA movable types is a crucial parameter determining the complexity of implementing an automatic machine to execute the write operations. Thus, adding more rounds of assembly is the preferred strategy to increase the storage capacity of the system. Through process optimization, it takes less than 5 min to accomplish one round of assembly, suggesting a fast data-writing process compared with DNA synthesis-based data writing. Importantly, the production of these DNA materials and the assembly-based data-writing process being completely biological processes, make the current system a green and sustainable DNA data-storage technique. Furthermore, this data-writing process can be easily parallelized. With the ongoing advancement of high-throughput automation solutions, the data write bandwidth of this system is expected to be improved substantially. In conclusion, the DNA movable-type storage system presents a promising route towards economical, sustainable, and environmentally friendly data storage solutions to meet the future demands of large data storage.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (2018YFA0900100), the Natural Science Foundation of Tianjin, China (19JCJQJC63300) and Tianjin University.

Compliance with ethics guidelines

Zi-Yi Gong, Li-Fu Song, Guang-Sheng Pei, Yu-Fei Dong, Bing-Zhi Li, and Ying-Jin Yuan declare that they have no conflict of interest or financial conflicts to disclose.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.eng.2022.05.023.

References

[1] Reinsel D, Gantz J, Rydning J. Data Age 2025: the evolution of data to life-critical. Framingham: International Data Corporation; 2017.
[2] Zhirnov V, Zadegan RM, Sandhu GS, Church GM, Hughes WL. Nucleic acid memory. Nat Mater 2016;15(4):366–70.
[3] Fontana Jr RE, Decad GM, Hetzler SR. Volumetric density trends (TB/in.^{3}) for storage components: TAPE, hard disk drives, NAND, and Blu-ray. J Appl Phys 2015;117(17):17E301.
[4] Davis LA. Clean energy perspective. Engineering 2017;3(6):782.
[5] Xie MH, Duan HB, Kang P, Qiao Q, Bai L. Toward an ecological civilization: China's progress as documented by the second national general survey of pollution sources. Engineering 2021;7(9):1336–41.
[6] Han MJ, Yoon DK. Advances in soft materials for sustainable electronics. Engineering 2021;7(5):564–80.
[7] Church GM, Gao Y, Kosuri S. Next-generation digital information storage in DNA. Science 2012;337(6102):1628.
[8] Rutten MGTA, Vaandrager FW, Elemans JAAW, Nolte RJM. Encoding information into polymers. Nat Rev Chem 2018;2(11):365–81.
[9] Ceze L, Nivala J, Strauss K. Molecular digital data storage using DNA. Nat Rev Genet 2019;20(8):456–66.
[10] Dong Y, Sun F, Ping Z, Ouyang Q, Qian L. DNA storage: research landscape and future prospects. Natl Sci Rev 2020;7(6):1092–107.
[11] Dickinson GD, Mortuza GM, Clay W, Piantanida L, Green CM, Watson C, et al. An alternative approach to nucleic acid memory. Nat Commun 2021;12(1):2371.
[12] Zhang Y, Kong L, Wang F, Li B, Ma C, Chen D, et al. Information stored in nanoscale: encoding data in a single DNA strand with Base64. Nano Today 2020;33:100871.
[13] Watson JD, Crick FHC. Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. Nature 1953;171(4356):737–8.
[14] Song LF, Deng ZH, Gong ZY, Li LL, Li BZ. Large-scale *de novo* oligonucleotide synthesis for whole-genome synthesis and data storage: challenges and opportunities. Front Bioeng Biotechnol 2021;9:689797.
[15] Goldman N, Bertone P, Chen S, Dessimoz C, LeProust EM, Sipos B, et al. Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. Nature 2013;494(7435):77–80.
[16] Grass RN, Heckel R, Puddu M, Paunescu D, Stark WJ. Robust chemical preservation of digital information on DNA in silica with error-correcting codes. Angew Chem Int Ed Engl 2015;54(8):2552–5.
[17] Erlich Y, Zielinski D. DNA Fountain enables a robust and efficient storage architecture. Science 2017;355(6328):950–4.
[18] Organick L, Ang SD, Chen YJ, Lopez R, Yekhanin S, Makarychev K, et al. Random access in large-scale DNA data storage. Nat Biotechnol 2018;36:242–8. Erratum in: Nat Biotechnol 2018;36:660.
[19] Song L, Geng F, Gong ZY, Chen X, Tang J, Gong C, et al. Robust data storage in DNA by de Bruijn graph-based *de novo* strand assembly. Nat Commun 2022;13(1):5361.
[20] Antkowiak PL, Lietard J, Darestani MZ, Somoza MM, Stark WJ, Heckel R, et al. Low cost DNA data storage using photolithographic synthesis and advanced information reconstruction and error correction. Nat Commun 2020;11(1):5345.
[21] Tabatabaei SK, Wang B, Athreya NBM, Enghiad B, Hernandez AG, Fields CJ, et al. DNA punch cards for storing data on native DNA sequences via enzymatic nicking. Nat Commun 2020;11(1):1742.
[22] Xu C, Ma B, Gao Z, Dong X, Zhao C, Liu H. Electrochemical DNA synthesis and sequencing on a single electrode with scalability for integrated data storage. Sci Adv 2021;7(46):eabk0100.
[23] Chen W, Han M, Zhou J, Ge Q, Wang P, Zhang X, et al. An artificial chromosome for data storage. Natl Sci Rev 2021;8(5):nwab028.
[24] Hao M, Qiao H, Gao Y, Wang Z, Qiao X, Chen X, et al. A mixed culture of bacterial cells enables an economic DNA storage on a large scale. Commun Biol 2020;3(1):416.
[25] Koch J, Gantenbein S, Masania K, Stark WJ, Erlich Y, Grass RN. A DNA-of-things storage architecture to create materials with embedded memory. Nat Biotechnol 2020;38(1):39–43.
[26] Lin KN, Volkel K, Tuck JM, Keung AJ. Dynamic and scalable DNA-based information storage. Nat Commun 2020;11(1):2981.
[27] Matange K, Tuck JM, Keung AJ. DNA stability: a central design consideration for DNA data storage systems. Nat Commun 2021;12(1):1358.
[28] Ren Y, Zhang Y, Liu Y, Wu Q, Su J, Wang F, et al. DNA-based concatenated encoding system for high-reliability and high-density data storage. Small Methods 2022;6(4):2101335.
[29] Song L, Zeng AP. Orthogonal information encoding in living cells with high error-tolerance, safety, and fidelity. ACS Synth Biol 2018;7(3):866–74.
[30] Meier MAR, Barner-Kowollik Christopher BK. A new class of materials: sequence-defined macromolecules and their emerging applications. Adv Mater 2019;31(26):1806027.
[31] Boukis AC, Meier MAR. Data storage in sequence-defined macromolecules via multicomponent reactions. Eur Polym J 2018;104:32–8.
[32] Martens S, Landuyt A, Espeel P, Devreese B, Dawyndt P, Du Prez F. Multifunctional sequence-defined macromolecules for chemical data storage. Nat Commun 2018;9(1):4451.
[33] Anavy L, Vaknin I, Atar O, Amit R, Yakhini Z. Data storage in DNA with fewer synthesis cycles using composite DNA letters. Nat Biotechnol 2019;37:1229–36. Correction in: Nat Biotechnol 2019;37:1237.
[34] Meiser LC, Antkowiak PL, Koch J, Chen WD, Kohll AX, Stark WJ, et al. Reading and writing digital data in DNA. Nat Protoc 2020;15(1):86–101.
[35] Yang M, Liu M, Cheng J, Wang H. A movable type bioelectronics printing technology for modular fabrication of biosensors. Sci Rep 2021;11(1):22323.
[36] Ping Z, Ma D, Huang X, Chen S, Liu L, Guo F, et al. Carbon-based archiving: current progress and future prospects of DNA-based data storage. GigaScience 2019;8(6):giz075.
[37] Palluk S, Arlow DH, de Rond T, Barthel S, Kang JS, Bector R, et al. *De novo* DNA synthesis using polymerase-nucleotide conjugates. Nat Biotechnol 2018;36(7):645–50.