Cost-effective DNA storage with DNA movable type

Chenyang Wang*[1,2], Di Wei*[1], Zheng Wei*[3], Dongxin Yang*[4], Jing Xing*[3], Yunze Wang*[1,2], Xiaotong Wang*[1], Pei Wang[4], Guannan Ma[1], Xinru Zhang[2,3], Haolan Li[4], Chuan Tang[4], Pengfei Hou[4], Jie Wang[1,2], Renjun Gao[5], Guiqiu Xie[5], Cuidan Li[1], Yingjiao Ju[1,2], Peihan Wang[1,2], Liya Yue[1], Yongliang Zhao[1], Yongjie Sheng[5], Jingfa Xiao[1], Haitao Niu[6], Sihong Xu[7], Bo Duan[†2,3,4], Dongbo Bu[†2,3,4,9], Guangming Tan[†2,3,4], Fei Chen[†1,2,6,8]

[1]*Beijing Institute of Genomics, Chinese Academy of Sciences, China National Center for Bioinformation, Beijing, China;*

[2]*University of Chinese Academy of Sciences, Beijing, China;*

[3]*SKLP, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China;*

[4]*Western Institute of Computing Technology, Chongqing, China;*

[5]*Key Laboratory for Molecular Enzymology and Engineering of Ministry of Education, School of Life Sciences, Jilin University, Changchun, China;*

[6]*Key Laboratory of Viral Pathogenesis & Infection Prevention and Control (Jinan University), Ministry of Education, Guangzhou, China;*

[7]*Division II of In Vitro Diagnostics for Infectious Diseases, Institute for In Vitro Diagnostics Control, National Institutes for Food and Drug Control, Beijing, China;*

[8]*State Key Laboratory of Pathogenesis, Prevention and Treatment of High Incidence Diseases in Central Asia, Clinical Medicine Institute, The First Affiliated Hospital of Xinjiang Medical University, Urumqi, Xinjiang, China;*

[9]*Central China Institute for Artificial Intelligence, Zhengzhou, China;*

*These authors contributed equally to this work

†Correspondence to: *chenfei@big.ac.cn* (Fei Chen), *tgm@ict.ac.cn* (Guangming Tan), *dbu@ict.ac.cn* (Dongbo Bu), *duanbo@ncic.ac.cn* (Bo Duan).

Summary

In the face of exponential data growth, DNA-based storage offers a promising solution for preserving big-data. However, most existing DNA storage methods, akin to traditional block printing, require costly chemical synthesis for each individual data file, adopting a sequential, one-time-use synthesis approach. To overcome these limitations, we introduce a novel, cost-effective "DNA-Movable-Type Storage" system, inspired by movable type printing. This system utilizes pre-fabricated DNA movable types—short, double-stranded DNA oligonucleotides encoding specific payload, address, and checksum data. These DNA-MTs are enzymatically ligated/assembled into cohesive sequences, termed "DNA movable type blocks", streamlining the assembly process with the automated BISHENG-1 DNA-MT inkjet printer. Using BISHENG-1, we successfully printed, assembled, stored and accurately retrieved 43.7 KB of data files in diverse formats (text, image, audio, and video) *in vitro* and *in vivo*, using only 350 DNA-MTs. Notably, each DNA-MT, synthesized once (2 *OD*), can be used up to 10,000 times, reducing costs to 121.57 $/MB—outperforming existing DNA storage methods. This innovation circumvents the need to synthesize entire DNA sequences encoding files from scratch, offering significant cost and efficiency advantages. Furthermore, it has considerable untapped potential to advance a robust DNA storage system, better meeting the extensive data storage demands of the big-data era.

Introduction

In the era of big data, diverse data, such as text, image, audio, and video, are being generated at an unprecedented rate, with predictions that global data will exceed 380 zettabytes by 2028[1]. However, current storage media (hard disk, magnetic tape, *etc*) are unable to meet the storage demands due to their large physical size, limited lifetime (up to around 30 years), and high energy consumption[2], resulting in an estimation that less than 2% of this data will be preserved[1]. Consequently, new techniques suitable for large-size data storage are highly desired.

DNA-based data storage presents a promising solution to big data storage, offering high density, longevity, and low energy consumption[3,4,5,6,7]. For instance, the human genome, containing over 3 giga base pairs, weighs only 3 picograms[8]; moreover, successful sequencing genomes of ancient animals like the mammoth's[9], underscores DNA's potential for preserving data over millennia with minimal energy. These remarkable attributes make DNA-based storage an ideal candidate for addressing large-scale and long-term data storage challenges.

Since 2010, advancements in DNA-based data storage technology have rapidly accelerated, driven by breakthroughs in DNA sequencing and synthesis technologies[10,11]. Notable milestones include the introduction of "bit-to-base" encoding by Church's group in 2012, which facilitates the translation of digital data into DNA sequences[12], the development of error-correcting codes by Goldman in 2013 to enhance data integrity[13], and the implementation of Erlich's fountain codes in

increasing storage density[14]. The establishment of the "DNA Data Storage Alliance" in 2020 represents a significant step towards an interoperable storage framework based on DNA data storage[15].

Despite these advances, it remains experimental due to the high DNA synthesis cost, substantially constraining its practicality[11,16]. Extensive researches are ongoing to reduce costs by improving synthesis efficiency through column- or microarray-based methods and exploring enzyme-mediated synthesis[17,18,19,20,21,22]. Additionally, innovative encoding schemas like fountain code[14], yin–yang codec system[23], and composite DNA letters[24] have been developed to improve storage density towards Shannon information capacity[25]. Although these improvements have significantly lowered the costs of DNA storage, the current expenses associated with DNA-based storage remain too high for practical application.

A closer look reveals that most existing DNA storage methods resemble woodblock printing, necessitating costly and time-consuming chemical synthesis for each data file from scratch. This sequential (base-by-base), one-time-use synthesis approach significantly hinders its practicality, highlighting an urgent need for more efficient and reusable strategies.

Addressing the bottlenecks, our study draws inspiration from movable type printing to innovatively propose "DNA movable type (DNA-MT) storage". Unlike traditional one-time-use approaches prevalent in existing DNA storage technologies, this method employs reusable, pre-fabricated "DNA movable types (DNA-MTs)" that are assembled via a one-step enzymatic ligation process. This innovation mirrors the

historical transition from block to movable type printing, offering significant cost and efficiency advantages. It allows for the selection and reuse of pre-made DNA-MTs, which significantly lowers DNA synthesis and storage costs. Additionally, the enzymatic ligation process is faster than existing single-use chemical synthesis methods, thereby enhancing the writing efficiency in DNA storage.

To implement DNA-MT storage, we developed a new system centered on a suite of pre-made double-stranded oligonucleotides, termed “DNA-MTs”, each encoding specific payload or address data. Additionally, we introduced column- and row-checksum mechanism for error detection and correction. Using a DNA-MT code table, we then selected the appropriate pre-made DNA-MTs for payloads, addresses, and checksums of target files, and enzymatically ligated them into cohesive sequences termed “DNA-MT blocks”. To automate the block assembly process, we engineered a DNA-MT inkjet printer, “BISHENG-1”, which handles the selection, injection, and ligation of DNA-MTs. Following assembly, these blocks are stored *in vitro* or *in vivo* in *E. coli.*

Utilizing this system, we successfully achieved DNA-MT printing, storage (both *in vitro* and *in vivo*), and 100% accurate decoding of 43.7 KB of text, images, audio, and video data (Extended Data Table 1) using only 350 DNA-MTs. This technology, in conjunction with the "BISHENG-1" system, demonstrate cost advantages in data writing, enabling the DNA-MTs to be used/printed up to 10,000 times after a single synthesis (2 *OD*). Currently, with a throughput of 556 and a printing speed of 4 bytes/second, the cost can be reduced to 121.57 $/MB, surpassing all reported DNA

storage technologies[16]. Additionally, BISHENG-1 printed 43.7 KB in ~5 hours, showcasing its efficiency advantage over sequential one-time-use synthesis methods (~100 hours). Looking ahead, this technology holds significant untapped potential for further cost and efficiency improvements, promising to evolve a more robust DNA storage ecosystem to meet the rapidly growing storage demands of the big data era.

Results

DNA Movable Type storage principle

DNA-MT storage employs a comparable technique for data encoding and storage storage with inspiration from ancient movable type printing, exemplified by *Shakespeare's Sonnet 12* (Fig. 1).

Historically, the replacement of woodblock printing by movable type printing greatly facilitated the spread of culture and the advancement of civilization. This was mainly due to its significant advantages in cost and efficiency over woodblock printing. Movable type printing is based on reusable pre-fabricated wooden or metal types for each character, allowing for their free selection and assembly into a frame to print texts (Fig. 1a, left panel). This approach enables the efficient assembly of pages using pre-made characters, significantly reducing the costs and labor required for printing. In contrast, woodblock printing is a single-use, costly, and labor-intensive method that requires a unique carved block for each page. This method lacks the reusability and flexibility of movable type printing, making it more expensive and time-consuming.

DNA-MTs, serving as the cornerstone of this method, are pre-synthesized, short, double-stranded DNA fragments (20-nt, termed “information carriers”), each flanked by sticky ends (4-nt, termed “linkers”) to facilitate easy linkage. Each DNA-MT encodes a byte of payload/address/checksum information. The method enables the selective ligation of multiple DNA-MTs into a cohesive, extended sequence, forming a “DNA-MT block” with a distinct address and checksum, thus improving the DNA storage’s precision. These blocks are then directly cloned into the optimized *pUC19* plasmid (termed “DNA-MT-plasmid”) (details in Methods section) using *Eco*RI and *Hin*dIII restriction sites on the outermost DNA-MTs, enabling preservation either *in vitro* within the plasmid or *in vivo* through *E. coli* transformation. Similar to the reusability and flexibility of movable type printing, the DNA-MT storage method offers significant cost and efficiency advantages.

Fig. 1a illustrates five payload DNA-MTs (PMTs) within a DNA-MT block, encoding the word “white” from *Sonnet 12*. The first PMT, encoding “w”, contains an “information carrier” sequence “TCACC…CCTCC/AGTGG…GGAGG”, with its 3’-end linker “*ttca*” pairing with the 5’-end linker “*aagt*” of the second PMT for “h”. Similarly, the 3’-end linker of the second type pairs with the 5’-end linker of the third one (for “i”), continuing this pattern to yield a DNA sequence “*atag*TCACCAATCACACCTCCTCC*aagt*…CGAATGGGTGGAGTGTAGGG” that represents “white”. In addition to these PMTs, four address DNA-MTs (AMTs) and a checksum DNA-MT (CMT) are integrated for precise localization and error correction (Fig. 1b). Employing T4 DNA ligase, these ten DNA-MTs are ligated via

nine linkers into a cohesive double-stranded oligonucleotide block (DNA-MT block) in a single step, encoding the word "white" at the 340th character address (notated as "0, 0, 0, 98") with a checksum of 5 at the end. Notably, the two block's terminal linkers, *Eco*RI and *Hin*dIII restriction sites, facilitate direct cloning the block into the DNA-MT-plasmid for *in vitro* storage or transformation into *E. coli* for *in vivo* storage. Moreover, our DNA-MT storage technology exhibits good versatility and scalability, as evidenced by our successful encoding and storage of files with various formats including text, images, audio, and video (detailed in Extended Data Table 1).

Four-phase workflow of DNA-MT storage

DNA-MT storage comprises four key steps: encoding, assembling/printing, storage, and decoding, as detailed in Fig. 2.

(1) Encoding the target file into DNA-MTs (Encoding):

Fragmentation, stripe formation, and binary conversion: First, we divide the target file (*Sonnet 12*) into 100-byte/character fragments, subsequently splitting each fragment into 20 addressed data strips to form a stripe. Each strip contains five bytes/characters (payload) and the address of the first character. Fig. 2 displays 20 consecutive data strips (row 0 – 19) of the fourth stripe, spanning payloads from the 300th to 399th character ("ed up...with white...do I question ma"). The first data strip encodes "ed up", beginning at the 300th character, and the last strip encodes "on ma", starting at the 395th character. The word "white", introduced in the previous section, is in the 8th data strip, beginning at the 340th character position. The binary

representations of these payload and address data (Fig. 6b) are ultimately transformed into DNA-MT IDs presented in decimal format.

Error detection and correction: To address sequencing errors like base insertions, deletions and substitutions, and incorrect connections from enzymatic ligation failures, we introduce additional strips and DNA-MTs for error detection and correction. Utilizing the Reed-Solomon algorithm[26], ten column-checksum strips (row 20 – 29) are generated from the previously mentioned twenty data strips (row 0 – 19) to identify and rectify column errors. Further, a row-checksum DNA-MT for each DNA-MT block (highlighted in orange) is calculated using the XOR algorithm[27], enabling row-wise error detection (Fig. 2, "Encoding" panel).

Translation to DNA-MT blocks: Employing a DNA-MT code table, we transform each byte from data strips (including payload, address and checksum) into its corresponding DNA-MT, further forming DNA-MT blocks. In the example in Fig. 2, each block comprises four AMTs (labeled as A_1, A_2, A_3, A_4), five PMTs (labeled as $P_1, P_2, \ldots, P_5$), and one CMT. Notably, each AMT (8 bits) represents one of 256 different values, therefore four AMTs in one block allow for the assignment of a unique address to each of the 256^4 possible addresses. This configuration yields a storage capacity of up to 20 gigabytes (256^4 strips × 5 bytes/strip). Certainly, the system's storage capacity can be expanded with more AMTs and PMTs per block. Detailed encoding procedures are provided in the Methods section.

(2) Assembling DNA-MT blocks through enzymatic ligation (Assembling): To construct a DNA-MT block, we enzymatically connect the ten DNA-MTs into a

cohesive long oligonucleotide by joining the 11 linkers of the DNA-MTs together using T4 ligase. Taking the 8th strip from the fourth stripe as an example, we linked four AMTs (A_1-0, A_2-0, A_3-0, A_4-98), five PMTs (P_1-119, P_2-104, P_3-105, P_4-116, P_5-101), and one CMT (C-5), yielding a cohesive 242 bp oligonucleotide sequence "…atagTCACCAATCACACCTCCTCCaagt…aggaCGAATGGGTGGAGTGTAGG Gtgtg…" (Detailed across Fig. 1 and 2's "Encoding panel" and fully in Methods section).

To automate assembling process, we developed "BISHENG-1," a specialized DNA-MT inkjet printer equipped with 556 cartridges. BISHENG-1 operates similarly to a traditional inkjet printer but employs cartridges filled with DNA-MTs rather than ink, injecting these DNA-MTs into EP tubes instead of injecting ink onto paper. Each EP tube is designated to collect all DNA-MTs for a specific block, e.g., the 8th tube for the 8th DNA-MT block's ten DNA-MTs (Fig. 2, "Printing" panel). These movable types will be injected into the tube when it passes by the cartridges holding these movable types. The process, including tube movement and the DNA-MT selection and injection, is fully automated. Two basic solutions (pre-digested vector and T4 DNA ligase) are then automatically injected into these tubes for enzymatic ligation, ultimately cloning the DNA-MTs into an DNA-MT-plasmid. A demonstrative video (Supplementary Video) offers an in-depth view of BISHENG-1's operation.

(3) *In vitro* and *in vivo* storage of DNA-MT blocks (Storage): The recombinant plasmids from the above process are storable in liquid or lyophilized form for *in vitro* storage. These plasmids can be further introduced into *E. coli* for *in vivo* storage (Fig.

2, “Storage” panel). *E. coli*’s rapid growth attribute ensures quick, reliable duplication of DNA-MT blocks in the recombinant plasmids through generations, providing a cost-effective and stable method for data replication and storage.

(4) Decoding DNA-MT stored file with error correction (Decoding): To retrieve original files encoded within the plasmids, we use high-throughput sequencing to read the DNA-MT blocks. The sequences are then decoded corresponding movable types based on the DNA-MT code table. For example, a 250 nt DNA read “…atagTCACCAATCACACCTCCTCCaagt…aggaCGAATGGGTG…” is decoded into ten DNA-MTs. This includes four AMTs (A_1-0, A_2-0, A_3-0, A_4-98), five PMTs (P_1-119, P_2-104, P_3-105, P_4-116, P_5-101) and one CMT (C-5), reconstructing the original data strip “white” at starting address 340 (Fig. 2, “Decoding” panel). The decoded blocks are then organized by address to recover the original file. Errors are detected and corrected using column-checksum strips and row-checksum DNA-MTs (details in Methods section).

DNA-MT design and library construction

We began by screening candidates for linkers and information carriers from pools of 20-nt and 4-nt random sequences. To minimize errors in synthesis and sequencing and enhance ligation efficiency, we applied stringent criteria, including GC content, homopolymer runs, secondary structures, and sequence editing distance[28], ensuring the functionality and distinctiveness of carriers and linkers (Fig. 3a and Methods section). Previous studies have reported that high GC content, extensive homopolymer runs, and complex secondary structures could elevate synthesis and

sequencing errors[29] , and small editing distances among similar sequences increase missequencing risks.

Employing these criteria, we first selected optimal linkers with high specificity to ensure precise pairings with their complementary partners. Non-specific linkers often lead to incorrect assemblies among DNA-MTs, like the erroneous linkage of A_1 with A_3, instead of the correct A_2, causing disorganized DNA-MT blocks. Here, we opted for 4-nt linkers based on their optimal balance between affinity and specificity. From a comprehensive screening of 256 possible 4-nt combinations (4^4), we identified 200 candidate 4-base linkers (details in Supplementary Table 2) that met our stringent requirements for high specificity and accurate assembly.

To determine the appropriate number of linkers for assembling DNA-MT blocks, we conducted preliminary experiments using blocks composed of 4 to 12 movable types (details in Extended Data Fig. 1). We observed that as the number of DNA-MTs increased, the assembly accuracy of the blocks decreased. Taking into consideration both the assembly accuracy and the storage density, we ultimately selected blocks containing 10 DNA-MTs for our subsequent storage scheme. In this scheme, each DNA-MT block consists of 10 DNA-MTs connected by 11 linkers. Specifically, *Hin*dIII and *Eco*RI restriction sites were utilized as fixed linkers at both ends of each block (*agct*, *aatt*). The remaining nine linkers were selected from a pool of 200 previously identified candidate 4-base linkers. Here we employed a scoring system along with experimental validation to confirm the effectiveness of these linkers, ensuring coherent assembly of DNA-MTs within each block (Extended Data Fig. 2; scoring principles in Methods).

After careful selection of the 11 4-nt linkers, we integrated these predetermined linkers with selected 20-nt information carrier candidates to complete the design of

our DNA-MTs, based on the established criteria (detailed in the Methods section). Each DNA-MT in our scheme includes two linkers at the 3' and 5' ends, and a central information carrier encoding 1 byte of data, designated to encode data for payload, address, and checksum information. Through this comprehensive integration and screening process, we have compiled a library of 2,560 DNA-MTs (Supplementary Table 1), each uniquely characterized by its information carriers and linkers.

Following the design of individual DNA-MTs, we proceeded to the design and assembly of DNA-MT blocks, each comprising 10 DNA-MTs, strategically configured to enhance data storage and retrieval efficiency. This configuration includes 5 PMTs encoding 5 bytes of data, 4 AMTs encoding 4 bytes of address information to support a total DNA-MT storage capacity of ~2 GB, and 1 CMT encoding 1 byte of checksum for error correction. The blocks are constructed by linking the 5' and 3' ends of these DNA-MTs (Fig. 3). To ensure sequential ligation of DNA-MTs in a specified order $A_1, A_2, A_3, A_4, P_1, P_2, P_3, P_4, P_5, C$, two strategic constraints were implemented (Fig. 3b): i) All A_1 DNA-MTs share identical linkers, and so do DNA-MTs in A_2 to C; ii) The 3'-linker of an A_1 DNA-MTs are designed to specifically pair with the 5'-linker of an A_2 DNA-MTs, and so on.

Employing pre-fabricated DNA-MTs significantly lowers the need for *de novo* oligonucleotide synthesis from scratch. For instance, the word "white" occurs twice in *Sonnet 12* at addresses 185 and 340 (Fig. 3c). Traditional "base-by-base" synthesis method would require repeating oligonucleotide synthesis for such case. However, with DNA-MT blocks, three AMTs and five PMTs do not need to be synthesized from the scratch.

DNA-MT inkjet printer for automating block assembly: BISHENG-1

To automate DNA-MT storage process, we developed BISHENG-1, a DNA-MT inkjet printer capable of constructing DNA-MT blocks with high-throughput. BISHENG-1 consists of five main components (Fig. 4): multiple cartridges for holding DNA-MTs, pre-digested vectors (*Hin*dIII and *Eco*RI) and T4 ligase, a conveyor belt for tube transport, an automated tube loading and discharging device, a printing monitor and a programable control system. Each cartridge carries a specific DNA-MT/digested vector/ligase. Each tube on the conveyor is designed to collect ten DNA-MTs of a block plus the digested vector and ligase for subsequent enzymatic ligation into a long oligonucleotide.

BISHENG-1 is equipped with 556 cartridges. To maximize storage density and address space, we devised an optimized cartridge allocation strategy (details in Methods section), enabling BISHENG-1 to cover an address space of 16×32^3 blocks and offer a storage capacity of ~2 megabytes with a theoretical logic density of 0.4 byte/MT ($16 \times 32^3 \times 4/1024^2$).

BISHENG-1 operates like standard inkjet printers, executing a cycle of "*loading empty tubes, printing DNA-MTs into tubes, and discharging the filled tubes*" (Fig. 4a; see Supplementary Video). Initially, empty tubes are placed onto the conveyor, moving past various cartridges. As tubes pass the cartridges for designated DNA-MTs, pre-digested vector and ligase, BISHSENG-1 commands the corresponding cartridge to inject a drop (~6 μl) of each into the tubes. Once the tube has passed all cartridges, it has collected all necessary components: assigned DNA-MTs, digested vector, and ligase. Upon collection completion, the filled tubes are discharged, making room for a

new cycle (loading, printing, and discharging), thereby streamlining the DNA storage process.

However, issues like cartridge nozzle blockages or incorrect droplet sizes may occur. To address these errors, BISHENG-1 is equipped with a video camera and image analysis software for monitoring malfunctioning nozzles (Fig. 4b). Detected errors trigger automatic reprinting of affected tubes, ensuring a high success rate.

BISHENG-1 successfully printed and stored 43.7 KB data (16,830 DNA-MT blocks) in four file (text, JPEG, MP3, and MP4) within ~5 hours (4 bytes/second), utilizing just 350 DNA-MTs. Except *in vitro* storage, these recombinant plasmids with DNA-MT blocks were also transformed into *E. coli* for *in vivo* storage, demonstrating data preservation in living organisms.

Cost analysis of DNA-MT storage

The cost of storing data using DNA-MTs is quantified by the formula:

$$C = \frac{S}{V \times \rho} + L.$$

Here, C (\$/byte) represents the cost per byte of data storage; S (\$) denotes the synthesis cost of all DNA-MTs; V (μL) represents the total volume of DNA-MTs solution; ρ (byte/μL) represents the information density (how many bytes of data can be encoded per μL of solution); L (\$/byte) represents the cost of T4 DNA ligase per byte. Based on BISHENG-1's data, we define these parameters as follows:

Synthesis cost S and solution volume V: S is 122.20 \$ for synthesizing 470 OD of 5'-phosphorylated DNA-MTs at a commercial rate (~260 \$ for 1000 *OD*,

General Biol (Anhui) Co. Ltd). The solution's total volume is 2.4 ×10^7 μL, setting $V = 2.4 \times 10^7$ μL.

Information density $\boldsymbol{\rho}$: Assembling a DNA-MT block requires 60 μ l solution (6 μl/block), and 30 blocks encode 80 bytes, yielding $\rho = 80/(60 \times 30) = 0.044$ bytes/μl.

Ligation cost L: Utilizing in-house produced T4 DNA ligase at approximately 0.43 /g, the cost to ligate 1 MB is about 0.23 $ based on enzyme activity estimation[30]. Thus, $L = 2.19 \times 10^{-7}$ $/byte.

Applying these parameter, the DNA-MT storage cost is estimated at $C = \frac{122.20}{2.4\times10^7\times0.044} + 2.19 \times 10^{-7} = 1.16 \times 10^{-4}$ $/byte (121.57 $/MB), significantly lower than existing DNA storage technologies–530 $/MB with *Antkowiak* et al.'s report and 2,260 $/MB with *Erlich* and *Zielinski*'s report[14,16] (Fig. 5a). The cost benefits of DNA-MTs become more pronounced for larger datasets, with costs diminishing to 35.28 $/MB for 1 GB and 30.56 $/MB for 1 TB (detailed in Extended Data Table 2), thereby demonstrating the cost efficiency of DNA-MT storage through reusability and enzymatic ligation.

Discussion

Our study introduces an innovative concept in DNA storage, drawing inspiration from the ancient movable type printing to develop "DNA-MT storage". Unlike traditional DNA storage methods that resemble block printing—simple yet costly—our system significantly reduces writing costs and time. Specifically, we have developed a comprehensive system (details in Extended Data Fig. 3) that includes DNA-MT libraries (payload, address, and checksum), a coding-decoding system (comprising DNA-MT codebooks and DNA-MT block assemblies), and an automatic DNA-MT inkjet printer, BISHENG-1. This system enables efficient storage and 100% accurate recovery of 43.7 KB diverse data, using only 350 DNA-MTs. Our approach offers cost and efficiency advantages described as follows:

Cost advantage: The cost efficiency of our DNA-MT storage system is its most significant benefit over traditional DNA storage technologies. One-time use approach precludes reusing the synthesized DNA fragments, leading to high synthesis expenses. Conversely, our system permits the repeated access and reuse of pre-made DNA-MTs, significantly reducing DNA synthesis costs. The BISHENG-1 printer showcases this cost-effective capability, where DNA-MTs, synthesized once (2 *OD*), can be used up to 10,000 times. This innovation lowers DNA storage costs to as low as 121.57 $/MB, making it the most cost-effective DNA storage solution reported thus far[16].

Time advantage: Another major advantage of DNA-MT storage is its significant time efficiency. Unlike traditional methods that rely on time-consuming chemical synthesis from scratch, our approach uses biological enzyme linking, a much quicker process that enables high-throughput data writing by parallel construction of multiple DNA-MT blocks within ~10 minutes. For example, the BISHENG-1 printer completed

the writing of 43.7 KB data in ~5 hours (4 bytes/second). In contrast, traditional DNA storage methods using chemical synthesis, such as the Dr. Oligo 192 high-throughput synthesizer (~6 minutes per base addition, 192 channels), are estimated to take about 100 hours to synthesize the same amount of data from scratch, given the DNA storage's capacity limit of 1.83 bits/nt[14]. This not only highlights significant time and labor savings but also enhances the practicality of DNA-MT storage.

Targeted retrieval advantage: DNA-MT storage excels in targeted data retrieval. Each file, during the encoding and storage process, can be marked with a unique AMT for future specific retrieval. This enables the specific amplification and sequencing of all DNA-MT blocks associated with the file (employing the AMT sequence as a forward primer and a universal reverse primer from the 3' end of the *pUC19* plasmid). This method ensures specific access to and reading of designated files, underscoring the system's data management efficiency.

Advantage in data duplication: Our DNA-MT storage system enables rapid, accurate, and cost-effective data duplication through the exponential amplification of stored DNA-MT files on plasmids during bacterial culturing.

Potential and Prospects: Since its inception in 2017, our DNA-MT storage technology has undergone several iterations. Although its current cost and efficiency are not yet suitable for commercial application, it holds considerable potential for further enhancements in both cost and time efficiency, as detailed below:

Further potential in cost reduction: Our approach has significant untapped potential in cost reduction, including:

i. Reducing droplet volume: A key improvement can be made by reducing the droplet volume used in printing each DNA-MT. Currently, the BISHENG-1 operates at a microliter scale, but transitioning to a smaller, nanoliter scale could significantly

enhance cost efficiency. As demonstrated in Fig. 5b, reducing the droplet volume to 6 nl can decrease the cost of DNA-MT storage to 0.12 $/MB. We have employed the ECHO 650 instrument (Labcyte) to decrease droplet volume to 50 nl and successfully achieved the ligation of ten DNA-MTs (Extended Data Fig. 4a), underscoring the potential for further reductions in droplet volume.

ii. Increasing encoding density: In the current scheme, each DNA-MT block comprises 10 DNA-MTs. Doubling the DNA-MTs to 20 could elevate the capacity to 15 bytes/block, thus enhancing the encoding density from 0.40 byte/MT to 0.75 byte/MT. We have already achieved 20 DNA-MT ligation successfully, as shown in Extended Data Fig. 4b. Simultaneously, if we can also reduce the droplet volume to 6 nl, it would lower the costs to 0.06 $/MB. Furthermore, encoding advancements, like integrating high-density fountain codes and refining checksum DNA-MTs, could further improve density and cost-effectiveness, marking a promising direction for the evolution of DNA-MT storage.

iii. DNA-MT biosynthesis factory: The primary cost in our current DNA-MT storage originates from DNA-MTs' chemical pre-synthesis. Future plans involve establishing a biosynthesis factory for DNA-MTs, requiring only a single initial chemical synthesis before being inserted into plasmids for mass, cost-effective replication in bacteria. This approach would considerably lower the synthesis and production costs of pre-fabricated DNA-MTs. Furthermore, scaling and automating the biosynthesis factory could further reduce expenses, demonstrating our technology's substantial commercial potential.

Additionally, the current cost of pre-made DNA-MT is based on commercial DNA synthesis quotations. By conducting the synthesis in-house, we could achieve substantial cost reductions. Furthermore, scaling up will spread out the expenses

associated with pre-fabricated DNA-MTs and enzymes, thereby significantly lowering overall costs and amplifying savings. This strategy will render our method more economically viable.

Efficiency potential: Our current first-gen DNA-MT injector printer, BISHENG-1, with 556 parallel channels and a 4 bytes/second speed, has significant untapped potential. We are developing the next-gen desktop DNA-MT printer, BISHENG-2, designed to substantially increase throughput and speed, aiming to enhance DNA storage efficiency and move towards industrialization and commercialization.

In summary, storage cost and efficiency are pivotal for DNA storage's practical application. DNA-MT storage holds immense untapped potential in both aspects, awaiting further exploration and utilization. Unlocking this potential could create a more robust DNA storage ecosystem, better serving the massive data storage needs of the big data era.

References

1 Wright, A. Worldwide IDC global datasphere forecast, 2024-2028: AI everywhere, but upsurge in data will take time. (International Data Corporation, 2024). https://www.idc.com/getdoc.jsp?containerId=US52076424.

2 Bornholt, J. *et al*. A DNA-Based Archival Storage System. *SIGARCH Comput. Archit*. News 44, 2 (May 2016), 637–649.

3 Extance, A. How DNA could store all the world's data. *Nature* **537**, 22-24 (2016).

4 Matange, K., Tuck, J. M. & Keung, A. J. DNA stability: a central design consideration for DNA data storage systems. *Nat. Commun.* **12**, 1358 (2021).

5 Zhirnov, V., Zadegan, R. M., Sandhu, G. S., Church, G. M. & Hughes, W. L. Nucleic acid memory. *Nat. Mater.* **15**, 366-370 (2016).

6 Ceze, L., Nivala, J. & Strauss, K. Molecular digital data storage using DNA. *Nat. Rev. Genet.* **20**, 456-466 (2019).

7 Rutten, M. G. T. A., Vaandrager, F. W., Elemans, J. A. A. W. & Nolte, R. J. M. Encoding information into polymers. *Nat. Rev. Chem.* **2**, 365-381 (2018).

8 Piovesan, A. *et al.* On the length, weight and GC content of the human genome. *BMC Res. Notes.* **12**, 106 (2019).

9 van der Valk, T. *et al.* Million-year-old DNA sheds light on the genomic history of mammoths. *Nature* **591**, 265-269 (2021).

10 LeProust, E. M. *et al.* Synthesis of high-quality libraries of long (150 mer) oligonucleotides by a novel depurination controlled process. *Nucleic Acids Res.* **38**, 2522-2540 (2010).

11 Tucker *et al.* Massively Parallel Sequencing: The Next Big Thing in Genetic Medicine. *Am. J. Hum. Genet*. **85**, 142-54 (2009).

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

13 Goldman, N. *et al.* Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. *Nature* **494**, 77-80 (2013).

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

15 DNA Storage Alliance. Preserving our digital legacy: an introduction to DNA

data storage. https://dnastoragealliance.org/dev/wp-content/uploads/2021/06/DNA-Data-Storage-Alliance-An-Introduction-to-DNA-Data-Storage.pdf.

16 Antkowiak, P. L. *et al*. Low cost DNA data storage using photolithographic synthesis and advanced information reconstruction and error correction. *Nat. Commun.* **11**, 5345 (2020).

17 Eisenstein, M. Enzymatic DNA synthesis enters new phase. *Nat. Biotechnol.* **38**, 1113-1115 (2020).

18 Palluk, S. *et al.* De novo DNA synthesis using polymerase-nucleotide conjugates. *Nat. Biotechnol.* **36**, 645-650 (2018).

19 Lee, H. *et al.* Photon-directed multiplexed enzymatic DNA synthesis for molecular digital data storage. *Nat. Commun.* **11**, 5246 (2020).

20 Lee, H. H., Kalhor, R., Goela, N., Bolot, J. & Church, G. M. Terminator-free template-independent enzymatic DNA synthesis for digital information storage. *Nat. Commun.* **10**, 2383 (2019).

21 Lee, J. M. *et al*. Semiautomated synthesis of sequence-defined polymers for information storage. *Sci. Adv.* **8**, eabl8614 (2022).

22 Nguyen, B. H. *et al.* Scaling DNA data storage with nanoscale electrode wells. *Sci. Adv.* **7**, eabi6714 (2021).

23 Ping, Z. *et al.* Towards practical and robust DNA-based data archiving using the yin-yang codec system. *Nat. Comput. Sci.* **2**, 234-242 (2022).

24 Anavy, L., Vaknin, I., Atar, O., Amit, R. & Yakhini, Z. Data storage in DNA

with fewer synthesis cycles using composite DNA letters. *Nat. Biotechnol.* **37**, 1229-1236 (2019).

25 MacKay, D. J. *Information theory, inference and learning algorithms*. (Cambridge university press, 2003).

26 Plank, J. S. A tutorial on Reed-Solomon coding for fault-tolerance in RAID-like systems. *Software Pract. Exper.* **27**, 995-1012 (1997).

27 Patterson, D. A., Gibson, G. & Katz, R. H. A case for redundant arrays of inexpensive disks (RAID). *SIGMOD Rec.* **17**, 109–116 (1988).

28 Navarro, G. A guided tour to approximate string matching. *ACM Comput. Surv.* **33**, 31-88 (2001).

29 Shin, S. & Park, J. Characterization of sequence-specific errors in various next-generation sequencing systems. *Mol. BioSyst.* **12** (2016).

30 Units of Enzyme Activity. *Eur. J. Org. Chem.* **97**, 319-320 (1979).

Methods

1. Design and construction of DNA-MTs and related library

This section outlines the design of linkers and information carriers for DNA-MTs, as well as the construction of the DNA-MT library, detailed as follows:

1.1 Linker design for DNA-MTs

To ensure high specificity and prevent unintended ligation of DNA-MTs, we selected linkers from a pool of 256 randomly generated 4-base sequences (4^4). The selection was based on the following criteria: (1) Base diversity: Each linker must contain at least two different types of bases; (2) Homopolymer Limit: No more than three consecutive identical bases are allowed; (3) GC Content Range: 20-80%. Using these criteria, we initially screened 200 viable 4-base linker candidates.

Subsequently, we randomly selected 9 linkers from this pool, supplemented by two fixed enzymatic restriction sites (*Hin*dIII: agct; *Eco*RI: aatt), yielding 55 potential linker pairs (C_{11}^2). A scoring system was developed to evaluate the specificity of these pairs: (1) If a linker's sequence (forward, reverse, complementary, or reverse complementary) matches another linker's sequence, that set of 11 linkers is discarded; (2) If a linker shares more than or equal to three consecutive or cumulative paired bases with another linker in either the forward sequence/reverse complementary sequence, and the two linkers are situated at opposite ends of the same movable type, then that combination of 11 linkers is discarded; (3) If a linker shares less than or equal to three consecutive or cumulative paired bases with another linker in either the forward sequence or reverse complementary sequence, and the two linkers are not located at opposite ends of the same movable type, then 1 point is added to this pair.

This simulation was repeated 5,000 times to determine optimal combinations of

nine linkers. A series of nine-linker combinations achieved high specificity scores (above 50), indicating a low likelihood of incorrect bindings and thereby enhancing the reliability of DNA-MT block assembly. Ultimately, validation experiments confirmed an optimal combination of nine linkers with a specificity score of 52 (for detailed sequences of these linkers see Supplementary Table 2).

1.2 Design of information carriers for DNA-MTs

To design effective DNA-MTs, we selected for a 20-nt length for the information carriers, balancing sequencing and synthesis costs with the need for sequence diversity. Shorter sequences lack sufficient diversity for encoding adequate information, whereas longer sequences increase both synthesis and sequencing costs. These carriers were randomly generated and coupled with the previously optimized 11 specifically optimized linkers at both 5' and 3' ends, generating a pool of 24/25-nt DNA-MTs (those at the ends being 25-nt in length).

Several constraints were imposed on these sequences to ensure functionality and specificity: (1) Nucleotide Diversity: Each carrier must include at least three types of nucleotide bases; (2) Homopolymer Runs: Homopolymers of A/T must be less than 5, and G/C less than 4; (3) GC content: 40-60%; (4) Restriction Site Avoidance: Exclusion of the *Eco*RI and *Hin*dIII restriction site sequences; (5) Secondary Structures: No more than 4-bp hairpin structures; (6) Self-Dimers: Consecutive self-dimers < 3, and interval self-dimers < 8; (7) Editing Distance: A minimum editing distance of 6 was required between any two DNA-MTs to avoid cross-hybridization.

Following these design principles, we successfully designed 2,560 unique DNA-MTs with high specificity, including 1,024 AMTs (4×256), 1,280 PMTs (5×256), and 256 CMTs. The detailed sequences of these carriers are systematically cataloged in a DNA-MT table for ease access.

1.3 Construction of DNA-MT library

The single-stranded DNAs corresponding to the designed DNA-MTs were custom-synthesized and subjected to 5'-phosphorylation modifications (Shanghai Biotech Co., Ltd). These synthesized single-stranded DNA-MT oligonucleotides were then equimolarly annealed (50 μM) to form double-stranded DNA-MT solutions. Subsequently, these pre-made DNA-MTs were allocated into three sub-libraries according to the described proportions (AMT: 1,024; PMT: 1,280; and CMT: 256), ready for the assembly of DNA-MT blocks.

2. Encoding target files with DNA-MTs

This section outlines the procedures for converting digital files into DNA-MT storage files, emphasizing fragmentation, data stripe formation, binary conversion, error detection and correction, and final encoding into DNA-MTs in blocks. These steps ensure accurate data encoding and robust error handling to optimize the integrity of the DNA-MT stored files.

Fragmentation, stripe formation, and binary conversion: To encode various types of files (e.g., text, images, audio, and video) into DNA-MT storage files, each target file is initially segmented into 100-byte fragments. Each fragment is further subdivided into twenty 5-byte data strips. Subsequently, each strip is appended with 4 bytes of address information to facilitate mapping within the storage file, and then subjected to binary conversion. This address allocation is designed to support a total DNA-MT storage capacity of ~2 GB.

For example, as illustrated in Fig. 2 (Encoding panel) and 6a-b, the fourth stripe from

Shakespeare's Sonnet 12, “ed up...with white...do I question ma”, spanning from the 300th to 399th byte/character, is divided into twenty sequential 5-character data strips (from stripe 3-0 to stripe 3-19). The first strip, 3-0, encodes the payload “ed up”, starting at the 300th character, and the last strip, 3-19, encodes "on ma", beginning at the 395th character. Notably, the word “white” is located in the 3-8 data strip, beginning at the 340th character. Following binary conversion, the strip 3-8's address, 340, is represented as “0, 0, 0, 98,” and its payload “white” is translated into the ASCII codes “119, 104, 105, 116, 101”. The sequence of the DNA-MT block representing “white”: agcttATCTTACAACCTCCAACCCGctccATCCCTTGCCGCTTATTCCCactaAGGG TGAAGGTGCTGATGATggtaGAAGATGTGGCGTGAGGGTGatagTCACCAATCA CACCTCCTCCaagtTTTGATTCGCTGTGATGGTGctgaCGCAAGAAAGGAAAAC AACGcaacAACACAAAAGGCAACACGGCaggaCGAATGGGTGGAGTGTAGGG tgtgAGTCCGTATGCCTTTTGCTGg.

Error detection and correction (Fig. 2: Encoding panel, and Fig. 6b)**:** Employing the Reed-Solomon algorithm, ten column-checksum strips (rows 20 – 29) are calculated from the twenty addressed data strips (rows 0 – 19) in each fragment. Additionally, a byte of row-checksum data, calculated using the XOR algorithm, is then appended to each of the thirty strips (twenty data strips and ten column-checksum strips). These dual-layer checksum system allows for the accurate restoration of the original data even if up to ten out of the thirty strips are compromised.

Encoding binary data into DNA-MTs: Using the DNA-MT code table, we translate the binary data from each strip into corresponding DNA-MTs. Based on the

designated DNA-MT storage codes for digital data files, these DNA-MTs are organized into 30 blocks per fragment, with each fragment encoding 100 bytes of data. Excluding ten blocks designated for checksums, each of the remaining 20 data blocks consists of four AMTs, five PMTs and one row CMT, arranged in the specified order: $A_1, A_2, A_3, A_4, P_1, P_2, P_3, P_4, P_5, C$.

3. Assembly and storage of DNA-MT blocks

3.1 Enzymatic assembly and storage of DNA-MT blocks

Following the previous DNA-MT encoding, we selected designated pre-made DNA-MTs (~0.06 μM) from three DNA-MT libraries and distributed them into various EP tubes to construct a series of DNA-MT blocks. For each tube, along with the selected DNA-MTs, we added T4 ligase (875 U, 2011B, *TAKARA*) and 10X reaction buffer, as well as pre-digested *pUC19* modified vector (termed DNA-MT vector, 150 ng, digested with *Hin*dIII and *Eco*RI), and incubated the mixture at 25 degrees Celsius for 30 minutes, 16 degrees Celsius for 30 minutes to facilitate the sequential ligation of multiple DNA-MTs into blocks and cloned into the DNA-MT vector. Following this, the resulting DNA-MT blocks contained within the recombinant plasmids can be stored in either liquid or lyophilized form for *in vitro* storage. Furthermore, these recombinant plasmids can also be introduced into *E. coli* for *in vivo* storage. These blocks can be stored individually or combined into a single tube to form a DNA-MT stored file, representing a single digital file such as a text, image, audio, or video file.

To minimize sequencing costs, the *pUC19* plasmid was modified to include Illumina sequencing adaptors and primers (Extended Data Fig.5 and Table 3). Specifically, we added the P5 adaptor, index 1, and Read 1 sequencing primer adjacent to the 5’ end of

the *Hin*dIII restriction site. Similarly, at the 3' end of the *Eco*RI site, the P7 adaptor, index 2, and Read 2 sequencing primer were incorporated. The modified plasmid, now termed DNA-MT plasmid, enables the direct sequencing of the DNA-MT blocks on the Illumina platform, streamlining the sequencing process.

3.2 DNA-MT inkjet printing for automating block assembly

BISHENG-1, a DNA-MT inkjet printer, was developed to automate DNA-MT block assembly with high-throughput. BISHENG-1 comprises five main components: multiple cartridges, a conveyor belt, an automatic tube loading and discharging device, a programmable control system, and a printing monitor (Fig. 4).

Preparation of printing components: Initially, the BISHENG-1 cartridges were filled with specific DNA-MTs (PMTs, AMTs, and CMTs), pre-digested DNA-MT vectors, and T4 DNA ligase, along with its reaction buffer. This preparation ensured all required components for DNA-MT printing and block assembly to ready for automated processing. To maximize the efficiency and capacity of the system, 556 cartridges were strategically allocated to optimize address space and storage density, enabling BISHENG-1 to encode 16×32^3 blocks, equivalent to ~2 megabytes of data with a theoretical logic density of 0.4 byte/MT. The printed file types included TXT (0.4 KB), JPEG (10.9 KB), MP3 (14.2 KB), and MP4 (18.2 KB).

Cartridge allocation used by BISHENG-1: We noticed that the simple sequential arrangement of cartridges always leads to overuse of certain cartridges and thereafter lowering their lifetime. To avoid this issue, we shuffle the cartridges by remapping the data and column-checksum strips (Extended Data Fig. 6).

AMT cartridges: A total of 160 cartridges were distributed among the AMTs as follows: 64 cartridges for A_1, including 16 cartridges for address information and 48

cartridges borrowed by payload information (Extended Data Fig. 6), and 32 cartridges each for A_2, A_3, and A_4.

PMT cartridges: Each PMT from P_1 to P_5 was assigned 64 cartridges, totaling 320 cartridges, to ensure sufficient encoding capacity for the data blocks.

CMT cartridges: 64 cartridges were allocated for CMTs. This allocation is based on the requirement that the row-checksum calculated from these PMTs will have 64 possible values, ensuring robust error correction capabilities.

Loading, printing, and discharging cycle: BISHENG-1 operates similarly to a standard inkjet printer but is designed to handle DNA-MTs. The process involves:

Loading: Empty tubes are automatically loaded onto a conveyor belt, moving past various cartridges.

Printing: As tubes pass under the designated cartridges, BISHENG-1's control system injects the corresponding DNA-MTs, digested vectors, and ligases into the tubes.

Discharging: Once the tubes have collected all necessary components for a DNA-MT block, they are automatically discharged from the conveyor, making room for a new cycle. This systematic approach streamlines the assembly of DNA-MT blocks.

Printing monitoring: To address potential issues like cartridge nozzle blockages or incorrect droplet sizes, BISHENG-1 is equipped with a video camera and image analysis software. Detected errors trigger automatic reprinting of affected tubes to maintain a high success rate in DNA block assembly.

Post-printing processing and storage: After printing, the tubes containing the DNA-MT blocks undergo enzymatic ligation at 25 degrees Celsius for 30 minutes, 16 degrees Celsius for 30 minutes. The completed blocks are then pooled together to form

a DNA-MT stored file, corresponding to a digital file such as text, image, audio and video. This DNA-MT stored file can be directly stored *in vitro* or transformed into *E coli* for preservation within living organisms.

4. Sequencing, decoding and reconstructing target files from DNA-MT stored files

4.1 Sequencing of the DNA-MT stored files

We initially sequenced DNA-MT stored files (stored *in vivo* or *in vitro*) to reconstruct target digital files. For *in vivo* samples, recombinant plasmids containing the stored information are first isolated and purified using a plasmid extraction kit (DP107-02, TIANGEN). Subsequently, the DNA-MT blocks within these recombinant plasmids are PCR amplified (R050Q, TAKARA) using primers Seq-F/Seq-R (details in Extended Data Table 3). The PCR products are then purified through agarose gel electrophoresis, followed by sequencing on the Illumina NovaSeq6000 platform using PE250 paired-end sequencing technology (Novagene Inc).

4.2 Alignment and mapping

Using the Burrows-Wheeler Alignment Tool (BWA)[31], we aligned the DNA-MT sequences from the DNA-MT code table with paired reads, translating them into corresponding DNA-MT IDs that include four AMTs, five PMTs, and one CMT.

4.3 Decoding and assembly

After mapping, we decoded the blocks according to the DNA-MT code table, and then converted to original information using remapping rules (Extended Data Fig. 6). This process involved translating into address and payload information from each read. If paired reads yielded conflicting results, both were discarded to ensure data integrity. Blocks were then assembled into groups of 30, with each block's position determined

by its address information. The data were reconstructed by aligning these sequences in a structured order: starting from individual strips, progressing to fragments, and finally assembling all fragments to restore the complete digital files.

4.4 Error correction and data recovery

Upon translating all reads, we conducted error correction by performing XOR calculations on the four AMTs and five PMTs within each block. If the XOR result matched the CMT ID, the row check passed; otherwise, the sequence was discarded as erroneous.

The output data from the row verification step are grouped based on address information into predefined sets, each consisting of 30 DNA-MT blocks. The first 20 blocks contain and store original data information, while the remaining 10 blocks are designated for redundancy column verification. If all blocks within a group are intact, column verification is then performed to identify any errors. If this verification passes, the data are confirmed for writing to the recovery file. In cases of failure, we simulate the loss of up to 10 blocks at any positions within the 30-block fragment and recalculate until the correct data is identified for writing to the recovery file. If errors persisted (for example, if there are more than 10 erroneous blocks or other unforeseen issues), that data group was recorded as erroneous in the recovery file. Finally, all payload information is written to the recovery file according to its address information (Extended Data Fig. 7).

Methods references

31 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).

Data availability statement

The raw data from the PE250 paired-end sequencing conducted on the Illumina NovaSeq6000 platform, have been deposited in the Genome Sequence Archive (Genomics, Proteomics & Bioinformatics, 2021) within the National Genomics Data Center (Nucleic Acids Research, 2022), hosted by the China National Center for Bioinformation / Beijing Institute of Genomics, Chinese Academy of Sciences. The datasets are publicly accessible at NGDC GSA. Specific dataset identifiers are CRA016200, corresponding to the raw data of the 4-12 DNA-MT one-step enzymatic ligation assay, and CRA016195, corresponding to the raw data of in *vitro* and in *vivo* storage of flies (TXT, JPEG, MP3, and MP4).

Code availability statement

The code used to replicate the entire DNA movable storage process described in this study is available at ChenF Lab on GitHub. https://github.com/ChenFLab/DNAMTStorage.

Acknowledgements

This work was supported by the National Key Research and Development Program of China (2020YFA0907000), and the National Natural Science Foundation of China (T2125013, 31971389, 32370657, 32271297) for providing financial supports for this study and publication charges. The numerical calculations in this study were supported by ICT Computer-X center, and the InforSuperBahn OneAINexus platform.

Author contributions

Conceptualization: F.C., G.M.T, D.B.B, B.D. Methodology: C.Y.W., D.W., Z.W., D.X.Y, J.X., Y.Z.W., X.T.W., P.W., Writing—original draft: C.Y.W., D.W., Z.W., D.X.Y,

J.X., Y.Z.W., X.T.W. Writing—review and editing: G.N.M., X.R.Z.,H.L.L., C.T., P.F.H., J.W., R.J.G., G.Q.X., C.D.L., Y.J.J., P.H.W., L.Y.Y., Y.L.Z., Y.J.S., J.F.X., H.T.N., S.H.X., F.C., G.M.T, D.B.B, B.D. Funding acquisition: F.C., D.B.B. Supervision: F.C., G.M.T, D.B.B, B.D. All authors read and approved the final manuscript.

Competing financial interests

The authors declare that they are inventors on several patent applications related to the DNA movable type storage technology and the automated DNA-MT inkjet printer described in this manuscript. All correspondence regarding this article should be directed to F.C. at chenfei@big.ac.cn.

Additional information

Supplementary information is available for this paper.

Figure legends

Fig. 1

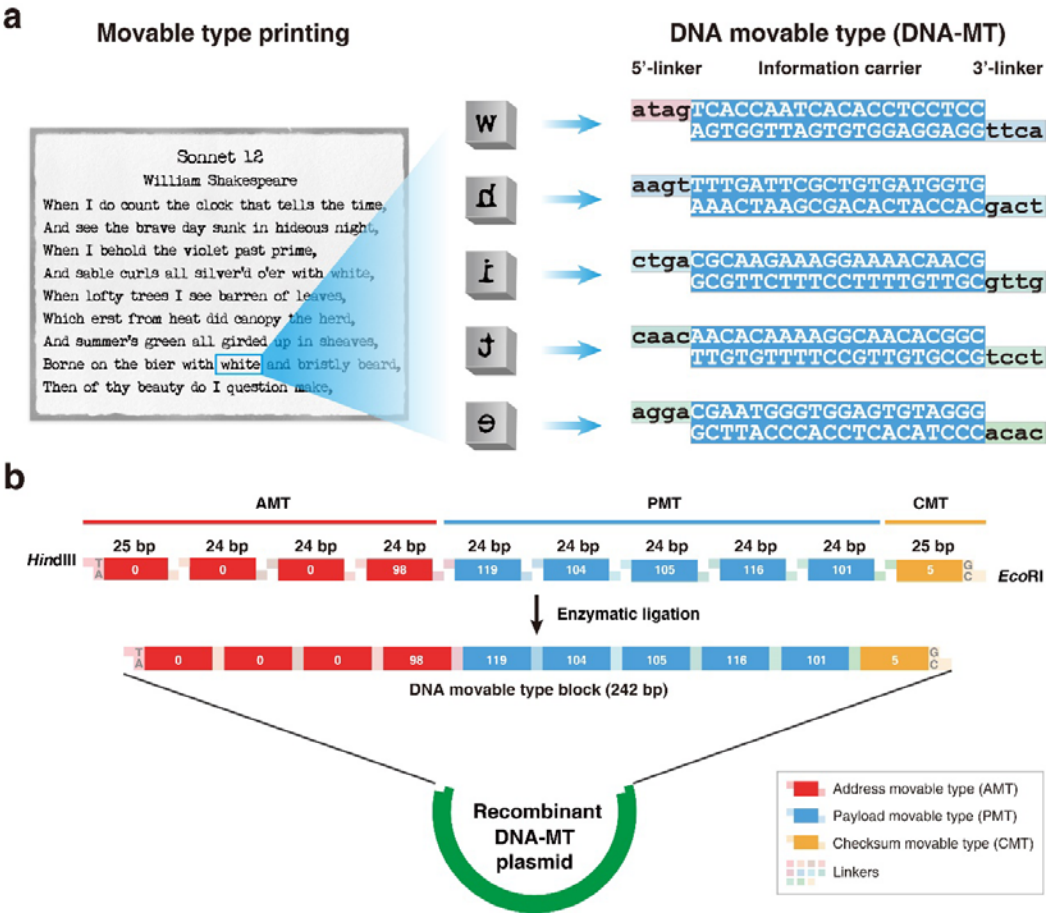


Fig. 1 | DNA-MT storage principle and structure. a, Mapping relationship between movable type printing and DNA-MT storage exemplified by the word “white” from *Shakespeare’s Sonnet 12*. The process of assembling pre-made wooden types in a frame to print text (left panel) is compared to the assembly of pre-synthesized DNA-MT, where each MT corresponds to a character (right panel). The word “white” from *Shakespeare’s Sonnet 12* is used to illustrate the method. Each character in the word "white" is represented by a DNA-MT consisting of a 20-bp double-stranded DNA information carrier with two 4-nt sticky ends that facilitate sequential assembly to form

in a cohesive DNA sequence that encodes “white”. **b**, Structure of DNA-MT block designed to encode the word "white". This block contains five Payload DNA-MTs (PMTs) encoding ASCII codes for “white” (119, 104, 105, 116, 101), four Address DNA-MTs (AMTs) specifying a four-byte address (0, 0, 0, 98) indicative of the 340th character position, and a checksum DNA-MT (CMT) for error detection (set at 5). These ten DNA-MTs are enzymatically sequentially linked using T4 DNA ligase through 11 linkers to form a cohesive DNA-MT block, with *Eco*RI and *Hin*dIII linkers at the ends for cloning into DNA-MT plasmid, enabling both *in vitro* and *in vivo* storage.

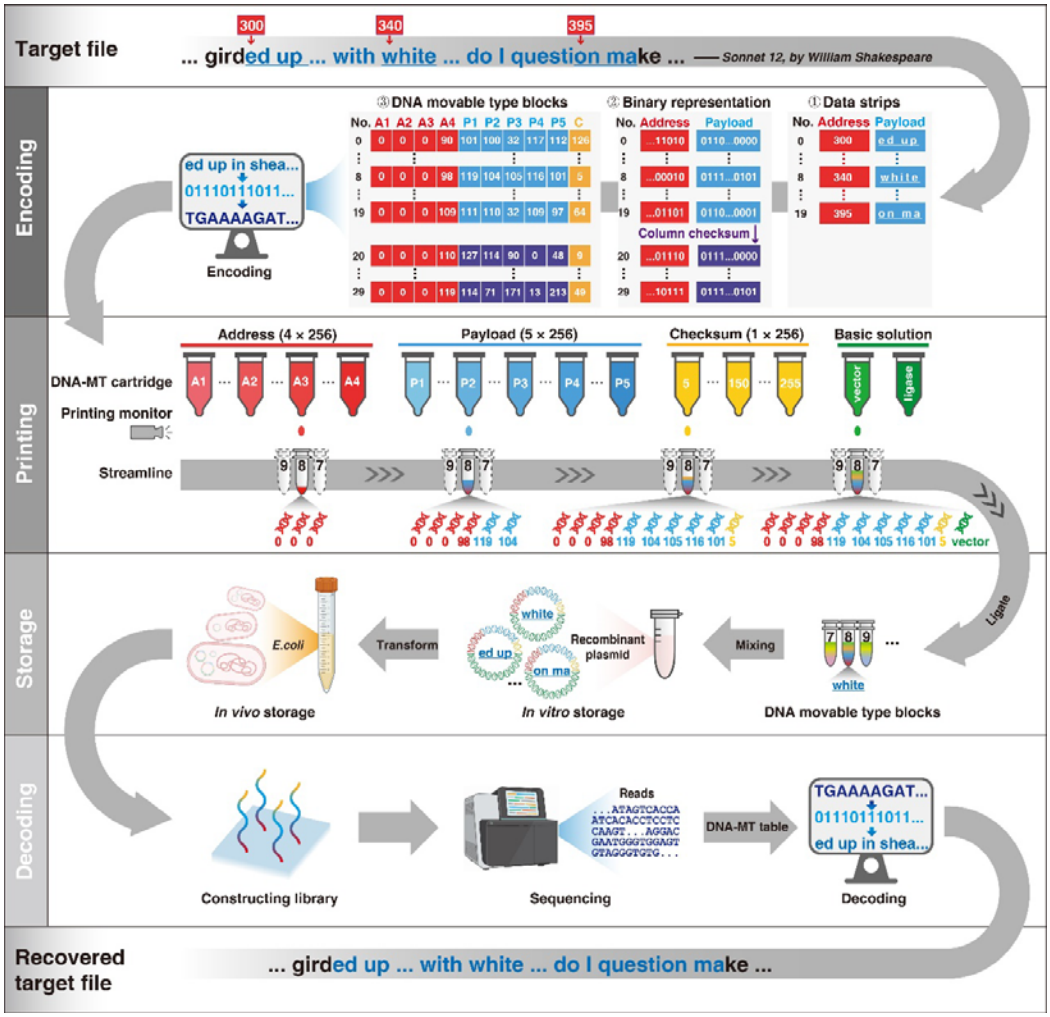


Fig. 2

Fig. 2 | Workflow of DNA-MT storage. The main steps include: **(1) Encoding**: The target file, specifically *Sonnet 12*, is divided into 100-byte/character fragments. Each fragment is subdivided into 20 data strips, and each strip comprises a five-byte payload and a four-byte address. The figure illustrates 20 sequential data strips (rows 0 – 19) of the fourth stripe, covering payloads from the 300th to the 399th character ("ed up...question ma"), with the 8th strip encoding "white" at the 340th position. Additional error detection and correction are provided through column (rows 20 – 29) and row checksums (column 9). **(2) Printing**: Utilizing BISHENG-1 DNA-MT inkjet

printer, ten specific DNA-MTs (4 AMTs, 5 PMTs, and 1 CMT), along with ligase and pre-digested vector are printed into each tube for enzymatic ligation into a DNA-MT block within plasmids. **(3) Storage**: The assembled recombinant plasmids are preserved either in liquid/lyophilized forms for *in vitro* storage or transformed into *E. coli* for long-term *in vivo* storage. **(4) Decoding**: Through next-generation sequencing, DNA-MT blocks are sequenced. The resulting DNA sequences are then decoded back into movable types according to the code table, thereby reconstructing the original data (e.g., decoding a 250 nt sequence enables the recovery of the word “white” at position 340). The decoded PMTs are arranged by their AMTs to restore the original file.

Fig. 3

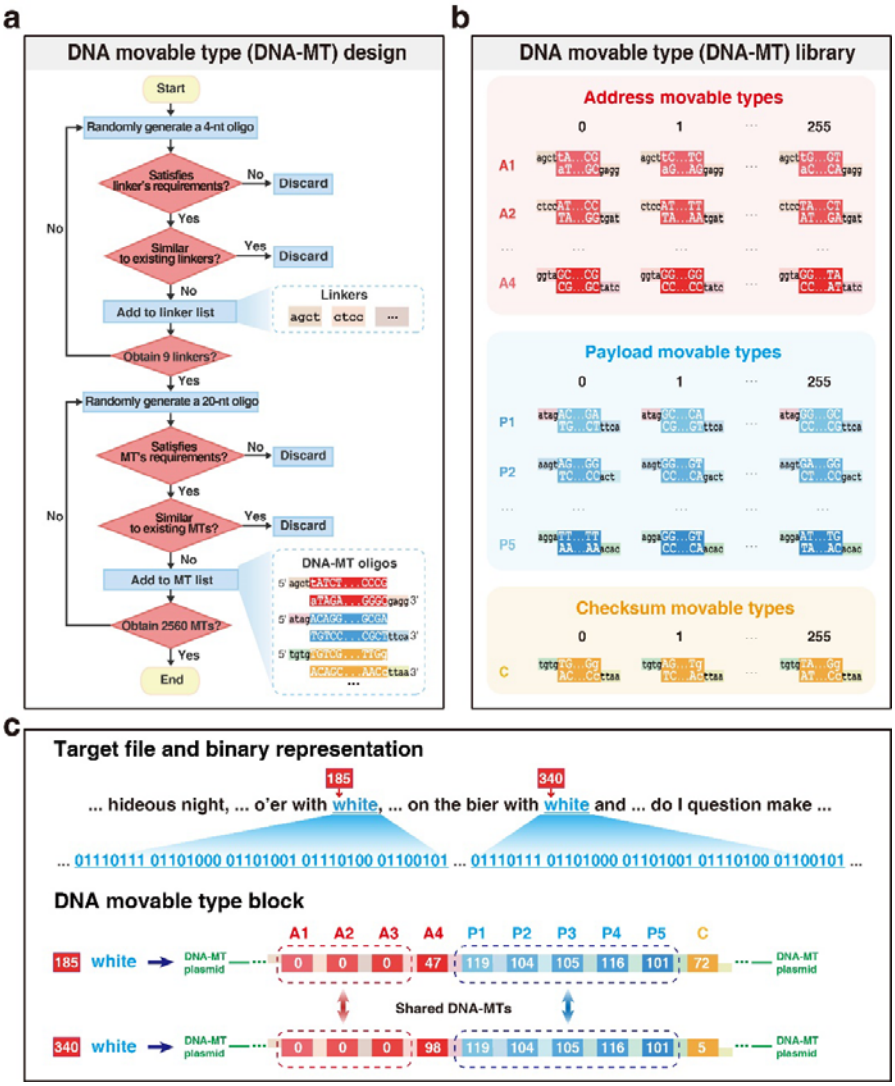


Fig. 3 | DNA-MT design and library construction. a, Sequence selection for DNA-MTs. 4-nt and 20-nt random oligonucleotides, as candidate linkers and information carriers for DNA-MTs, are subjected to a series of filtering criteria (such as GC content and homopolymer runs) to select optimal sequences, ensuring DNA-MTs' functionality and specificity; **b,** DNA-MT Library Composition. The DNA-MT library is divided into three sub-libraries: AMT sub-library (1024 AMTs: 4×256), PMT sub-library (1280 PMTs: 5×256), and CMT sub-library (256 CMTs); **c,**

Example of Pre-Made DNA-MTs for Reusability: Illustrates two DNA-MT blocks encoding the word “white” at positions 185 and 340. These blocks share three AMTs and five PMTs, exemplifying the reusability of pre-made DNA-MTs for reducing costs by minimizing the need for synthesizing new oligonucleotides for each occurrence.

Fig. 4

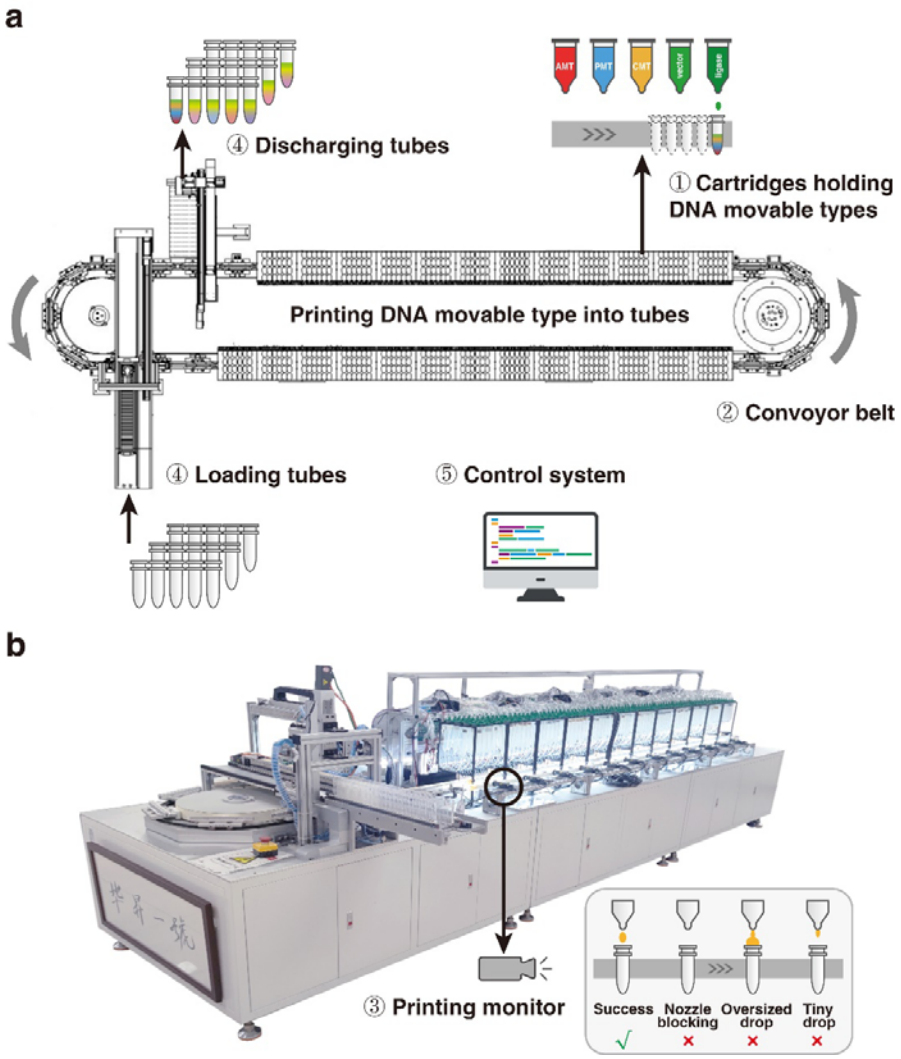


Fig. 4 | DNA movable type inkjet printer BISHENG-1 for automated DNA-MT block assembly. a, Architecture of BISHENG-1. The printer includes: i) multiple cartridges holding DNA movable types, pre-digested vectors, and T4 ligase, essential for DNA-MT assembly; ii) a conveyor belt for carrying and transporting tubes; iii) a printing monitor for real-time feedback and error detection; iv) devices for automatic tube loading and unloading; v) a programmable control system. Each tube on the conveyor is designed to collect all DNA-MTs of a block plus the digested vector and

ligase for subsequent one-step enzymatic ligation into a cohesive DNA-MT block. **b,** A photograph of BISHENG-1 with a focus on the printing monitor.

Fig. 5

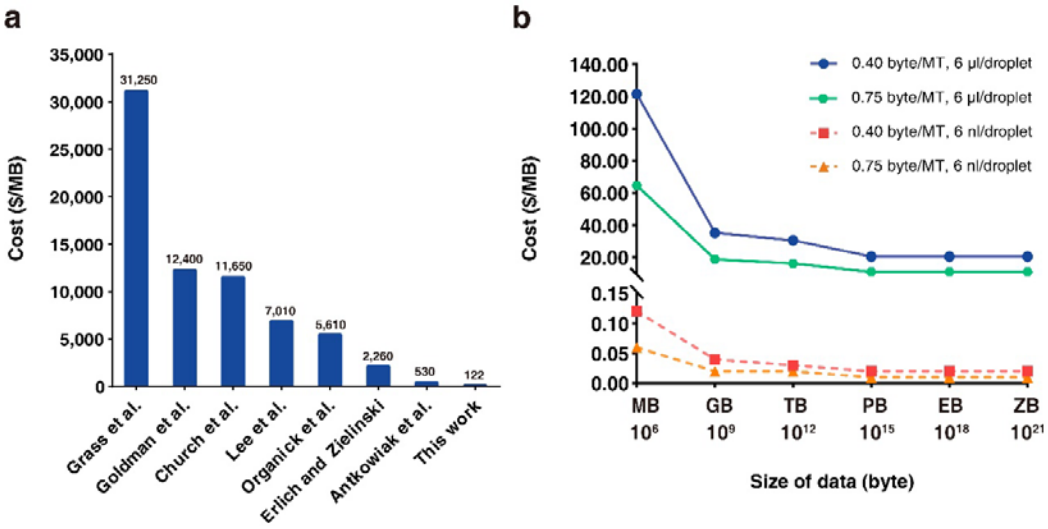


Fig. 5 | Cost-efficiency analysis of DNA-MT storage technology. a, Cost comparison between DNA-MT storage and other prevalent DNA storage technologies. Our method's cost is estimated at 121.57 $/MB, in comparison to higher costs reported by other studies, such as 530 $/MB by *Antkowiak* et al., and 2,260 $/MB by *Erlich and Zielinski*. **b,** Potential for cost reduction with improvements in data size, reaction volume, and encoding density. Using the current system with 10 DNA-MTs per block and a 6 µL droplet volume, the cost of storage is 121.57 $ per megabyte (MB). Scaling up to store 1 gigabyte (GB) of data reduces the cost to 35.28 $/MB, and increasing to 1 terabyte (TB) further lowers it to 30.56 $/MB. By adopting a system with 20 DNA-MTs per block and maintaining a 6 µL droplet volume, the cost per MB can be reduced to 64.75 $/MB. Reducing the droplet volume to 6 nanoliters (nL) with 10 DNA-MTs per block decreases the cost per MB to 0.12 $/MB. Similarly, using 20 DNA-MTs per block with a 6 nL droplet volume further reduces the cost per MB to 0.06 $/MB. Further details on these cost calculations can be found in Extended Data Table 2.

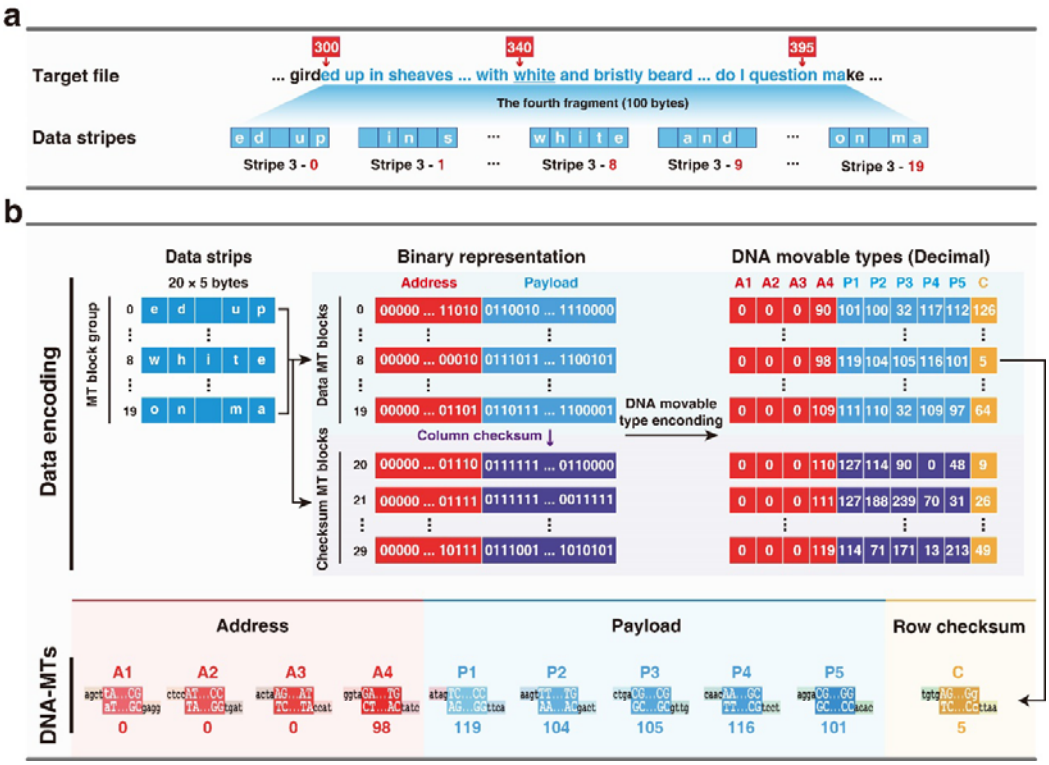


Fig. 6

Fig. 6 | Conversion of target files into DNA-MT blocks. a, File fragmentation and data strip formation. The target file is divided into 100-byte/character fragments. As illustrated, the fragment "ed up...with white...do I question ma" (spanning characters 300 to 399) is subdivided into twenty sequential 5-character data strips. The first strip begins with the characters "ed up" at position 300, while the last strip encodes "on ma" beginning at position 395. The word "white" is highlighted in the 8th data strip, beginning at character 340. **b,** Conversion to DNA-MT blocks. Each of the twenty data strips (rows 0-19) is appended with four-byte address information to form addressed data strips, which is further translated into binary representations. For error detection and correction, ten column-checksum strips (rows 20 – 29) and a row-checksum DNA-MT (column 9) are incorporated. Utilizing a code table, characters from the strips are mapped to their corresponding DNA-MTs, leading to

the assembly of DNA-MT blocks. Specifically, the DNA-MT block encoding the word “white” in the 8th strip comprises four AMTs (A_1-0, A_2-0, A_3-0, A_4-98), five PMTs (P_1-119, P_2-104, P_3-105, P_4-116, P_5-101) and one CMT (C-5).

Extended Data Table 1:

File format	Movable type blocks	Size (bytes)	Content Description
TXT	180	434	Chinese Tang Poetry: “SHUI DIAO GE TOU”
JPEG	4,200	11,128	Image of a panda
MP3	5,460	14,490	Song: "Departure"
MP4	6,990	18,605	Video: “Horse”

Extended Data Table 1: Details of four printed DNA movable type stored files using BISHENG-1. *This table provides* information *on four files stored using DNA-MTs, including their format, size, number of DNA-MT blocks required, and a brief description of the content.*

Extended Data Table 2:

Scheme	Data size	*S* ($)	*V* (ul)	*ρ* (byte/ul)	*L* ($/byte)	*C* ($/MB)
6 µl/droplet 0.4 byte/MT	MB	1.21E+02	2.36E+07	0.04	2.19E-07	121.57
	GB	3.51E+04	2.36E+10	0.04	2.19E-07	35.28
	TB	3.03E+07	2.36E+13	0.04	2.19E-07	30.56
	PB	2.02E+10	2.36E+16	0.04	2.19E-07	20.45
	EB	2.02E+13	2.36E+19	0.04	2.19E-07	20.45
	ZB	2.02E+16	2.36E+22	0.04	2.19E-07	20.45
6 µl/droplet 0.75 byte/MT	MB	6.48E+01	1.26E+07	0.08	3.81E-08	64.75
	GB	1.87E+04	1.26E+10	0.08	3.81E-08	18.73
	TB	1.62E+07	1.26E+13	0.08	3.81E-08	16.22
	PB	1.08E+10	1.26E+16	0.08	3.81E-08	10.83
	EB	1.08E+13	1.26E+19	0.08	3.81E-08	10.83
	ZB	1.08E+16	1.26E+22	0.08	3.81E-08	10.83
6 nl/droplet 0.4 byte/MT	MB	1.29E-01	2.50E+04	44.44	2.19E-10	0.12
	GB	3.71E+01	2.50E+07	44.44	2.19E-10	0.04
	TB	3.21E+04	2.50E+10	44.44	2.19E-10	0.03
	PB	2.14E+07	2.50E+13	44.44	2.19E-10	0.02
	EB	2.14E+10	2.50E+16	44.44	2.19E-10	0.02
	ZB	2.14E+13	2.50E+19	44.44	2.19E-10	0.02
6 nl/droplet 0.75 byte/MT	MB	7.71E-02	1.50E+04	83.33	3.81E-11	0.06
	GB	2.23E+01	1.50E+07	83.33	3.81E-11	0.02
	TB	1.93E+04	1.50E+10	83.33	3.81E-11	0.02
	PB	1.29E+07	1.50E+13	83.33	3.81E-11	0.01
	EB	1.29E+10	1.50E+16	83.33	3.81E-11	0.01
	ZB	1.29E+13	1.50E+19	83.33	3.81E-11	0.01

Extended Data Table 2*: *Cost calculation for DNA movable type storage. *This table provides detailed cost calculations related to fig.5b, showing cost reductions with increases in reaction volume, encoding density and data size improvements.*

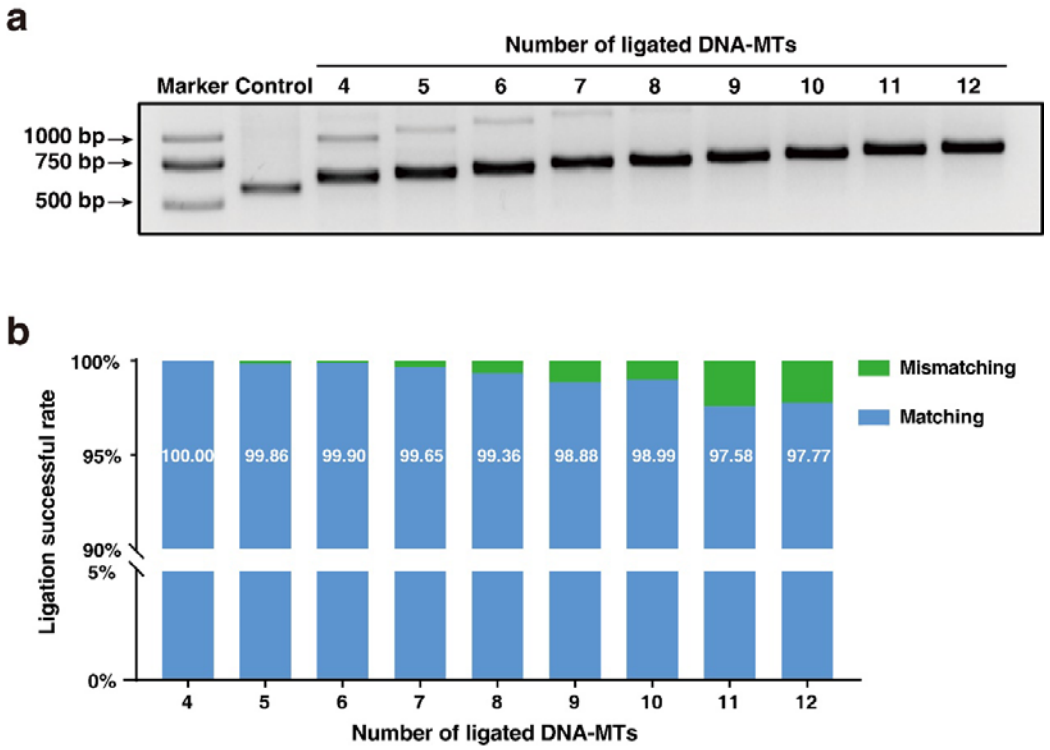
Extended Data Table 3:

Name	Sequence (5'-3')	Function
Test-F	CATTAGGCACCCCAGGCTT	Plasmid identification & Sanger sequencing
Test-R	GCACAGATGCGTAAGGAGAAAAT	
Seq-F	AATGATACGGCGACCACC	Next-generation sequencing library construction
Seq-R	CAAGCAGAAGACGGCATACG	
HF	AGCTAGCTT*AATGATACGGCGACCACCGAG ATCTACAC***AGTGGTCA**TCGTCGGCAGCGTC AGATGTGTATAAGAGACAGA	Insert fragment (P5 adaptor + index1 + read1 sequencing primer)
HR	AGCTTCTGTCTCTTATACACATCTGACGCT GCCGACGA**TGACCACT***GTGTAGATCTCGGT GGTCGCCGTATCATT*AAGCT	
EF	AATTC*CTGTCTCTTATACACATCTCCGAGC CCACGAGAC***TATCAGCA**ATCTCGTATGCCG TCTTCTGCTTGGAATT	Insert fragment (P7 adaptor + index2 + read2 sequencing primer)
ER	AATTAATTCCAAGCAGAAGACGGCATACGA GAT**TGCTGATA***GTCTCGTGGGCTCGGAGAT GTGTATAAGAGACAG*G	

Note: In the sequences, italic indicates the adaptor sequence, bold indicates the index sequence, and underlined text denotes the read sequencing primer sequence.

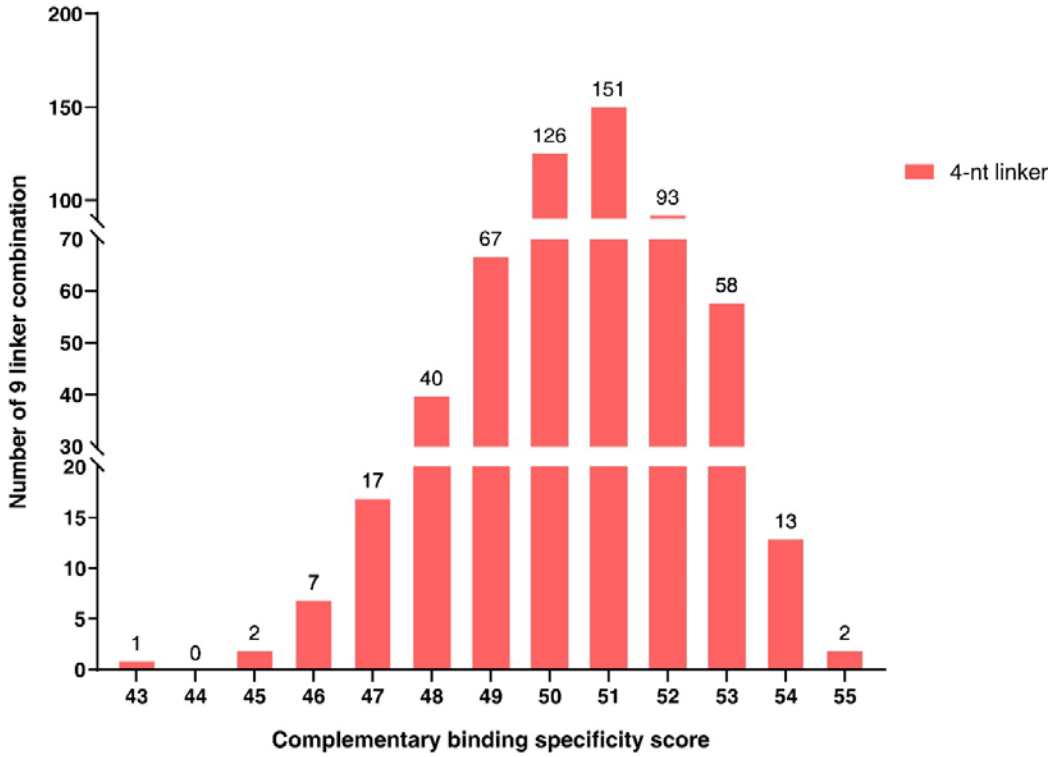
Extended Data Table 3: Primer sequences for constructing DNA movable type block storage plasmids (Extended Data Fig. 5). *The table lists the primer sequences used to construct DNA-MT plasmids. HF + HR primers insert the P5 adaptor, index 1, and Read 1 sequencing primer at the 5' end of the HindIII site. EF + ER primers add the P7 adaptor, index 2, and Read 2 sequencing primer at the 3' end of the EcoRI site. Test-F/Test-R primers are used to amplify DNA-MT blocks in recombinant plasmids for Sanger sequencing. Seq-F/Seqt-R primers are used for PCR amplification of DNA-MT blocks in recombinant plasmids, enabling the direct construction of next-generation sequencing libraries for DNA-MT blocks.*

Extended Data Fig. 1:



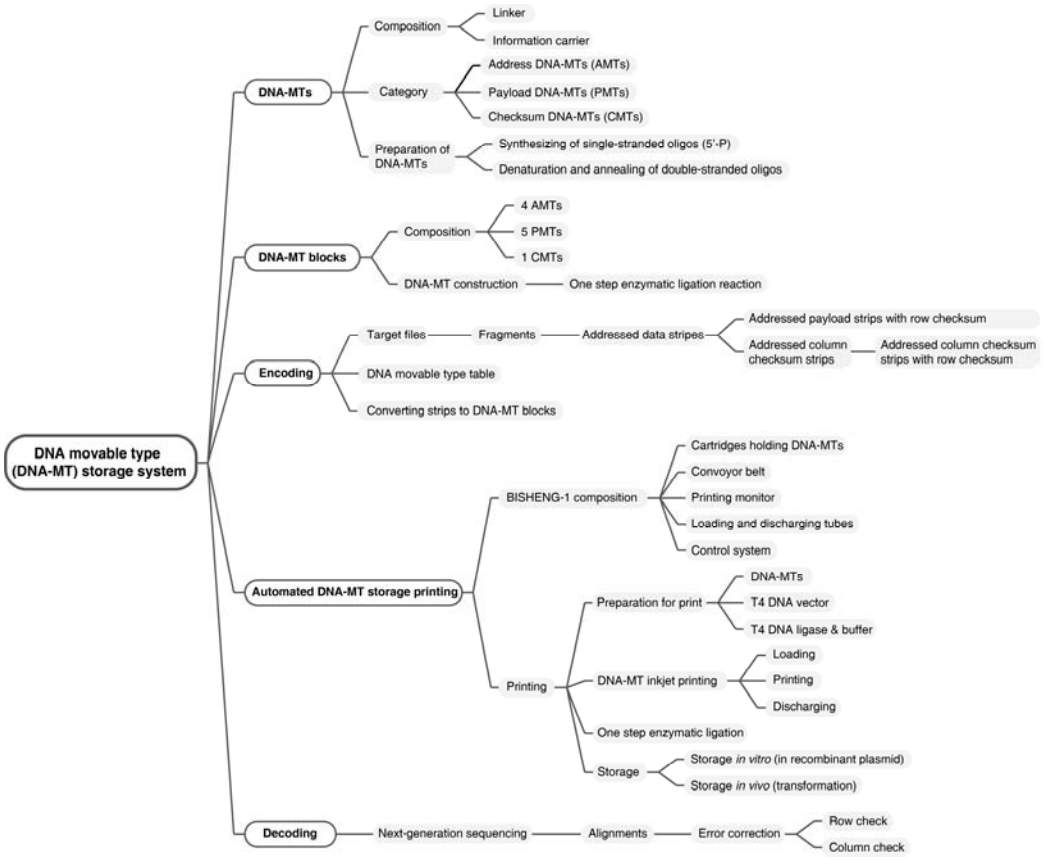
Extended Data Fig. 1: Characterization of manually assembly of DNA movable type blocks using one-step enzymatic ligation. a, Agarose gel electrophoresis of DNA-MT block products from enzymatic ligation of 4 to 12 DNA-MTs. *Nine one-step ligation reactions were performed to assemble 4 to 12 DNA-MTs. The inserted DNA-MT blocks in recombinant plasmids were PCR (Test-F/Test-R) amplified, and the PCR products were run on a 1.5% agarose gel. Empty plasmids without any inserts served as controls, showing a 514 bp band. Inserts with 4 to 12 DNA movable types displayed distinct amplification bands at 561 bp, 585 bp, 609 bp, 633 bp, 657 bp, 681 bp, 705 bp, 729 bp, and 753 bp, respectively. Successful ligation was then confirmed by Sanger sequencing (Applied Biosystems 3730xl DNA Analyzer), details in Supplementary Data 1.* ***b, Sequencing of ligation products for assembled DNA-MT blocks using next-generation sequencing (Illumina NovaSeq6000).*** *Recombinant plasmids containing 4 to 12 DNA-MT inserts in E. coli were extracted for sequencing. The sequencing results showed that almost all sequences exactly matched the desired DNA-MT blocks, with only a very small number of sequences missing at least one DNA-MT.*

Extended Data Fig. 2:



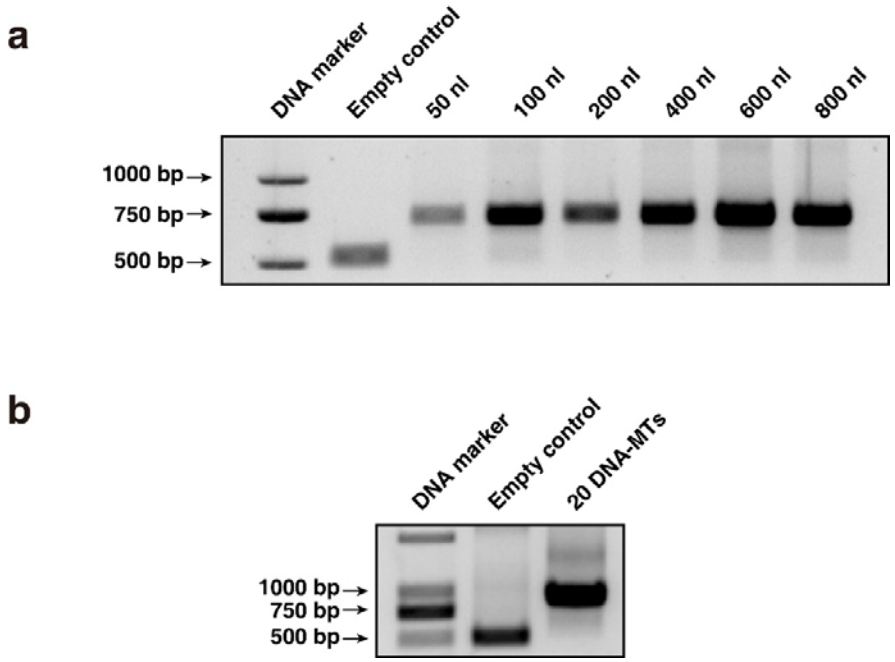
Extended Data Fig. 2: Complementary binding specificity scores of 9 linker combinations.

Extended Data Fig. 3:



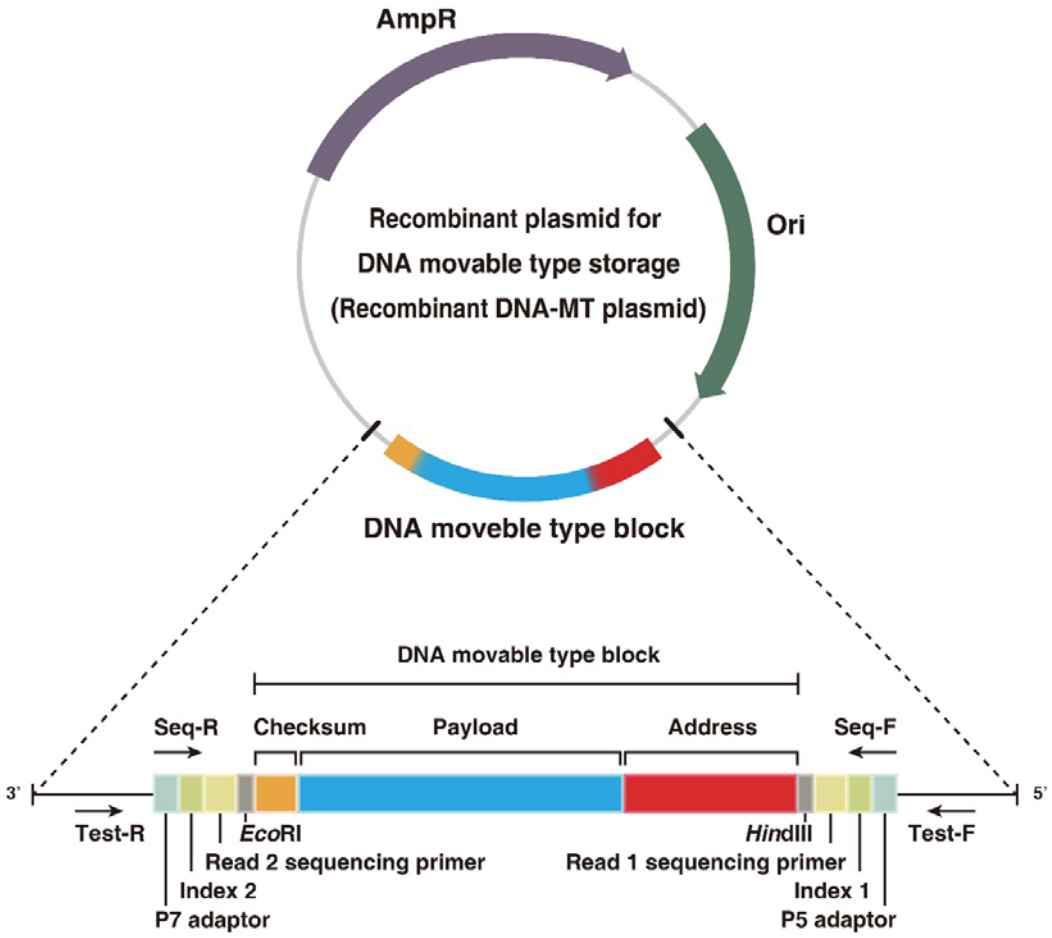
Extended Data Fig. 3: Brain atlases of DNA movable type storage. *This map provides a comprehensive overview of the system, including the design of DNA-MTs, DNA-MT libraries (payload, address, and checksum), a coding-decoding system (comprising DNA-MT table and DNA-MT block assemblies), and an automatic DNA-MT inkjet printer, BISHENG-1.*

Extended Data Fig. 4:



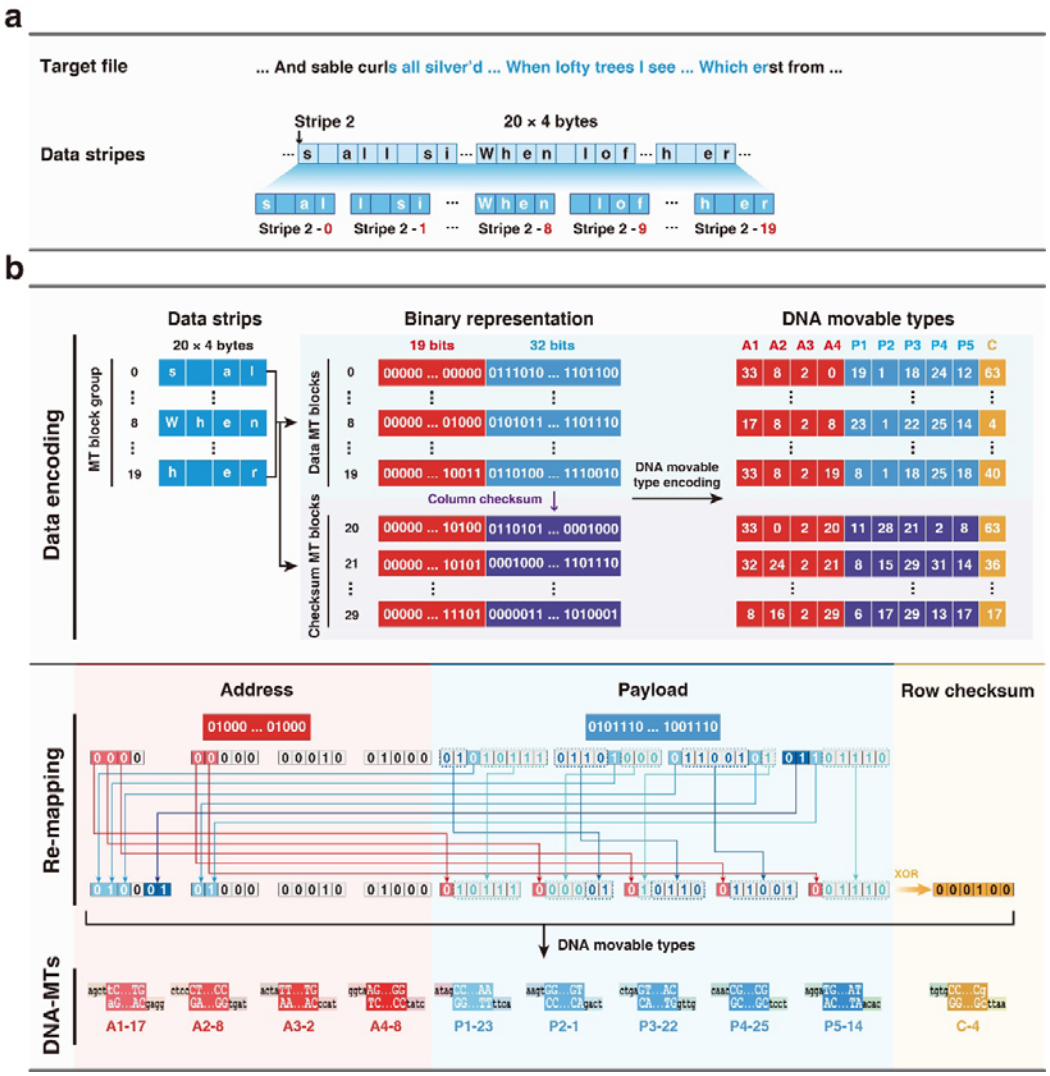
Extended Data Fig. 4: Scalability testing of DNA-MT blocks assembly using one-step enzymatic ligation. a. Assembly of 10 DNA-MT blocks at nano-scale. *We employed acoustic droplet ejection (ADE) technology (ECHO650, Labcyte) to decrease the volume of DNA-MT droplets from 50 nl to 800 nl. The ligation products of 10 DNA-MT blocks assembled in recombinant plasmids were amplified via PCR using Test-F/Test-R primers and run on an 1% agarose gel. The PCR fragments measured 705 bp for the ligation product and 514 bp for the empty control. The accuracy of the assembly was subsequently verified by Sanger sequencing of the PCR-amplified fragments, details in Supplementary Data 2.* ***b, Assembly of 20 DNA-MT blocks.*** *The ligation product of 20 DNA-MTs (shown in Supplementary Table 3) was assembled in recombinant plasmids and initially transformed into E.coli for long-term preservation. These blocks were then PCR amplified using Test-F/Test-R primers and run on an 1% agarose gel. The PCR fragments measured 945 bp for the ligation product and 514 bp for the empty control. The assembly accuracy was then verified by Sanger sequencing of the PCR amplification fragments, details in Supplementary Data 3.*

Extended Data Fig. 5:



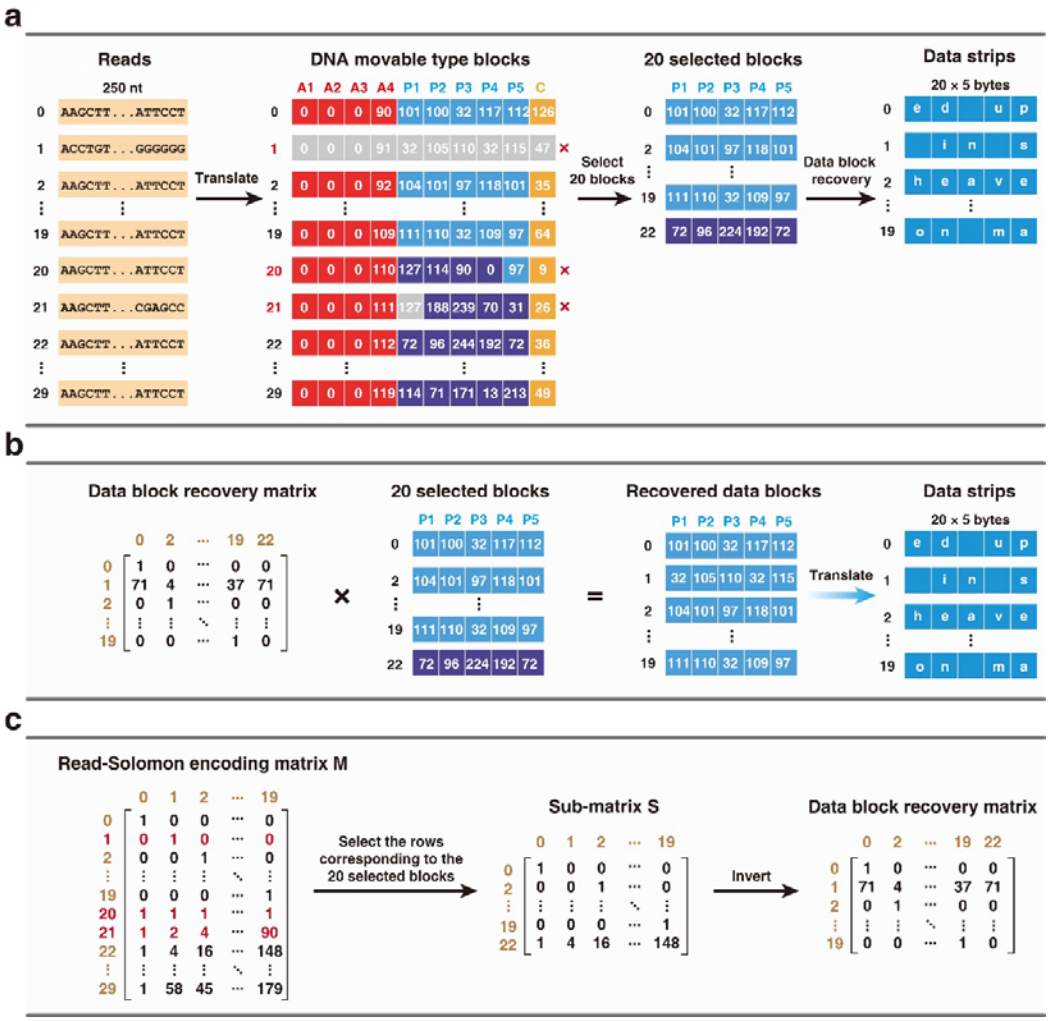
Extended Data Fig. 5: Construction of DNA-MT plasmid. *This plasmid has been engineered to incorporate Illumina sequencing sequences to minimize sequencing costs (Extended Data Table 3). Specifically, P5 adaptor, index 1, and Read 1 sequencing primer sequences were added adjacent to the 5' end of the HindIII site, and P7 adaptor, index 2, and Read 2 sequencing primer sequences were added at the 3' end of the EcoRI site. Amplification primers (Seq-F/Seq-R) were designed for direct library construction and sequencing of DNA-MT blocks on the Illumina platform.*

Extended Data Fig. 6:



Extended Data Fig. 6: Encoding the target file of Shakespeare's Sonnet 12 into DNA-MT blocks using BISHENG-1. a, Splitting the target file into data stripes. *We first split the target file into 80-character fragments and then divide each fragment into twenty 4-character data strips.* ***b, Encoding data strips into DNA movable type blocks.*** *For every 20 data strips, we generate 10 column-checksum strips are generated using the Reed-Solomon algorithm for error detection and correction. The original arrangement of BISHENG-1's cartridges often result in the overuse of certain cartridges, thereby reducing their lifespan. To avoid this, we shuffle the cartridges by remapping the data and column-checksum strips. This remapping process includes exchanging the top 3 bits of the highest order AMT and the top 2 bits of the second highest order AMT with the top 1 bit of 5 PMTs. This strategy is designed to achieve a balanced use of cartridges across the library. A row-checksum is then calculated and appended to each strip before translating each strip into a DNA movable type block.*

Extended Data Fig. 7:



Extended Data Fig. 7: Reconstructing original data from DNA movable type stored file. a, Decoding sequencing reads into data strips. *Sequencing reads are decoded into corresponding DNA movable types based on a DNA movable type code table (Supplementary Table 1), which are then compiled into data strips, further forming a data fragment (comprising 20 × 5 = 100 bytes). The missing data strips can be recovered using column-checksum DNA-MTs. In this case, the row 22 data strip, randomly selected, is used for recovery of row 1 data strip.* ***b, Recovering missing data blocks.*** *A set of 20 data blocks from a fragment (from 19 data strips and the row 22 column-checksum data strip) is selected and multiplied with a pre-calculated data block recovery matrix, translating recovered data blocks into the original 20 data strips (100 bytes).* ***c, Calculating the data block recovery matrix.*** *A 20×20 sub-matrix is extracted from a pre-calculated Reed-Solomon encoding matrix, and its inverse is calculated to serve as the data block recovery matrix.*