ChemComm

FEATURE ARTICLE

View Article Online
View Journal | View Issue

Cite this: *Chem. Commun.*, 2025, **61**, 2891

Enzymes as green and sustainable tools for DNA data storage

Xutong Liu, Enyang Yu, Qixuan Zhao, Haobo Han* and Quanshun Li*

Received 30th November 2024,
Accepted 13th January 2025

DOI: 10.1039/d4cc06351a

rsc.li/chemcomm

DNA is considered as an ideal supramolecular material for information storage with high storage density and long-term stability. Enzymes, as green and sustainable tools, offer several unique advantages for DNA-based information storage. These advantages include low cost and reduced generation of hazardous wastes during DNA synthesis, as well as the improvements in data reading speed and data recovery accuracy. Moreover, enzymes could achieve scalable data steganography. In this review, we introduced the exciting application strategies of enzymatic tools in each step of DNA information storage (writing, storing, retrieval and reading). We further address the challenges and opportunities associated with enzymatic tools for DNA information storage, aiming at developing new techniques to overcome these obstacles.

1. Introduction

Enzymatic polymerization has been considered as a green and effective route for polymer synthesis and promotes the controlled polymerization of artificial monomers by isolated enzymes *via* non-biosynthetic pathways.[1–4] To date, various polymeric materials have been successfully synthesized by enzymatic polymerization methods including lipase-catalyzed ring-opening polymerization and polycondensation and chemoenzymatic polymerization.[5–7] For instance, polymers with ultrahigh molecular weight could be synthesized by reversible addition–fragmentation chain transfer polymerization, using formate oxidase and horseradish peroxidase as catalysts.[8] In comparison to chemical synthesis, enzymatic polymerization possesses unique characteristics of mild conditions, high enantio-, chemo- and regio-selectivities, and no trace residues of metal catalysts.[9,10] Thus, it represents a green approach for polymer synthesis, which will greatly drive the polymer synthesis in an environmentally friendly and sustainable manner. DNA, as a natural supramolecular polymer, can regulate the biological reaction in living organisms through programmed monomer polymerization.[11–13] Recently, DNA has been considered as a potential alternative to solve the challenges in traditional information storage methods due to its characteristics of environmental friendliness, storage robustness and scalable information density.[14–17] Thus, it is necessary to address the issue of enzymatic DNA synthesis, especially in data storage.

Conventional data storage devices, such as magnetic, optical and solid-state devices, have several drawbacks that limit their

Key Laboratory for Molecular Enzymology and Engineering of Ministry of Education, School of Life Sciences, Jilin University, Changchun 130012, China. E-mail: quanshun@jlu.edu.cn, hanhaobo@jlu.edu.cn; Fax: +86-431-85155200; Tel: +86-431-85155200

Xutong Liu

Xutong Liu is a PhD student at the School of Life Sciences, Jilin University, China. She completed her BS degree in Biotechnology at Jilin University in 2020. Her research focuses on the exploration of new enzymatic tools for DNA information storage, particularly the application of DNA polymerases.

Enyang Yu

Enyang Yu is a master student at the School of Life Sciences, Jilin University, China. She completed her BS degree at Jilin University in 2023. Her research interests focus on DNA information storage techniques.

future applications.[18] Notably, the rapid growth of global digital data production is driven by the increasing shift of industrial networking demands to the cloud and the proliferation of internet-connected smart devices.[19] As data grow exponentially, conventional storage devices approach their physical limitations and struggle to keep pace with digital storage requirements. Moreover, the storage of digital data on a zettabyte scale requires the use of significant physical space. However, datacenters for centralized physical storage have huge electricity consumption and require additional energy for thermal balance and cooling, which in turn drives up the long-term maintenance costs.[20] Furthermore, silicon-based information storage systems have limited data retention time, which is also a great challenge for existing storage devices.[21] Thus, there is an urgent need for development of alternative storage media to bridge this ever-widening gap. Besides, the great demand for information storage has facilitated the integration of biotechnology and information technology, driving the advancement of DNA-based information storage.

The process of DNA information storage involves four key steps. Digital information is converted into DNA sequences by DNA synthesis (writing). DNA sequences are either encapsulated in silica for long-term storage *in vitro* or integrated into bacterial genomes or plasmids *in vivo* to ensure their stability (storing). To extract data from the stored DNA, the target DNA sequences are selectively accessed through PCR to separate from oligonucleotide pools (retrieval). Subsequently, the DNA molecules are sequenced to convert them back into digital data (reading). The major concern is the chemical DNA synthesis, which is often used in the writing of digital information. However, the process is tedious and time-consuming with toxic by-products, and meanwhile an exponential increase in cost will occur as the storage capacity scales up.[22–25] In contrast, enzymatic DNA synthesis provides a sustainable, ecofriendly and cost-effective strategy.[26] In addition, enzymes have gained attraction in the field of DNA information storage, which includes writing,[27,28] rewriting,[28–30] retrieval,[31,32] and data steganography.[33–35] With the development of molecular biology, various enzymes have been utilized as tools to ensure efficient biological catalysis for DNA processing. For instance, engineered DNA polymerases could be used in the synthesis of unnatural nucleic acids for long-term data storage and steganography.[36,37] In *in vivo* DNA data storage, enzymes play a critical role in DNA manipulation at the molecular level, which drives the advancement of DNA-based data storage towards further commercialization.

In this feature article, the key role of enzyme-mediated processes was highlighted in DNA-based information storage systems, especially for the four key steps (writing, storing, retrieval and reading). The critical challenges and opportunities were also discussed for use of enzyme molecules as tools for the scalability and sustainability of DNA information storage.

2. Overview of DNA information storage

Considering its long lifespan and scalable physical density, DNA is an excellent choice for information storage. DNA-based information storage is performed in the form of adenine (A), guanine (G), cytosine (C), and thymine (T) to encode different

Qixuan Zhao

Qixuan Zhao is a master student at the School of Life Sciences, Jilin University, China. He completed his BS degree in Biotechnology at Hebei University in 2023. He is currently working on the exploration and design of enzymes for their application in DNA information storage.

Haobo Han

Haobo Han is an Associate Professor at the School of Life Sciences, Jilin University. He earned his PhD from Jilin University in 2019 under the supervision of Professor Quanshun Li. From 2017 to 2019, he was a visiting scientist at Tufts University, USA. His research focuses on synthetic biology technologies, particularly the application of enzymes in DNA information storage and gene therapy for autoimmune diseases.

Quanshun Li

Quanshun Li received his PhD degree from Jilin University in 2009 and is currently a Distinguished Professor of Biochemistry and Molecular Biology at Jilin University. His research group focuses on the catalytic mechanism and molecular design of enzymes, and the enzymatic synthesis of polymeric materials for drug/gene delivery and DNA data storage. He has published more than 90 papers, obtained 5 patents and won the Second Prize of the Natural Science Award of Jilin Province.

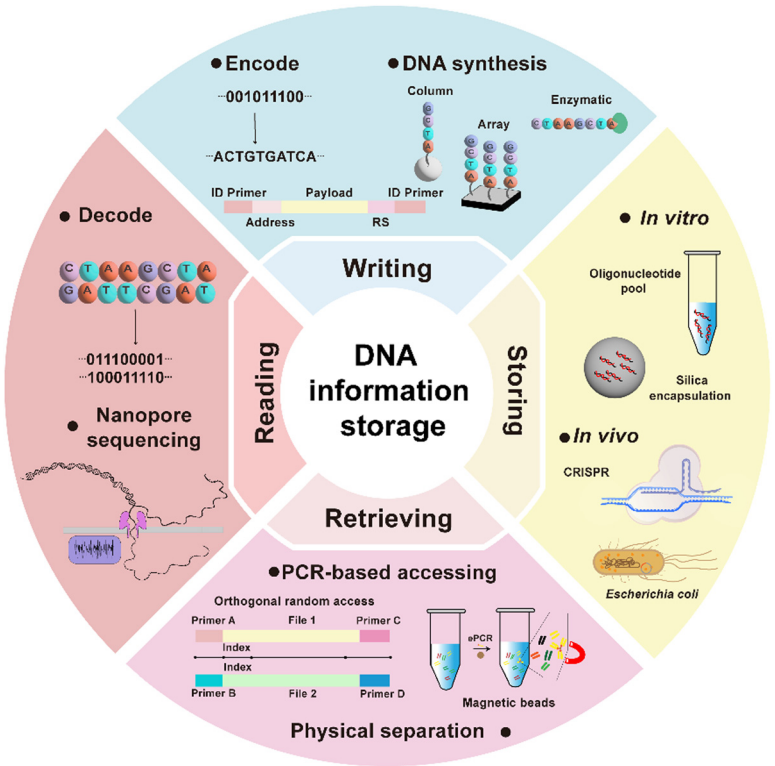

Fig. 1 Schematic illustration of the basic process of DNA information storage. The process consists of four steps: converting digital information into DNA sequences (writing), physically encapsulating in silica to protect DNA molecules against the breakage *in vitro* or integrating DNA fragments into bacterial genomes or plasmids *in vivo* for long-term storage (storing), selectively accessing the target DNA sequence (retrieving), and sequencing DNA molecules to convert them back into digital data (reading).

digital data. Norbert Wiener and Mikhail Neiman first discussed the idea of using DNA as "genetic memory" instead of magnetic devices in the 1960s.[38,39] With the advancement of DNA synthesis and sequencing platforms, DNA as a data storage medium has attracted increasing attention to gradually make it commercially viable.

The technological process of a synthetic DNA-based storage system consists of four steps: writing, storing, retrieval and reading. An overview of the steps involved in DNA information storage is illustrated in Fig. 1. Conversion of digital data from binary information (bits) into DNA sequences is initially achieved using error-correction codes,[40,41] such as 00 for A, 01 for C, 10 for G and 11 for T.

The encoding process introduces physical redundancy, which serves to prevent the data loss during storage. Following the encoding of DNA, oligonucleotides are synthesized by solid-phase phosphoramidite chemical synthesis or enzymatic synthesis.[42,43] Nevertheless, the production of toxic waste is inherent in chemical DNA synthesis, thereby rendering the enzymatic DNA synthesis to be a sustainable alternative.[44] Moreover, the stability of DNA is another critical feature in DNA information storage. The DNA strands can be preserved to shield the stored information from environmental degradation by freezing DNA molecules in solution, drying the DNA samples, or encapsulating DNA molecules in silica nanoparticles.[45] In addition, the insertion of synthetic DNA into plasmids or the integration into the genome of living cells could be potentially used in the intracellular DNA data storage.[46,47] Furthermore, selectively accessing the target DNA sequence is referred to as random access, which is more challenging in DNA-based data storage than in digital storage media. Random access in DNA data storage mainly focuses on PCR utilization.[48] This process can be supported by selective methods such as magnetic bead extraction with probes mapped to data blocks or PCR using primers associated with data blocks.[49,50] Thus, the target DNA is selectively retrieved and amplified. Subsequently, DNA molecules are read to recover digital data by sequencing techniques, such as Illumina and nanopore sequencing.[51] The success of this step depends on the sequencing coverage and the error rate experienced throughout the decode process.[52] Taken together, emerging technologies in synthetic biology that involve the above processes will be pivotal in transforming DNA data storage into a commercially available technology. This transformation will be facilitated by the advent of enzymatic DNA synthesis, digital microfluidics-based random access, and high-throughput sequencing.

3. Writing

3.1 Chemical DNA synthesis

To date, solid-phase phosphoramidite chemical synthesis has been widely used in all major DNA synthesis platforms.[53–55]

Due to the labour-intensive nature of chemical synthesis workflow, the DNA synthesis capability has been concentrated within specialized reagent manufacturers. Leading vendors such as Agilent Technologies, GenScript, Integrated DNA Technologies, Twist Bioscience and others offer custom DNA (and RNA) synthesis service on demand in a range of formats. Furthermore, commercial oligonucleotide synthesizers could be purchased to produce DNA for laboratory service.

The typical solid-phase synthesis of oligonucleotides *via* phosphoramidite chemistry is described in Fig. 2. Current technologies employ a solid support to build up a sequence, nucleotide by nucleotide, through a four-step synthesis cycle.[56–59] The initial step is the reaction of nucleotides with the protected active group, which has been pre-attached to the solid phase carrier CPG, in the presence of trichloroacetic acid. This process serves to expose the 5′-hydroxyl group, which will be utilized in the subsequent coupling step. In the coupling step, the raw material for DNA synthesis is the phosphoramidite-protected nucleotide monomer. The monomer is mixed with the activator tetrazolium, leading to the formation of the nucleoside phosphite activation intermediate. The intermediate is activated at the 3′-end, while the 5′-hydroxyl group remains to be dimethoxytrityl (DMT)-protected. Subsequently, the compound undergoes a condensation reaction with the free 5′-hydroxyl group of the nucleotide attached to CPG in solution. In the third step, known as capping, the unreacted 5′-hydroxyl group attached to the CPG needs to be closed to prevent the extension in subsequent cycles. Acetylation is commonly used to close the hydroxyl group once the coupling reaction completes. In the fourth step, oxidation occurs where the phosphinylidene form is transformed into a more stable phosphotriester. This conversion takes place in the presence of iodine, which is dissolved in tetrahydrofuran acting as an oxidant.

The major drawbacks of chemical synthesis of DNA are the use of hazardous chemicals and the inability to synthesize oligonucleotide sequences longer than 200 bp.[60,61] Meanwhile, the accuracy of the synthesized oligonucleotides decreases with the elongation of DNA chains. The length limitation of chemical DNA synthesis affects the storage of large-scale DNA information.[62] Therefore, it is necessary to encode information in DNA by physically dividing the encoded digital information into different blocks to form short DNA fragments, or using DNA assembly and ligation to form long DNA fragments. These practical issues may hinder further industrialization of chemical DNA synthesis for data storage.

3.2 Enzymatic DNA synthesis using terminal deoxynucleotidyl transferase (TdT)

Enzymatic DNA synthesis is an attractive approach with scalable, stereospecific and environmentally friendly characteristics.[63] DNA polymerases can promote the DNA synthesis, including the amplification of existing DNA sequences or the generation of entirely novel DNA sequences through *de novo* synthesis.

Terminal deoxynucleotidyl transferase (TdT) is a low-fidelity DNA polymerase of the X Family,[64] which has been explored for the application in DNA information storage and gene synthesis. Enzymatic DNA synthesis using TdT has been considered as a natural candidate method, which can catalyze the template-free addition of nucleotides on the 3′-end of a DNA strand. The synthesis of novel DNA strands was controlled by the composition of nucleotide substrates. In 1959, Bollum first described that TdT could be used as a ssDNA polymerase for the template-free *de novo* DNA synthesis.[65] Before this research, DNA polymerases were always used in the amplification of dsDNA, which requires templates and primers. Afterwards, Bollum revealed

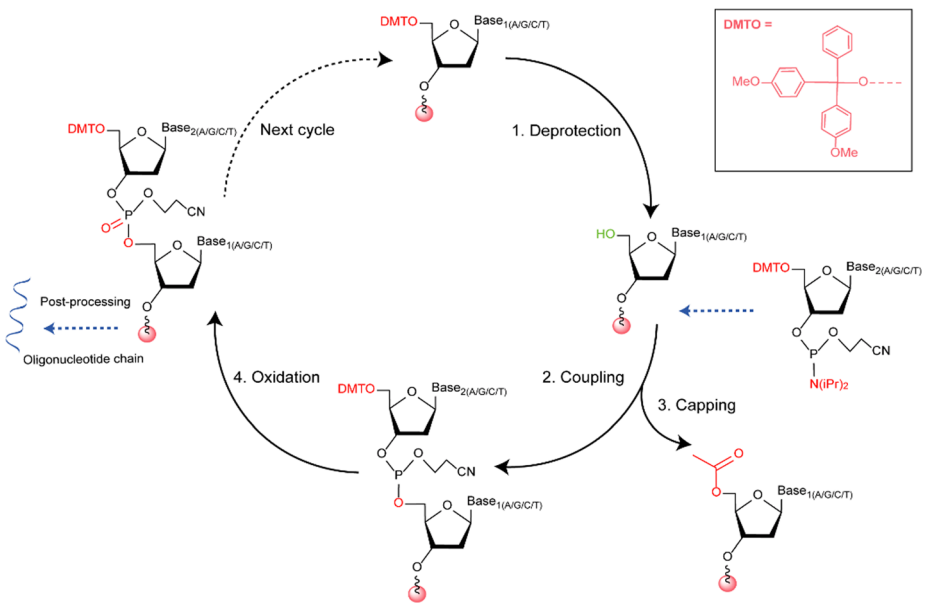

Fig. 2 DNA synthesis using a solid-phase phosphoramidite chemistry method. The synthetic cycle involves four steps: deprotection, base coupling, capping and oxidation.

the enzymatic mechanism of TdT-mediated DNA synthesis.[66] Reversible 3′-protecting groups are required to prevent further addition of nucleotides and control strand elongation. However, natural TdT is difficult to recognize the substrate of 3′-blocked dNTPs. To overcome the limitations of TdT with 3-blocked dNTPs, Palluk *et al.* designed a strategy based on the use of TdT–dNTP conjugates to control the synthesis of oligonucleotides.[67] As shown in Fig. 3A, TdT was conjugated to a dNTP molecule *via* a cleavable linker, and the tethered dNTP can be synthesized on the ssDNA primer to block the further elongation owing to steric hindrance. The linker can be cleaved to drive the deprotection for subsequent extension. The reversible termination of chain extension by TdT–dNTP conjugates can realize the enzymatic *de novo* synthesis of a 10-mer oligonucleotide. In addition, Lu *et al.* constructed an engineered TdT from *Zonotrichia albicollis* (ZaTdT), which could solve the issue of the incompatibility between 3′-ONH_2-dNTPs and the catalytic cavity of ZaTdT.[68] The engineered ZaTdT showed a 1000-fold higher catalytic activity with 3′-ONH_2-dNTPs than the commonly used MmTdT.

To overcome the natural promiscuity of TdT, Lee *et al.* realized the controlled TdT extension activity with apyrase by degrading free nucleotides,[69] as shown in Fig. 3B. This strategy could be used in the synthesis of short homopolymeric blocks, which then encode digital information by the transition between nucleotides. To further improve the efficiency of TdT-based enzymatic *de novo* DNA synthesis, Lee *et al.* designed a multiplexed synthesis method using photolithography to selectively control the extension activity of TdT in an array (Fig. 3C).[70] Co^{2+}, the essential metal cofactor of TdT, was trapped with the photocleavable caging molecule, which was cleaved by UV light to be released. Thus, the catalytic activity of TdT can be regulated in the multiplexed oligonucleotide synthesis in the array by a mask pattern of UV light. This method can individually synthesize 12 unique oligonucleotides, which could simultaneously encode 110 bits of digital data on the

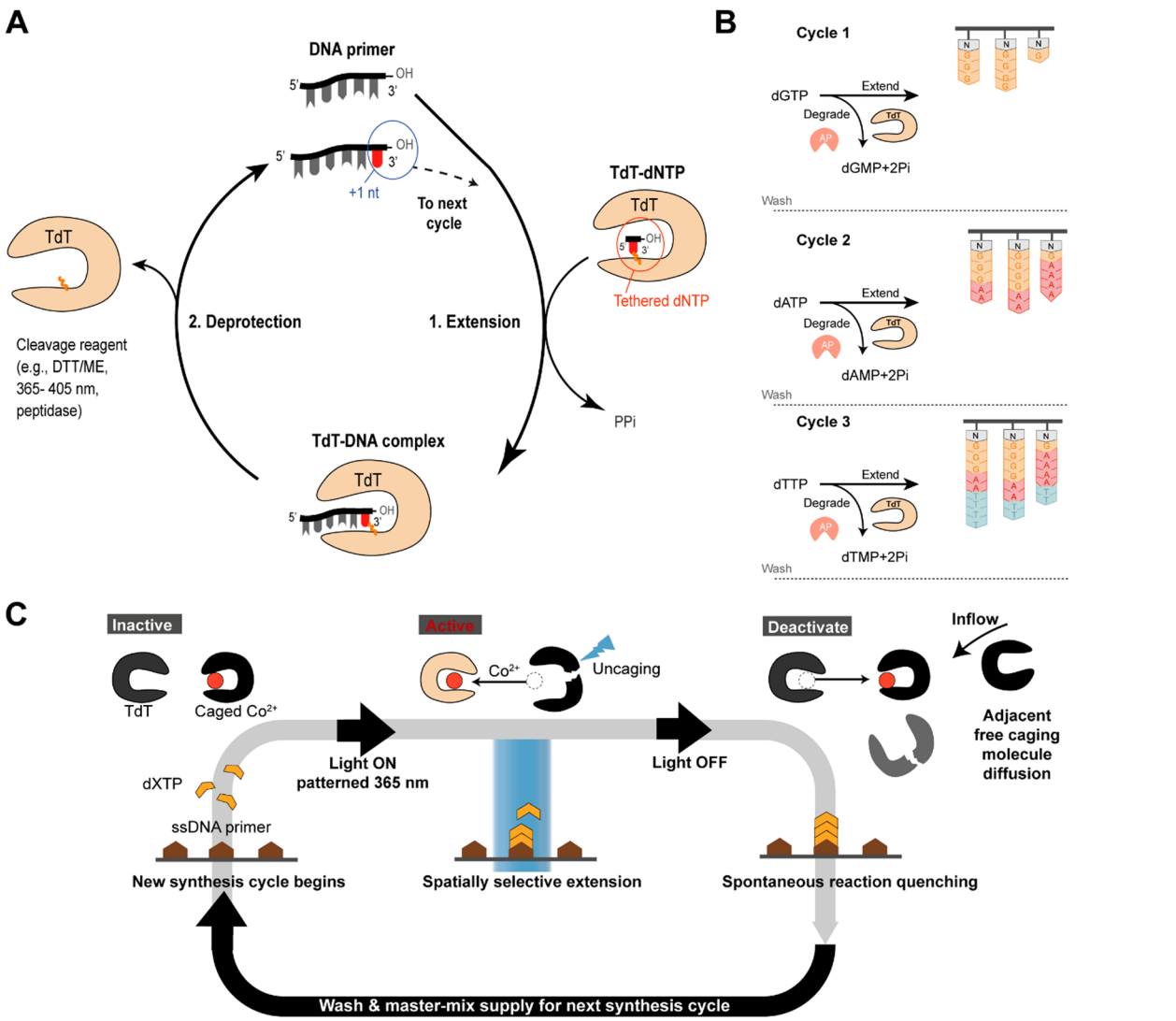

Fig. 3 The TdT-mediated DNA synthesis. (A) Scheme for two-step oligonucleotide extension using TdT–dNTP conjugates. (B) Design of enzymatic synthesis consisting of TdT and apyrase (AP). (C) Overview of a photon-directed parallelized enzymatic DNA synthesis system.

array surface. TdT has been proved to be a far better candidate in enzymatic oligonucleotide synthesis compared to its predecessors. Nevertheless, the TdT-based enzymatic DNA synthesis still requires optimization to accurately address the on-demand synthesis of oligonucleotides and reduce the generation of failure strands.

3.3 Enzymatic DNA synthesis using engineered DNA polymerases

Natural DNA polymerases find it difficult to execute the synthesis of DNA molecules with unique structures, such as modified DNA, xenobiotic nucleic acids (XNA) and L-DNA.[71] The synthesized unnatural DNA has a significant advantage of evading the degradation by nucleases, exhibiting great potential in long-term data storage. Moreover, encoded information that can exclusively be synthesized or read by engineered DNA polymerase may potentially present a viable approach for data steganography.[72–76] Shroff *et al.* reported an error-correcting reverse transcriptase that could read 2′-*O*-methyl (2′-Ome)-modified templates.[72] The designed variant could reversely transcribe a template with 44 sequential 2′-Ome bases (Fig. 4A). To demonstrate the utility of the designed variant in DNA information storage, 24 kb digital information was encoded in 2′-Ome RNA and reversed into sequencing-compatible DNA. Encoded DNA and 2′-Ome RNA were mixed for amplification and decoding. Commercial DNA polymerase only recognized the encoded DNA sequences. In contrast, the designed variant can recover both the sequence of natural DNA and 2′-Ome RNA,

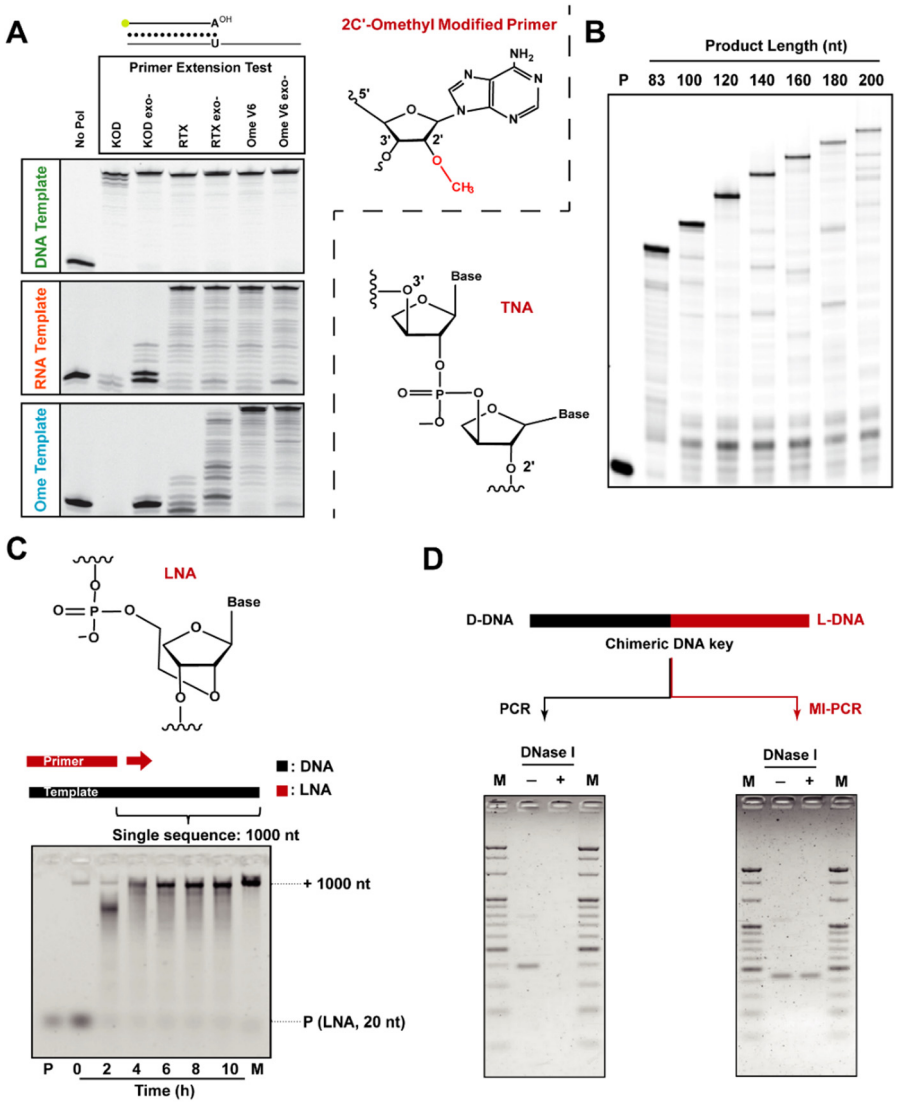

Fig. 4 Enzymatic synthesis of modified DNA and XNA by engineered DNA polymerases. (A) Primer extension of engineered DNA polymerase on DNA, RNA, and 2′-*O*-methyl templates. Reproduced from ref. 72 with permission from the American Chemical Society, copyright 2021. (B) TNA oligonucleotides amplified from DNA templates for data storage against the nuclease exposure. Reproduced from ref. 73 with permission from the American Chemical Society, copyright 2020. (C) One kilobase LNA synthesis by engineered DNA polymerase. Reproduced from ref. 74 with permission from the American Chemical Society, copyright 2020. (D) Mirror-image DNA information storage for chiral steganography. Reproduced from ref. 76 with permission from Springer Nature, copyright 2021.

indicating that the engineered polymerase provided a viable option to hide message storage among normal information.

Furthermore, due to the degradation resistance of nucleic acid analogues, engineered DNA polymerases can copy DNA templates into XNAs with the potential for long-term DNA data storage. Yang *et al.* developed an engineered family B DNA polymerase with the ability to synthesize α-L-threofuranosyl nucleic acid (TNA), resulting in the faithful reading of information encoded in TNA,[73] as shown in Fig. 4B. The 22 349 bytes of digital information were encoded in 7451 unique DNA oligonucleotides of 83 nt and written into TNA. The engineered DNA polymerase transferred the information between DNA and TNA in a write–store–read cycle to meet the sequencing demands. Through the information writing in TNA, storage files could be completely recovered from the nuclease exposure, revealing that TNA is a biologically stable system for long-term information storage. Meanwhile, engineered DNA polymerase could produce TNA strands of up to 200 nt, providing a viable option to improve the storage capacity. To further increase the capability of XNA synthesis, Hoshino *et al.* developed the variants of KOD DNA polymerase with high efficiency and fidelity allowing the synthesis of 1000-nt locked nucleic acid (LNA),[74] as shown in Fig. 4C.

L-DNA, the enantiomer of natural D-DNA, is an ideal nucleic acid analogue because it is completely resistant to nuclease, but it has identical kinetic and thermodynamic properties to D-DNA.[75] Furthermore, L-DNA is unable to form contiguous Watson–Crick base pairs with D-DNA. To overcome the inability of natural DNA polymerases to recognize L-DNA, Fan *et al.* developed an effective mirror-image DNA information system by chemically synthesizing a mirror-image Pfu DNA polymerase,[76] as shown in Fig. 4D. The mirror-image DNA polymerase could achieve the synthesis of a 1500-nt L-DNA strand. The information stored in L-DNA is more resistant to the biodegradation in the natural environment than that in D-DNA. Moreover, chimeric D-DNA/L-DNA molecules were designed to transmit false and secret messages. Using mirror-image Pfu DNA polymerase, the L-DNA sequence in the chimeric DNA key could be successfully amplified. Thus, mirror-image DNA polymerase provides a better solution for data steganography in DNA information storage to encrypt key data.

3.4 Enzymatic DNA storage using nicking endonucleases

Except for the base composition of DNA sequences, DNA information storage could be realized on the structure of nucleic acids. Nicking endonucleases were induced to cut the 450-bp dsDNA backbone, which were amplified from genomic DNA as DNA registers containing designated nicking positions.[77] The encode of digital data was related to the nick in dsDNA, with the value 1 corresponding to a nick and the value 0 corresponding to the absence of a nick. The nick-based storage system is a cost-effective method and suitable for the repeated utilization. Nevertheless, the use of nick-based DNA data storage resulted in a reduction of storage density. In addition, it was challenging to scale up or perform other functions, such as generating copies by PCR.

4. Storage

4.1 *In vitro* DNA data storage

In *in vitro* DNA data storage, encoded information was synthesized as a single-stranded DNA and often stored as freezing DNA solution or dehydrated DNA for long-term storage.[78,79] However, the synthesized DNA is prone to be degraded in the presence of light, water and oxygen, making the stored information difficult to recover. Liu *et al.* stored DNA in cellulose paper by electrostatic adsorption to reduce the presence of water, thereby keeping the samples away from those damaging factors.[80]

Furthermore, the method of physical encapsulation was proposed to improve the stability of DNA data storage. Koch *et al.* designed a DoT storage architecture that encapsulated DNA in silica beads and then used it as a material for 3D printing,[81] as shown in Fig. 5A. The 0.3% weight of DNA from 3D-printed Stanford Bunny that contained a 45-kb digital DNA blueprint was amplified from the previous generation and encapsulated into the next generation for multi-round replication. The potential for long-term storage was demonstrated by creating five successive generations of rabbits without the loss of information. Additionally, encoded DNA was loaded in polymethyl methacrylate, which was cast in the shape of a lens. This approach provided a novel strategy of physical steganography, which could secretly store digital information in daily objects. Antkowiak *et al.* developed an approach for long-term data storage, in which the dehydrated DNA was encapsulated in silica nanoparticles and added on the surface of glass to protect it from degradation (Fig. 5B).[82] The accelerated aging experiment revealed that the encapsulated DNA was more stable than unprotected DNA, which could not realize the information recovery. Although the *in vitro* strategies improved the stability of DNA information storage, rewriting and retrieval of large-scale DNA information storage remain a challenge that needs to be addressed.

4.2 *In vivo* DNA data storage

With the rapid development of synthetic biology, it is possible to store digital data *in vivo* and achieve the rewriting of intracellular DNA.[83] The technical challenges of developing the system are large-scale genomic integration and accurate rewriting of information without the off-target effect. Integrase and clustered regularly interspaced short palindromic repeats (CRISPR)-Cas systems provide a feasible option for *in vivo* DNA information storage to overcome these challenges.

Serine integrases are capable of catalyzing the recombination between att sites on both linear and circular DNA substrates.[84,85] The outcome of this process depends on the specific position and orientation of the att sites, which can result in the integration, excision or inversion of DNA. Sun *et al.* developed a recombinase-based site-specific genome engineering toolbox, which can assemble the synthesized 3–8 kb DNA fragments into 51-kb DNA fragments and integrate into the bacterial genomes,[86] as shown in Fig. 6A. After a continuous passage of 2000 generations, complete DNA information could still be successfully retrieved.

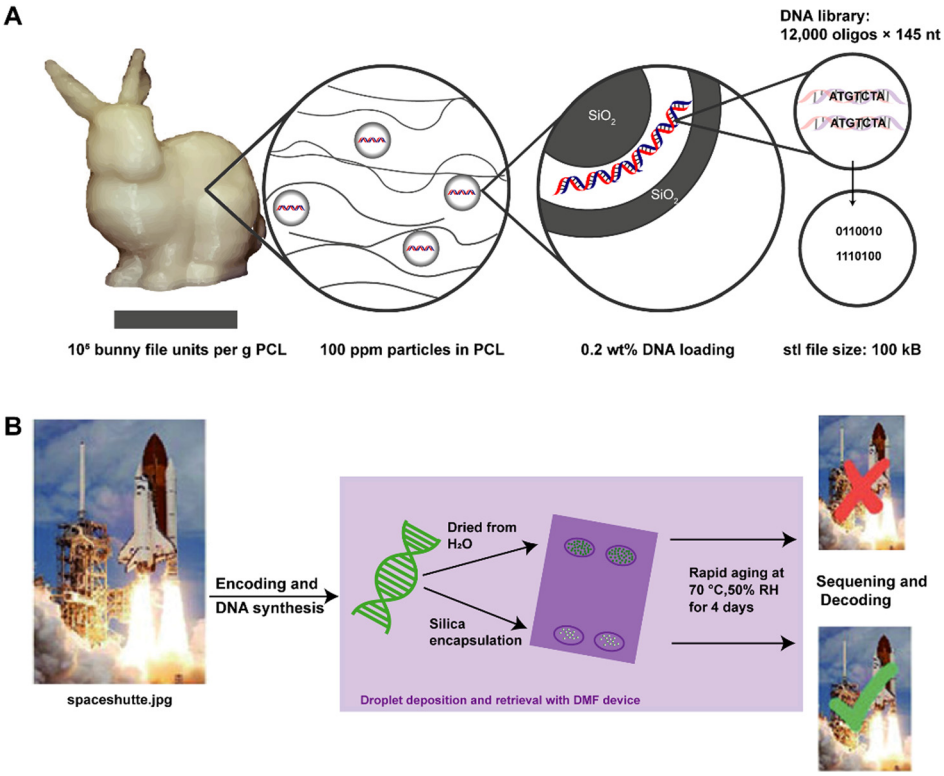

Fig. 5 Strategies of *in vitro* long-term DNA data storage. (A) Schematic architecture of the 3D-printed Stanford Bunny, which contained the encapsulated DNA library. Reproduced from ref. 81 with permission from Springer Nature, copyright 2020. (B) Overview of digital microfluidics (DMF)-compatible silica nanoparticles encapsulating the DNA library for DNA data storage. Reproduced from ref. 82 with permission from Wiley, copyright 2022.

This strategy revealed that the integration of stored information into the bacterial genome environment was stable and error-proof for replication. In contrast to the storage of oligonucleotide pools, the cell growth could automatically regenerate data, thereby avoiding the data loss caused by the long-term storage and frequent retrieval.

Other strategies for *in vivo* DNA data storage mainly focused on the CRISPR-Cas system.[87–91] As an adaptive immune system for bacteria, the CRISPR-Cas system could protect the bacteria by acquiring invader DNA and integrating it into the CRISPR array. The system exhibits a significant impact on genetic engineering and has been successfully used in the data rewriting process. Farzadfard *et al.* designed a dual-plasmid system based on CRISPR-Cas12a-λRed to rewrite the information *in vivo*,[90] as shown in Fig. 6B. The info-plasmid included crRNA and the encoded information sequence, and the help plasmid was used to express Cas12a and λRed. The Cas12a guided by crRNA could selectively cleave the target DNA sequence and then the λRed replaced the target DNA fragment to recombine the info- plasmid *via* homologous arms for information rewriting. The ratio of rewriting cells and the accuracy of rewritten information could reach 94%. Cas9 is another Cas protein that has been used in the intracellular rewriting of DNA information storage.[87] The CRISPR/Cas9 system was developed to rewrite the image with the same gRNA in yeast cells (Fig. 6C). The targeted cells could be selectively removed by counterselection operation. The *Trp1* and *Ura3* genes of untargeted cells were eliminated, preventing their growth on the synthetic medium without tryptophan. Moreover, gRNA activated the targeted cells to completely acquire the *Ura3* gene, which were non-survivable in the medium containing 5-fluoro-orotic acid (5-foa) to promote the counterselection. Thus, the targeted cells were completely removed from cell pools. Finally, the cells containing new information were added into the cell pools to achieve the rewriting of specific information. The dCas9, a variant of Cas9, still recognized and bound to specific DNA targets, with no cleavage activity. This property of dCas9 could be exploited in cooperation with the mutagenic protein APOBEC3A to achieve the programmable system of information rewriting,[91] as shown in Fig. 6D. When dCas9 bound the targeted DNA to form a nucleotide R-loop structure, APOBEC3A efficiently facilitated the mutation of dC to dT in the displaced strand of the R-loop. Subsequently, the rewriting DNA sequence could be read *via* sequencing. These enzymatic tools provided a useful strategy for rewriting or editing data rather than chemically resynthesizing DNA, thus reducing the cost of storing information in DNA.

5. Retrieval

To scale up the DNA information storage, selectively reading data from large-scale data pools is a critical process, referred to random access in conventional digital storage.[92] This process eliminates the need to read all the stored information to access

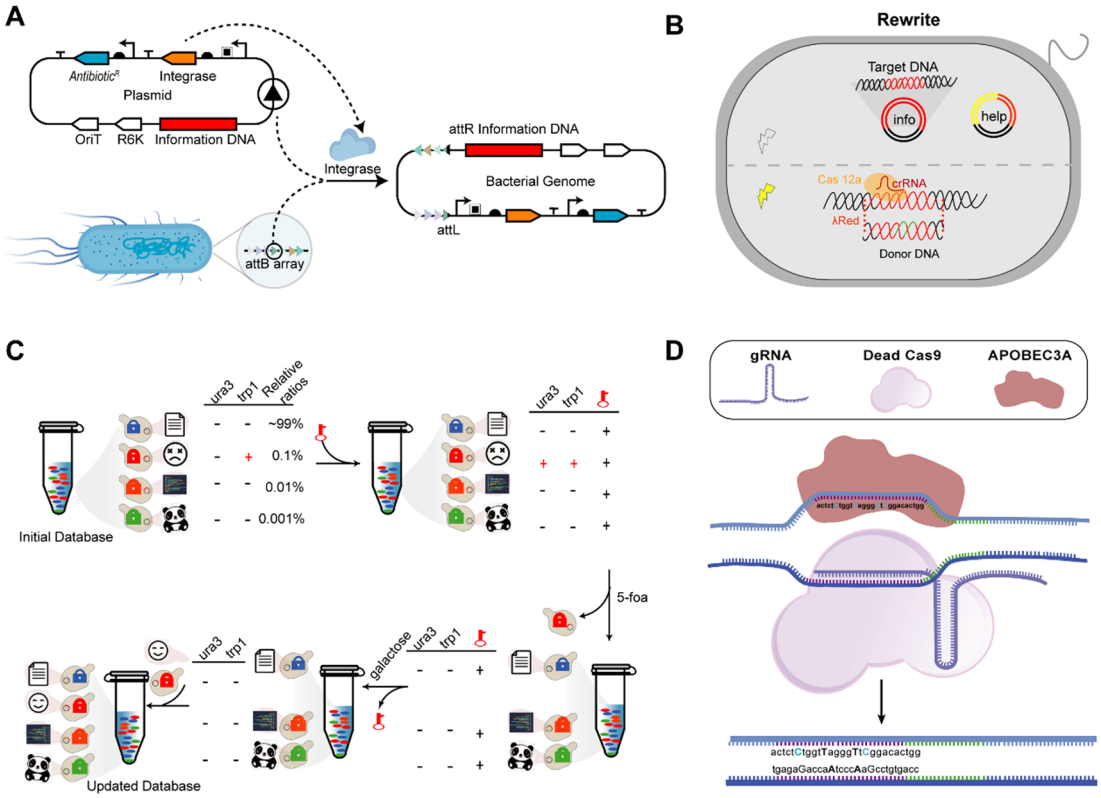

Fig. 6 The *in vivo* DNA information storage. (A) Schematic integration of information DNA into bacterial genomes. (B) Schematic illustration of DNA information storage and rewriting within living cells. (C) The process of randomly rewriting the digital data from the cell pool. Reproduced from ref. 87 with permission from Wiley, copyright 2024. (D) Schematic view of writing digital information in the form of precise DNA sequence edits on pre-made DNA molecules.

a specific file. Different methods are used to selectively retrieve the data, such as DNA polymerase-based PCR amplification, magnetic bead extraction, or fluorescence-activated sorting.[93–95]

Designing unique primers for every data file is a widely used method for random access by DNA polymerase-based PCR amplification.[92] The challenge of this method is designing primers that do not conflict with the payloads. As shown in Fig. 7A, 35 files were encoded and segmented into more than 13 million oligonucleotides with a unique file ID as primers.[93] Each file could be retrieved to be recovered with no errors. Furthermore, nested PCR has been proposed to increase the number of file addresses in storage systems. Since the probability of potential off-target molecular interaction increases with the capacity of the system, the addresses must be sufficiently different from each other in sequences so that the number of addresses is limited and thus restricts the total capacity of the system. The purification with magnetic beads was integrated with PCR primers to solve the challenge in the random access of large-scale DNA information storage,[94] as shown in Fig. 7B. A 9-kb file was selectively retrieved from a 5-TB database through a unique PCR primer. The primer was chemically modified and bound to the functionalized magnetic beads and then physically separated from unbound oligonucleotides by emulsion PCR for future reuse. In addition, the nested PCR primer was designed to expand the number of possible addresses and combined with functionalized magnetic beads to further improve the capacity of the database. This strategy could be used to store and access individual files containing at least GBs of data.

To improve the throughput of file retrieval, fluorescence-activated sorting is also a potential approach for data access.[95] The plasmid DNA was encapsulated by positively charged silica *via* electrostatic interaction (Fig. 7C). The orthogonal ssDNA barcodes, describing key features of image for file selection, were modified on the surface of silica capsules. Fluorescent labelled oligonucleotide probes were bound to the barcode on the surface of silica particles by annealing, enabling the sorting of the target file from a pool of 10^6 data. To further avoid the PCR crosstalk, a strategy based on the temperature-dependent semipermeable microcompartment was designed for repeated PCR-based access from complex file pools,[96] as shown in Fig. 7D. Semipermeable microcompartments were constructed using protein–polymer conjugates, enabling the localization of biotinylated DNA files in proteinosomes. The proteinosomes were thermoresponsive with reversible temperature-controlled membrane permeability, releasing amplified dsDNA at PCR temperature, thereby significantly reducing the PCR crosstalk. Magnetic particles were incorporated into proteinosomes to retrieve target DNA files by magnetic separation, allowing reliable repeated access to DNA-encoded data.

Erasing is another feature to store different sets of data by DNA polymerase-based PCR.[97,98] The specific files were addressed by

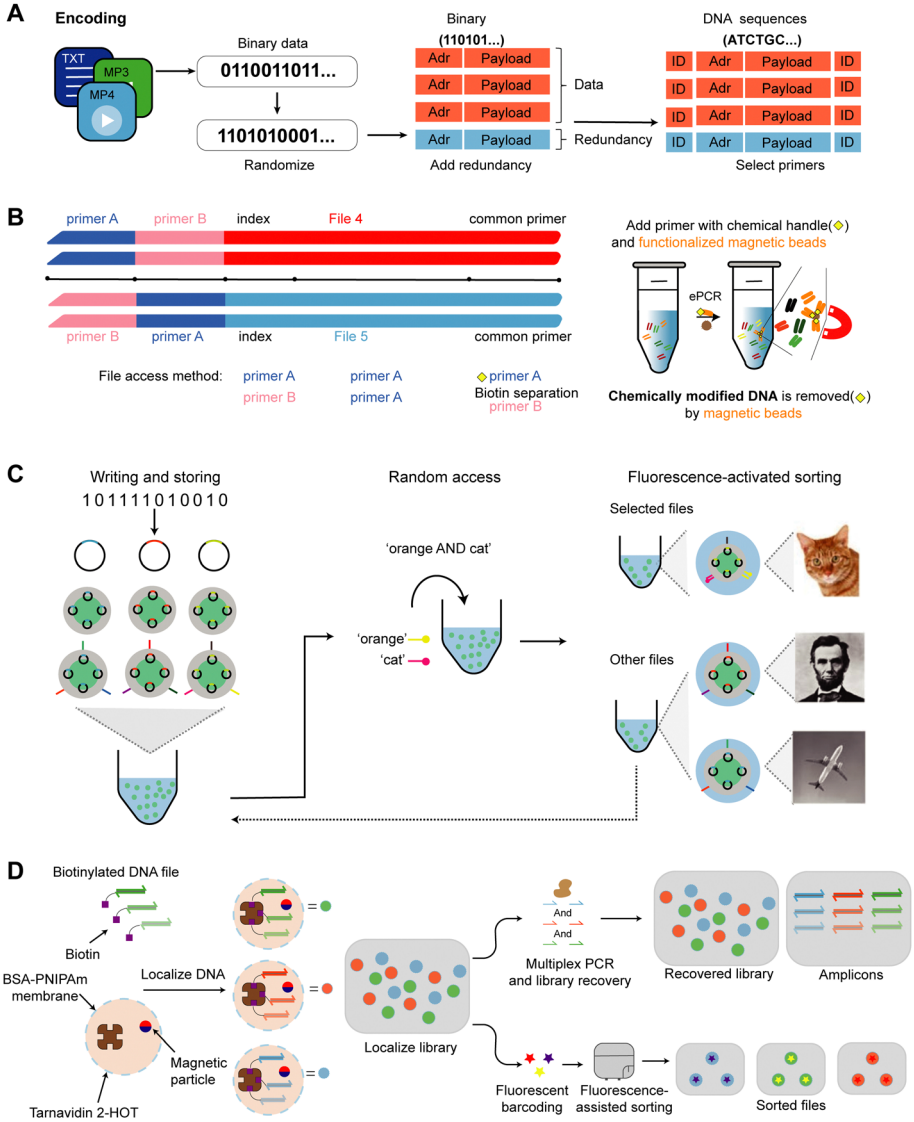

Fig. 7 Random access in DNA data storage. (A) Design a primer library for PCR-based random access. (B) Using nested, hierarchical primer for addresses. (C) Silica capsules with dye-labelled orthogonal barcodes used to select the specific data by fluorescence-activated sorting. Reproduced from ref. 95 with permission from Springer Nature, copyright 2021. (D) Thermoresponsive proteinosomes using fluorescence-assisted sorting for repeated access in DNA data storage.

PCR and deleted by the cleavage of restriction endonuclease. Thus, specific files were successfully erased, and new files were repeatedly loaded into sets of data. However, the chances of off-target interaction are most likely to occur between files that have a higher similarity to the address sequences of the desired file. Orthogonal barcode design and chemical modification of probes could be helpful to achieve and extend to achieve a higher capability of data access in the future.

6. Reading

6.1 Detection of ionic currents using nanopore proteins

DNA sequencing information could be decoded using well-established approaches such as Sanger sequencing, Illumina sequencing, and nanopore sequencing. Nanopore sequencing, as a powerful tool, has the advantages of long reads and high throughput compared to Sanger sequencing and Illumina sequencing.[99–101] Meanwhile, nanopore sequencing is a comparatively more portable and real-time sequencing technique.[100] However, nanopore sequencing has higher error rates and requires a higher sample quality compared to other sequencing methods.[102] There are two main types of nanopores used in DNA sequencing: biological nanopores and solid-state nanopores.[103] Biological nanopores are protein molecules embedded in phospholipid bilayers, while solid-state nanopores are made of materials such as silicon nitride (SiNx), carbon nanotubes and graphene. Biological nanopores, as bio-inspired nanodevices, have a specific pore size, flexible chemical or biological modification, low cost and high sensitivity.[104,105]

Alpha-hemolysin (α-HL) and *Mycobacterium smegmatis* porin A (MspA) have been extensively utilized in biological pores.[106]

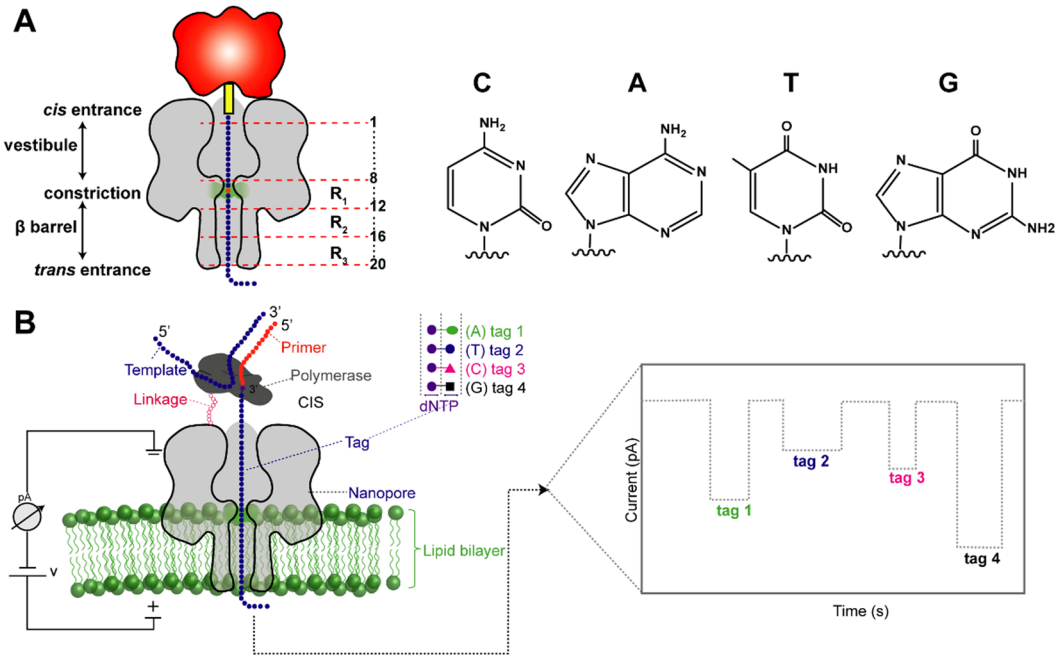

Fig. 8 Nanopore-based DNA sequencing. (A) The structure of α-HL nanopores. R1, R2 and R3 represent the three base-recognition sites within the β-barrel domain. (B) The principle of nanopore-based single-molecule DNA sequencing using nucleotides with different polymer tags.

α-HL is a membrane channel protein forming 1.4 nm internal-diameter β-barrel transmembrane pores.[107,108] It has been widely used in the detection of single-stranded nucleic acid molecules due to its ability to form rigid nanopores with consistent diameters. The transmembrane β-barrel of an engineered α-HL pore contains three recognition sites that can be used to identify all four DNA bases in an immobilized single-stranded DNA molecule,[109] as shown in Fig. 8A. Stoddart *et al.* initially considered that two recognition sites (R1 and R2) in the transmembrane region might be favorable for sequencing, because each base is read twice, first at R1 and second at R2.[109,110] The built-in proof-reading mechanism could improve the overall sequencing quality. However, more than two recognition sites could not be practical as it is difficult to assess the ionic current generated by three recognition sites from electrical noise. To address this challenge, Stoddart *et al.* proposed to modify the R1 recognition site in α-HL by mutagenesis to enhance a nucleotide-detecting induction site.[110] The hydrophobic and bulky side chains provide steric barriers to ion flow, which improve the discrimination of nucleobases at R1 to yield accurate signal recording. Ayub *et al.* attempted to reduce the number of recognition sites in the α-HL pore by using truncated pores.[111] By deleting and mutating amino acids on the β-barrel, a pore could be created with just two recognition sites. Compared to the wild-type α-HL with 5-nm long β-barrel, the pores with shortened β-barrels were proved to be more suitable for high-resolution nucleotide sensing. They could bind the positively charged β-cyclodextrin, permitting the continuous recognition of individual nucleoside monophosphates.

Moreover, Stranges *et al.* constructed a nanopore-based sequencing-by-synthesis (Nanopore-SBS) approach, using a set of nucleotides with polymer tags to allow the discrimination of nucleotides in a biological nanopore,[112] as shown in Fig. 8B. A high-throughput sequencing platform was built using nanopore sensors, enabling parallel sequencing of multiple DNA templates at the single molecular level. This approach provided real-time single-molecule electronic DNA sequencing data with single-base resolution.

α-HL is a structurally stable nano-detection device, but the limited pore size (~1.4 nm) has restricted its application in the analysis of ssDNA, RNA or small molecules. Additionally, the 5-nm long cylindrical β-barrel of α-HL presents a structural limitation in the accurate sequencing that dilutes the ion current specific to individual nucleotides and yields minor differences between nucleotides.[113]

6.2 Controlling translocation speed using motor proteins

To overcome the structural defects of the β-barrel of α-HL, *Mycobacterium smegmatis* porin A (MspA) was identified for a variety of bio-nanotechnological applications, including DNA and RNA sequencing.[114,115] It is a conical-shaped octameric pore-forming protein composed of eight tightly interconnected monomeric subunits. The pore channel has a short and narrow constriction estimated to be 1.2 nm in diameter and only 0.6 nm thick near the *trans* mouth of the pore. The pore channel is relatively small and narrow compared to the α-HL nanopore, thus improving the spatial resolution of ssDNA sequencing. Thus, the difference in the ionic current of bases through MspA nanopores is significantly greater than that for α-HL.

Despite MspA having a better recognition performance than α-HL, a common challenge in nanopore sequencing is the rapid

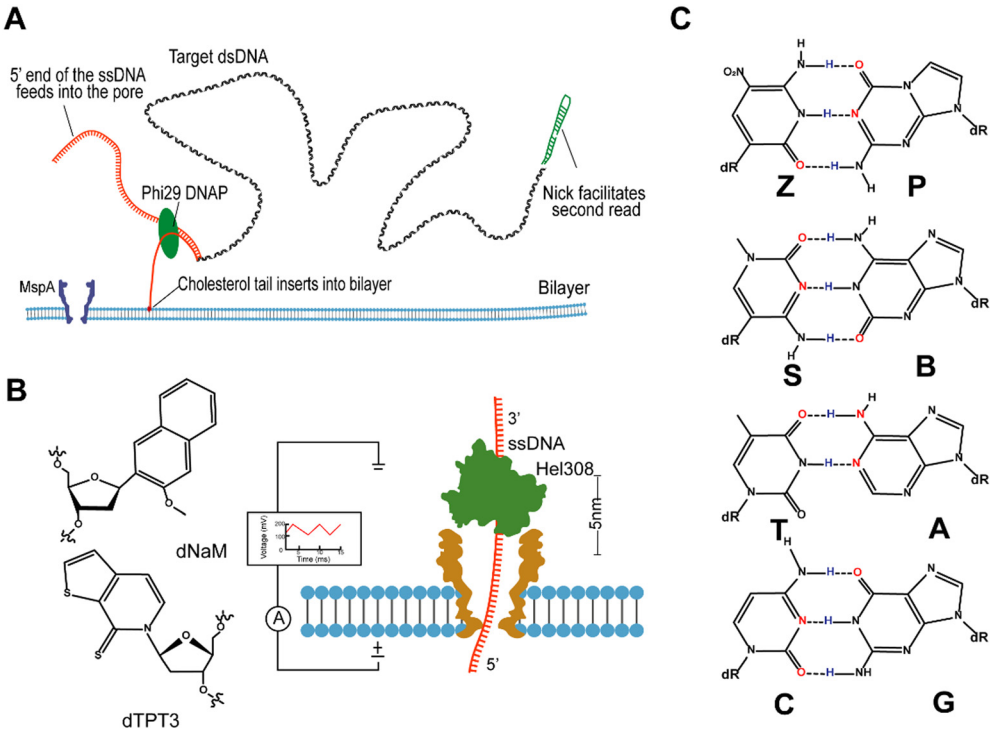

Fig. 9 Controlling translocation speed using motor proteins. (A) Method of adapting dsDNA for long nanopore sequencing. (B) The dNaM-dTPT3 unnatural base pairs and the MspA/Hel308 nanopore system. The Hel308 helicase (green) draws ssDNA through the porin when a variable voltage is applied across the membrane, and meanwhile the current is measured. (C) The genetic system of eight nucleotides consists of four standard DNA bases (A, T, C and G) and four additional nonstandard bases (P, Z, B and S).

DNA translocation,[116] with speeds exceeding 1 nucleotide per microsecond in both α-HL and MspA. Phi29 DNA polymerase was demonstrated to possess superior performance in ratcheting DNA through the nanopore to slow the rate of translocation.[117] An engineered MspA mutant combining phi29 DNA polymerase was elucidated to disassemble ionic currents in single-stranded DNA molecules into individual nucleotide signals.[118] In contrast to previous DNA translocation tests that were poorly controlled, the addition of motor protein reduced the fluctuation in translocation kinetics, thus improving the data quality.[105] The MspA nanopores can accurately sequence the phiX174 genome up to 4500 bases in length by the above methods,[119] as shown in Fig. 9A.

One of the biggest advantages of nanopores over Illumina in terms of data output is single-molecule sequencing of the extended alphabet or the ability to sequence not only natural nucleotides but also chemically modified nucleotides.[120] Ledbetter *et al.* used the nanopore sequencing system based on MspA/Hel308 DNA helicase to evaluate the replication fidelity of six nucleotides consisting of four natural letters, dTPT3, and dNaM by time-varying voltage,[121] as shown in Fig. 9B. The nanopore moved DNA with two steps per nucleotide to produce two distinct ion-current segments. Thus, this nanopore sequencing is sensitive to nucleotide modifications and the unique structure of unnatural nucleotides. Moreover, Thomas *et al.* achieved extended nanopore sequencing of four synthetic DNAs using MspA nanopores to evaluate the signal range (Fig. 9C).[122] The nanopore system of MspA combined with Hel308 DNA helicase could detect and accurately differentiate all eight different nucleotides, and the conductance signals of the four synthetic DNA occupied a larger dynamic range than those of the standard letters. Thus, the application of extended alphabet could further improve the density in DNA data storage.

7. Conclusions and perspectives

Digital revolution has led the society into the era of information explosion. DNA, as a natural carrier of genetic information, has great potential in information storage. DNA-based information storage can overcome the limitation of logic density and storage capacity of current media by devising new encoding schemes. To date, great progress has been achieved to store 200 MB digital data within 13.4 million DNA sequences of 150 nucleotides. Moreover, the technique can reduce the carbon emission produced by the large data storage centers. In addition, the scalability and durability of DNA information storage provide a better option for long-term data storage.

Despite its great potential, DNA information storage still faces many challenges owing to the limitation of the physical technology platform. Storing 1 TB of data requires the synthesis of billions of oligonucleotides. However, the current DNA synthesis throughput could not meet the demands in DNA data storage. When oligonucleotides are synthesized over 100 nt, their

purity and accuracy will gradually decrease, thus limiting the commercialization of DNA information storage. Another major impact of DNA data storage is the inability to synthesize long DNA sequences by chemical routes. Recent research mainly focuses on the synthesis of oligonucleotides of <200 nt, which could be equally cleaved from long DNA fragments for massively parallel DNA synthesis. Additionally, each oligonucleotide must possess addressing information to ensure data reconstruction. Thus, the shorter the oligonucleotide, the more the oligonucleotides required, resulting in a higher proportion of addressing information and a reduction in information density.

Enzymes provide effective tools for addressing these issues in DNA information storage. For instance, TdT is a particularly promising enzyme for *de novo* DNA synthesis, overcoming the contamination of organic reagents and the length limitation in the progress of phosphoramidite chemical synthesis. To date, DNA chains with $\sim$8 kb could be successfully synthesized *via* the TdT-based strategy. The superior synthesis length and speed are far beyond the reach of phosphoramidite chemistry synthesis. The scalability of DNA polymerase to recognize unnatural nucleic acid substrates provides new tools for DNA steganography and cryptography. Additionally, the exploration of key proteins in nanopores has the potential to build DNA sequencing technology of natural or unnatural nucleic acids for information decode. Meanwhile, DNA polymerase selectively amplifies target DNA fragments by designing primers, and thus the DNA polymerase-based PCR technology facilitates the data retrieval from large scale data pools. Besides, integrase and CRISPR-Cas nuclease enable the *in vivo* rewriting and erasing of stored data.

Although these enzymes showed great potential in DNA information storage, there are still some issues to be solved. The main challenge in applying TdT to the programmed DNA synthesis is to control the ordered polymerization of nucleotides. Controllable strategies for polymerization should be established *via* the management of reaction steps such as the degradation of substrates and the release of divalent ion cofactors. In addition, DNA polymerase with high fidelity is an efficient low-error tool to expand the data storage capacity and preservation stability in natural or novel DNA storage architectures. Moreover, the resolution of single nucleobase and the stability of long reads are important directions for future nanopore sequencing technology. Improving the reading accuracy of base sequencing technologies is also an important direction to facilitate the data steganography in DNA information storage. The challenges faced by each step in DNA data storage are interconnected. Thus, it is crucial to integrate the enzymes that play key roles in the different processes of DNA information storage. The aim is to ensure these enzymes with suitable efficiency, boost their synergistic effects, and prevent any adverse impacts on their activity. Directed evolution and *de novo* design of proteins have been demonstrated to be effective strategies for engineering enzymes with desired properties or improved functionality. Deep learning and artificial intelligence methods trained on large scale sequence and structure datasets provide researchers with powerful assistance in "writing" proteins from scratch and creating proteins with entirely new shapes and molecular functions. We believe that the enzymes could be tailored to promote the high-throughput automated DNA data storage, thereby providing a robust and scalable platform to fulfill the industrial requirements.

Author contributions

All authors contributed to the completion of the manuscript.

Data availability

No primary research results, software or code have been included and no new data were generated or analysed as part of this review.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

This work was supported by the National Key R&D Program of China (2020YFA0907003), the National Natural Science Foundation of China (U24A20365, 32271319 and 32471315), the Science and Technology Department of Jilin Province (20240402035GH), the Development and Reform Commission of Jilin Province (2023C015 and 2024C013-8) and the Fundamental Research Funds of the Central Universities, China (2024-JCXK-11).

Notes and references

1 S. Shoda, H. Uyama, J. Kadokawa, S. Kimura and S. Kobayashi, *Chem. Rev.*, 2016, **116**, 2307–2413.
2 J. Zhang, D. Wu, H. Shi, Z. Xing, A. Zhang, Y. Yang and Q. Li, *Process Biochem.*, 2014, **49**, 797–806.
3 J. Yang, Y. Liu, X. Liang, Y. Yang and Q. Li, *Macromol. Biosci.*, 2018, **18**, 1800131.
4 R. Li, W. Kong and Z. An, *Angew. Chem., Int. Ed.*, 2022, **61**, e202202033.
5 Y. Yang, Y. Yu, Y. Zhang, C. Liu, W. Shi and Q. Li, *Process Biochem.*, 2011, **46**, 1900–1908.
6 Y. Yu, D. Wu, C. Liu, Z. Zhao, Y. Yang and Q. Li, *Process Biochem.*, 2012, **47**, 1027–1036.
7 Y. Yang, J. Zhang, D. Wu, Z. Xing, Y. Zhou, W. Shi and Q. Li, *Biotechnol. Adv.*, 2014, **32**, 642–651.
8 R. Li, S. Zhang, Q. Li, G. G. Qiao and Z. An, *Angew. Chem., Int. Ed.*, 2022, **61**, e202213396.
9 K. Muthusamy, K. Lalitha, Y. S. Prasad, A. Thamizhanban, V. Sridharan, C. U. Maheswari and S. Nagarajan, *ChemSusChem*, 2018, **11**, 2453–2463.
10 A. Das, S. Ghosh, A. Mishra, A. Som, V. B. Banakar, S. S. Agasti and S. J. George, *J. Am. Chem. Soc.*, 2024, **146**, 14844–14855.
11 A. E. Fazary, Y. H. Ju and H. S. M. Abd-Rabboh, *Int. J. Biol. Macromol.*, 2017, **101**, 862–881.
12 J. Lu, P. Hu, L. Cao, Z. Wei, F. Xiao, Z. Chen, Y. Li and L. Tian, *Angew. Chem., Int. Ed.*, 2021, **60**, 5377–5385.
13 X. Wang, C. C. Wong, H. Chen, K. Fu, L. Shi, H. Su, S. Guo, H. Gou, X. Hu, L. Zhang, J. Ji and J. Yu, *Cell Rep.*, 2023, **42**, 112279.
14 Y. Hao, Q. Li, C. Fan and F. Wang, *Small Struct.*, 2021, **2**, 2000046.
15 G. M. Church, Y. Gao and S. Kosuri, *Science*, 2012, **337**, 1628.

16 R. N. Grass, R. Heckel, M. Puddu, D. Paunescu and W. J. Stark, *Angew. Chem., Int. Ed.*, 2015, **54**, 2552–2555.
17 L. Organick, Y. J. Chen, S. D. Ang, R. Lopez, X. Liu, K. Strauss and L. Ceze, *Nat. Commun.*, 2020, **11**, 616–623.
18 K. Goda and M. Kitsuregawa, *Proc. IEEE*, 2012, **100**, 1433–1440.
19 D. Panda, K. A. Molla, M. J. Baig, A. Swain, D. Behera and M. Dash, *3 Biotech*, 2018, **8**, 239.
20 J. D. Robin, A. T. Ludlow, R. LaRanger, W. E. Wright and J. W. Shay, *Sci. Rep.*, 2016, **6**, 24067.
21 V. Zhirnov, R. M. Zadegan, G. S. Sandhu, G. M. Church and W. L. Hughes, *Nat. Mater.*, 2016, **15**, 366–370.
22 S. L. Beaucage and M. H. Caruthers, *Tetrahedron Lett.*, 1981, **22**, 1859–1862.
23 S. P. Adams, K. S. Kavka, E. J. Wykes, S. B. Holder and G. R. Galluppi, *J. Am. Chem. Soc.*, 1983, **105**, 661–663.
24 H. Tsunoda, T. Kudo, A. Ohkubo, K. Seio and M. Sekine, *Molecules*, 2010, **15**, 7509–7531.
25 G. Roelfes, *Mol. BioSyst.*, 2007, **3**, 126–135.
26 P. Knyphausen, L. Lindenburg and F. Hollfelder, *Trends Biotechnol.*, 2021, **39**, 861–865.
27 C. Xu, B. Ma, X. Dong, L. Lei, Q. Hao, C. Zhao and H. Liu, *ACS Appl. Mater. Interfaces*, 2023, **15**, 24097–24108.
28 K. Hoff, M. Halpain, G. Garbagnati, J. S. Edwards and W. Zhou, *ACS Synth. Biol.*, 2020, **9**, 283–293.
29 R. U. Sheth, S. S. Yim, F. L. Wu and H. H. Wang, *Science*, 2017, **358**, 1457–1461.
30 S. L. Shipman, J. Nivala, J. D. Macklis and G. M. Church, *Nature*, 2017, **547**, 345–349.
31 P. P. Pillai, S. Reisewitz, H. Schroeder and C. M. Niemeyer, *Small*, 2010, **6**, 2130–2134.
32 C. Plesa, A. M. Sidore, N. B. Lubock, D. Zhang and S. Kosuri, *Science*, 2018, **359**, 343–347.
33 S. Fan, D. Wang, J. Cheng, Y. Liu, T. Luo, D. Cui, Y. Ke and J. Song, *Angew. Chem., Int. Ed.*, 2020, **59**, 12991–12997.
34 R. Adee and H. Mouratidis, *Sensors*, 2022, **22**, 1109.
35 S. Y. Li, J. K. Liu, G. P. Zhao and J. Wang, *ACS Synth. Biol.*, 2018, **7**, 1174–1178.
36 D. Whitaker and M. W. Powner, *Nat. Chem.*, 2022, **14**, 766–774.
37 E. J. Yik, V. A. Maola and J. C. Chaput, *Methods Enzymol.*, 2023, **691**, 29–59.
38 N. Wiener, *U.S. News World Rep.*, 1964, **56**, 84–86.
39 M. S. Neiman, *Radiotekhnika*, 1965, **6**, 1–8.
40 Y. Erlich and D. Zielinski, *Science*, 2017, **355**, 950–954.
41 M. Puddu, W. J. Stark and R. N. Grass, *Adv. Healthcare Mater.*, 2015, **4**, 1332–1338.
42 A. C. Pease, D. Solas, E. J. Sullivan, M. T. Cronin, C. P. Holmes and S. P. Fodor, *Proc. Natl. Acad. Sci. U. S. A.*, 1994, **91**, 5022–5026.
43 J. Loc'h, S. Rosario and M. Delarue, *Structure*, 2016, **24**, 1452–1463.
44 L. F. Song, Z. H. Deng, Z. Y. Gong, L. L. Li and B. Z. Li, *Front. Bioeng. Biotechnol.*, 2021, **9**, 689797.
45 L. Ceze, J. Nivala and K. Strauss, *Nat. Rev. Genet.*, 2019, **20**, 456–466.
46 M. Hao, H. Qiao, Y. Gao, Z. Wang, X. Qiao, X. Chen and H. Qi, *Commun. Biol.*, 2020, **3**, 416.
47 J. Bonnet, P. Subsoontorn and D. Endy, *Proc. Natl. Acad. Sci. U. S. A.*, 2012, **109**, 8884–8889.
48 C. Winston, L. Organick, D. Ward, L. Ceze, K. Strauss and Y. J. Chen, *ACS Synth. Biol.*, 2022, **11**, 1727–1734.
49 K. N. Lin, K. Volkel, J. M. Tuck and A. J. Keung, *Nat. Commun.*, 2020, **11**, 2981.
50 S. M. Yazdi, Y. Yuan, J. Ma, H. Zhao and O. Milenkovic, *Sci. Rep.*, 2015, **5**, 14138.
51 A. Zee, D. Deng, M. Adams, K. D. Schimke, R. Corbett-Detig, S. L. Russell, X. Zhang, R. J. Schmitz and C. Vollmers, *Genome Res.*, 2022, **32**, 2092–2106.
52 R. Heckel, G. Mikutis and R. N. Grass, *Sci. Rep.*, 2019, **9**, 9663.
53 M. H. Caruthers, *Biochem. Soc. Trans.*, 2011, **39**, 575–580.
54 B. I. Andrews, F. D. Antia, S. B. Brueggemeier, L. J. Diorazio, S. G. Koenig, M. E. Kopach, H. Lee, M. Olbrich and A. L. Watson, *J. Org. Chem.*, 2021, **86**, 49–61.
55 R. Obexer, M. Nassir, E. R. Moody, P. S. Baran and S. L. Lovelock, *Science*, 2024, **384**, eadl4015.
56 S. Kosuri and G. M. Church, *Nat. Methods*, 2014, **11**, 499–507.
57 R. A. Hughes and A. D. Ellington, *Cold Spring Harbor Perspect. Biol.*, 2017, **9**, a023812.
58 M. Yu, X. Tang, Z. Li, W. Wang, S. Wang, M. Li, Q. Yu, S. Xie, X. Zuo and C. Chen, *Chem. Soc. Rev.*, 2024, **53**, 4463–4489.
59 H. Li, Y. Huang, Z. Wei, W. Wang, Z. Yang, Z. Liang and Z. Li, *Sci. Rep.*, 2019, **9**, 5058.
60 N. Goldman, P. Bertone, S. Chen, C. Dessimoz, E. M. LeProust, B. Sipos and E. Birney, *Nature*, 2013, **494**, 77–80.
61 M. H. Caruthers, *J. Biol. Chem.*, 2013, **288**, 1420–1427.
62 Q. Zhang, K. Xia, M. Jiang, Q. Li, W. Chen, M. Han, W. Li, R. Ke, F. Wang, Y. Zhao, Y. Liu, C. Fan and H. Gu, *Angew. Chem., Int. Ed.*, 2023, **62**, e202212011.
63 I. Sarac and M. Hollenstein, *ChemBioChem*, 2019, **20**, 860–871.
64 M. Delarue, J. B. Boulé, J. Lescar, N. Expert-Bezançon, N. Jourdan, N. Sukumar, F. Rougeon and C. Papanicolaou, *EMBO J.*, 2002, **21**, 427–539.
65 F. J. Bollum, *J. Biol. Chem.*, 1959, **234**, 2733–2734.
66 F. J. Bollum, *J. Biol. Chem.*, 1962, **237**, 1945–1949.
67 S. Palluk, D. H. Arlow, T. de Rond, S. Barthel, J. S. Kang, R. Bector, H. M. Baghdassarian, A. N. Truong, P. W. Kim, A. K. Singh, N. J. Hillson and J. D. Keasling, *Nat. Biotechnol.*, 2018, **36**, 645–650.
68 X. Lu, J. Li, C. Li, Q. Lou, K. Peng, B. Cai, Y. Liu, Y. Yao, L. Lu, Z. Tian, H. Ma, W. Wang, J. Cheng, X. Guo, H. Jiang and Y. Ma, *ACS Catal.*, 2022, **12**, 2988–2997.
69 H. H. Lee, R. Kalhor, N. Goela, J. Bolot and G. M. Church, *Nat. Commun.*, 2019, **10**, 2383.
70 H. Lee, D. J. Wiegand, K. Griswold, S. Punthambaker, H. Chun, R. E. Kohman and G. M. Church, *Nat. Commun.*, 2020, **11**, 5246.
71 J. C. Chaput and P. Herdewijn, *Angew. Chem., Int. Ed.*, 2019, **58**, 11570–11572.
72 R. Shroff, J. W. Ellefson, S. S. Wang, A. A. Boulgakov, R. A. Hughes and A. D. Ellington, *ACS Synth. Biol.*, 2022, **11**, 554–561.
73 K. Yang, C. M. McCloskey and J. C. Chaput, *ACS Synth. Biol.*, 2020, **9**, 2936–2942.
74 H. Hoshino, Y. Kasahara, M. Kuwahara and S. Obika, *J. Am. Chem. Soc.*, 2020, **142**, 21530–21537.
75 A. M. Kabza and J. T. Sczepanski, *Molecules*, 2020, **25**, 947.
76 C. Fan, Q. Deng and T. F. Zhu, *Nat. Biotechnol.*, 2021, **39**, 1548–1555.
77 S. K. Tabatabaei, B. Wang, N. Athreya, B. Enghiad, A. G. Hernandez, C. J. Fields, J. P. Leburton, D. Soloveichik, H. Zhao and O. Milenkovic, *Nat. Commun.*, 2020, **11**, 1742–1752.
78 T. Knebelsberger and I. Stöger, *Methods Mol. Biol.*, 2012, **858**, 311–338.
79 T. J. Anchordoquy and M. C. Molina, *Cell Preserv. Technol.*, 2007, **5**, 180.
80 Q. Liu, Y. Wei, Z. Wang, D. P. Song, J. Cui and H. Qi, *Small Methods*, 2023, **7**, e2201610.
81 J. Koch, S. Gantenbein, K. Masania, W. J. Stark, Y. Erlich and R. N. Grass, *Nat. Biotechnol.*, 2020, **38**, 39–43.
82 P. L. Antkowiak, J. Koch, B. H. Nguyen, W. J. Stark, K. Strauss, L. Ceze and R. N. Grass, *Small*, 2022, **18**, e2107381.
83 W. Chen, M. Han, J. Zhou, Q. Ge, P. Wang, X. Zhang, S. Zhu, L. Song and Y. Yuan, *Natl. Sci. Rev.*, 2021, **8**, nwab28.
84 N. Roquet, A. P. Soleimany, A. C. Ferris, S. Aaronson and T. K. Lu, *Science*, 2016, **353**, aad8559.
85 C. A. Merrick, J. Zhao and S. J. Rosser, *ACS Synth. Biol.*, 2018, **7**, 299–310.
86 F. Sun, Y. Dong, M. Ni, Z. Ping, Y. Sun, Q. Ouyang and L. Qian, *Adv. Sci.*, 2023, **10**, 2206201.
87 Z. Hou, W. Qiang, X. Wang, X. Chen, X. Hu, X. Han, W. Shen, B. Zhang, P. Xing, W. Shi, J. Dai, X. Huang and G. Zhao, *Adv. Sci.*, 2024, **11**, e2305921.
88 F. Farzadfard, N. Gharaei, R. J. Citorik and T. K. Lu, *Cell Syst.*, 2021, **12**, 860–872.
89 S. S. Yim, R. M. McBee, A. M. Song, Y. Huang, R. U. Sheth and H. H. Wang, *Nat. Chem. Biol.*, 2021, **17**, 246–253.
90 Y. Liu, Y. Ren, J. Li, F. Wang, F. Wang, C. Ma, D. Chen, X. Jiang, C. Fan, H. Zhang and K. Liu, *Sci. Adv.*, 2022, **8**, eabo7415.
91 A. Sadremomtaz, R. F. Glass, J. E. Guerrero, D. R. LaJeunesse, E. A. Josephs and R. Zadegan, *Nat. Commun.*, 2023, **14**, 6472–6482.
92 K. N. Lin, K. Volkel, C. Cao, P. W. Hook, R. E. Polak, A. S. Clark, A. San Miguel, W. Timp, J. M. Tuck, O. D. Velev and A. J. Keung, *Nat. Nanotechnol.*, 2024, **19**, 1654–1664.
93 L. Organick, S. D. Ang, Y. J. Chen, R. Lopez, S. Yekhanin, K. Makarychev, M. Z. Racz, G. Kamath, P. Gopalan, B. Nguyen,

C. N. Takahashi, S. Newman, H. Y. Parker, C. Rashtchian, K. Stewart, G. Gupta, R. Carlson, J. Mulligan, D. Carmean, G. Seelig, L. Ceze and K. Strauss, *Nat. Biotechnol.*, 2018, **36**, 242–248.
94 K. J. Tomek, K. Volkel, A. Simpson, A. G. Hass, E. W. Indermaur, J. M. Tuck and A. J. Keung, *ACS Synth. Biol.*, 2019, **8**, 1241–1248.
95 J. L. Banal, T. R. Shepherd, J. Berleant, H. Huang, M. Reyes, C. M. Ackerman, P. C. Blainey and M. Bathe, *Nat. Mater.*, 2021, **20**, 1272–1280.
96 B. W. A. Bögels, B. H. Nguyen, D. Ward, L. Gascoigne, D. P. Schrijver, A. Makri Pistikou, A. Joesaar, S. Yang, I. K. Voets, W. J. M. Mulder, A. Phillips, S. Mann, G. Seelig, K. Strauss, Y. Chen and T. F. A. de Greef, *Nat. Nanotechnol.*, 2023, **18**, 912–921.
97 J. Kim, J. H. Bae, M. Baym and D. Y. Zhang, *Nat. Commun.*, 2020, **11**, 5008.
98 K. J. Tomek, K. Volkel, E. W. Indermaur, J. M. Tuck and A. J. Keung, *Nat. Commun.*, 2021, **12**, 3518.
99 S. L. Amarasinghe, S. Su, X. Dong, L. Zappia, M. E. Ritchie and Q. Gouil, *Genome Biol.*, 2020, **21**, 30.
100 A. L. McNaughton, H. E. Roberts, D. Bonsall, M. de Cesare, J. Mokaya, S. F. Lumley, T. Golubchik, P. Piazza, J. B. Martin, C. de Lara, A. Brown, M. A. Ansari, R. Bowden, E. Barnes and P. C. Matthews, *Sci. Rep.*, 2019, **9**, 7081.
101 Y. Choi, T. Ryu, A. C. Lee, H. Choi, H. Lee, J. Park, S. H. Song, S. Kim, H. Kim, W. Park and S. Kwon, *Sci. Rep.*, 2019, **9**, 6582.
102 F. A. Ferreira, K. Helmersen, T. Visnovska, S. B. Jorgensen and H. V. Aamot, *Microb. Genomics*, 2021, **7**, 000557.
103 T. Ding, J. Yang, V. Pan, N. Zhao, Z. Lu, Y. Ke and C. Zhang, *Nucleic Acids Res.*, 2020, **48**, 2791–2806.
104 R. F. Purnell and J. J. Schmidt, *ACS Nano*, 2009, **3**, 2533–2538.
105 Y. Wang, Y. Zhao, A. Bollas, Y. Wang and K. F. Au, *Nat. Biotechnol.*, 2021, **39**, 1348–1365.
106 J. Wilson, K. Sarthak, W. Si, L. Gao and A. Aksimentiev, *ACS Sens.*, 2019, **4**, 634–644.
107 L. Song, M. R. Hobaugh, C. Shustak, S. Cheley, H. Bayley and J. E. Gouaux, *Science*, 1996, **274**, 1859–1866.
108 S. L. Cockroft, J. Chu, M. Amorin and M. R. Ghadiri, *J. Am. Chem. Soc.*, 2008, **130**, 818–820.
109 D. Stoddart, A. J. Heron, E. Mikhailova, G. Maglia and H. Bayley, *Proc. Natl. Acad. Sci. U. S. A.*, 2009, **106**, 7702–7707.
110 D. Stoddart, A. J. Heron, J. Klingelhoefer, E. Mikhailova, G. Maglia and H. Bayley, *Nano Lett.*, 2010, **10**, 3633–3637.
111 M. Ayub, D. Stoddart and H. Bayley, *ACS Nano*, 2015, **9**, 7895–7903.
112 P. B. Stranges, M. Palla, S. Kalachikov, J. Nivala, M. Dorwart, A. Trans, S. Kumar, M. Porel, M. Chien, C. Tao, I. Morozova, Z. Li, S. Shi, A. Aberra, C. Arnold, A. Yang, A. Aguirre, E. T. Harada, D. Korenblum, J. Pollard, A. Bhat, D. Gremyachinskiy, A. Bibillo, R. Chen, R. Davis, J. J. Russo, C. W. Fuller, S. Roever, J. Ju and G. M. Church, *Proc. Natl. Acad. Sci. U. S. A.*, 2016, **113**, 6749–6756.
113 M. Yu, W. Si, T. Zeng, C. Chen, X. Lin, Z. Ji, F. Guo, Y. Li, J. Sha and Y. Dong, *J. Phys. Chem. Lett.*, 2021, **12**, 9132–9141.
114 I. M. Derrington, T. Z. Butler, M. D. Collins, E. Manrao, M. Pavlenok, M. Niederweis and J. H. Gundlach, *Proc. Natl. Acad. Sci. U. S. A.*, 2010, **107**, 16060–16065.
115 M. Faller, M. Niederweis and G. E. Schulz, *Science*, 2004, **303**, 1189–1192.
116 Z. L. Hu, M. Z. Huo, Y. L. Ying and Y. T. Long, *Angew. Chem., Int. Ed.*, 2021, **60**, 14738–14749.
117 G. M. Cherf, K. R. Lieberman, H. Rashid, C. E. Lam, K. Karplus and M. Akeson, *Nat. Biotechnol.*, 2012, **30**, 344–348.
118 E. A. Manrao, I. M. Derrington, A. H. Laszlo, K. W. Langford, M. K. Hopper, N. Gillgren, M. Pavlenok, M. Niederweis and J. H. Gundlach, *Nat. Biotechnol.*, 2012, **30**, 349–353.
119 A. H. Laszlo, I. M. Derrington, B. C. Ross, H. Brinkerhoff, A. Adey, I. C. Nova, I. M. Craig, K. W. Langford, J. M. Samson, R. Daza, K. Doering, J. Shendure and J. H. Gundlach, *Nat. Biotechnol.*, 2014, **32**, 829–833.
120 G. Hu, H. Yan, G. Xi, Z. Gao, Z. Wu, Z. Lu and J. Tu, *IET Nanobiotechnol.*, 2023, **17**, 257–268.
121 M. P. Ledbetter, J. M. Craig, R. J. Karadeema, M. T. Noakes, H. C. Kim, S. J. Abell, J. R. Huang, B. A. Anderson, R. Krishnamurthy, J. H. Gundlach and P. E. Romesberg, *J. Am. Chem. Soc.*, 2020, **142**, 2110–2114.
122 C. A. Thomas, J. M. Craig, S. Hoshika, H. Brinkerhoff, J. R. Huang, S. J. Abell, H. C. Kim, M. C. Franzi, J. D. Carrasco, H. J. Kim, D. C. Smith, J. H. Gundlach, S. A. Benner and A. H. Laszlo, *J. Am. Chem. Soc.*, 2023, **145**, 8560–8568.