



# Large-Scale *de novo* Oligonucleotide Synthesis for Whole-Genome Synthesis and Data Storage: Challenges and Opportunities

Li-Fu Song<sup>1,2†</sup>, Zheng-Hua Deng<sup>1,2†</sup>, Zi-Yi Gong<sup>1,2</sup>, Lu-Lu Li<sup>3</sup> and Bing-Zhi Li<sup>1,2\*</sup>

<sup>1</sup> Frontiers Science Center for Synthetic Biology and Key Laboratory of Systems Bioengineering (Ministry of Education), Tianjin University, Tianjin, China, <sup>2</sup> School of Chemical Engineering and Technology, Tianjin University, Tianjin, China, <sup>3</sup> LC-BIO Technologies Co., Ltd., Hangzhou, China

## OPEN ACCESS

### Edited by:

Yuan Lu,  
Tsinghua University, China

### Reviewed by:

Shuobo Shi,  
Beijing University of Chemical  
Technology, China  
Shigeki Sasaki,  
Nagasaki International University,  
Japan

### \*Correspondence:

Bing-Zhi Li  
bzli@tju.edu.cn

† These authors have contributed  
equally to this work

### Specialty section:

This article was submitted to  
Synthetic Biology,  
a section of the journal  
Frontiers in Bioengineering and  
Biotechnology

**Received:** 01 April 2021

**Accepted:** 27 May 2021

**Published:** 22 June 2021

### Citation:

Song L-F, Deng Z-H, Gong Z-Y,  
Li L-L and Li B-Z (2021) Large-Scale  
*de novo* Oligonucleotide Synthesis  
for Whole-Genome Synthesis  
and Data Storage: Challenges  
and Opportunities.  
Front. Bioeng. Biotechnol. 9:689797.  
doi: 10.3389/fbioe.2021.689797

Over the past decades, remarkable progress on phosphoramidite chemistry-based large-scale *de novo* oligonucleotide synthesis has been achieved, enabling numerous novel and exciting applications. Among them, *de novo* genome synthesis and DNA data storage are striking. However, to make these two applications more practical, the synthesis length, speed, cost, and throughput require vast improvements, which is a challenge to be met by the phosphoramidite chemistry. Harnessing the power of enzymes, the recently emerged enzymatic methods provide a competitive route to overcome this challenge. In this review, we first summarize the status of large-scale oligonucleotide synthesis technologies including the basic methodology and large-scale synthesis approaches, with special focus on the emerging enzymatic methods. Afterward, we discuss the opportunities and challenges of large-scale oligonucleotide synthesis on *de novo* genome synthesis and DNA data storage respectively.

**Keywords:** oligonucleotide synthesis, DNA synthesis, synthetic biology, whole-genome synthesis, data storage in DNA, DNA-based data storage

## INTRODUCTION

*De novo* oligonucleotide synthesis (oligo synthesis) is the synthesis of defined sequences of single-strand nucleic acids. Its early efforts could be found out in the 1950s, followed by large-scale automation in the 1980s (Kosuri and Church, 2014). Starting in the early 1990s, the microchip-based large-scale oligo synthesis methods were developed (Fodor et al., 1991; Pease et al., 1994; Singh-Gasson et al., 1999; Gao et al., 2001; Ghindilis et al., 2007; Ma et al., 2012; Kosuri and Church, 2014). To clarify, “large-scale oligo synthesis” here refers to the parallel synthesis of massive oligos with various sequence contents. The massive and cheap oligos produced by large-scale oligo synthesis have a wide range of applications. For examples, the oligos can be used as molecular tools for the construction of large cell populations with various genotypes (Kosuri and Church, 2014). It also can be used for understanding and engineering of regulatory elements, proteins, genetic networks and metabolic pathways (Kosuri and Church, 2014). Due to the wide application potentials, large-scale oligo synthesis have greatly helped with the life science studies in many aspects, e.g., synthetic biology, protein engineering, genome engineering, metabolic engineering and so on (Ma et al., 2012; Kosuri and Church, 2014; Casini et al., 2015). Among these applications, *de novo* whole-genome synthesis and DNA data storage are drawing more

and more attention (Cello et al., 2002; Church et al., 2012; Gibson, 2014; Kosuri and Church, 2014; Shen et al., 2017; Wu et al., 2017; Xie et al., 2017; Zhang et al., 2017; Kohman et al., 2018; Lee et al., 2018). Despite the remarkable progress that large-scale oligo synthesis has achieved in the past decades, substantial further improvements are still highly demanded for emerging wide applications. However, due to the natural limitations of chemical reactions, it is a challenge to further improve the phosphoramidite chemistry-based methods in terms of accuracy and length dramatically. The recently emerged enzymatic methods with advantages of mild conditions, fast coupling, no hazardous waste generation, show potentials in the realization of next-generation oligo synthesis technologies (Hoff et al., 2019; Lee et al., 2019). In this mini-review, we discuss the current status of large-scale oligo synthesis technologies with special focus on the recent innovations of the enzymatic methods. Then we discuss its applications in *de novo* whole-genome synthesis and DNA-based data storage. The different development routes to meet the future requirements of the two distinct applications will be discussed respectively.

## CURRENT STATUS OF PHOSPHORAMIDITE CHEMISTRY-BASED OLIGO SYNTHESIS

In the past few decades, the solid-phase phosphoramidite method has been the primary choice for most commercial oligo synthesizers (Roy and Caruthers, 2013; Kosuri and Church, 2014). This method was developed by Beaucage and Caruthers in 1980s (Beaucage and Caruthers, 1981; Caruthers, 1985). The four-step synthesis cycle of this method is illustrated in **Figure 1**. First, the dimethoxytrityl group (DMT) of the nucleotide immobilized on the solid phase is detached by acid catalysis. Second, the 3' hydroxyl group of the nucleotide to be added in is activated by phosphoramidite and mixed with the tetrazole activator, and the obtained nucleoside-phosphite activator is 5' hydroxyl-activated nucleotides undergo condensation. Third, a small number of 5' hydroxyl-activated nucleotides not involved in the condensation reaction are prevented from participating in the reaction by acetylation. Fourth, oxidation of trivalent phosphotriester to pentavalent phosphotriester using an iodine solution. After addition of all nucleosides in series from 3' to 5', the obtained oligo is released. Currently, the length of oligos synthesized by the phosphoramidite chemistry-based methods is limited within 200 nucleotides in general and cannot exceed 300 nucleotides theoretically (Palluk et al., 2018). For construction of long DNAs, DNA assembly is required to assemble the short oligos into long DNAs (Casini et al., 2015).

### Column-Based Oligo Synthesis

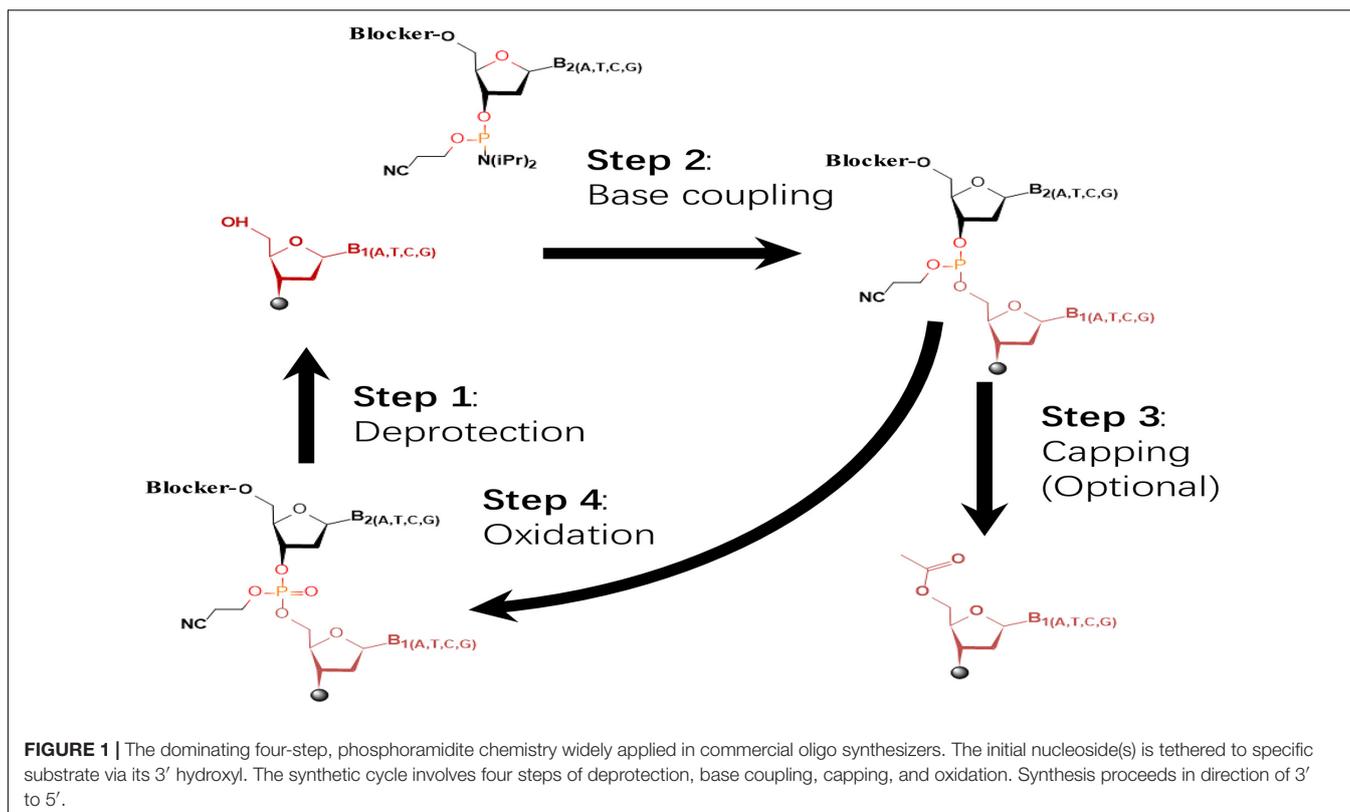
Column-based oligo synthesis is the first-generation technology based on phosphoramidite chemistry. In this technology, oligo synthesis is carried out separately in the columns. The reagents are pumped through the columns, enabling the iterative addition of nucleotides in a programmable way. Over the past decades, advances in materials, automation, procedure, and purification

have led to the synthesis of 100 nt at cost of \$0.05 to \$0.15 per nt, with an error rate of 1/600 or less (Kosuri and Church, 2014). The commercially available column-based oligo synthesizer can synthesize 96–768 oligonucleotides each containing 10 nmol to 2  $\mu$ mol at the same time. Column-synthesized oligonucleotides have been used as modules for DNA constructions by assembly methods in previous studies (Ma et al., 2012). However, column-based oligonucleotide synthesis cannot satisfy the requirements of large-scale DNA synthesis in the era of synthetic biology, due to the limitations of low throughput and high cost (Ma et al., 2012).

### Microchip-Based Oligo Synthesis

Starting in the 1990s, oligo synthesis gradually developed into a higher throughput manner by parallel synthesis on a silica surface (Fodor et al., 1991; Pease et al., 1994). This microchip-based (also called 'microarray-based') large-scale oligo synthesis technologies, provides an inexpensive source of oligo building blocks for various applications (Ma et al., 2012; Kosuri and Church, 2014). Surpassing the column-based synthesis in terms of throughput and cost, microchip-based oligo synthesis has received wide attention. The commercially available microchip-based synthesizers are based on principle of phosphoramidite chemistry with slight modifications. The major differences between various microchip-based methods are the different mechanisms that applied in steps of deprotection and base coupling (Ma et al., 2012). At present, there are light control, electrochemical, and inkjet printing methods.

The light control methods use light to control the deprotection process, including photolabile 5' protecting groups (PPG, NimbleGen/Affymetrix) and photo-generated acid deprotection (LC Science). The basic strategy for PPG-based synthesis is to illuminate the surface of a PPG modified solid support with a mask to generate free hydroxyl groups for the coupling of the 3'-O-phosphoramidite activated deoxynucleoside. After coupling and capping, the surface of the solid support is illuminated through a new mask, exposing the next active hydroxyl group and coupling with the second 3'-O-phosphoramid-activated deoxynucleoside (Fodor et al., 1991; Pease et al., 1994). The expensive masks applied and long synthesis time limits its applications. Later, Singh-Gasson et al. (1999) reported a mask less array synthesis method in which the expensive masks were replaced with Digital Micromirror Device (DMD). The DMD forms an ultraviolet image on the surface of a glass support, enabling selective deprotection. This DMD-based method avoids the need for photolithographic masks, reducing the cost and time (Singh-Gasson et al., 1999). Differently, Gao et al. (2001) used photo-generated acids (PGAs) to activate deprotection reaction of the 5'-OH group to achieve microchip-based synthesis. The 5'-OH group is formed in the subsequent chain extension reaction and coupled with the introduced monomer (Gao et al., 2001). The electrochemical methods (CombiMatrix/CustomArray) use the adjacent anode and cathode electrodes to produce a deprotected active material, which is coupled to a silicon plate for the synthesis reaction in order to selectively deprotect the DMT protection group to control the synthesis of the desired oligonucleotide in specific positions (Egeland and Southern, 2005; Ghindilis et al., 2007). Compared to the PPG and PGA-based methods, more



precise and complexed semiconductor fabrication is essential. In addition to the methods as described, several other extensions and changes in microfluidics have also been reported, but have not been widely available or commercialized (Ma et al., 2012). The inkjet printing method was developed by Agilent. In this method, a commercial inkjet printer head was utilized to deliver the phosphoramidite monomer to a specific location on a silicon surface, enabling programmable and parallel synthesis of massive oligos.

Although microchip-based oligo synthesis is more prone to errors due to the heterogeneity and edge effects of the microchip, it has enabled high-fidelity synthesis of oligo pools ~300-mer after procedure optimization (Ma et al., 2012; Kosuri and Church, 2014). In general, the costs of microchip-based oligo synthesis are 2–4 orders of magnitude cheaper than the column-based oligo synthesis. The cost per nucleotide is between \$0.00001 and 0.001 (Kosuri and Church, 2014).

## THE EMERGING ENZYMATIC OLIGO SYNTHESIS METHODS

Enzymatic *de novo* synthesis of oligonucleotide can be dated back to 1955 (Grunberg-Manago et al., 1955). However, due to technical limitations, the earlier studies on enzymatic methods are quite limited (Grunberg-Manago et al., 1955; Severo Ochoa, 1961; Bollum, 1962; Mackey and Gilham, 1971). Since 2016, several studies proved the feasibility of enzymatic oligo synthesis using template-independent or dependent polymerases

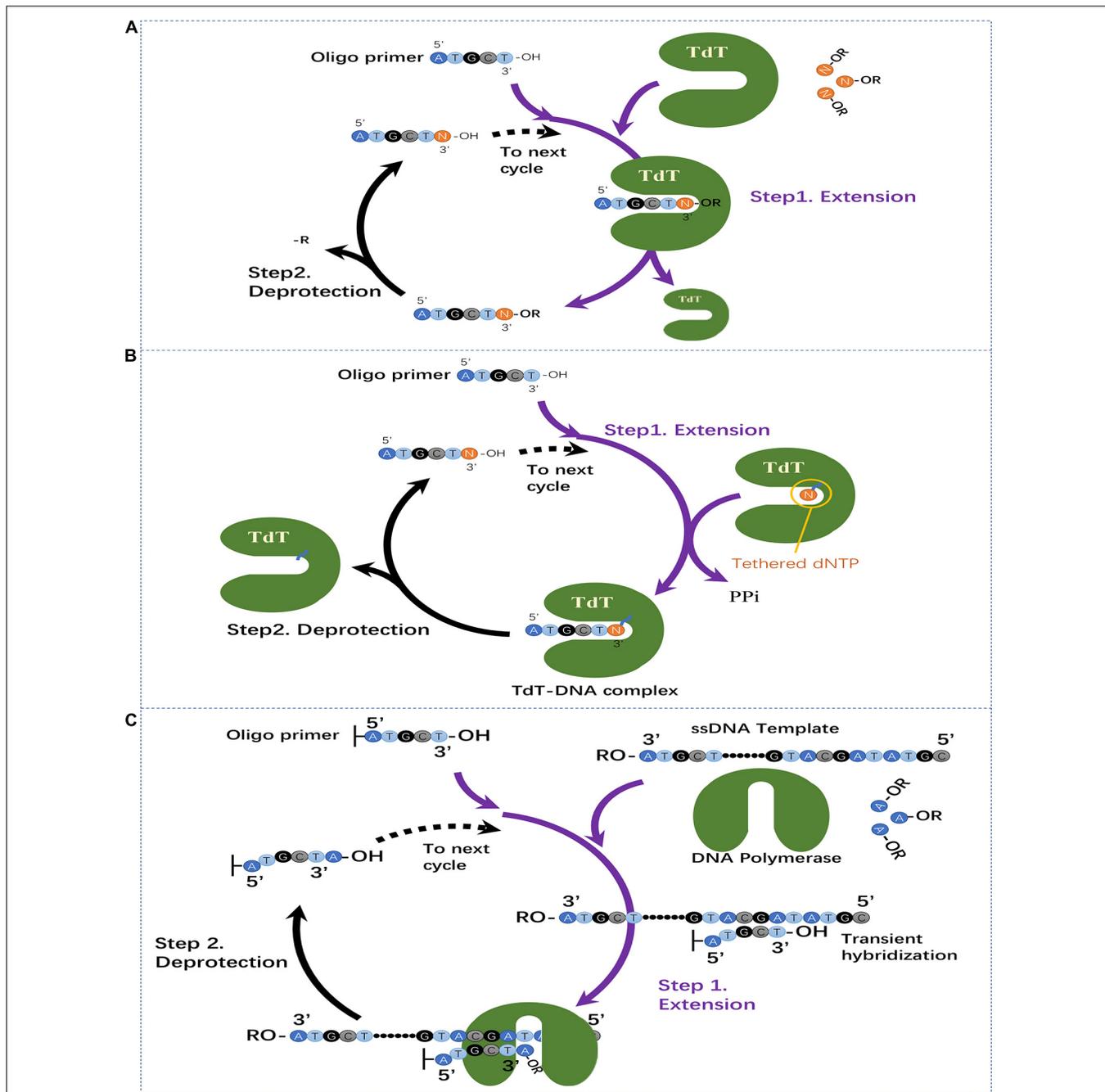
(Mathews et al., 2016, 2017; Palluk et al., 2018; Hoff et al., 2019). Harnessing the power of enzymes, there are several advantages of enzymatic oligo synthesis methods that compared with chemical synthesis: (1) The enzymatic synthesis reaction is carried out under hydrated and mild conditions, coupled with the specificity of the enzyme, the enzymatic synthesis can reduce the formation of by-products and the depurination of DNA and other damages, so that longer oligonucleotides can be synthesized directly; (2) Synthesis can start from natural DNA (i.e., DNA without protecting groups); and (3) Protein engineering (such as directed evolution or rational protein design) can be used to optimize the system, which cannot be achieved using organic chemistry alone (Mathews et al., 2016, 2017; Palluk et al., 2018; Hoff et al., 2019). Here, we review the recent emerged enzymatic oligo synthesis methods that utilizes either template-independent or dependent polymerases.

## Template Independent Polymerase-Based Methods

Utilization of template-independent polymerase for oligo synthesis can be dated back to 1992 (Hyman, 1992). Ramadan et al. (2004) verified that several DNA polymerases, and TdT showed template-independent DNA polymerase activities (Ramadan et al., 2004). The crystal structure of TdT was solved by Delarue et al. (2002). The predominant activity of TdT is addition of deoxynucleotide triphosphates to the 3' end of DNA. This makes TdT an ideal candidate for usage in oligo synthesis (Motea and Berdis, 2010; Palluk et al., 2018). Although the

natural features of TdT are favorable for the implementation of the enzymatic oligo synthesis method, no concrete method using TdT was established until 2017 (Palluk et al., 2018). Early efforts focused on a strategy of using 3'-O-modified nucleotides with blocking groups, which have been successfully applied in DNA sequencing by synthesis strategy (Ruparel et al., 2005; Wu et al., 2007; Bentley et al., 2008; Mathews et al., 2016, 2017).

By developing the methods for synthesis and purification of 3'-O-caged 2'-deoxyribonucleoside triphosphates, Mathews et al. (2016, 2017) demonstrated the feasibility of light-mediated deprotection for enzymatic oligo synthesis using TdT. As shown in **Figure 2A**, the proposed synthesis cycle using TdT and dNTPs with 3'-blocked groups contains two steps. Step 1 – Extension: the solid-phase immobilized oligo primers are mixed with TdT



**FIGURE 2 |** Three recently emerged enzymatic oligo synthesis methods. **(A)** Two-step extension synthesis of DNA oligo using TdT and dNTPs with reversible blocking groups. **(B)** Two-step extension synthesis of DNA oligo using TdT-dNTP conjugates for reversible termination of oligo elongation. **(C)** Template-dependent polymerase mediated oligo synthesis by transient hybridization and chemically blocked substrates.

and specified dNTP with 3'-O-blocking groups, then the TdT will catalyze the coupling of dNTP with the oligo primers. Due to the existence of the 3'-O-blocking groups, the elongation will stop after coupling of one base. Step 2 – Deprotection: the 3'-O-blocking groups of the solid-phase immobilized oligo primers are removed for extension of next base. The coupling time was indicated to be around 60 min and a 4 mer oligonucleotide was synthesized using this method (stepwise yields not reported) (Mathews et al., 2016, 2017). This method suffers from the obstacle that 3'-O-modified nucleotides cannot be efficiently incorporated by TdT since there is almost no room for 3'-OH modifications in the nucleotide position of TdT structure (Palluk et al., 2018). Although it has been proposed that protein engineering with TdT can potentially improve its incorporation efficiency (Efcavitch and Sylvester, 2015), no further efforts have been reported.

To overcome the limitations of TdT with 3'-blocked dNTPs, Palluk et al. (2018) established a novel strategy by using TdT-dNTP conjugates for suspending of oligo elongation in each synthesis cycle. In this strategy, a TdT molecule and a dNTP molecule are conjugated to each other, and the dNTP molecule can be added to the primer by the conjugated TdT. When the bound dNTP is added, the 3' end of the primer is still covalently bound to TdT, blocking the random addition of other TdT-dNTP molecules. This blocking effect is then released by breaking the bond between the TdT and the nucleotide allowing subsequent extension (Palluk et al., 2018). As shown in **Figure 2B**, the synthesis process contains two major steps. Step 1 – Extension: the oligo primer is exposed to excess TdT-dNTP conjugates and the tethered nucleotide is incorporated into the 3' end. The covalently attached conjugate can block the extension of other TdT-dNTP molecules. Step 2 – Deprotection: The cleavage reagent cleaves the bond between the incorporated nucleotide and TdT, and the primer is released for subsequent extension. After optimization, the coupling time of this method was around 10–20 s.

## Template Dependent Polymerase-Based Methods

Until very recently, Hoff et al. (2019) reported a novel enzymatic oligo synthesis method using template-dependent polymerase including DNA polymerases and reverse transcriptase. They achieved the single-base extension using oligo that can instantaneously hybridizes to adjacent strands using as few as two binding bases. When multiple transient hybridizations templating different bases are possible, DNA polymerase and reverse transcriptase can then extend the DNA strand. The oligo sequence to be synthesized then can be controlled by adding the desired base. As shown in **Figure 2C**, specific single strand DNAs (ssDNA) serve as templates and the surface-bound oligo is extended by sequential polymerase-based incorporation of 3'-reversibly blocked nucleotides. Specifically, there are two steps in each extension cycle for the synthesis of arbitrary oligo sequences: Step 1 – Extension: the ssDNA template, polymerase and desired dNTP with 3'-blocking groups are added and the polymerase will extend one base by transient hybridization of the surface-bound oligonucleotide to ssDNA template. Step 2 – Deprotection:

the 3'-blocking groups of surface-bound oligonucleotide is removed for next base extension; A 20-base oligonucleotide was successfully synthesized utilizing this approach with stepwise efficiency more than 98% and coupling time around 1 min.

## Future Developments of Enzymatic Oligo Synthesis

Currently, enzymatic oligo synthesis methods are still in their methodology developing stage. The capacities of currently available enzymatic oligo synthesis methods reported in peer-reviewed articles were summarized in **Table 1**, which includes coupling time, stepwise yield, maximal length achieved and substrate requirements, *etc.* Despite preliminary results from proof-of-concept studies, the coupling time of enzymatic methods was shown to be much shorter than the chemical methods. As preliminary results of proof-of-concept studies, the stepwise yield and maximal length of enzymatic methods are still below the phosphoramidite method. It needs to be clarified that these are results reported in peer-reviewed papers. Several enzymatic oligo synthesis start-up companies, including DNA Script, Molecular Assemblies, Nuclera, Ansa Biotechnologies, Camena Bioscience and Kern Systems also have reported their progress on enzymatic oligo synthesis in public media (Eisenstein, 2020). DNA Script claimed to achieve a step-wise yield of 99.7% and successful synthesis of a 280-mer oligo. Camena even stated that they can produce 300-mer oligo with a step-wise yield greater than 99.9%. Currently, no microchip-based enzymatic *de novo* oligo synthesis has been reported either in peer-reviewed articles or in public media. Nevertheless, the knowledge obtained during optimization of the chemical methods can be very helpful to speed up the development process of microchip-based enzymatic synthesis.

## EMERGING APPLICATIONS

The massive and cheap oligos produced by large-scale oligo synthesis can be used for many purposes, enabling numerous interesting applications. For example, the oligos can be used for the constructions of large cell populations or DNA variations, which are beneficial in engineering regulatory elements, genetic networks, metabolic pathways, and DNA origami (Kosuri and Church, 2014). This list is only growing with the future developments and innovations relevant to this technology. A previous review has discussed these applications in detail (Kosuri and Church, 2014). Here, we focus on the two attractive applications of *de novo* synthesis of whole genomes and DNA-based data storage.

### *De novo* Synthesis of Whole Genomes

“What I cannot create, I do not understand,” as implied by the famous quote by Richard Feynman, our ability to create and build arbitrary DNA constructs can greatly boost our capacity in understanding the mechanisms of biological systems. Genomes are the whole blueprints of all living matter on earth. The progress of whole-genome synthesis technology has fundamental impacts on life science in many aspects (Cello et al., 2002; Gibson, 2014; Shen et al., 2017; Wu et al., 2017; Xie et al., 2017;

**TABLE 1** | Current capacities of enzymatic oligo synthesis methods reported in literatures in comparison with the dominating phosphoramidite chemistry methods.

	Coupling time (per base)	Step-wise Yield	Maximal length achieved	Substrates	Enzyme	Template requirements
Reversible terminator	60 min	Not available	4 mer	Purified 3'-blocked dNTP	TdT	No
TdT-dNTP conjugator	10~20 s	97.7%	10 mer	Purified TdT-dNTP complex	TdT	No
Transient hybridization	1 min	98.4%	20 mer	Purified 3'-blocked dNTP	Template dependent DNA polymerase	Yes
Phosphoramidite chemistry	~4~10 min	99.5%	300 mer	Purified 5'-blocked dNTP	NA	NA

To clarify, these results are based on peer-reviewed articles. Step-wise yield greater than 99.9% and maximal length around 300 mer have been reported by start-up companies in public media. NA, not applicable.

Zhang et al., 2017). We will briefly review the history of whole-genome synthesis, and outline the key technologies and workflow involved. We then discuss the future developments to meet the requirements of whole-genome synthesis.

### Brief History of *de novo* Whole-Genome Synthesis

*De novo* synthesis of genomes offers the capability of complete control over the genetic code of an organism. Due to the small size of viruses and their important roles in the advancement of health and biotechnology, great progress has been made in viral genome synthesis. In 2002, Eckard Wimmer's research team was the first to generate infectious poliovirus by synthesizing complete cDNA (Cello et al., 2002). After that, humans have reconstructed dozens of RNA viruses through chemical methods, including the Spanish influenza virus in 1918 (Tumpey et al., 2005) and many others (Kosuri and Church, 2014). Several DNA bacteriophages were *de novo* synthesized as well (Smith et al., 2003; Liu et al., 2012). In addition to the viral genome, the Venter Institute has also designed, constructed, assembled and transplanted a fully synthetic bacterial genome to encode viable organisms. In 2008, the 583 kb genome of *Mycoplasma genitalium*, the smallest prokaryotic genome in nature, was chemically synthesized (Gibson et al., 2008). In 2010, The 1.08 Mb *Mycoplasma* genome was artificially synthesized (Gibson et al., 2010). The first eukaryotic gene combination project (Sc2.0) started in 2011 and has now completed the synthesis and assembly of chromosomes 2, 3, 5, 6, 10, and 12 (Dymond et al., 2011; Annaluru et al., 2014; Mercy et al., 2017; Mitchell et al., 2017; Richardson et al., 2017; Shen et al., 2017; Wu et al., 2017; Xie et al., 2017; Zhang et al., 2017). Recently, Pelletier et al. (2021) constructed a synthetic minimal cell with only 480 genes, which can achieve normal division and proliferation.

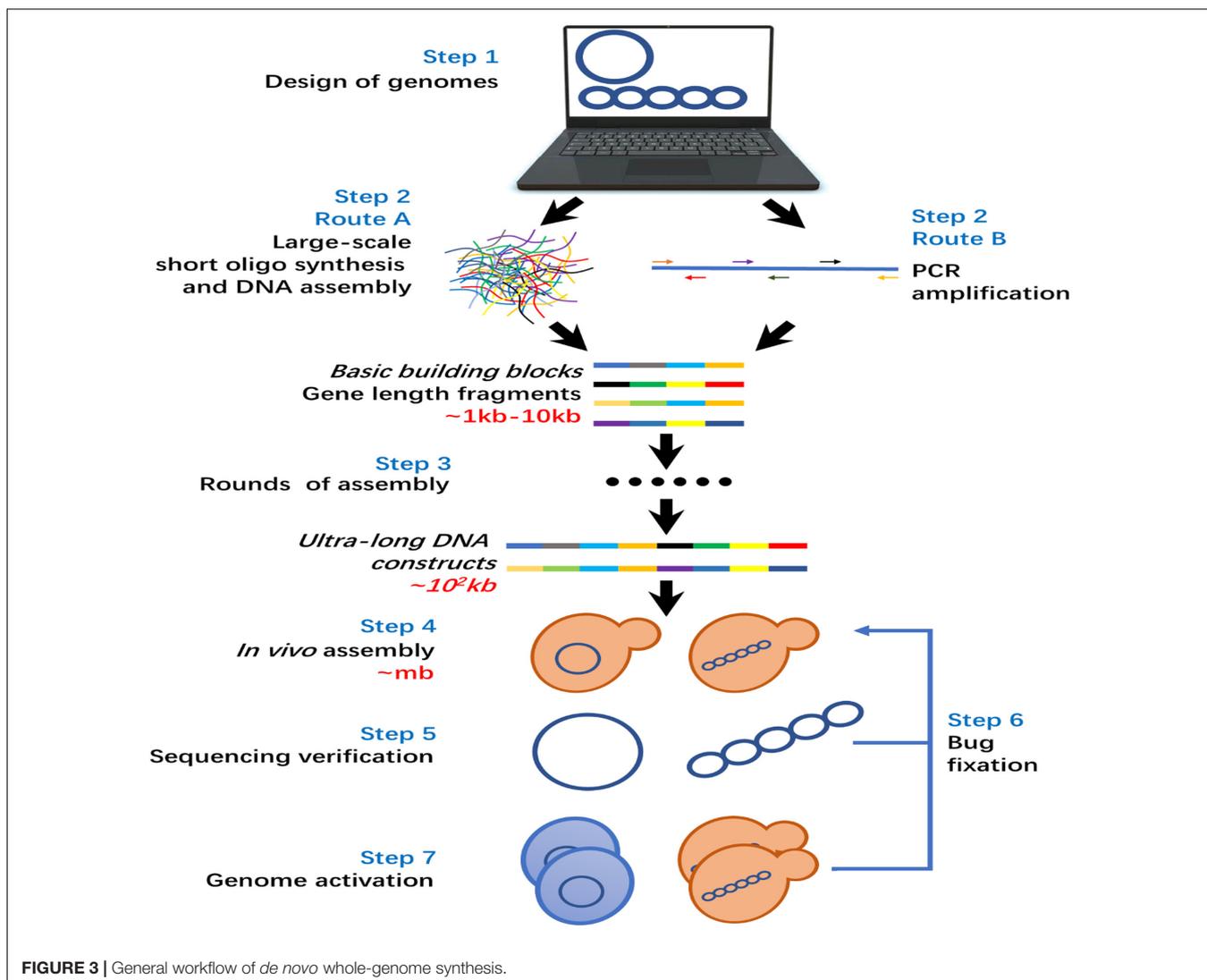
### General Workflow of Whole-Genome Synthesis

The general workflow of *de novo* genome synthesis is illustrated in **Figure 3**. Step 1, the genome sequence to be synthesized is designed on a computer with the help of various genome design tools (Lee et al., 2013). Step 2, generation of gene length builds blocks for genome assembly, and there are two optional routes in this step. In route A, the large-scale oligo synthesis is utilized for the generation of huge amounts of short DNA fragments as starting materials for the synthesis of a whole

genome. The short oligos are then assembled into gene-length DNA fragments by DNA assembly and ligation. For route B, target gene-length fragments are generated by PCR amplification in using primers produced by traditional column-based oligo synthesizer. In early studies of *de novo* genome synthesis, route B is preferred (Cello et al., 2002; Gibson et al., 2008, 2010). Owing to the fast developments in large-scale oligo synthesis, route A became more advanced in terms of cost and labor intensity and has been applied widely in recent studies (Shen et al., 2017; Wu et al., 2017; Xie et al., 2017). Step 3, the generated gene-length DNA fragments go through several rounds of DNA assembly in a trial-and-error way to generate ultra-long DNA fragments. The assembly rounds required mainly depend on the genome size and applied assembly methods. There is a general trend that the cost and efforts required rises exponentially with the increased size of the genome to be synthesized (Gibson, 2014; Casini et al., 2015; Kohman et al., 2018). Step 4, the ultra-long DNA fragments are assembled into the final genome-size. Due to the low efficiency of *in vitro* assembly of long DNA fragments, *in vivo* assembly is the general choice in most previous studies (Gibson, 2014; Shen et al., 2017; Wu et al., 2017; Xie et al., 2017). Step 5, the assembled genome was sequenced to check for errors/bugs. In case of critical errors/bugs occurring, a bug fixation step (Step 6) was applied to fix them. Step 7, the synthesized genome is activated in a proper host cell. If the activation fails, a bug fixation step is also required to remove the bugs until the synthesized genome was activated successfully. It is worth to emphasize that 'bug fixation', i.e., correction of the errors, is crucial for whole genome synthesis. The rapid developing CRISPR/Cas9 base editing tools, which has been widely used for correction of disease associated mutations, have promise in fixing the bugs (Ravindran, 2019).

### Future Developments of Large-Scale Oligo Synthesis for Whole-Genome Synthesis

Nowadays, all the techniques required to synthesize a whole-genome are already available as shown in **Figure 3**, especially those with small genome size such as bacteria genomes. However, the cost to synthesize a whole-genome is still very high due to the chemical reagent consumption during oligo synthesis, as well as sequencing validation and error correction steps during oligo assemble steps. Certainly, the large-scale oligo synthesis technology will be improved toward the aspects of low cost, high



**FIGURE 3** | General workflow of *de novo* whole-genome synthesis.

accuracy, and longer length. This will achieve a cheaper source of oligos as basic step in building blocks for genome synthesis. Meanwhile, the time and labor-intensive DNA assembly and testing process consumes a major part of the costs (Ma et al., 2012; Boles et al., 2017; Kohman et al., 2018) which cannot be benefited through advanced oligo synthesis technology. Thus, in order to reduce the total cost fundamentally, all the functions, e.g., oligo synthesis, DNA assembly, DNA analysis, that are required by whole genome synthesis should be integrated into one platform by using modern automation technologies. A pioneering study toward this direction shows dramatically reduced cost and time required to synthesize a genome (Boles et al., 2017).

## Dense and Long-Term Data Storage in DNA

The early efforts of storing artificial information in DNA can be found since 1996 (Davis, 1996). Limited by the high-cost and low-throughput oligo synthesis and sequencing technologies of

the time, only a minimal amount of information was stored (Davis, 1996; Clelland et al., 1999; Bancroft et al., 2001). Church et al. (2012) published a milestone study in using microchip-based oligo synthesis and next-generation sequencing for data 'writing' and 'reading' of digital data in DNA. This opens up a new window in using synthesized DNA as massive data storage media (Church et al., 2012). Indeed, with outstanding features of long-term, high-density, and low maintain cost, DNA is believed to be a viable and compelling alternative to traditional storage media and to be potential to solve the world crisis of digital data explosion (Church et al., 2012; Zhirnov et al., 2016; Ceze et al., 2019; Ping et al., 2019). The long-term, high-density and low maintenance cost features of data storage in DNA has drawn attention of companies, universities and research institutes world-wide (Goldman et al., 2013; Gabrys et al., 2015; Grass et al., 2015; Yazdi et al., 2015; Holmes, 2016; Erlich and Zielinski, 2017; Choi et al., 2018; Jensen et al., 2018; Organick et al., 2018; Song and Zeng, 2018; Anavy et al., 2019; Chen K. et al., 2019; Lee et al., 2019; Tomek et al., 2019).

Data storage in DNA is a technology of “storing information in polymers” which can be dated back to the 1980s when Kaempfer et al. (1987) proposed the concept. This technology was termed as “molecular data storage” in later studies and has been an active and challenging area since then (Boukiss and Meier, 2018; Martens et al., 2018; Ceze et al., 2019; Meier and Barner-Kowollik, 2019). Unlike the traditional planner media which requires a prepared surface for data writing, and molecular data storage requires precise polymer synthesis and sequencing methods for ‘writing’ and ‘reading’ of information instead. Furthermore, the synthesized polymers can be simply mixed to distribute the encoded information in a three-dimensional (3D) space. This extra dimension allows the molecular data storage to be an ultra-high-density storage technology outpacing the traditional planner media which are only two-dimensional (2D) (Boukiss and Meier, 2018; Martens et al., 2018; Song and Zeng, 2018; Ceze et al., 2019; Meier and Barner-Kowollik, 2019). DNA is the natural formed polymers which are utilized as information carrier by all living matter on earth. With tremendous availability in enzymes that can synthesize, replicate and even repair DNA molecules, DNA and its derivatives have significant advantages over other types of polymers for data storage purposes, at least for now. In this section, we will briefly summarize the recent studies on DNA data storage technology followed by discussions on the future developments of large-scale oligo synthesis technology for data storage purposes. For readers who are interested in other types of molecular data storage can refer to the recent review by Meier and Barner-Kowollik (2019).

### Principle of Data Storage in DNA

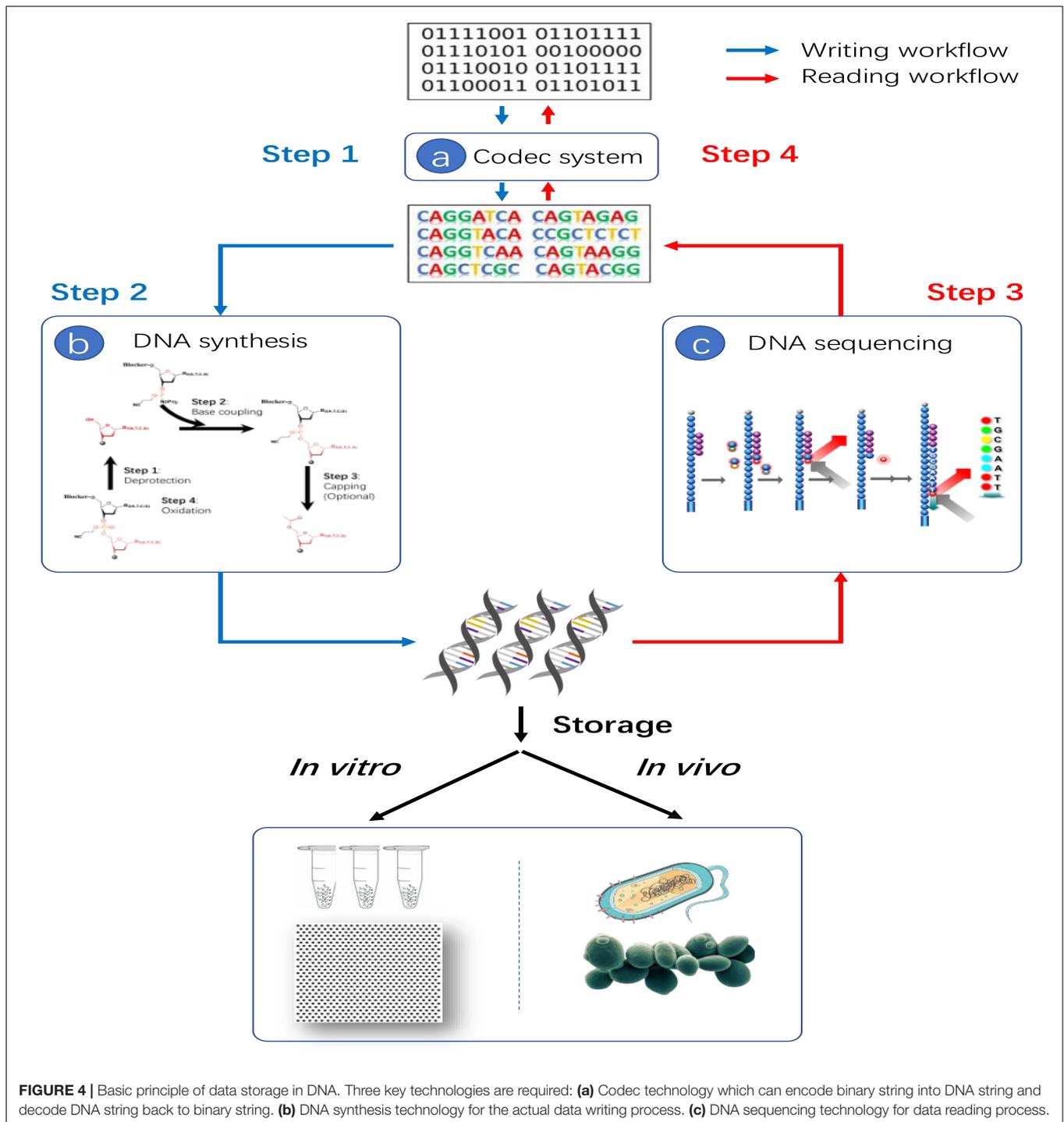
Although reliable data writing and reading with massive error-rich DNA molecules are complex and challenging, the basic principle of data storage in DNA is not complicated. As shown in **Figure 4**, the four steps are involved in the process of data storage in DNA generally. At first, the digital information, represented by string(s) of “0” and “1,” which is transformed into DNA string(s) of “A,” “T,” “G” and “C” with redundancy codes added by a well-designed codec system. Then, a DNA synthesis device is utilized to synthesize the DNA string(s) into actual DNA molecules, accomplishing the data-writing process. Next, the DNA molecules are sequenced by a DNA sequencer in order to read out the information in forms of DNA string(s). Finally, the DNA string(s) are decoded into the original binary information by the codec system with error correction codes corresponding to the applied redundancy codes in the first step. There are three key technologies required for data storage in DNA: (A) codec system, (B) DNA synthesis technology, (C) DNA sequencing technology as shown in **Figure 4**. In addition, depending on the storage vessel of actual DNA molecules, additional steps to modify and transfer the synthesized DNA molecules into a proper niche, i.e., tubes, plates or even living cells, may also be required.

### Recent Progress of Data Storage in DNA

The key achievements of data storage in DNA since 2012 is presented in **Figure 5**. Shortly after the proof concept of milestone study by Church et al. (2012), researchers of European Bioinformatics Institute developed a base-3 Huffman

code system that significantly increased the scale, density, and reliability of data storage in DNA (Goldman et al., 2013). In 2015, by storing data-encoding DNA strands in silica beads, Grass et al. (2015) achieved reliable data storage in DNA, which was estimated to be stable for more than 2000 years. This strategy was later improved in order to increase the storage density (Chen W. D. et al., 2019). Also in 2015, Yazdi et al. (2015) established a DNA-based storage architecture that enables random access of data blocks for the first time. The next year (2016), Tabatabaei Yazdi et al. (2016) applied Nanopore sequencer for reading of DNA encoded data and considering its low-cost and pocket-size advantages, proving the concept of portable DNA data storage. Later in 2017, by applying fountain codes, Erlich and Zielinski (2017) established an architecture which achieves ultra-high coding efficiency (bits per base) and enables PCR-based copy free DNA data storage. Organick et al. (2018) constructed a random access approach which can work on large data sets of more than 200M. However, this approach was proved to work on even larger and higher density data sets later (Organick et al., 2019). Takahashi et al. (2018) developed a complete end-to-end DNA data storage device. Song and Zeng (2018) established an *in vivo* coding system which enables ultra-long-term short message storage in the DNA of living cells. Lopez et al. (2019) established a DNA assembly strategy to overcome the problem in reading short DNA strands by Nanopore sequencer. Chen K. et al. (2019) developed a unique way of using Nanopore sequencer by encoding information in different DNA hairpins instead of nucleic acid bases. Newman et al. (2019) applied digital microfluidic to automatically retrieve data from high-density DNA data storage library. Anavy et al. (2019) implemented a high coding efficiency of 4.29 bits per base by using composite DNA letters. The recent review by Ping et al. (2019) has described these progress in detail which includes a comprehensive summary of the evolving error correction codes. In brief summary, great efforts have been focused on improving the accuracy, scale, coding efficiency and density of data storage in DNA by construction of error correction codes and integration with oligo synthesis and DNA sequencing technologies. The data scale has increased from the 0.66 MB (megabytes) in 2012 to 200 MB in 2018 (Church et al., 2012; Organick et al., 2018). The storage density has increased from 2.2 PB/g (petabytes per gram DNA) to 17,000 PB/g (Erlich and Zielinski, 2017; Organick et al., 2020). The coding efficiency has increased from 1 to 4.29 bits per base (Anavy et al., 2019). Meanwhile, the function of data copy (Erlich and Zielinski, 2017), random access (Yazdi et al., 2015; Organick et al., 2018), long-term stability (Grass et al., 2015; Chen W. D. et al., 2019) have also achieved outstanding outcomes.

Besides the rapid evolving codec systems, the novel storage modes and applications were also proposed. Koch et al. (2020) proposed a DNA-of-things storage architecture to create materials with embedded memory. Choi et al. (2020) proposed DNA micro-disks concept for the efficient management of DNA-based data storage by QR-coded micro-sized disks carrying data-encoded DNA molecules. Furthermore, several recent studies have also proved that the feasibility of massive data storage in DNA of living cells. Chen et al. (2021) encoded the information in a synthetic yeast artificial chromosome with four autonomously

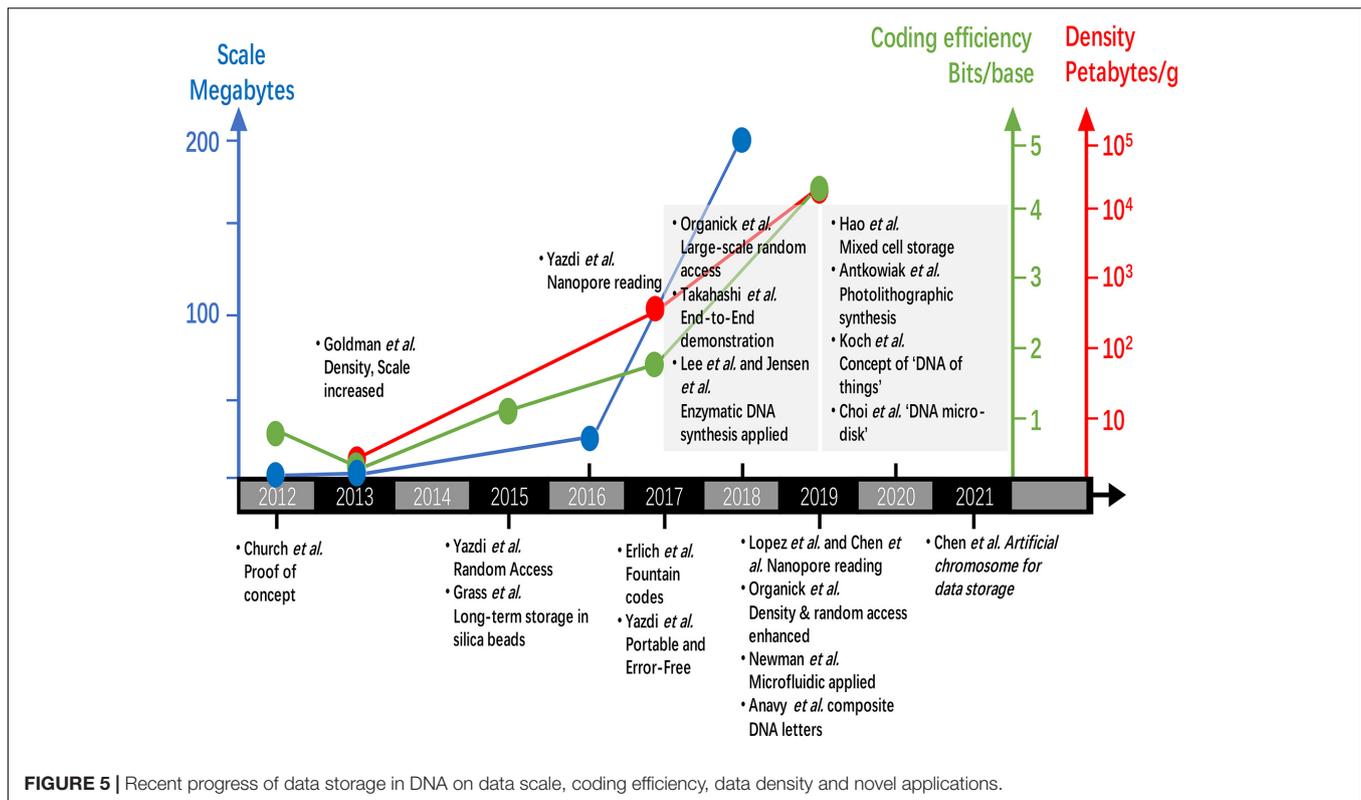


replicating sequences and maintained good stability during yeast replication. Yim et al. (2021) used electrical signals to regulate redox reactions in cells, thus causing changes in plasmid copy number, and then used it to control the trapping frequency of different sequences by Cas proteins and to achieve one-step storage of data in cells. Hao et al. (2020) assembled a large oligo pool into vector plasmids based on homology and introduced them into bacterial cells for mixed culture, then combining the

advantages of *in vivo* cell systems and *in vitro* data storage mediated by oligo pools to achieve high-fidelity replication of DNA molecules at low cost.

### Future Development of Large-Scale Oligo Synthesis for DNA-Based Data Storage

The development of cheaper, longer, and higher throughput oligo synthesis technologies would be necessary for DNA-based



storage. While previous studies achieved substantial progress on data scale, stability, random access and copy free, *etc.*, the cost, especially the synthesis cost, emerged as one of the key obstacles for practical data storage in DNA (Ceze *et al.*, 2019; Ping *et al.*, 2019). It was estimated that 7–8 orders magnitude decrease of the synthesis cost is essential for DNA data storage to outpace the currently used magnetic tape based on storage technology. And, another obstacle to current data storage technology is bandwidth (throughput of data writing and reading). Without an accordant writing and reading bandwidth, the storage capacity of DNA is not feasible and cannot be utilized to solve the world crisis of digital data explosion. There is a six orders-of-magnitude gap for DNA synthesis to catch up with mainstream cloud archival storage system. In order to support low-cost and high bandwidth data storage in DNA, the large-scale oligo synthesis technology needs to be dramatically improved in terms of cost, speed, and throughput. Driven by the demands of life science, the throughputs of DNA synthesis technology will be persistently improved via more parallel synthesis regardless of DNA data storage application. Also, increased parallelism could be carried out via more supporting microarray or plate area and smaller spot size. The smaller spot size can also save the reagent consumption proportionally as discussed previously (Ceze *et al.*, 2019). Although the throughput and cost gap seem immense, there are many aspects that can be optimized to reduce the cost of oligo synthesis technology for data storage purposes. The state-of-the-art oligo synthesis technologies are designed for life science, where the accuracy of synthesized DNA molecules is highly desired. Unlike life science, with the incorporation of

error-correcting codes, DNA data storage requires much lower accuracy of synthesized DNA strands. Thus, the expensive and labor-intensive validation and error correction steps could be avoided. The synthesizing steps can be simplified and the high purity reagents used can be replaced with low purity ones to reduce costs. For example, the capping step is applied to reduce the deletion errors and can be removed which may increase the synthesis speed, length limitation and reduce the cost in principle. High purity reagents are much more expensive than the low purity ones. The state-of-the-art large-scale oligo synthesis platforms use super high purity (>99.99%) dNTP reagents. For data storage purposes, the dNTP reagents with 99% purity or even lower may be acceptable which may greatly reduce the cost. Otherwise, the newly emerged enzymatic DNA synthesis methods using template-independent TdT or using template-dependent DNA polymerase can provide another route for development of high-throughput and low-cost oligo synthesis technology for data storage purposes (Palluk *et al.*, 2018; Hoff *et al.*, 2019). Indeed, two groups of Jensen *et al.* (2018) and Lee *et al.* (2018) have already applied TdT-based enzymatic DNA synthesis method to data storage applications. In their study, they applied a similar strategy of using terminator free TdT enzymes for DNA data storage purpose. Random identical bases were added sequentially and the transition of base types was utilized for data recording. It is worth mentioning that, in the study by Lee *et al.* (2019), the terminator free TdT-based DNA synthesis was proved to be a superior combination with Nanopore sequencer which is optimal for sequencing of long DNA strands with high error rate.

In addition to the fundamental technical developments regarding the cost and throughput, the oligo synthesis technology should be integrated with sequencing technology as well as other essential automatic functions to accomplish the full function of data writing, copying, reading, and random access in one device as the proof-of-concept study by Takahashi et al. (2018). For *in vivo* DNA data storage, additional function models for handling of cell cultivation, DNA assembly and transformation are also required. Finally, four new synthetic nucleic acids: 6-Amino-5-nitropyridin-2-one (Z), 5-Aza-7-deazaguanine (P), Isocytosine (S), and Isoguanine (B) are reported (Hoshika et al., 2019). Incorporation of these new bases would further increase the coding efficiency in the near future (Ping et al., 2019).

## CONCLUSION

Our chase to more advanced large-scale oligo synthesis technology in terms of cost and throughput is ‘endless’ which is similar to the endless demands of faster processors in the electronic industry. Despite the astonishing achievements on large-scale oligo synthesis that have been made in the past decades, our capacity to write DNA sequences still lags far behind our ability to read them. While the rapid development of DNA sequencing technology has heavily relied on the “sequencing by synthesis” method which can utilize the power of naturally formed polymerase for sequencing purpose, the oligo synthesis is dominated by chemical synthesis method with several limitations like limited synthesis length and environmentally hazardous, *etc.* The emerging enzymatic methods uncovers an exciting route for pushing large-scale oligo synthesis technology into next-generation by harnessing the power of naturally formed polymerases. Given that the numerous advances of mild conditions, no generation of hazardous waste, no requirements of protecting groups, *etc.*, we believe that the further optimization of enzymatic methods and its microchip-based parallelization will potentially drive the large-scale oligo synthesis technology into next-generation which is bound to be more accurate, cheaper, faster, and be able to synthesize longer sequences.

The future developments in large-scale oligo synthesis technology to whole-genome synthesis and data storage are both expected to be focused on the automation integration of required functions. For whole-genome synthesis, a fully automated device which can integrate large-scale oligo synthesis with DNA assembly and DNA analysis, and would dramatically reduce the cost and labor intensive process of genome assembly (Boles et al., 2017). For DNA-based data storage, the large-scale

oligo synthesis technology is desired to be integrated with DNA sequencing technology to achieve data “writing” and “reading” in one device. Furthermore, the functions of “random access,” “data copying” also needs to be implemented with modern automation technologies on this integrated platform. In detail, the future adaptations of large-scale oligo synthesis technology for genome synthesis and data storage are believed to have slightly different focus while they both strive for cost controls and more parallel oligo synthesis technique. For oligo synthesis, accuracy and cost are two contradictory factors which means increasing one factor will bring down another one when using the same technique. For *de novo* whole-genome synthesis, accuracy is always the priority focus. This is because the debugging cost is much higher than the benefits that we can get from an inaccurate synthesis process. Thus, the developments of large-scale oligo synthesis for genome synthesis, we must guarantee a high accurate process while improving throughput and reducing cost. At the same time, the DNA assembly and bug fixation technologies should also be developed. Differently, for the DNA-based data storage, the accuracy requirement is much lower compared to the genome synthesis application since the error correction codes could be integrated to correct the errors that emerge. Instead, data storage application requires much higher throughput and lower costs. Thus, there is a trade-off between the accuracy and the cost-throughput. The work of Lee et al. (2019) is a perfect case study of such adjustments between the cost and accuracy. More studies following this strategy in the future are expected to reduce the synthesis cost of data storage in DNA for several orders which should be greatly helpful to make the DNA-based storage a practical reality.

## AUTHOR CONTRIBUTIONS

L-FS and Z-HD wrote the first draft. B-ZL supervised the whole study. Z-YG, L-LL, and B-ZL edited, revised, and finalized the text. All authors contributed to the article and approved the submitted version.

## FUNDING

We gratefully acknowledge the financial support from the National Key Research and Development Program of China (2018YFA0900100) and the Tianjin Fund for Distinguished Young Scholars (19JCJQC63300).

## REFERENCES

- Anavy, L., Vaknin, I., Atar, O., Amit, R., and Yakhini, Z. (2019). Data storage in DNA with fewer synthesis cycles using composite DNA letters. *Nat. Biotechnol.* 37:0240–x. doi: 10.1038/s41587-019-0240-x
- Annaluru, N., Muller, H., Mitchell, L. A., Ramalingam, S., Stracquadanio, G., Richardson, S. M., et al. (2014). Total synthesis of a functional designer eukaryotic chromosome. *Science* 344, 55–58. doi: 10.1126/science.1249252
- Bancroft, C., Bowler, T., Bloom, B., and Clelland, C. T. (2001). Long-term storage of information in DNA. *Science* 293, 1763–1765. doi: 10.1126/science.293.5536.1763c
- Beaucage, S. L., and Caruthers, M. H. (1981). Deoxynucleoside phosphoramidites—A new class of key intermediates for deoxypolynucleotide synthesis. *Tetrahedron Lett.* 22, 1859–1862.
- Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., et al. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456, 53–59. doi: 10.1038/nature07517

- Boles, K. S., Kannan, K., Gill, J., Felderman, M., Gouvis, H., Hubby, B., et al. (2017). Digital-to-biological converter for on-demand production of biologics. *Nat. Biotechnol.* 35, 672–675. doi: 10.1038/nbt.3859
- Bollum, F. J. (1962). Oligodeoxyribonucleotide-primed reactions catalyzed by calf thymus polymerase. *J. Biol. Chem.* 237, 1945–1949.
- Boukris, A. C., and Meier, M. A. R. (2018). Data storage in sequence-defined macromolecules via multicomponent reactions. *Eur. Polymer J.* 104, 32–38. doi: 10.1016/j.eurpolymj.2018.04.038
- Caruthers, M. H. (1985). Gene synthesis machines: DNA chemistry and its uses. *Science* 230, 281–285. doi: 10.1126/science.3863253
- Casini, A., Storch, M., Baldwin, G. S., and Ellis, T. (2015). Bricks and blueprints: methods and standards for DNA assembly. *Nat. Rev. Mol. Cell Biol.* 16, 568–576. doi: 10.1038/nrm4014
- Cello, J., Paul, A. V., and Wimmer, E. (2002). Chemical synthesis of poliovirus cDNA: generation of infectious virus in the absence of natural template. *Science* 297, 1016–1018. doi: 10.1126/science.1072266
- Ceze, L., Nivala, J., and Strauss, K. (2019). Molecular digital data storage using DNA. *Nat. Rev. Genet.* 20, 456–466. doi: 10.1038/s41576-019-0125-3
- Chen, K., Kong, J., Zhu, J., Ermann, N., Predki, P., and Keyser, U. F. (2019). Digital Data Storage Using DNA Nanostructures and Solid-State Nanopores. *Nano Lett.* 19, 1210–1215. doi: 10.1021/acs.nanolett.8b04715
- Chen, W., Han, M., Zhou, J., Ge, Q., Wang, P., Zhang, X., et al. (2021). An artificial chromosome for data storage. *National Sci. Rev.* 2021:nwab028. doi: 10.1093/nsr/nwab028
- Chen, W. D., Kohll, A. X., Nguyen, B. H., Koch, J., Heckel, R., Stark, W. J., et al. (2019). Combining Data Longevity with High Storage Capacity—Layer-by-Layer DNA Encapsulated in Magnetic Nanoparticles. *Adv. Funct. Mater.* 29:1901672. doi: 10.1002/adfm.201901672
- Choi, Y., Bae, H. J., Lee, A. C., Choi, H., Lee, D., Ryu, T., et al. (2020). DNA Micro-Disks for the Management of DNA-Based Data Storage with Index and Write-Once-Read-Many (WORM) Memory Features. *Adv. Mater. Weinheim* 32:e2001249. doi: 10.1002/adma.202001249
- Choi, Y., Ryu, T., Lee, A. C., Choi, H., Lee, H., Park, J., et al. (2018). Addition of Degenerate Bases to DNA-based Data Storage for Increased Information Capacity. [preprint].
- Church, G. M., Gao, Y., and Kosuri, S. (2012). Next-generation digital information storage in DNA. *Science* 337:1628. doi: 10.1126/science.1226355
- Clelland, C. T., Risca, V., and Bancroft, C. (1999). Hiding messages in DNA microdots. *Nature* 399, 533–534. doi: 10.1038/21092
- Davis, J. (1996). Microvenus. *Art J.* 55:70.
- Dymond, J. S., Richardson, S. M., Coombes, C. E., Babatz, T., Muller, H., Annaluru, N., et al. (2011). Synthetic chromosome arms function in yeast and generate phenotypic diversity by design. *Nature* 477, 471–476. doi: 10.1038/nature10403
- Hyman, E. D. (1992). *Method for enzymatic synthesis of oligonucleotides*. Alexandria, VA: United States Patent and Trademark Office.
- Egeland, R. D., and Southern, E. M. (2005). Electrochemically directed synthesis of oligonucleotides for DNA microarray fabrication. *Nucleic Acids Res.* 33:e125. doi: 10.1093/nar/gni117
- Eisenstein, M. (2020). Enzymatic DNA synthesis enters new phase. *Nat. Biotechnol.* 38, 1113–1115. doi: 10.1038/s41587-020-0695-9
- Erlich, Y., and Zielinski, D. (2017). DNA Fountain enables a robust and efficient storage architecture. *Science* 355, 950–954. doi: 10.1126/science.aaj2038
- Fodor, S. P., Read, J. L., Pirrung, M. C., Stryer, L., Lu, A. T., and Solas, D. (1991). Light-directed, spatially addressable parallel chemical synthesis. *Science* 251, 767–773. doi: 10.1126/science.1990438
- Gabrys, R., Kiah, H. M., and Milenkovic, O. (2015). “Asymmetric Lee distance codes for DNA-based storage,” in *2015 IEEE International Symposium on Information Theory (ISIT)* (New York, NY: IEEE), 909–913.
- Gao, X., LeProust, E., Zhang, H., Srivannavit, O., Gulari, E., Yu, P., et al. (2001). A flexible light-directed DNA chip synthesis gated by deprotection using solution photogenerated acids. *Nucleic Acids Res.* 29, 4744–4750. doi: 10.1093/nar/29.22.4744
- Ghindilis, A. L., Smith, M. W., Schwarzkopf, K. R., Roth, K. M., Peyvan, K., Munro, S. B., et al. (2007). CombiMatrix oligonucleotide arrays: genotyping and gene expression assays employing electrochemical detection. *Biosens Bioelectr.* 22, 1853–1860. doi: 10.1016/j.bios.2006.06.024
- Gibson, D. G. (2014). Programming biological operating systems: genome design, assembly and activation. *Nat. Methods* 11, 521–526. doi: 10.1038/nmeth.2894
- Gibson, D. G., Benders, G. A., Andrews-Pfannkoch, C., Denisova, E. A., Baden-Tillson, H., Zaveri, J., et al. (2008). Complete chemical synthesis, assembly, and cloning of a *Mycoplasma genitalium* genome. *Science* 319, 1215–1220. doi: 10.1126/science.1151721
- Gibson, D. G., Glass, J. I., Lartigue, C., Noskov, V. N., Chuang, R.-Y., Algire, M. A., et al. (2010). Creation of a bacterial cell controlled by a chemically synthesized genome. *Science* 329, 52–56. doi: 10.1126/science.1190719
- Goldman, N., Bertone, P., Chen, S., Dessimoz, C., LeProust, E. M., Sipos, B., et al. (2013). Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. *Nature* 494, 77–80. doi: 10.1038/nature11875
- Grass, R. N., Heckel, R., Puddu, M., Paunescu, D., and Stark, W. J. (2015). Robust chemical preservation of digital information on DNA in silica with error-correcting codes. *Angew. Chem. Int. Ed. Engl.* 54, 2552–2555. doi: 10.1002/anie.201411378
- Grunberg-Manago, M., Oritz, P. J., and Ochoa, S. (1955). Enzymatic synthesis of nucleic acidlike polynucleotides. *Science* 122, 907–910. doi: 10.1126/science.122.3176.907
- Hao, M., Qiao, H., Gao, Y., Wang, Z., Qiao, X., Chen, X., et al. (2020). A mixed culture of bacterial cells enables an economic DNA storage on a large scale. *Commun. Biol.* 3:416. doi: 10.1038/s42003-020-01141-7
- Hoff, K., Halpain, M., Garbagnati, G., Edwards, J., and Zhou, W. (2019). Rapid and dynamic nucleic acid hybridization enables enzymatic oligonucleotide synthesis by cyclic reversible termination: A novel mechanism for enzymatic DNA synthesis. [Preprint].
- Holmes, I. (2016). Modular non-repeating codes for DNA storage. [Preprint].
- Hoshika, S., Leal, N. A., Kim, M.-J., Kim, M.-S., Karalkar, N. B., Kim, H.-J., et al. (2019). Hachimoji DNA and RNA: A genetic system with eight building blocks. *Science* 363, 884–887. doi: 10.1126/science.aat0971
- Jensen, M. A., Griffin, P. B., and Davis, R. W. (2018). Free-running enzymatic oligonucleotide synthesis for data storage applications. [Preprint].
- Kaempfer, G., Loewer, H., and Witman, M. W. (1987). Polymers as substrates and media for data storage. *Polymer Eng. Sci.* 27, 1421–1435.
- Koch, J., Gantenbein, S., Masania, K., Stark, W. J., Erlich, Y., and Grass, R. N. (2020). A DNA-of-things storage architecture to create materials with embedded memory. *Nat. Biotechnol.* 38, 39–43. doi: 10.1038/s41587-019-0356-z
- Kohman, R. E., Kunjapur, A. M., Hysolli, E., Wang, Y., and Church, G. M. (2018). From Designing the Molecules of Life to Designing Life: Future Applications Derived from Advances in DNA Technologies. *Angew. Chem. Int. Ed. Engl.* 57, 4313–4328. doi: 10.1002/anie.201707976
- Kosuri, S., and Church, G. M. (2014). Large-scale de novo DNA synthesis: technologies and applications. *Nat. Methods* 11, 499–507. doi: 10.1038/nmeth.2918
- Lee, B.-R., Cho, S., Song, Y., Kim, S. C., and Cho, B.-K. (2013). Emerging tools for synthetic genome design. *Mol. Cells* 35, 359–370. doi: 10.1007/s10059-013-0127-5
- Lee, H. H., Kalhor, R., Goela, N., Bolot, J., and Church, G. M. (2018). Enzymatic DNA synthesis for digital information storage. [Preprint].
- Lee, H. H., Kalhor, R., Goela, N., Bolot, J., and Church, G. M. (2019). Terminator-free template-independent enzymatic DNA synthesis for digital information storage. *Nat. Commun.* 10:2383. doi: 10.1038/s41467-019-10258-1
- Liu, Y., Han, Y., Huang, W., Duan, Y., Mou, L., Jiang, Z., et al. (2012). Whole-genome synthesis and characterization of viable S13-like bacteriophages. *PLoS One* 7:e41124. doi: 10.1371/journal.pone.0041124
- Lopez, R., Chen, Y.-J., Dumas Ang, S., Yekhanin, S., Makarychev, K., Racz, M. Z., et al. (2019). DNA assembly for nanopore data storage readout. *Nat. Commun.* 10:2933. doi: 10.1038/s41467-019-10978-4
- Ma, S., Tang, N., and Tian, J. (2012). DNA synthesis, assembly and applications in synthetic biology. *Curr. Opin. Chem. Biol.* 16, 260–267. doi: 10.1016/j.cbpa.2012.05.001
- Mackey, J. K., and Gilham, P. T. (1971). New approach to the synthesis of polyribonucleotides of defined sequence. *Nature* 233, 551–553. doi: 10.1038/233551a0
- Martens, S., Landuyt, A., Espeel, P., Devreese, B., Dawyndt, P., and Du Prez, F. (2018). Multifunctional sequence-defined macromolecules for chemical data storage. *Nat. Commun.* 9:4451. doi: 10.1038/s41467-018-06926-3

- Mathews, A. S., Yang, H., and Montemagno, C. (2016). Photo-cleavable nucleotides for primer free enzyme mediated DNA synthesis. *Org. Biomol. Chem.* 14, 8278–8288. doi: 10.1039/C6OB01371F
- Mathews, A. S., Yang, H., and Montemagno, C. (2017). 3'-O-Caged 2'-Deoxynucleoside Triphosphates for Light-Mediated, Enzyme-Catalyzed, Template-Independent DNA Synthesis. *Curr. Protoc. Nucleic Acid Chem.* 71, 13.17.1–13.17.38. doi: 10.1002/cpnc.41
- Meier, M. A. R., and Barner-Kowollik, C. (2019). A New Class of Materials: Sequence-Defined Macromolecules and Their Emerging Applications. *Adv. Mater. Weinheim* 31:e1806027. doi: 10.1002/adma.201806027
- Mercy, G., Mozziconacci, J., Scolari, V. F., Yang, K., Zhao, G., Thierry, A., et al. (2017). 3D organization of synthetic and scrambled chromosomes. *Science* 355:aaf4597. doi: 10.1126/science.aaf4597
- Mitchell, E. A. D., Mulhauser, B., Mulot, M., Mutabazi, A., Glauser, G., and Aebi, A. (2017). A worldwide survey of neonicotinoids in honey. *Science* 358, 109–111. doi: 10.1126/science.aan3684
- Motea, E. A., and Berdis, A. J. (2010). Terminal deoxynucleotidyl transferase: the story of a misguided DNA polymerase. *Biochim. Biophys. Acta* 1804, 1151–1166. doi: 10.1016/j.bbapap.2009.06.030
- Newman, S., Stephenson, A. P., Willsey, M., Nguyen, B. H., Takahashi, C. N., Strauss, K., et al. (2019). High density DNA data storage library via dehydration with digital microfluidic retrieval. *Nat. Commun.* 10:1706. doi: 10.1038/s41467-019-09517-y
- Organick, L., Ang, S. D., Chen, Y.-J., Lopez, R., Yekhanin, S., Makarychev, K., et al. (2018). Random access in large-scale DNA data storage. *Nat. Biotechnol.* 36, 242–248. doi: 10.1038/nbt.4079
- Organick, L., Chen, Y.-J., Ang, S. D., Lopez, R., Strauss, K., and Ceze, L. (2019). Experimental Assessment of PCR Specificity and Copy Number for Reliable Data Retrieval in DNA Storage. [Preprint].
- Organick, L., Chen, Y.-J., Dumas Ang, S., Lopez, R., Liu, X., Strauss, K., et al. (2020). Probing the physical limits of reliable DNA data retrieval. *Nat. Commun.* 11:616. doi: 10.1038/s41467-020-14319-8
- Palluk, S., Arlow, D. H., Rond, T., de Barthel, S., Kang, J. S., et al. (2018). De novo DNA synthesis using polymerase-nucleotide conjugates. *Nat. Biotechnol.* 36, 645–650. doi: 10.1038/nbt.4173
- Pease, A. C., Solas, D., Sullivan, E. J., Cronin, M. T., Holmes, C. P., and Fodor, S. P. (1994). Light-generated oligonucleotide arrays for rapid DNA sequence analysis. *Proc. Natl. Acad. Sci. U S A.* 91, 5022–5026. doi: 10.1073/pnas.91.11.5022
- Pelletier, J. F., Sun, L., Wise, K. S., Assad-Garcia, N., Karas, B. J., Deerinck, T. J., et al. (2021). Genetic requirements for cell division in a genomically minimal cell. *Cell* 2021:008. doi: 10.1016/j.cell.2021.03.008
- Ping, Z., Ma, D., Huang, X., Chen, S., Liu, L., Guo, F., et al. (2019). Carbon-based archiving: current progress and future prospects of DNA-based data storage. *Gigascience* 8:giz075. doi: 10.1093/gigascience/giz075
- Ramadan, K., Shevelev, I. V., Maga, G., and Hübscher, U. (2004). De novo DNA synthesis by human DNA polymerase lambda, DNA polymerase mu and terminal deoxynucleotidyl transferase. *J. Mol. Biol.* 339, 395–404. doi: 10.1016/j.jmb.2004.03.056
- Ravindran, S. (2019). Got mutation? 'Base editors' fix genomes one nucleotide at a time. *Nature* 575, 553–555. doi: 10.1038/d41586-019-03536-x
- Richardson, S. M., Mitchell, L. A., Stracquadanio, G., Yang, K., Dymond, J. S., DiCarlo, J. E., et al. (2017). Design of a synthetic yeast genome. *Science* 355, 1040–1044. doi: 10.1126/science.aaf4557
- Roy, S., and Caruthers, M. (2013). Synthesis of DNA/RNA and their analogs via phosphoramidite and H-phosphonate chemistries. *Molecules* 18, 14268–14284. doi: 10.3390/molecules18114268
- Ruparel, H., Bi, L., Li, Z., Bai, X., Kim, D. H., Turro, N. J., et al. (2005). Design and synthesis of a 3'-O-allyl photocleavable fluorescent nucleotide as a reversible terminator for DNA sequencing by synthesis. *Proc. Natl. Acad. Sci. U S A.* 102, 5932–5937. doi: 10.1073/pnas.0501962102
- Severo Ochoa, S. M. (1961). Enzymatic Synthesis of Polynucleotides. *J. Biol. Chem.* 236, 3303–3301.
- Shen, Y., Wang, Y., Chen, T., Gao, F., Gong, J., Abramczyk, D., et al. (2017). Deep functional analysis of synII, a 770-kilobase synthetic yeast chromosome. *Science* 355:aaf4791. doi: 10.1126/science.aaf4791
- Singh-Gasson, S., Green, R. D., Yue, Y., Nelson, C., Blattner, F., Sussman, M. R., et al. (1999). Maskless fabrication of light-directed oligonucleotide microarrays using a digital micromirror array. *Nat. Biotechnol.* 17, 974–978. doi: 10.1038/13664
- Smith, H. O., Hutchison, C. A., Pfannkoch, C., and Venter, J. C. (2003). Generating a synthetic genome by whole genome assembly: phiX174 bacteriophage from synthetic oligonucleotides. *Proc. Natl. Acad. Sci. U S A.* 100, 15440–15445. doi: 10.1073/pnas.2237126100
- Song, L., and Zeng, A.-P. (2018). Orthogonal Information Encoding in Living Cells with High Error-Tolerance, Safety, and Fidelity. *ACS Synth. Biol.* 7, 866–874. doi: 10.1021/acssynbio.7b00382
- Tabatabaei Yazdi, S. M. H., Gabrys, R., and Milenkovic, O. (2016). Portable and Error-Free DNA-Based Data Storage. *Sci. Rep.* 7:5011.
- Takahashi, C. N., Nguyen, B. H., Strauss, K., and Ceze, L. (2018). Demonstration of End-to-End Automation of DNA Data Storage. *Sci. Rep.* 9:4998.
- Tomek, K. J., Volkell, K., Simpson, A., Hass, A. G., Indermaur, E. W., Tuck, J. M., et al. (2019). Driving the Scalability of DNA-Based Information Storage Systems. *ACS Synth. Biol.* 8, 1241–1248. doi: 10.1021/acssynbio.9b01000
- Tumpey, T. M., Basler, C. F., Aguilar, P. V., Zeng, H., Solórzano, A., Swaine, D. E., et al. (2005). Characterization of the reconstructed 1918 Spanish influenza pandemic virus. *Science* 310, 77–80. doi: 10.1126/science.1119392
- Efcavitch, W. J., and Sylvester, E. J. (2015). *Modified template-independent enzymes for polydeoxynucleotide synthesis*. Alexandria, VA: United States Patent and Trademark Office.
- Wu, J., Zhang, S., Meng, Q., Cao, H., Li, Z., Li, X., et al. (2007). 3'-O-modified nucleotides as reversible terminators for pyrosequencing. *Proc. Natl. Acad. Sci. U S A.* 104, 16462–16467. doi: 10.1073/pnas.0707495104
- Wu, Y., Li, B.-Z., Zhao, M., Mitchell, L. A., Xie, Z.-X., Lin, Q.-H., et al. (2017). Bug mapping and fitness testing of chemically synthesized chromosome X. *Science* 355:aaf4706. doi: 10.1126/science.aaf4706
- Xie, Z.-X., Li, B.-Z., Mitchell, L. A., Wu, Y., Qi, X., Jin, Z., et al. (2017). "Perfect" designer chromosome V and behavior of a ring derivative. *Science* 355:aaf4704. doi: 10.1126/science.aaf4704
- Yazdi, S. M. H. T., Yuan, Y., Ma, J., Zhao, H., and Milenkovic, O. (2015). A Rewritable, Random-Access DNA-Based Storage System. *Sci. Rep.* 5:14138. doi: 10.1038/srep14138
- Yim, S. S., McBee, R. M., Song, A. M., Huang, Y., Sheth, R. U., and Wang, H. H. (2021). Robust direct digital-to-biological data storage in living cells. *Nat. Chem. Biol.* 17, 246–253. doi: 10.1038/s41589-020-00711-4
- Zhang, W., Zhao, G., Luo, Z., Lin, Y., Wang, L., Guo, Y., et al. (2017). Engineering the ribosomal DNA in a megabase synthetic chromosome. *Science* 355:aaf3981. doi: 10.1126/science.aaf3981
- Zhirnov, V., Zadegan, R. M., Sandhu, G. S., Church, G. M., and Hughes, W. L. (2016). Nucleic acid memory. *Nat. Mater.* 15, 366–370. doi: 10.1038/nmat4594

**Conflict of Interest:** L-LL is an employee at LC-BIO Technologies (Hangzhou) Co., Ltd.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2021 Song, Deng, Gong, Li and Li. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.