ACS SyntheticBiology

pubs.acs.org/synthbio

Research Article

Scaling the High-Yield Potential of Large-Scale DNA Data Storage with Cap-Free DNA Synthesis

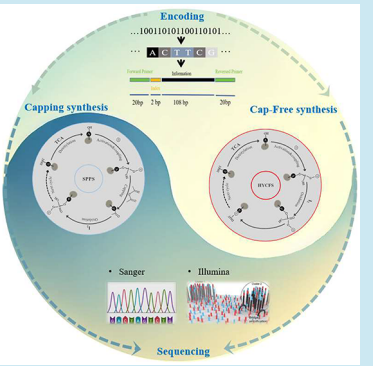
Weiming Lin, Haotian Yu, Weihao Li, Yemin Han, Manman Lv, Han Gao, Mengqing Cheng, Yan Huang, Kun Bi, Zuhong Lu, and Quanjun Liu*

Cite This: https://doi.org/10.1021/acssynbio.5c00175

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: DNA has emerged as a promising storage medium for addressing exponentially growing data storage demands, owing to its exceptional information density and chemical stability. However, current DNA synthesis techniques face significant limitations in achieving high-throughput data storage due to constrained synthesis yields. This study presents a novel high-yield cap-free synthesis (HYCFS) strategy that overcomes the limitations of conventional solid-phase phosphoramidite chemistry in both synthesis length and production yield. We established a theoretical product prediction model based on cap-free synthesis characteristics and systematically evaluated the strategy through standard DNA storage workflows. Under column-based synthesis conditions with high coupling efficiency (>99%), this approach demonstrated a 3-fold enhancement in effective sequence yield compared to traditional methods. Theoretical modeling predicts superior performance in array-based synthesis systems for large-scale data storage applications. HYCFS shows potential to enhance DNA-based data storage capacity by 2 orders of magnitude while reducing storage costs, thereby advancing the development of large-scale DNA data storage technologies.



KEYWORDS: *oligonucleotide synthesis, next-generation sequencing, error profiling, DNA data storage*

INTRODUCTION

Global data generation has been increasing exponentially, creating greater demands for data storage. Conventional storage methods dependent on magnetic and semiconductor-based media have reached their capacity limitations.[1] As the genetic information carrier in organisms, DNA possesses excellent information density and stability.[2−4] Given that each nucleotide base in DNA molecules measures approximately 0.34 nm,[5] DNA can achieve a storage density as high as 6 bits per nanometer. Furthermore, successful extraction of ancient DNA from fossils confirms its remarkable stability under proper preservation conditions.[6] These exceptional properties establish DNA as a highly competitive storage medium, demonstrating promising potential to meet growing data storage needs.[2,7−9] The fundamental processes of DNA data storage include encoding, synthesis, random access, and sequencing. Since Church et al.[10] established a groundbreaking milestone in DNA data storage, years of research have achieved notable progress in these key aspects.[11−13] However, significant challenges persist in read/write throughput and cost efficiency. Essentially, current DNA storage research focuses primarily on two fundamental processes: data input (DNA synthesis) and data output (DNA sequencing).

Current DNA synthesis techniques employ two principal methodologies: chemical synthesis and enzymatic synthesis.[14−16] Conventional chemical synthesis predominantly utilizes established solid-phase phosphoramidite synthesis (SPPS). However, inherent chemical reaction limitations result in restricted nucleotide coupling efficiency (<99.9%), inducing approximately 1% sequence deletion errors per nucleotide under standard synthesis conditions.[17,18] Furthermore, the acid-catalyzed deprotection process during synthesis may trigger depurination, causing DNA chain hydrolysis and subsequent reductions in product yield and purity.[19] As depurination damage can accumulate through successive synthetic cycles, severely constraining long-strand chemical synthesis yields.[20] The conventional solid-phase phosphoramidite chemistry comprises four cyclic steps: deprotection, coupling, capping, and oxidation. Each cycle completes the synthesis of a single nucleotide. Since the current nucleotide synthesis does not affect subsequent processes, each nucleotide incorporation can be treated as an independent event. Consequently, the final product yield follows a Bernoulli distribution, which also reflects the accumulation of synthesis errors. For example, the theoretical yield of 200-nt synthesis is only about 13% ($0.99^{200} \approx 0.134$), and most of the other product will be purified. While array synthesis theoretically

Received: March 7, 2025
Revised: May 30, 2025
Accepted: June 20, 2025
Published: July 4, 2025

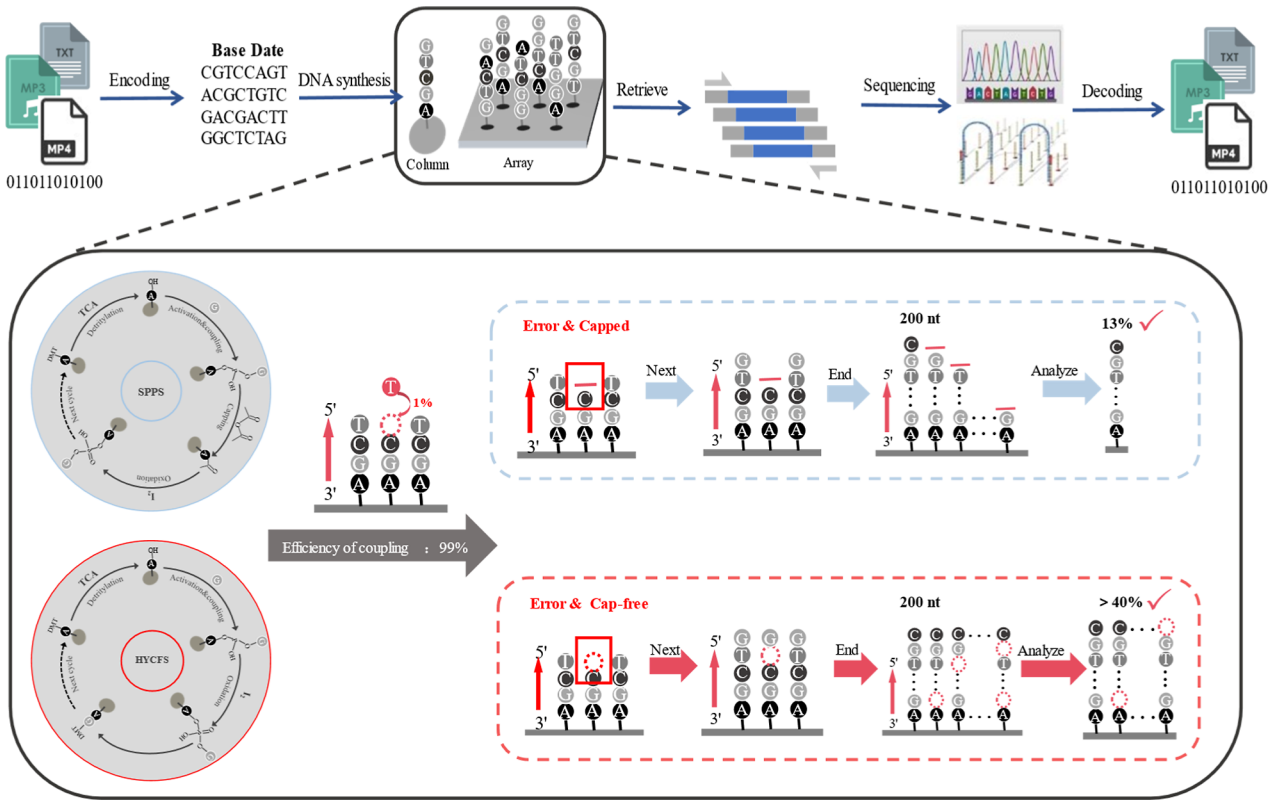


Figure 1. Schematic comparison of synthesis workflows and strand extension patterns between HYCFS and standard SPPS. The HYCFS strategy eliminates the capping step in conventional phosphoramidite chemistry, enabling continuous extension of DNA strands even after coupling failures. For 200-nt sequence synthesis at 99% coupling efficiency, SPPS produces only ~13% storage-effective sequences, whereas HYCFS predominantly generates full-length products and sequences with sparse deletions. The unique error profile of HYCFS allows systematic conversion of erroneous strands into functional sequences during data analysis, achieving >40% yield of useable sequences under single-base deletion assumptions.

offer enhanced data throughput through parallel processing mechanisms to better accommodate large-scale storage needs, empirical evidence from published studies demonstrates that both material deposition-based and electrode array-based implementations exhibit significantly lower synthesis efficiency compared to conventional column-based systems.[13,17,21,22] These limitations fundamentally restrict achievable strand lengths and production yields in array-based synthesis.

In conventional solid-phase chemical synthesis, capping steps are employed to prevent the extension of DNA chains that failed to couple correctly, thus avoiding the production of erroneous sequences. Although the capping step ensures the production of full-length DNA sequences, it turns out that these target products are only a small fraction of the total synthetic yield. It is important to note that although these error-containing products are ineffective for standard DNA applications, they may be repurposed for DNA data storage. By combining these fragments with information processing techniques, they could potentially be converted into valid sequences, thereby significantly improving yield. However, considering subsequent storage procedures including PCR-based random access and sequencing, the postcapping products cannot be directly converted into effective data storage sequences.

To address this limitation, we propose a high-yield capping-free synthesis (HYCFS) strategy. By modifying the workflow of conventional solid-phase phosphoramidite chemical synthesis and removing the capping step (as shown in Figure 1), this approach significantly improves the yield of synthesized sequences, which are then processed through subsequent data conversion protocols to generate valid storage sequences. In this experimental study, we established the theoretical model of HYCFS and implemented this synthetic strategy in both randomized sequences and practical storage cases. Through systematic analysis, we obtained general characteristics of high-yield capping-free synthesis and successfully validated the reliability and practicality of HYCFS-based DNA data storage. Furthermore, compatibility verification confirmed the strategy's effective operation across various synthesis instruments, including both column-based and array-based systems. Through simulation analysis, we further demonstrated the superiority of this strategy in high-throughput synthesis of long DNA sequences. Our findings highlight the potential of capping-free methodologies for achieving scalable and efficient data storage, proposing a transformative solution framework to address the escalating challenge of digital storage demands.

RESULTS

Bias Analysis of Random Sequences Based on HYCFS.

To establish general sequence characteristics for HYCFS implementation in DNA storage systems, we generated randomized DNA sequences ranging from conventional length (120-nt) to extended-length (200-nt) using Python code. In DNA-based data storage systems, biochemical constraints are typically imposed on single-stranded DNA sequences encoding information to minimize errors during chemical synthesis, PCR amplification, and sequencing. Specifically, the GC content is constrained between 40%−60%, and homopolymer lengths are

https://doi.org/10.1021/acssynbio.5c00175
ACS Synth. Biol. XXXX, XXX, XXX−XXX

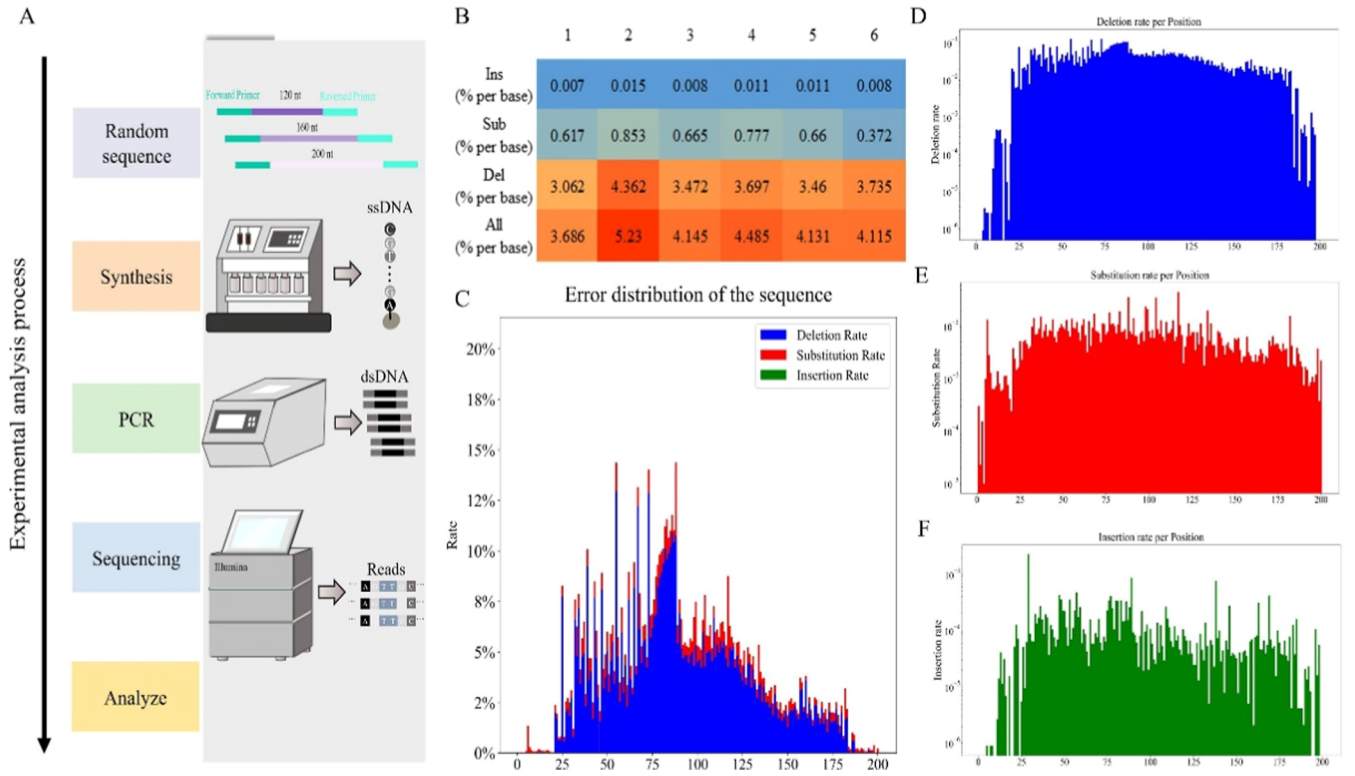


Figure 2. (A) Workflow for error profile analysis of HYCFS-synthesized random sequences, including sequence design, synthesis, amplification, sequencing, and bioinformatic analysis to simulate practical DNA storage applications. (B) Sequence libraries were constructed with three distinct lengths (120, 160, and 200 nucleotides), each containing two unique sequences (six total samples labeled 1−6). Error frequencies for insertions (Ins), substitutions (Sub), and deletions (Del) were quantified. (C) Position-dependent error analysis showing nucleotide-level error rates for (D) deletions, (E) substitutions, and (F) insertions. Extended data are provided in Note S4.

limited to no more than 3 nucleotides. Church et al.[10] first emphasized in their kilobyte-scale digital information storage experiments that avoiding extreme GC content, long repetitive sequences, and potential secondary structures is crucial for achieving reproducibility in DNA synthesis and sequencing. Excessively high or low GC content can promote stable secondary structure formation or reduce polymerase binding efficiency, thereby decreasing synthesis coupling efficiency and causing polymerase stalling and read termination during sequencing. Furthermore, extreme deviations in GC content during PCR amplification can lead to amplification efficiency disparities, resulting in uneven library coverage and sequence dropout.[23] Additionally, consecutive identical nucleotides exceeding 3 bases exacerbate coupling failures during synthesis and exceed the linear detection range of biofluorescence signals in sequencing platforms, inducing insertion or deletion errors.[24] Then we imposed general biochemical constraints on the sequences as restrictions, such as GC content ranging from 40% to 60%, and no excessive homopolymers, etc. During this synthesis process, starting from the 3′ end, the dimethoxytrityl (DMT)-blocked phosphoramidite monomers were deprotected, bases were coupled, and the DNA strands that failed to couple successfully were not capped. Then, the oligonucleotides were oxidized, and the next required monomer was added for the next synthesis cycle. The overall experiment is shown in Figure 2A, where the analysis of random sequences was accomplished through a complete DNA storage process and analysis method.

Previous studies[18,25,26] have demonstrated that chemical synthesis errors may occur during oligonucleotide synthesis. Technical limitations inherent to sequencing platforms also introduce systematic errors, while DNA sequences may undergo mutations during storage or retrieval processes. Collectively, these error-prone mechanisms across DNA data storage workflows (e.g., synthesis, amplification, and sequencing)[26−28] compromise data integrity during readout operations. To address this, modern error-correcting codes[29−31] are conventionally implemented. The HYCFS strategy inherently tolerates coupling failures during synthesis, implying comparable susceptibility to sequence-level errors in DNA storage applications. Given the fundamental similarity in error profiles, HYCFS-generated sequences remain theoretically compatible with existing error-correction paradigms, including error-correcting code integration.

To facilitate subsequent error mitigation, we quantified sequence errors in randomized HYCFS samples using next-generation sequencing (NGS). Deep sequencing via Illumina NovaSeq6000 generated 6,255,322 reads across six samples. Sequencing reads were aligned to reference sequences using Burrows−Wheeler Aligner (BWA), followed by systematic error characterization through a custom analytical pipeline. As anticipated, the observed errors exhibited multifactorial origins, with these results collectively reflecting HYCFS-based storage bias.

Following the established analytical pipeline, we calculated error frequencies (errors per base) and positional error rates (substitutions, insertions, or deletions) across synthesized samples, with intersample comparisons detailed in Figure S8. The results reveal consistent HYCFS-based storage bias patterns across sequences of varying lengths (Figure 2B). In

https://doi.org/10.1021/acssynbio.5c00175
ACS Synth. Biol. XXXX, XXX, XXX−XXX

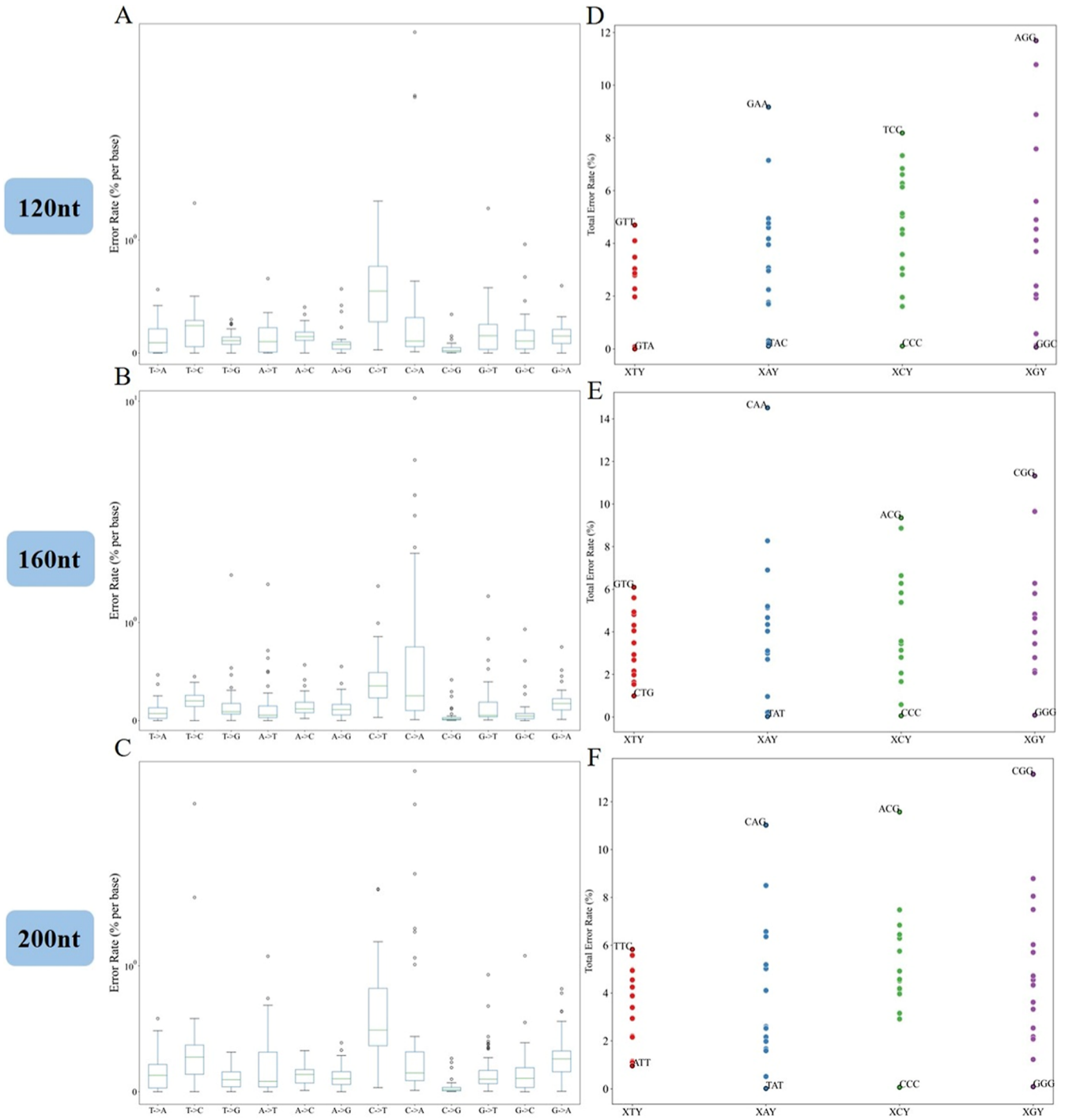


Figure 3. Analysis of sequence error patterns in relation to nucleotide context. (A−C) Substitution error profiles across 120-nt, 160-nt, and 200-nt sequences were statistically analyzed to determine intrinsic base-specific error propensities. (D−F) Correlation analysis of error rates with adjacent nucleotide contexts. Each base position's error profile was jointly analyzed with its flanking nucleotides (5′ and 3′ neighbors). Each data point corresponds to a 3-mer within payload regions, color-coded by central base identity (T: red, A: blue, C: green, G: purple). Error rates were calculated with weighted averaging to address unequal occurrence frequencies among 3-mer combinations. Results exhibited consistent patterns across all sequence lengths. Extended data sets are presented in Figures S9 and S10.

Figure 2C,D,E,F present identical data sets for deletion, insertion, and substitution errors, but emphasize different analytical perspectives. Figure 2C displays comprehensive error type analysis across 200-nt sequences, visually demonstrating the proportional distribution of deletions, insertions, and substitutions at various positions, revealing that deletions constitute the predominant error type in the HYCFS strategy. Furthermore, Figure 2D−F illustrate the positional distribution patterns of individual error types, elucidating the relationship between specific error categories and synthesis positions. As shown in Figure 2C, deletion (Del) errors dominated the error profile with an average rate of 3−4% per base, followed by substitution (Sub) errors at less than 1% per base. The insertion error rate demonstrated in Figure 2F remains below

https://doi.org/10.1021/acssynbio.5c00175
ACS Synth. Biol. XXXX, XXX, XXX−XXX

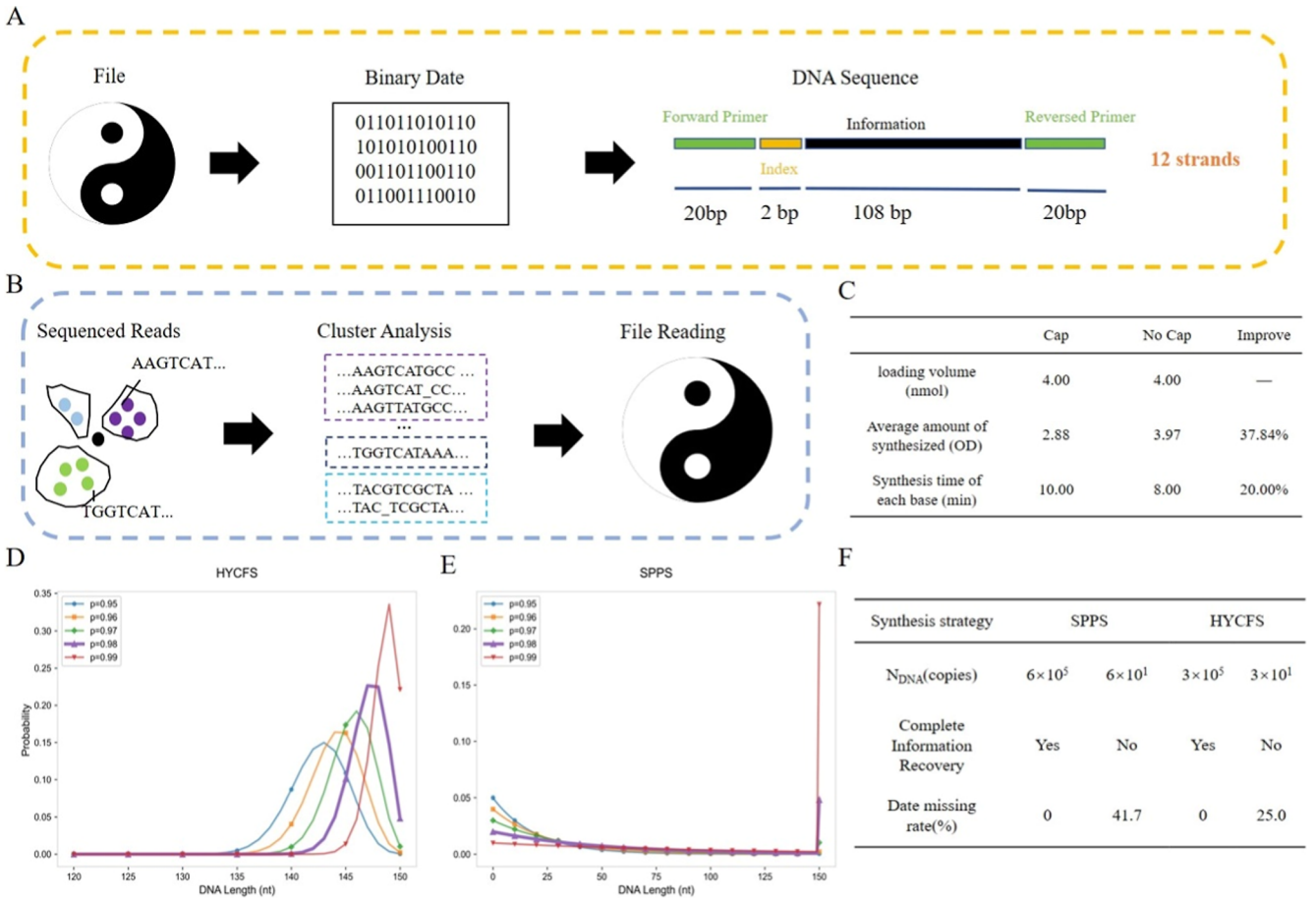


Figure 4. (A) File encoding workflow. Taiji image data were converted to binary format and subsequently encoded into nucleotide sequences. Designed sequences contain flanking 20-nt primer regions, a 2-nt index following the forward primer, and a 108-nt payload region for data storage. (B) File decoding workflow. Sequencing reads were clustered by similarity, with consensus sequences reconstructed from each cluster prior to decoding into original information. (C) Comparative synthesis of storage sequences via HYCFS and SPPS, with synthesis parameters recorded. Theoretical predictions of product length distributions for (D) HYCFS and (E) SPPS synthesized sequences based on established models. Extended analytical data are provided in Figure S2. (F) Analysis of sequence recovery for samples with different copy numbers.

0.1%, exhibiting orders-of-magnitude reduction compared to deletion and substitution error rates, contributing minimally to total errors (All).

This distribution contrasts with reported SPPS-based sequences,[30] where substitution errors predominate over deletions. This discrepancy may arise from postsynthesis purification in SPPS workflows, which shifts main error origins to amplification and sequencing phases. This occurs because in conventional SPPS applications for ssDNA synthesis, column-synthesized samples typically undergo postsynthesis purification processes (e.g., PAGE), which effectively eliminate most deletion errors, thereby predominantly retaining correct target sequences. Consequently, subsequent analyses primarily detect substitution errors introduced during amplification and sequencing stages. In contrast, the HYCFS strategy intentionally omits purification steps to better simulate large-scale array synthesis scenarios where high-precision purification is impractical. This approach preserves synthesis-derived deletion errors while maintaining comparable substitution error rates from amplification/sequencing processes, albeit with lower relative proportions compared to deletion errors. As demonstrated in Figure 2C, deletion errors emerge as the predominant error type under HYCFS conditions. This methodological design facilitates more realistic scalability analysis for HYCFS in large-scale storage applications. Analysis of error type versus sequence position (Figure 2D−F) further demonstrated reduced error rates at terminal regions compared to uniformly distributed errors across internal positions. This observation aligns with the HYCFS theoretical model, which predicts statistical independence between synthesis cycles, leading to uniform error propagation. Reduced terminal errors are likely attributable to primer region screening during amplification and sequencing, which partially eliminates errors at the ends.

Base-Level Characterization of HYCFS Samples. Consistent with observations from conventional SPPS synthesis, substitution errors remain a significant byproduct of chemical oligonucleotide synthesis, with error profiles exhibiting base-specific correlations. To precisely characterize HYCFS-specific storage bias, we performed base-level analyses (A, T, C, G) as illustrated in Figure 3A−C. Notably, cytosine (C) demonstrated the highest substitution error rates across all samples. Using the 160-nt sample as an example, cytosine exhibited substitution errors at an average rate of 0.46% per base (s.d. 1.16%), representing a 3.09-fold increase compared to thymine—the base with the lowest observed substitution frequency at 0.15% per base (s.d. 0.17%). Among C-substitution pathways, C to A transitions predominated

https://doi.org/10.1021/acssynbio.5c00175
ACS Synth. Biol. XXXX, XXX, XXX−XXX

(0.94% per base), potentially attributable to oxidative DNA damage during PCR amplification.[32] Secondary C to T substitutions (0.32% per base) likely originate from cytosine deamination to uracil, subsequently recognized as thymine by DNA polymerases.[33] Notably, G-to-A and T-to-C substitutions—which are prominently observed in standard SPPS-synthesized samples—exhibited frequencies below average values in HYCFS-derived sequences (0.20%, s.d. 0.14% and 0.19%, s.d. 0.11%, respectively). These substitutions may be attributed to guanine and thymine amination. In SPPS workflows, capping-phase acetic anhydride-mediated acylation reactions likely predominantly drive these substitutions.[34]

Furthermore, our analysis revealed that error rates may depend not only on nucleotide identity but also on adjacent base composition.[30] To investigate this, we conducted 3-mer-based analyses to assess error-state correlations with flanking nucleotides. As shown in Figure 3D−F, error rate variations across distinct neighboring base contexts demonstrated statistically significant increases in mononucleotide repeat regions compared to heterogeneous triplet configurations. This observation supports the necessity of implementing homopolymer constraints during sequence design to mitigate error propagation.

HYCFS-Based DNA Storage Application Case Profiling. Having established the universal sequence bias patterns in HYCFS, we proceeded to validate its practical implementation by storing an SVG-format Taiji diagram within DNA sequences synthesized through HYCFS. The complete encoding and decoding workflow is depicted in Figure 4A,B, with further details described in Note S5. Utilizing the Storage-D platform, we first translated the Taiji image into base data. Additional identical 20-nt primer sequences were appended to both termini of the encoded DNA constructs. A unique 2-nt index was incorporated immediately downstream of the forward primer to enable precise sequence reconstruction during decoding, with technical specifications provided in the Section Methods and Table S3. All designed sequences were synthesized using both SPPS and HYCFS strategies to generate comparative sample sets. Throughout this process, HYCFS-based synthesis demonstrated a 20% reduction in average per-base synthesis time compared to SPPS, coupled with a 37.84% increase in mean sequence yield (Figure 4C), findings that align with our theoretical projections. Additionally, random sampling of the synthesized products was performed and characterized via capillary electrophoresis. The experimental results closely aligned with theoretical simulations, with detailed profiles provided in Figure S11.

The same optical density (OD) value of DNA was then selected from the two synthetic products as experimental samples. Subsequently, identical dilution procedures were applied to obtain sample solutions with multiple concentration gradients. We systematically selected DNA aliquots with varying copy numbers for NGS-based data retrieval. This experimental phase specifically evaluated HYCFS's intrinsic error resilience—defined as the capacity to reconstruct stored information under observed error profiles—through comparative analysis against SPPS-synthesized controls. The decoding process exclusively employed data available during actual storage retrieval operations, meaning that no additional information—including encoded reference sequences—was utilized beyond the sequencing-obtained data.

To enable precise quantitative analysis, we implemented a length distribution model for mass-weighted correction of synthetic products, thereby obtaining accurate relative molecular mass values to calculate sequencing sample copy numbers. This approach allowed for a more precise determination of the relative molecular mass, which in turn facilitated an accurate calculation of the copy number of the sequencing samples. Analysis of the sequencing results for the synthetic samples revealed that, when higher concentrations of samples (with copy numbers on the order of 10^6) were utilized for sequencing, the HYCFS synthetic samples successfully read all information sequences and restored the correct data. This outcome was comparable to that achieved with SPPS samples (Figure 4F), further validating the feasibility of employing HYCFS in DNA data storage.

In addition, we conducted experiments with low-concentration samples (with copy numbers on the order of 10) to simulate scenarios with limited synthesis yields in high-throughput array fabrication. Under these conditions, we evaluated the data recovery capabilities of both synthesis methods. Our observations revealed sequence loss in both HYCFS- and SPPS-synthesized samples. Notably, the HYCFS-derived samples exhibited less severe data loss compared to SPPS-synthesized counterparts, primarily due to the error-scattered products that remained analyzable through clustering algorithms. These residual sequences could still generate consensus sequences through alignment, thereby enabling accurate sequence retrieval. This result indicates superior data recovery performance of HYCFS-based samples over SPPS-based counterparts at equivalent copy numbers. Quantitative analysis demonstrated that SPPS-synthesized samples required approximately 3-fold higher quantities than HYCFS samples to achieve comparable data retrieval capacity. These experimental results align with theoretical predictions from our model (Figure 4D,E), which estimated effective yield rates of 22% for SPPS versus 56% for HYCFS. The consistency between theoretical projections and experimental outcomes confirms the enhanced performance of HYCFS methodology under low-concentration conditions. These results collectively suggest that the HYCFS strategy exhibits greater advantages in low-concentration environments, particularly in array-based synthesis applications.

DISCUSSION

In this study, HYCFS-synthesized samples—including both randomly designed sequences for bias analysis and information-encoded storage sequences—demonstrated accurate sequence reconstruction through sequencing, confirming the feasibility of applying HYCFS to DNA data storage for reliable data encoding and retrieval. Notably, this reconstruction was achieved using exclusively sequencing-derived data through cluster-based alignment to generate consensus sequences matching target templates, consistent with predictions from our theoretical model. The HYCFS strategy essentially involves repeated independent Bernoulli trials during synthesis. As shown in the theoretical product length distribution (Figure S2), synthesized products primarily consist of full-length sequences and the sequences with a minority containing scattered deletions uniformly distributed across positions. This predictable error pattern enables accurate sequence recovery through intersequence comparison. Given that these recoverable sequences constitute the majority of product, the self-correction capability remains functional even with low-concentration HYCFS samples during decoding. We have validated HYCFS compatibility with conventional phosphor-

https://doi.org/10.1021/acssynbio.5c00175
ACS Synth. Biol. XXXX, XXX, XXX−XXX

amidite chemistry instruments through successful implementation on two distinct column-based DNA synthesizers. Notably, as HYCFS represents an optimization of fundamental phosphoramidite synthesis principles, it theoretically maintains applicability across all phosphoramidite-based platforms, including but not limited to high-throughput array systems and microfluidic synthesis devices. However, practical performance may vary depending on system-specific material properties and substrate handling protocols, necessitating further empirical validation for different implementation scenarios.

Furthermore, the presented data provide accurate error analysis profiles for HYCFS-based sequences. We observed that the error types fundamentally match those in standard SPPS modes, indicating that error-correction methods from previous studies remain applicable. Moreover, these analytical results can guide encoding design optimization. For instance, given the predominance of deletion errors in HYCFS, additional error-correction redundancy specifically effective for deletion correction could be incorporated during sequence design. Concurrently, by integrating base-state analysis results, nucleotide usage can be strategically balanced during sequence design, with error-prone patterns serving as additional filtration criteria. Although this study quantified the overall error rate, certain variations in storage bias were not adequately explained, such as the differential performance of C-to-A and C-to-T substitution types across sequences of varying lengths. We preliminarily attribute these discrepancies to potential associations with sequence-specific characteristics, including nucleotide frequency and adjacent base-pair interactions. Furthermore, specific error sources and coverage bias were not analyzed. While these limitations do not affect the comparative conclusions drawn in this paper, systematic investigation of error rates and coverage bias in HYCFS-based DNA storage workflows could facilitate the development of encoding systems better adapted to HYCFS, thereby further enhancing the reliability and practical utility of HYCFS strategies. Such optimizations would reduce error frequencies and improve decoding efficiency, thereby enhancing HYCFS's practical utility.

We validated in column-based DNA synthesis that HYCFS exhibits superior performance in DNA storage applications, with significant improvements in terms of time cost and total synthesis yield, and notably, the yield of effective sequences increased by approximately 3-fold, consistent with the predictions of the theoretical model. Furthermore, with regard to array synthesis, which is more widely applied in large-scale data storage, our theoretical model predictions (Note S3) indicate that the advantages of HYCFS will become even more pronounced. Considering that synthesis efficiency on array surfaces is often below 98% and further decreases with increasing synthesis length,[35] for example, in the synthesis of 300-nt long DNA strands, the effective product yield achieved using the conventional SPPS strategy is less than 0.23%, whereas the HYCFS strategy can sustain effective sequence yields between 40% and 60%, corresponding to an improvement factor ranging from 4-fold to several hundred-fold. From another perspective, the enhanced yield of effective products indicates that the application of the HYCFS strategy allows for achieving the same data retrieval performance with a reduced synthesis quantity.

Specifically, during array synthesis, it enables the use of smaller array spot sizes for data storage, with an anticipated capacity of up to 25×10^8 DNA strands per square centimeter.[13] Given that storing 1 TB of data requires approximately 10^{10} 150-nt oligonucleotide strands,[11] the use of HYCFS for synthesizing longer DNA sequences (>200-nt) could enable TB-scale storage capacity per square centimeter. HYCFS strategy holds broad application prospects, with the potential to enhance storage capacity by 2 orders of magnitude under the current storage system, reduce storage costs, and further advance the development of large-scale DNA data storage technology.

METHODS

The Establishment of Theoretical Model. In standard DNA synthesis processes, each base coupling step can be modeled as an independent Bernoulli trial, expressed probabilistically as $X \sim B(1, p)$ where p denotes the base coupling efficiency. Throughout our experimental investigation, we employed average coupling efficiency values to simplify calculations by replacing step-specific parameters.

When synthesizing DNA strands via SPPS, the capping of failure sequences fundamentally alters synthesis dynamics—only successfully coupled strands continue elongation. The resultant length distribution model postsynthesis is formulated as shown in eq SI, where Y represents final strand length, L denotes target length, and k corresponds to the number of successful coupling steps.

When employing the HYCFS strategy for DNA sequence synthesis, coupling failure events do not terminate the reaction process. This synthesis mode can be modeled as L independent Bernoulli trials, following a binomial distribution $X \sim B(L, p)$. The corresponding DNA strand length distribution model is shown in eq SII, where Y represents the final synthesized chain length, L denotes the theoretical target length, and k indicates the actual number of successfully synthesized bases.

These two equations not only characterize the length distribution patterns of synthesis products under each methodology but also describe the profiles of storage-effective sequences. Notably, while nontarget products from both synthesis strategies exhibit deletion errors at their core, their manifestation modes differ fundamentally (Figure 1). In SPPS workflows, uncapped sequences with coupling failures are excluded from subsequent synthesis cycles, resulting in terminal-truncated deletions concentrated at the 5′-end of nontarget sequences. Conversely, HYCFS permits full-length synthesis of all oligonucleotides regardless of coupling efficiency, leading to stochastically distributed deletions across sequence positions. This mechanistic divergence directly impacts downstream processing outcomes. Storage sequences typically incorporate primer regions at both termini for amplification and sequencing accessibility. SPPS-based nontarget products with terminal-truncated deletions show impaired primer binding capacity during PCR amplification, substantially increasing sequence dropout risks. In contrast, HYCFS-generated sequences with scattered deletions retain functional primer regions, enabling successful sequence retrieval and subsequent conversion into storage-effective sequences for data reconstruction.

Data Encoding and Decoding. During the design of randomized sequences and primers, we implemented generalized constraint parameters (GC content 40%−60%, forbidden excessive homopolymer) to simulate universal DNA storage conditions. We generated a series of different sequences (samples 1−6). The corresponding algorithmic architectures

https://doi.org/10.1021/acssynbio.5c00175
ACS Synth. Biol. XXXX, XXX, XXX−XXX

were developed using Python programming assistance, with full code implementations documented in Note S1, and detailed sequence information is provided in Table S2.

For comprehensive evaluation of HYCFS in DNA data storage applications, we conducted digital-to-base conversion of Taiji diagram files (2216 bits) through the Storage-D platform[36] (http://storage.dailab.xyz:16666/). Employing the Wukong algorithm with enhanced coding density configurations (sequence length = 108-nt, GC range = 39%−61%, Codec Pin = default), we successfully generated 12 optimized DNA sequences (samples TJ1−TJ12), detailed in Table S3.

DNA Sample Preparation. Random sequence samples were synthesized using a Syn-HCY-192P DNA synthesizer (Beijing Tsingke Biotech Co., Ltd.) on 1 μmol-scale controlled pore glass (CPG) solid carriers. Storage case samples were prepared using a YB-192/768 synthesizer (Shanghai Dynegene Biotech Co., Ltd.) with an initial synthesis scale of 4 nmol. All DNA specimens were synthesized through execution of two distinct phosphoramidite-based synthesis protocols. The standard method involved synthesis cycles comprising deprotection, coupling, capping, and oxidation steps. In contrast, the cap-free method eliminated capping steps, retaining only deprotection, coupling, and oxidation cycles.

Following synthesis via both protocols, all oligonucleotides underwent standardized postsynthesis processing: cleavage from CPG solid supports through ammonia vapor treatment at ambient temperature, followed by desalting of mixtures. The resulting products were subsequently implemented experimentally without additional purification. The synthesized products were quantified using the Implen NanoPhotometer N60 spectrophotometer.

Selectively Amplifying DNA. All synthesized DNA samples were reconstituted in RNase-free H_2O according to manufacturer specifications. Serial 10-fold dilutions were performed as required. PCR amplifications were conducted using the Gentier 96R system (Xi'an Tianlong Science and Technology Co., Ltd.) with the following thermal profile: initial denaturation at 95 °C for 4 min, 30 cycles of denaturation (95 °C, 30 s), annealing (60 °C, 30 s), and extension (74 °C, 30 s). Cycle numbers were adjusted based on experimental requirements. Each PCR reaction mixture had a final volume of 20 μL, with detailed compositions provided in Tables S4 and S5. All amplification primers synthesized by Dynegene and their sequences are cataloged in Table S1.

Library Preparation and High Throughput Sequencing. Library preparation was performed by two step PCR. First round PCR reaction was set up as follows: DNA (10 ng/μL) 2 μL; amplicon PCR forward primer mix (10 μM) 1 μL; amplicon PCR reverse primer mix (10 μM) 1 μL; 2 × PCR Ready Mix 15 μL (total 25 μL) (Kapa HiFi Ready Mix). The plate was sealed and PCR performed in a thermal instrument (BIO-RAD, T100TM) using the following program: 1 cycle of denaturing at 98 °C for 5 min, then 8 cycles of denaturing at 98 °C for 30 s, annealing at 60 °C for 30 s, elongation at 72 °C for 30 s, and a final extension at 72 °C for 5 min. Finally hold at 4 °C. The PCR products were checked using electrophoresis in 1% (w/v) agarose gels in TBE buffer (Tris, boric acid, EDTA) stained with SYBR Green I and visualized under UV light. Then we used AMPure XP beads to purify the amplicon product.

After that, the second round PCR was performed. PCR reaction was set up as follows: DNA (10 ng/μL) 2 μL; universal P7 primer with index (10 μM) 1 μL; P5 primer with index (10 μM) 1 μL; 2× PCR Ready Mix 15 μL (total 30 μL) (Kapa HiFi Ready Mix). The plate was sealed and PCR performed in a thermal instrument (BIO-RAD, T100TM) using the following program: 1 cycle of denaturing at 98 °C for 3 min, then 5 cycles of denaturing at 94 °C for 30 s, annealing at 55 °C for 20 s, elongation at 72 °C for 30 s, and a final extension at 72 °C for 5 min. Then we used AMPure XP beads to purify the amplicon product. The libraries were then quantified and pooled. Paired-end sequencing of the library was performed on the NovaSeq6000 with PE150 model (Illumina, San Diego, CA).

Raw Sequencing Data Preprocessing. We have deposited the relevant raw data in the NCBI database under the accession number PRJNA1269461 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1269461) for review. Raw reads were filtered according to three steps: (1) Removing adaptor sequence if reads contains by cutadapt (v 1.2.1), the main procedures and parameters are as follows: cutadapt -a {adapter}10 -o trimed.fastq raw.fastq, (2) removing low quality bases from reads 3′ to 5′ ($Q < 20$) by PRINSEQ-lite (v 0.20.3), (3) removing chimras sequence by usearch software (v11.0.667) with de novo mode by default parameter, the main procedures and parameters are as follows: usearch -uchime3_-denovo uniq.fasta -nonchimeras non_chimeras.fasta.

Error Rate Statistics and Data Visualization. Following the data processing steps, custom Python programs were used to identify target sequences through reference sequence matching and perform multiple sequence alignment. Based on the alignment results, four types of sequence were quantified: Match, Deletion, Insertion, and Substitution. For each position, the count of each error type was divided by the total read count and multiplied by 100% to compute error rates. Visualization plots were subsequently generated using custom Python programs according to these statistical results. Additional details are provided in Note S4.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acssynbio.5c00175.

The comprehensive design rationale and detailed information on all primer and oligonucleotide sequences; the amplification system composition and thermocycling parameters for DNA samples; elucidation of the theoretical framework with corresponding analytical outcomes; detailed error profile analysis of experimental specimens; capillary electrophoretic characterization results for storage samples (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Quanjun Liu − *State Key Laboratory of Digital Medical Engineering, School of Biological Science and Medical Engineering, Southeast University, Nanjing 210096, China; Southeast University Shenzhen Research Institute, Shenzhen 518063, China*; Email: lqj@seu.edu.cn

Authors

Weiming Lin − *State Key Laboratory of Digital Medical Engineering, School of Biological Science and Medical Engineering, Southeast University, Nanjing 210096, China*; orcid.org/0009-0002-3931-3330

https://doi.org/10.1021/acssynbio.5c00175
ACS Synth. Biol. XXXX, XXX, XXX−XXX

Haotian Yu − *State Key Laboratory of Digital Medical Engineering, School of Biological Science and Medical Engineering, Southeast University, Nanjing 210096, China*
Weihao Li − *State Key Laboratory of Digital Medical Engineering, School of Biological Science and Medical Engineering, Southeast University, Nanjing 210096, China*
Yemin Han − *State Key Laboratory of Digital Medical Engineering, School of Biological Science and Medical Engineering, Southeast University, Nanjing 210096, China*
Manman Lv − *State Key Laboratory of Digital Medical Engineering, School of Biological Science and Medical Engineering, Southeast University, Nanjing 210096, China*
Han Gao − *State Key Laboratory of Digital Medical Engineering, School of Biological Science and Medical Engineering, Southeast University, Nanjing 210096, China*
Mengqing Cheng − *State Key Laboratory of Digital Medical Engineering, School of Biological Science and Medical Engineering, Southeast University, Nanjing 210096, China*
Yan Huang − *State Key Laboratory of Digital Medical Engineering, School of Biological Science and Medical Engineering, Southeast University, Nanjing 210096, China*
Kun Bi − *State Key Laboratory of Digital Medical Engineering, School of Biological Science and Medical Engineering, Southeast University, Nanjing 210096, China*
Zuhong Lu − *State Key Laboratory of Digital Medical Engineering, School of Biological Science and Medical Engineering, Southeast University, Nanjing 210096, China*

Complete contact information is available at: https://pubs.acs.org/10.1021/acssynbio.5c00175

Author Contributions

Q.L. and Z.L. provided conceptual idea which is possible to apply to DNA data storage. Weiming Lin designed the study and performed all experiments. Weihao Li, Yemin Han, and M.L. established protocol of PCR amplification and NGS library preparation. Haotian Yu, H.G., M.C. participated in data analysis and discussions. Yan Huang and K.B. examined all experimental data and sequencing results. All authors reviewed and edited the manuscript.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors acknowledge the Beijing Tsingke Biotech Co., Ltd. and Dynegene Biotech Co., Ltd. for its support with DNA synthesis during this work. This work was supported by the National Natural Science Foundation of China (Grant Nos. 61827814 and 62201141), National key research and development project (2020YFA0712104) and Shenzhen Science and Technology Program (award number: JCYJ20230807114612024 and JCYJ20220530160416036).

■ REFERENCES

(1) Kim, S. J.; Jung, W. B.; Jung, H. S.; Lee, M. H.; Heo, J.; Horgan, A.; Godron, X.; Ham, D. J. M. b. The bottom of the memory hierarchy: Semiconductor and DNA data storage. *MRS Bull.* **2023**, *48*, 547−559.
(2) Wang, S.; Mao, X.; Wang, F.; Zuo, X.; Fan, C. Data Storage Using DNA. *Adv. Mater.* **2024**, *36*, No. e2307499.
(3) Sun, F.; Dong, Y.; Ni, M.; Ping, Z.; Sun, Y.; Ouyang, Q.; Qian, L. Mobile and Self-Sustained Data Storage in an Extremophile Genomic DNA. *Adv. Sci.* **2023**, *10*, No. e2206201.
(4) Zhirnov, V.; Zadegan, R. M.; Sandhu, G. S.; Church, G. M.; Hughes, W. L. J. N. M. Nucleic acid memory. *Nat. Mater.* **2016**, *15*, 366−370.
(5) Chi, Q.; Wang, G.; Jiang, J. The persistence length and length per base of single-stranded DNA obtained from fluorescence correlation spectroscopy measurements using mean field theory. *Phys. A* **2013**, *392*, 1072−1079.
(6) Allentoft, M. E.; Collins, M.; Harker, D.; Haile, J.; Oskam, C. L.; Hale, M. L.; Campos, P. F.; Samaniego, J. A.; Gilbert, M. T.; Willerslev, E.; Zhang, G.; Scofield, R. P.; Holdaway, R. N.; Bunce, M. The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils. *Proc. R. Soc. B* **2012**, *279*, 4724−4733.
(7) Meiser, L. C.; Nguyen, B. H.; Chen, Y. J.; Nivala, J.; Strauss, K.; Ceze, L.; Grass, R. N. Synthetic DNA applications in information technology. *Nat. Commun.* **2022**, *13*, 352.
(8) Organick, L.; Chen, Y. J.; Dumas Ang, S.; Lopez, R.; Liu, X.; Strauss, K.; Ceze, L. Probing the physical limits of reliable DNA data retrieval. *Nat. Commun.* **2020**, *11*, 616.
(9) Extance, A. How DNA could store all the world's data. *Nature* **2016**, *537*, 22−24.
(10) Church, G. M.; Gao, Y.; Kosuri, S. Next-Generation Digital Information Storage in DNA. *Science* **2012**, *337*, 1628.
(11) Weng, Z.; Li, J.; Wu, Y.; Xiu, X.; Wang, F.; Zuo, X.; Song, P.; Fan, C. Massively parallel homogeneous amplification of chip-scale DNA for DNA information storage (MPHAC-DIS). *Nat. Commun.* **2025**, *16*, 667.
(12) Zhang, C.; Wu, R.; Sun, F.; Lin, Y.; Liang, Y.; Teng, J.; Liu, N.; Ouyang, Q.; Qian, L.; Yan, H. Parallel molecular data storage by printing epigenetic bits on DNA. *Nature* **2024**, *634*, 824−832.
(13) Nguyen, B. H.; Takahashi, C. N.; Gupta, G.; Smith, J. A.; Rouse, R.; Berndt, P.; Yekhanin, S.; Ward, D. P.; Ang, S. D.; Garvan, P.; Parker, H. Y.; Carlson, R.; Carmean, D.; Ceze, L.; Strauss, K. Scaling DNA data storage with nanoscale electrode wells. *Sci. Adv.* **2021**, *7*, No. eabi6714.
(14) Lee, H. H.; Kalhor, R.; Goela, N.; Bolot, J.; Church, G. M. Terminator-free template-independent enzymatic DNA synthesis for digital information storage. *Nat. Commun.* **2019**, *10*, 2383.
(15) Hoose, A.; Vellacott, R.; Storch, M.; Freemont, P. S.; Ryadnov, M. G. DNA synthesis technologies to close the gene writing gap. *Nat. Rev. Chem.* **2023**, *7*, 144−161.
(16) Kosuri, S.; Church, G. M. Large-scale de novo DNA synthesis: technologies and applications. *Nat. Methods* **2014**, *11*, 499−507.
(17) Lietard, J.; Leger, A.; Erlich, Y.; Sadowski, N.; Timp, W.; Somoza, M. M. Chemical and photochemical error rates in light-directed synthesis of complex DNA libraries. *Nucleic Acids Res.* **2021**, *49*, 6687−6701.
(18) LeProust, E. M.; Peck, B. J.; Spirin, K.; McCuen, H. B.; Moore, B.; Namsaraev, E.; Caruthers, M. H. Synthesis of high-quality libraries of long (150mer) oligonucleotides by a novel depurination controlled process. *Nucleic Acids Res.* **2010**, *38*, 2522−2540.
(19) An, R.; Jia, Y.; Wan, B.; Zhang, Y.; Dong, P.; Li, J.; Liang, X. Non-enzymatic depurination of nucleic acids: factors and mechanisms. *PLoS Ones* **2014**, *9*, No. e115950.
(20) Septak, M. Kinetic studies on depurination and detritylation of CPG-bound intermediates during oligonucleotide synthesis. *Nucleic Acids Res.* **1996**, *24*, 3053−3058.
(21) Xu, C.; Ma, B.; Gao, Z.; Dong, X.; Zhao, C.; Liu, H. Electrochemical DNA synthesis and sequencing on a single electrode with scalability for integrated data storage. *Sci. Adv.* **2021**, *7*, No. eabk0100.
(22) Egeland, R. D.; Southern, E. M. Electrochemically directed synthesis of oligonucleotides for DNA microarray fabrication. *Nucleic Acids Res.* **2005**, *33*, No. e125.
(23) Aird, D.; Ross, M. G.; Chen, W.-S.; Danielsson, M.; Fennell, T.; Russ, C.; Jaffe, D. B.; Nusbaum, C.; Gnirke, A. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol.* **2011**, *12*, R18.

(24) Ross, M. G.; Russ, C.; Costello, M.; Hollinger, A.; Lennon, N. J.; Hegarty, R.; Nusbaum, C.; Jaffe, D. B. Characterizing and measuring bias in sequence data. *Genome Biol.* **2013**, *14*, R51.
(25) Ma, X.; Shao, Y.; Tian, L.; Flasch, D. A.; Mulder, H. L.; Edmonson, M. N.; Liu, Y.; Chen, X.; Newman, S.; Nakitandwe, J.; Li, Y.; Li, B.; Shen, S.; Wang, Z.; Shurtleff, S.; Robison, L. L.; Levy, S.; Easton, J.; Zhang, J. Analysis of error profiles in deep next-generation sequencing data. *Genome Biol.* **2019**, *20*, 50.
(26) Heckel, R.; Mikutis, G.; Grass, R. N. A Characterization of the DNA Data Storage Channel. *Sci. Rep.* **2019**, *9*, 9663.
(27) Gimpel, A. L.; Stark, W. J.; Heckel, R.; Grass, R. N. A digital twin for DNA data storage based on comprehensive quantification of errors and biases. *Nat. Commun.* **2023**, *14*, 6026.
(28) Chen, Y.-J.; Takahashi, C. N.; Organick, L.; Bee, C.; Ang, S. D.; Weiss, P.; Peck, B.; Seelig, G.; Ceze, L.; Strauss, K. Quantifying molecular bias in DNA data storage. *Nat. Commun.* **2020**, *11*, 3264.
(29) Erlich, Y.; Zielinski, D. DNA Fountain enables a robust and efficient storage architecture. *Science* **2017**, *355*, 950−954.
(30) Organick, L.; Ang, S. D.; Chen, Y. J.; Lopez, R.; Yekhanin, S.; Makarychev, K.; Racz, M. Z.; Kamath, G.; Gopalan, P.; Nguyen, B.; Takahashi, C. N.; Newman, S.; Parker, H. Y.; Rashtchian, C.; Stewart, K.; Gupta, G.; Carlson, R.; Mulligan, J.; Carmean, D.; Seelig, G.; Ceze, L.; Strauss, K. Random access in large-scale DNA data storage. *Nat. Biotechnol.* **2018**, *36*, 242−248.
(31) Grass, R. N.; Heckel, R.; Puddu, M.; Paunescu, D.; Stark, W. J. Robust chemical preservation of digital information on DNA in silica with error-correcting codes. *Angew. Chem., Int. Ed. Engl.* **2015**, *54*, 2552−2555.
(32) Yeom, H.; Kim, N.; Lee, A. C.; Kim, J.; Kim, H.; Choi, H.; Song, S. W.; Kwon, S.; Choi, Y. Highly Accurate Sequence- and Position-Independent Error Profiling of DNA Synthesis and Sequencing. *ACS Synth. Biol.* **2023**, *12*, 3567−3577.
(33) Kamiya, H. Mutagenic potentials of damaged nucleic acids produced by reactive oxygen/nitrogen species: approaches using synthetic oligonucleotides and nucleotides: survey and summary. *Nucleic Acids Res.* **2003**, *31*, 517−531.
(34) Masaki, Y.; Onishi, Y.; Seio, K. Quantification of synthetic errors during chemical synthesis of DNA and its suppression by non-canonical nucleosides. *Sci. Rep.* **2022**, *12*, 12095.
(35) McGall, G. H.; Barone, A. D.; Diggelmann, M.; Fodor, S. P. A.; Gentalen, E.; Ngo, N. The Efficiency of Light-Directed Synthesis of DNA Arrays on Glass Substrates. *J. Am. Chem. Soc.* **1997**, *119*, 5081−5090.
(36) Huang, X.; Cui, J.; Qiang, W.; Ye, J.; Wang, Y.; Xie, X.; Li, Y.; Dai, J. Storage-D: A user-friendly platform that enables practical and personalized DNA data storage. *iMeta* **2024**, *3*, No. e168.

NOTE ADDED AFTER ASAP PUBLICATION

This paper was published ASAP on July 4, 2025, with the wrong TOC/abstract graphic. The corrected version was reposted on July 8, 2025.

https://doi.org/10.1021/acssynbio.5c00175
ACS Synth. Biol. XXXX, XXX, XXX−XXX