ACS Synthetic Biology

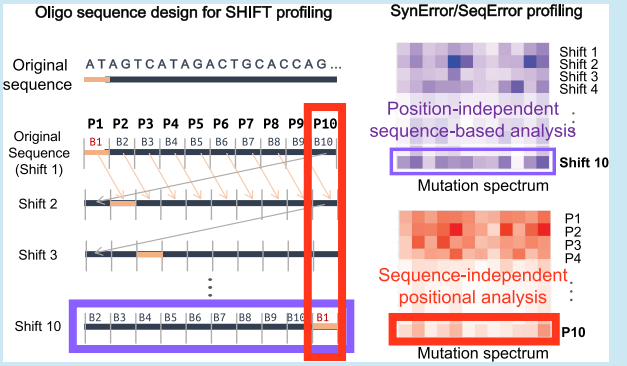
pubs.acs.org/synthbio
Research Article

Highly Accurate Sequence- and Position-Independent Error Profiling of DNA Synthesis and Sequencing

Huiran Yeom,* Namphil Kim, Amos Chungwon Lee, Jinhyun Kim, Hamin Kim, Hansol Choi, Seo Woo Song, Sunghoon Kwon, and Yeongjae Choi*

Cite This: ACS Synth. Biol. 2023, 12, 3567−3577

ABSTRACT: A comprehensive error analysis of DNA-stored data during processing, such as DNA synthesis and sequencing, is crucial for reliable DNA data storage. Both synthesis and sequencing errors depend on the sequence and the transition of bases of nucleotides; ignoring either one of the error sources leads to technical challenges in minimizing the error rate. Here, we present a methodology and toolkit that utilizes an oligonucleotide library generated from a 10-base-shifted sequence array, which is individually labeled with unique molecular identifiers, to delineate and profile DNA synthesis and sequencing errors simultaneously. This methodology enables position- and sequence-independent error profiling of both DNA synthesis and sequencing. Using this toolkit, we report base transitional errors in both synthesis and sequencing in general DNA data storage as well as degenerate-base-augmented DNA data storage. The methodology and data presented will contribute to the development of DNA sequence designs with minimal error.



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

INTRODUCTION

DNA data storage has emerged as an innovative digital data storage medium owing to its durability and storage density.[1−4] As DNA molecules, which have a length per base of 0.34 nm, consist of four different bases (A, T, G, and C), the theoretical information density of DNA molecules greatly surpasses that of conventional electrical media.[5] However, an ideal encoding rate of 2 bits per nucleotide cannot be achieved in practice owing to several limitations, particularly those originating from errors in current DNA writing and reading technologies.[6,7] Numerous errors can arise from both the DNA synthesis and sequencing process, which lead to erroneous insertions, deletions, and substitutions of bases in the data-encoded molecules or decoded data.[8] Synthesis errors usually arise from chemical oligonucleotide (oligo) synthesis technologies based on phosphoramidite chemistry, which have a limited monomer coupling efficiency of <99.9% per nucleotide.[9,10] In other words, state-of-the-art chemical oligo synthesis can only reliably manufacture molecules shorter than 150 nucleotides with an indel error of ∼1% per nucleotide.[11,12] While alternative modes of synthesis such as those utilizing the enzyme terminal deoxynucleotidyl transferase (TdT) have emerged, they introduce their own form of sequence-specific biases and errors.[13−15] Another major constraint is the sequencing error, the rate of which is approximately 0.1% per nucleotide and occurs during the sequencing by synthesis stage in the state-of-the-art next-generation sequencing (NGS) platform.[16,17] These error rates are extremely high compared to those of the conventional data storage medium, which are less than 10^{-6}. Therefore, error minimization is necessary for the practical use of DNA-based data storage.

Errors in DNA synthesis and sequencing processes are fatal when attempting to decode information accurately from DNA and ultimately compromise storage density.[18] Therefore, several error-correction methods have been developed by considering each error pattern in oligo synthesis[19] and NGS.[20] For example, a method for purifying oligos with the single-base resolution was introduced to exclude erroneous strands in oligo synthesis.[21] This method was developed to correct the indels of DNA, which is the major source of synthesis error patterns in oligo pools. In addition, novel binary-to-DNA encoding algorithms, such as the DNA fountain code, impose restrictions on the designed sequences, such that sequences with certain patterns are avoided during encoding.[6] These methods can reduce the amount of redundancy required to

Received: May 15, 2023
Revised: November 1, 2023
Accepted: November 1, 2023
Published: November 14, 2023

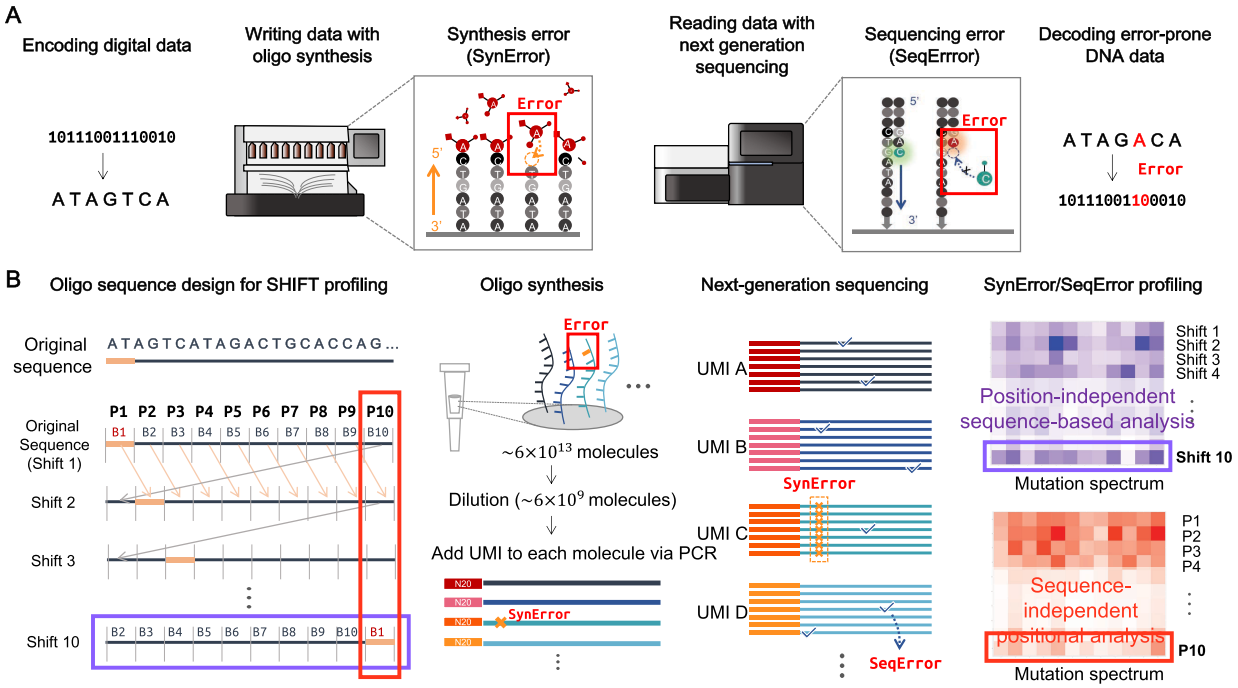


Figure 1. Workflow of DNA sequencing (SeqError) and synthesis error (SynError) profiling using unique molecular identifiers for SHIFT. (A) The DNA data storage workflow consists of encoding digital data and writing the data with oligo synthesis, where SynError occurs; reading the DNA data, where SeqError occurs; and decoding it to digital data. (B) SHIFT was developed to simultaneously profile SeqError and SynError with high accuracy. For sequence design in SHIFT, we shifted 10 base pairs from the original sequence to generate 10 different shifted sequences, or shifts, for position- and sequence-independent error profiling. We synthesized the oligos with the original sequence as well as those with the nine additional shifts. Through tagging each synthesized shift with UMI, amplifying, and sequencing, SeqError and SynError were separated from the sequencing result. SynError was defined as the error occurring at the same position with high frequency within the same UMI group, while SeqError was defined as the error occurring randomly along the sequence with low frequency within the same UMI group.

design a reliable DNA data storage system by utilizing previously reported error-prone characteristics of DNA, such as high GC content and homopolymers.[22] However, previous studies[23−25] have mainly focused solely on one factor, namely either oligo synthesis or DNA sequencing, despite the synthesis and sequencing process being conducted serially and errors accumulating within the same oligo strands mutually[26] (Figure 1A). Potential sources of error must be profiled to determine whether they originate from oligo synthesis or sequencing.[27] Therefore, errors during both encoding and decoding can be minimized by discovering the source of errors, such as specific sequence patterns or positions, according to synthesis or sequencing.

In this regard, both synthesis and sequencing errors must be profiled with high accuracy at the single-base level for each oligo strand,[8,18,28] given that error bias is position- and sequence-dependent in both cases. In terms of position, synthesis errors tend to increase toward the 5′-end, whereas sequencing errors increase toward the 3′-end.[9] In terms of sequences, previous studies[16,29] have reported that certain transition patterns of sequences cause higher sequencing error rates. Thus, whether oligo synthesis exhibits similar or different sequence-dependent patterns must be investigated. In this regard, considering the error pattern according to the position and transition of sequences will increase the accuracy of the DNA data storage system[30] when encoding the digital data to the quaternary data storage system.[31] However, to characterize both positional and base-transitional errors in one sequence, it is important to consider whether the two factors affect each other and to investigate them independently. To determine such parameters, a solution for simultaneously delineating oligo synthesis and sequencing errors with high accuracy from the same oligo strand is needed.

In this study, we present a sequence/position-independent highly accurate error profiling toolkit (SHIFT) that can simultaneously profile DNA synthesis and sequencing errors. The procedure for SHIFT consists of oligo sequence design, oligo synthesis, sequencing, and error analysis (Figure 1A,B). SHIFT designs 10 nucleotide blocks, 'B*N*,' and 10 arrays of block-shifted sequence, "shift *N*" (Figure 1B). Given that the total length of DNA is 100 nt, these 10 shifts cover all cases in which each block is located in different positions from P1 to P10. Therefore, it enables the analysis of error patterns for all positions and sequences in a sequence- and position-independent manner, respectively. Herein, after synthesizing and sequencing the designed oligos, we defined the variant frequency to delineate synthesis and sequencing errors within one oligo. Unique molecular identifiers (UMIs)[32] were utilized to classify each error as a synthesis error (SynError), which appears in the same position in multiple sequences with the same UMI, or a sequencing error (SeqError), which appears sporadically within the UMI group (Figure 1B). In addition, SynError and SeqError can be profiled in a sequence- and position-independent manner using the SHIFT. We analyzed 621,553 oligo strands in total and verified each error pattern, focusing on the transitional sequence and positions along the sequence independently. The findings presented herein provide a comprehensive guide for the design of algorithms and architectures for highly accurate DNA data storage (Supplementary Note).

https://doi.org/10.1021/acssynbio.3c00308
ACS Synth. Biol. 2023, 12, 3567−3577

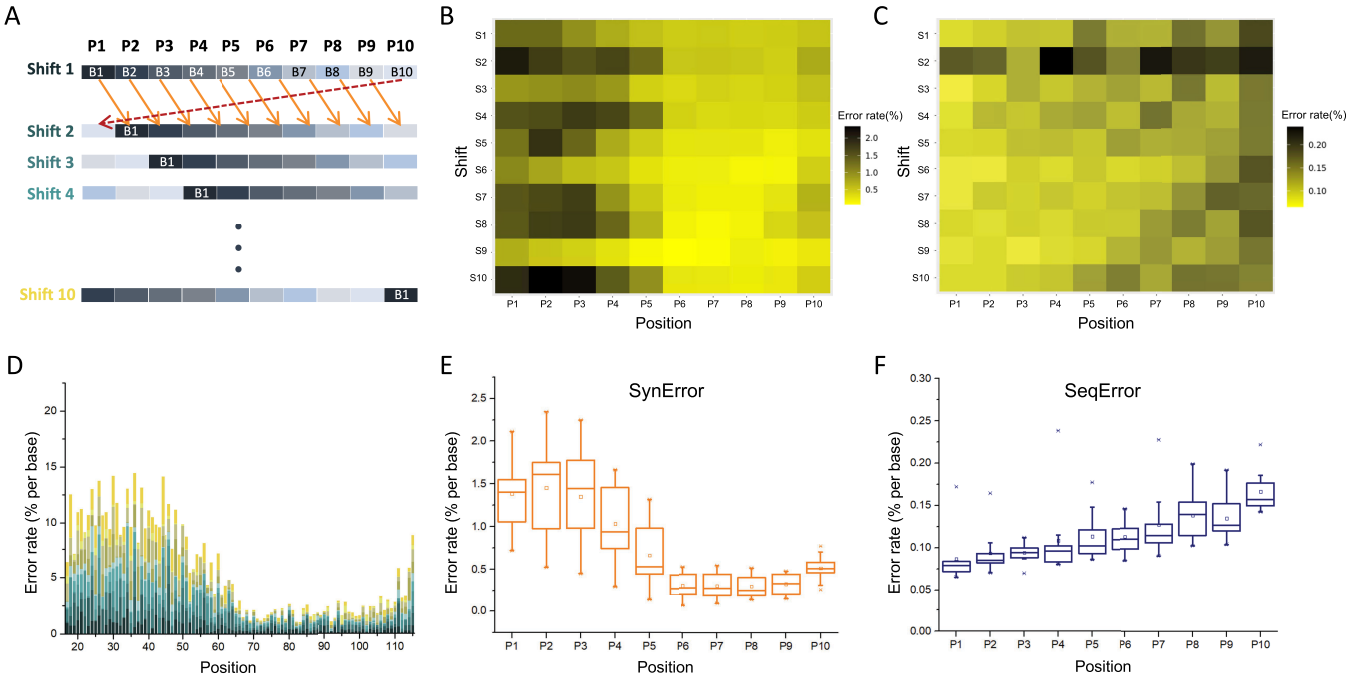


Figure 2. Positional SynError and SeqError were profiled using SHIFT. (A) Sequence design to analyze sequence-independent positional errors profiling. Ten shifted sequences with 10 blocks consisting of 10 nucleotides were shifted to the left by one block. Each position included 10 bases and the position from P1 to P10 is from the region at the 5′-end and to 3′-end within the oligo sequences. SynError (B) and SeqError (C) profiling of the blocks along the position (62,153 UMIs per each shift on average). (D) Accumulated positional error rate of 10 shifted sequences. The errors were distributed along the sequences and accumulated according to shifts 1−10. (E, F) Average error rates of each block of shifts were plotted along the position. The heatmap describes each position error of SynError (E) and SeqError (F) along the sequence for each sequence. Average values: SynError, 0.77% per base (s.d. 0.25%); SeqError, 0.21% (s.d. 0.05%).

RESULTS

Simultaneous Profiling of DNA Synthesis and Sequencing Errors. To differentiate between errors originating from oligo synthesis and sequencing using NGS data, each oligo strand was tagged with UMIs in the primers for PCR.[28] Oligo synthesis begins with dimethoxytrityl (DMT)-blocked phosphoramidite monomers from the 3′-end, followed by deblocking the monomer, coupling the bases, capping the uncoupled bases, oxidizing the oligo, and adding the next desired monomer for the next cycle of synthesis. The synthesis errors are mainly caused by phosphoramidite monomer coupling failure, which leads to deletion errors. In addition, the acid solutions used for deblocking or oxidation lead to substitution errors because of damage, such as depurination.[9] In contrast, sequencing errors, especially for Illumina platforms, originate from dephasing when detecting fluorescence signals in monoclonal DNA clusters by elongation failure by polymerase or incomplete removal of the terminator.[33] We assumed that the synthesis errors could be amplified during PCR and would exist in the same location within each sequence that shares its UMI in the sequencing results. This allowed us to count the synthesis errors by creating a consensus sequence within each UMI family and finding errors with a high variant frequency.

To add UMI to the synthesized oligonucleotides, we considered the ratio between the number of synthesized oligo strands and UMI molecules. The initial number of molecules of the synthesized oligos was 6×10^{13}, and the length of the UMI bases was 20, which led to a theoretical diversity of 4^{20} or approximately 1×10^{12} unique molecules (Figure 1B). To avoid pairing the synthesized oligos with the same UMI, we diluted the oligos such that approximately 1×10^{7} copies of oligos were present in the reaction tube. We then applied two-cycle PCR with UMI primers to minimize PCR errors, which included the partial sequence of the synthesized oligo for hybridization and Illumina adapters so that the resulting product can be used for sequencing. To ensure that each UMI was read with 400× sequencing reads on average, DNA strands originating from 621,553 of the 10 different sequences were read with 937,940,515 sequencing reads (Figure S1).

We then determined the variant frequency threshold to distinguish SynError from SeqError (Figure 1b). To this end, we created a consensus sequence according to each UMI sequence, followed by the designed sequences and used the indel error rate distribution as a positive control (Figure S3). Indel errors usually occur during oligo synthesis but are rare in NGS.[9,16] We discarded the reads corresponding to the number of family reads under 100 to remove noise error considering the sequencing error that occurred in the UMI region (Figure S2). After aligning these filtered reads to the reference sequence, we discarded the reads at each position level given that it had less than 100 reads (Figures S4 and S5). We then plotted the indel error rate, and two normal distributions appeared at low and high frequencies that correspond to sequencing and synthesis errors, respectively. Therefore, the variant frequency threshold was determined to be 75% considering that the two peaks can be separated. In addition, the value of the threshold ensures that PCR errors incurred in the earliest cycles via erroneous incorporation or oxidation are not falsely labeled as synthesis errors. Even in the case of jackpot mutations, which occur during the very first cycle of

https://doi.org/10.1021/acssynbio.3c00308
ACS Synth. Biol. 2023, 12, 3567−3577

Table 1. Sequence Designed To Analyze Sequence-Independent Positional Error Profiling[a]

	5′ → 3′
shift_1	GAGGTCACTACGACGgtgatgaacagcctgtcagaacgattcaaccttaattaacaccctaccgatgcatgtcaggccatagatagtgccaatccagccagatcaccaggcaacaGGGTATCATGGAGCC
shift_2	GAGGTCACTACGACGgcctgtcagaacgattcaaccttaattaacaccctaccgatgcatgtcaggccatagatagtgccaatccagccagatcaccaggcaacagtgatgaacaGGGTATCATGGAGCC
shift_3	GAGGTCACTACGACGacgattcaaccttaattaacaccctaccgatgcatgtcaggccatagatagtgccaatccagccagatcaccaggcaacagtgatgaacagcctgtcagaGGGTATCATGGAGCC
shift_4	GAGGTCACTACGACGcttaattaacaccctaccgatgcatgtcaggccatagatagtgccaatccagccagatcaccaggcaacagtgatgaacagcctgtcagaacgattcaacGGGTATCATGGAGCC
shift_5	GAGGTCACTACGACGaccctaccgatgcatgtcaggccatagatagtgccaatccagccagatcaccaggcaacagtgatgaacagcctgtcagaacgattcaaccttaattaacGGGTATCATGGAGCC
shift_6	GAGGTCACTACGACGtgcatgtcaggccatagatagtgccaatccagccagatcaccaggcaacagtgatgaacagcctgtcagaacgattcaaccttaattaacaccctaccgaGGGTATCATGGAGCC
shift_7	GAGGTCACTACGACGgccatagatagtgccaatccagccagatcaccaggcaacagtgatgaacagcctgtcagaacgattcaaccttaattaacaccctaccgatgcatgtcagGGGTATCATGGAGCC
shift_8	GAGGTCACTACGACGgtgccaatccagccagatcaccaggcaacagtgatgaacagcctgtcagaacgattcaaccttaattaacaccctaccgatgcatgtcaggccatagataGGGTATCATGGAGCC
shift_9	GAGGTCACTACGACGagccagatcaccaggcaacagtgatgaacagcctgtcagaacgattcaaccttaattaacaccctaccgatgcatgtcaggccatagatagtgccaatccGGGTATCATGGAGCC
shift_10	GAGGTCACTACGACGccaggcaacagtgatgaacagcctgtcagaacgattcaaccttaattaacaccctaccgatgcatgtcaggccatagatagtgccaatccagccagatcaGGGTATCATGGAGCC

[a]The inner sequence of 100 nt was divided into 10 blocks consisting of 10 nucleotides (30 nt at both ends is the primer site for PCR). The sequence was opted to have GC content of 50%.

PCR, their variant frequencies cannot exceed 50% and are therefore classified as sequencing errors.

Positional SynError and SeqError Profiling by SHIFT. To profile positional errors while not being affected by the sequence, we designed 10 blocks, each of which contained 10 bases, from 100 bases (Figure 2A). The nucleotide sequences were designed randomly to maintain a GC content of 50% (Table 1).[34] The array of blocks was shifted toward the 3′-end. The original array is denoted as "shift 1." If one block was shifted, the shifted array of blocks was denoted as "shift 2." If two blocks were shifted, the array was denoted as "shift 3." Following this process, we generated 10 arrays of blocks or 10 shifted sequences. The positions of the array of blocks were labeled from P1 to P10 to form a 10 × 10 matrix of blocks and P*N*. Matrix generation of the shifted sequences enabled us to analyze the error from identical sequences of blocks. We averaged the error from 10 blocks at the same position. We analyzed both SynError and SeqError of the sequences from the 16th base to the 115th base to exclude the primer region. The results showed that high error rates of over 1% per nucleotide mostly accumulated from the start to the 65th base, which correspond to the 5′-end of the sequences (Figure 2D). We then attempted to delineate the cause of the high error rates between the synthesis and sequencing.

According to the defined variant frequency threshold of 75% per base, we analyzed the SynError and SeqError distributions along their sequences. It was observed that SynError appeared from Position (P) 1 to P4 with high error rates of 1.41, 1.61, 1.45, and 0.94% of the median value per base, respectively, and their average error rate was 3.76 times greater than that from the median values of P5 to P10 (Figure 2B,E). This was also mainly affected by position rather than sequence. Errors tended to accumulate near the 5′-end, as expected, due to the errors incurred in the coupling failure in phosphoramidite chemistry, which start from the 3′-end to the 5′-end of the sequence. In contrast, SeqError was relatively sporadic based on position, except for the errors that accumulated in P10 or the 3′-end within the sequences (Figure 2C). Two positions had a high error rate, namely, P4 and P7, in the oligos of "shift 2," which was caused by a substitution error that occurred 10.5 and 74.8 times, respectively, more than indel errors (Figure 2F). In P4, a T-to-A substitution occurred at 1.64% per base, which is unusual given that C-to-A substitution was dominant in the overall SeqError (Figure S6). In contrast, P7 included a C-to-A substitution. In addition, "shift 9" was found to have the lowest error rate on both sides, with a SynError of 0.29% and a SeqError of 0.10% per base, respectively. This phenomenon was then applied to "shift 6," which had a low error rate in both. These results indicate that high-quality oligos generated high-quality sequencing results. For the sequence-dependent analysis using sequence blocks, errors in both the SynError and SeqError analyses appeared randomly compared to the positional analysis.

Base Transitional Syn/Seq Error Profiling and Mutation Spectrum of 10 Shifted Sequences. To identify base transitional error patterns in SynError and SeqError, we focused on single sequence transitions (AT, AG, AC, TA, TC, TG, GA, GT, GC, CA, CT, and CG) and homopolymers (AA, TT, GG, CC, and CCC). Using SHIFT, we analyzed the Syn/Seq Error according to each shift and position in a position- and sequence-independent manner. Overall, the frequency of errors was 10 times higher in SynError than in SeqError. Therefore, we compared the error patterns within the same error type, SynError, or SeqError. All transitional error patterns of SynError depended on the position (Figure 3E,G) and had a higher error rate at the 5′-end (P1 to P4) regardless of their sequences (Figure 3A,C). However, in SeqError profiling, the nearer the 3′-end (P10) bases were located, the more the errors appeared (Figure 3F), especially in homopolymer sequences (Figure 3H).

In addition, we observed a transitional error pattern that had a lower error rate in the transition of GA and a higher error rate of AC, GC, and TC (Figure 3B,F). This was distinguished from SynError (Figure 3A) and exhibited the same pattern in both sequence- and position-dependent profiling. In single transitional errors, SeqError had a higher error rate in AC and TC transition bases, namely 0.16 and 0.21%, which is 1.3 and 1.7 times higher than the average of single-base transitional error (Figure 3B). However, the error rate of SynError was lower, namely, 0.36 and 0.57%, which is 0.49 and 0.77 times

https://doi.org/10.1021/acssynbio.3c00308
ACS Synth. Biol. 2023, 12, 3567−3577

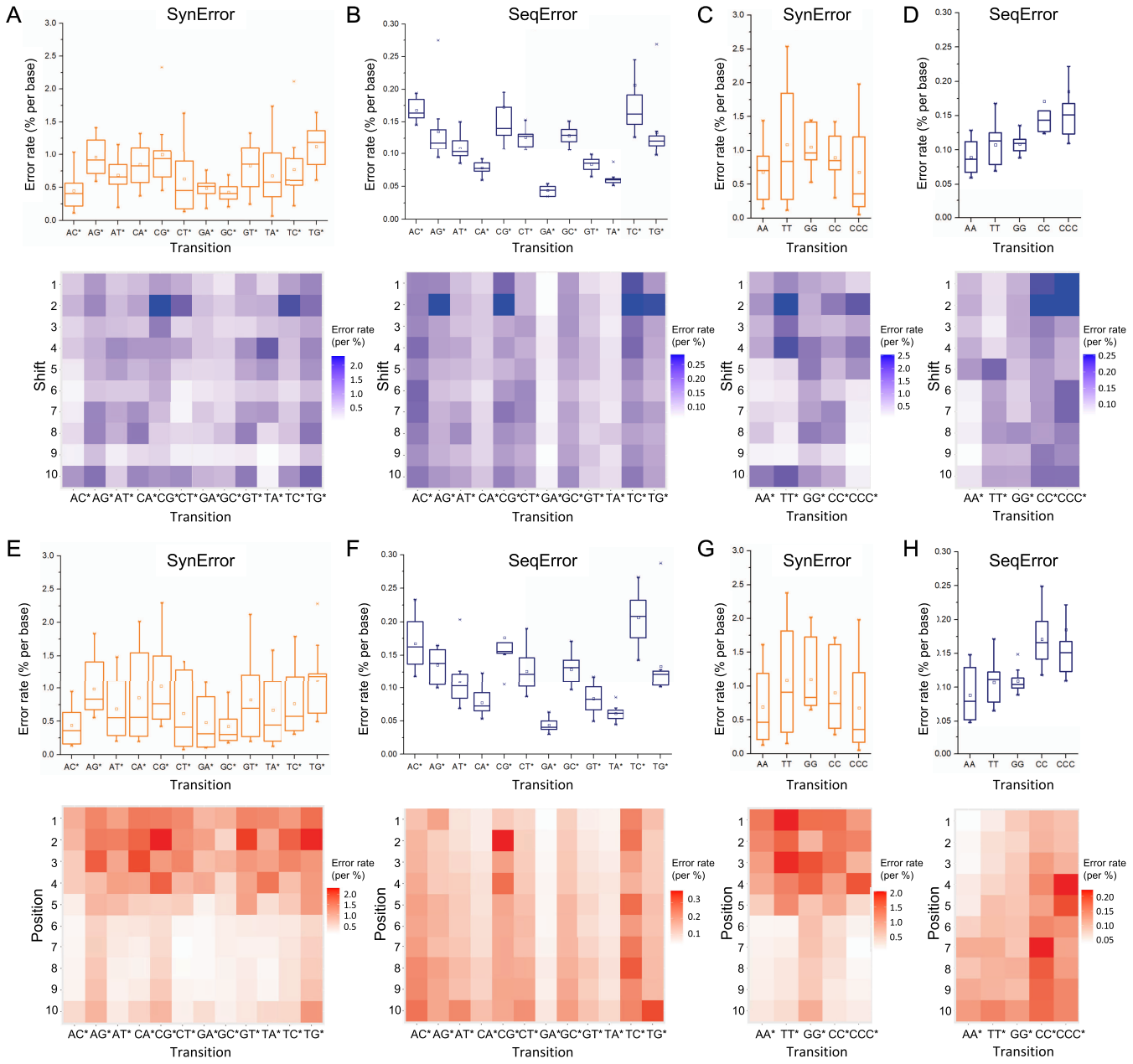


Figure 3. Base transitional SynError and SeqError were profiled using SHIFT. Position-independent single-base transition error profiling of SynError (A) and SeqError (B) corresponding to the 10 sequences. Position-independent homopolymer sequence transition error profiling of SynError (C) and SeqError (D) corresponding to the 10 sequences. Sequence-independent single-base transition error profiling of SynError (E) and SeqError (F) for each position. P1 is the 5′-end and P10 is the 3′-end in the oligo strands. Sequence-independent homopolymer sequences transition error profiling of SynError (G) and SeqError (H) for each position.

lower than the average of the single-base transitional error (Figure 3A). In homopolymer sequence transitions, SeqError and SynError depended on the position (Figure 3G,H) more than the sequences of each oligo (Figure 3)C,D, and SeqError had a relatively higher error rate in the CC and CCC sequences (Figure 3D,H). In this particular design, there was only one instance of a homopolymer with three consecutive bases. To account for this, we also explored sequences with different types of homopolymers (Table S1 and Figures S11−S15). The findings revealed a phenomenon in which characteristics induced by homopolymers become dominant in addition to the common traits of SynError and SeqError. Consequently, this highlights the potential variability in transitional errors and mutation spectra based on the nucleotide sequence type.

In addition, we analyzed error types, such as substitution, indel, and mutation spectrum (i.e., A to T, A to G, A to C, T to A, T to C, T to G, G to A, G to T, G to C, C to A, C to T, and C to G). First, we found that SynError had an average indel error rate of 0.73% per base, which was 6.1 times higher than the substitution error rate (Figure 4A), whereas SeqError contained an average indel error rate of 0.0083% per base, which was 15 times lower than the substitution error rate. The profiles of both synthesis and sequencing errors in the context of its position were obtained while disregarding its sequence. For SynError, the G-to-A substitution was prominent, with an

https://doi.org/10.1021/acssynbio.3c00308
ACS Synth. Biol. 2023, 12, 3567−3577

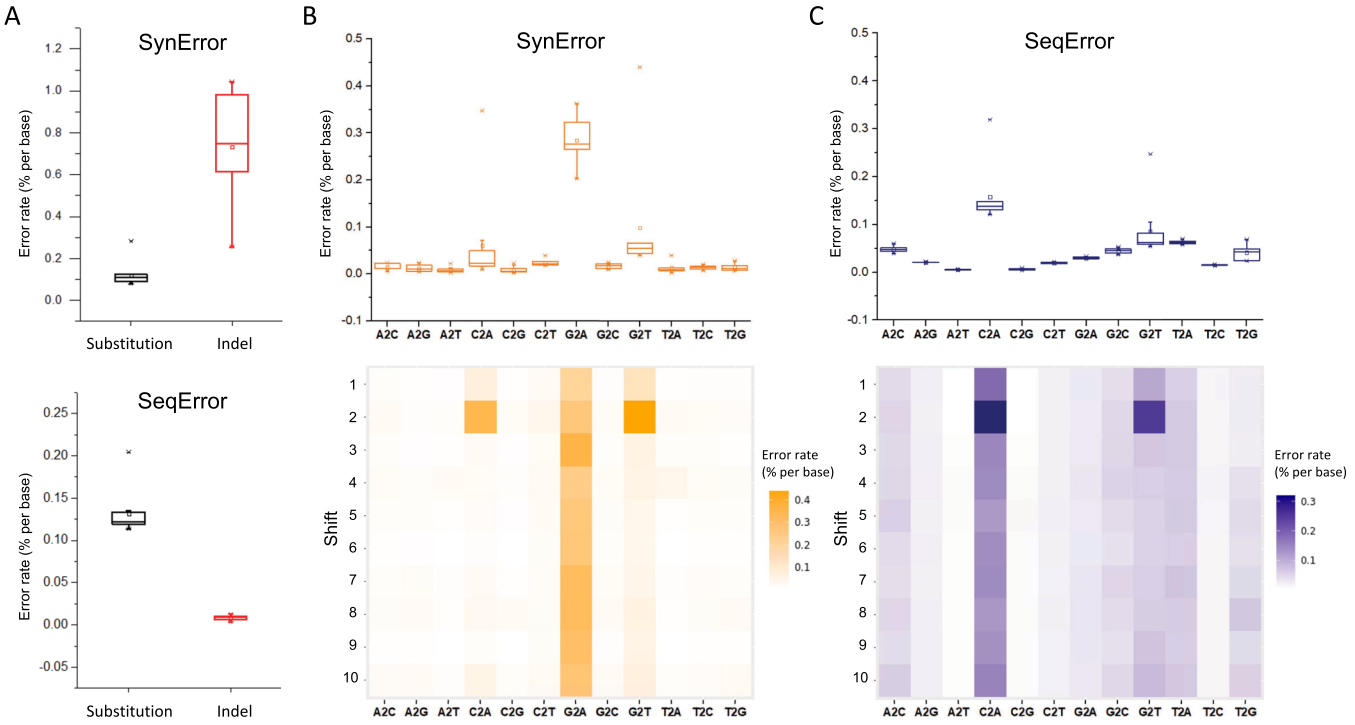


Figure 4. Substitution and indel error in SynError and SeqError were profiled using SHIFT. (A) Substitution and indel error rate in SynError and SeqError. SynError contained an average substitution error rate of 0.12% (s.d. 0.06%) per base and an indel error rate of 0.73% (s.d. 0.03%) per base. SeqError contained a substitution error rate of 0.13% per base (s.d. 0.013%) and an indel error rate of 0.008% per base (s.d. 0.002%). (B, C) Substitution error spectrum in SynError (B) and SeqError (C). G-to-A substitution was dominant in SynError (mean value of 0.28%) while C-to-A substitution was dominant in SeqError (mean value of 0.46%).

Table 2. Sequence Designed To Investigate the Synthesis Bias of Degenerate Bases

	5' → 3'
Front	gtgatgaacagcctgNNNNNNNNtcagaacgattcaaccttaattaacaccttatgtcaggccatagatagtgccaatccagccagatcaccaggcaaca
End	gtgatgaacagcctgtcagaacgattcaaccttaattaacaccttatgtcaggccatagatagtgccaatccagccaNNNNNNNNgatcaccaggcaaca

average value of 0.28% per base, followed by the G-to-T substitution, with an average value of 0.098% per base (Figures 4B and S7). The G-to-A substitution can be related to capping failure during oligo synthesis, as reported in previous studies.[9,35] In the case of SeqError, the C-to-A substitution was the most prominent with an average value of 0.46% per base. This can be attributed to DNA damage due to oxidation during PCR amplification[16] (Figure 4C).

SHIFT Profiling for Positional Synthesis Bias of Degenerate Bases. Profiling oligo synthesis and sequencing errors simultaneously become more sophisticated when DNA data storage is augmented with more units using degenerate bases.[36] Degenerate bases are an important tool in biological studies for generating randomized DNA sequences, and their usefulness has recently been extended to DNA data storage.[37] Previous studies on the augmentation of quaternary DNA data storage to higher-order DNA data storage have reported DNA data storage systems with higher storage densities.[36] Researchers have used degenerate bases to expand the molecular alphabet of DNA from four to 15 by encoding information to combinations of bases, such as A and C, T and G, or even all four bases of A, T, G, and C (denoted by the International Union of Biochemistry as M, K, and N). In ideal cases, each base in each combination would be equivalent in abundance; however, a thorough investigation has not been performed on the subject. Such biases can lead to the corruption of information because the entire process, from DNA synthesis to sequencing, comprises multiple sampling steps. An unbalanced initial distribution can ultimately lead to the complete loss of less abundant DNA strands in the final sequencing output. Prior knowledge regarding these potential biases can be used as a guide for designing methods that can minimize such circumstances.

To characterize the synthesis bias of the degenerate bases according to oligo strand positions, we synthesized oligonucleotides that have eight degenerate bases in the front region from the 16th to 23rd base and the back region from the 108th to 115th base, denoted as "front" and "end," respectively (Table 2). Theoretically, the length of the degenerate bases is 8, resulting in a sequence variety of 65,536 ($=4^8$). From the sequencing results, we found that the front had 64,879 degenerate sequences and the end had 65,432 sequences (Figure 5A). For the front, the maximum number of NGS reads, including those with a single read, was 535 with a mean value of 15.01 reads. In the case of the end, the maximum number of reads was 3542, with a mean value of 14.09 reads. To measure their bias, we calculated the Gini coefficient with the number of each unique degenerate sequence and found that the front had 0.51 and the end had 0.41, which indicate that the end had more bias during degenerate base synthesis

https://doi.org/10.1021/acssynbio.3c00308
ACS Synth. Biol. 2023, 12, 3567−3577

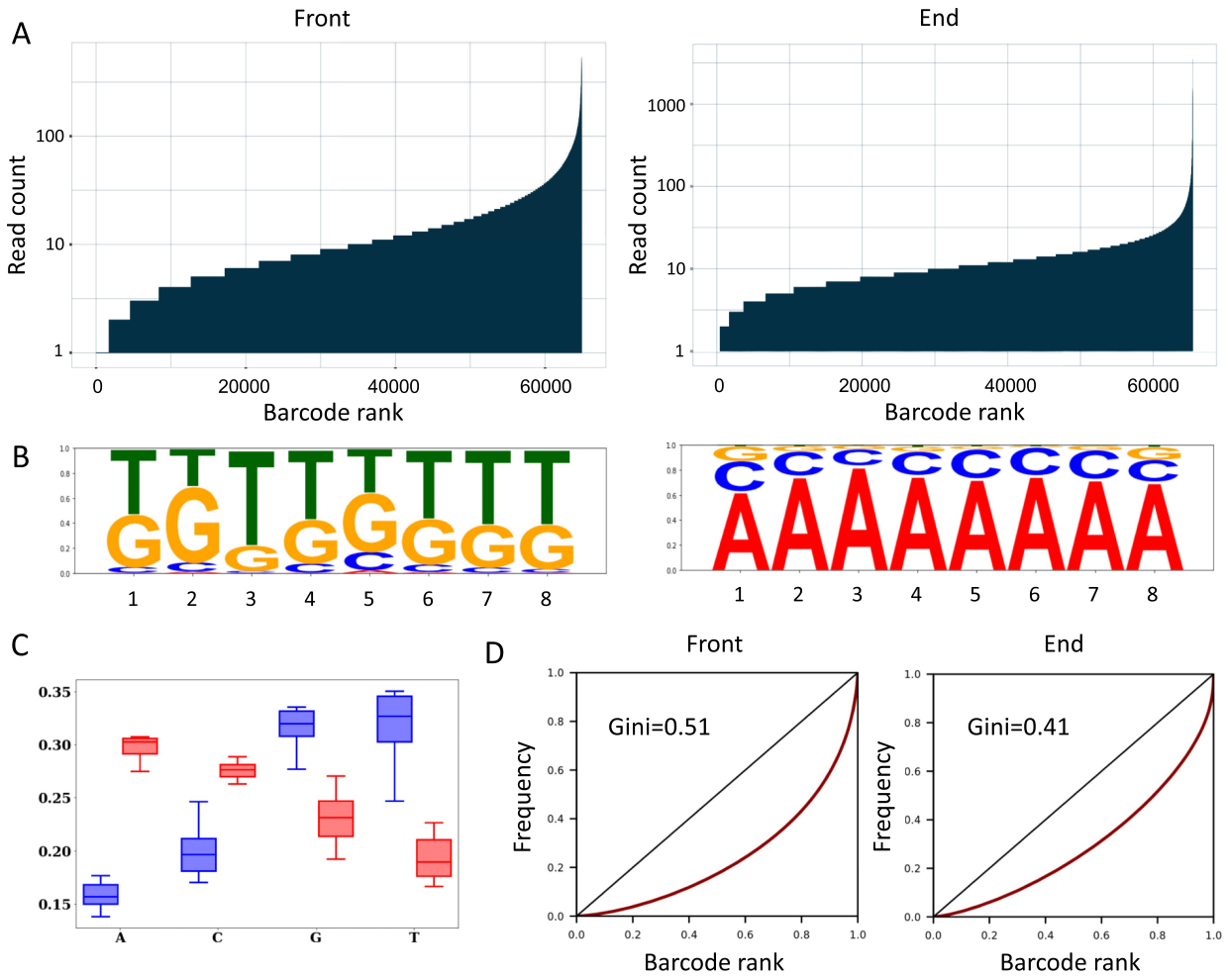


Figure 5. Synthesis bias of degenerate bases according to positional information. (A) Read count distribution along barcodes. (Left: "front", the degenerate bases are located near the 5′-end, from the 16th to 23rd base; right: "end," the degenerate bases are located near the 3′-end, from the 108th to 115th base). (B) Sequence logo of the degenerate bases at each position. (Left: "front"; right: "end"). (C) Boxplot of the relative abundance of each base in the eight base degenerate barcodes. The blue box indicates the values from the front, while the red box indicates values from the end. (D) Lorenz curve and Gini coefficient related to the degenerate bases bias.

(Figure 5D). In the front, the percentage of reads including a deletion error was 91.66%, followed by 7.58% with perfect length and 0.76% with an insertion. In the end, the percentages of reads including deletion, perfect length, and insertion were 70.05, 21.47, and 8.48%, respectively.

To visualize the bias pattern in sequences with a high number of read replicates, we investigated the distribution of each nucleotide in more than 100 replicates. In the front, the nucleotide "T" was found to be dominant over all the positions of the degenerate bases, while, in the end, "A" was the most dominant (Figure 5B,C). This tendency also occurred in tandem for sequences with more than 10 reads (Figures S9 and S10).

DISCUSSION

In this study, we simultaneously profiled DNA synthesis and sequencing errors within the same oligos using SHIFT. Approximately 100,000 molecules were generated using the 1B reads raw data set. The use of SHIFT allowed us to delineate the synthesis and sequencing errors by accurately defining the variant frequency threshold. Accurate error patterns were profiled because SHIFT enables the profiling of both synthesis and sequencing errors in a position- and sequence-independent manner. We found that the quality of oligos could affect sequencing quality; for base transition error, the error rate of AC and TC transition in SynErrors was low, while the error rate of SeqError was high. Thus, SHIFT is a useful tool for standardizing the simultaneous error profiling in oligo synthesis and sequencing.

SHIFT is compatible not only with the state-of-the-art DNA synthesizers and sequencers presented in this study but also with different DNA synthesis and/or sequencing methods. For example, enzymatic DNA synthesis[14] is highly advantageous over chemical DNA synthesis but is not free from enzymatic errors. Because the SHIFT protocol begins with presynthesized oligonucleotides, we can apply it regardless of such synthesis methods and generate their respective error profiles. Furthermore, SHIFT is compatible with several sequencing methodologies, such as nanopore sequencing, PacBio, IonTorrent, MGI, and other novel technologies that generate the sequences of synthesized oligos.

In addition, the data set presented provides accurate DNA synthesis and sequencing error profiling data with 10 shifted sequences, which provides sequence-independent positional and position-independent base transitional analyses. We also investigated degenerate synthesis bias with oligomers of the

https://doi.org/10.1021/acssynbio.3c00308
ACS Synth. Biol. 2023, 12, 3567−3577

designed sequences in which the degenerate bases of 8N were profiled positionally. Different bias patterns were observed in terms of the dominant base population at each sequence position, such as the finding that "T" and "A" were dominant in degenerate bases for the front and end, respectively. Although more empirical data generation is necessary to verify whether this pattern is general for all synthesis conditions, these data demonstrate that synthesis bias to synthesize degenerate bases occurred during the synthesis process. Further studies will need to consider the synthesis bias of degenerate bases in terms of sequence loss when designing the encoding scheme to use degenerate bases in oligo synthesis.

In DNA data storage, the error patterns reported here can immediately be applied to encoding algorithms such as the DNA fountain code.[6] The DNA fountain encoding scheme combines random sampling of data blocks and rejection of formed sequences based on predetermined filters to achieve near-ideal coding rates while using a simple quaternary conversion of bits to bases. The rejection stage originally limits only the GC content range and homopolymer length, but the synthesis and sequencing error-prone patterns reported in this work can be added as additional filtering conditions to ensure the accurate synthesis and sequencing of DNA molecules. In addition, using the shifted sequences, the optimal length to overlap in the Goldman encoding/decoding method can be calculated.[7] Finally, because the error characteristics were categorized based on the position within each DNA strand, functional sequences such as indices, seed values, and parity or Reed-Solomon bases for most encoding algorithms can be positioned in regions that are less likely to be subjected to errors.[38] This will aid in reducing collateral errors that may be caused by the decoding failure of such key information. In this regard, the DNA synthesis and sequencing error profiling reported in this study can help make DNA data storage systems more practical by guiding accurate encoding/decoding strategies.

METHODS

Oligo Preparation. The oligos with the shifted sequences (shift 1 ∼ shift 10) were synthesized by Macrogen Korea and were individually diluted to a concentration of 100 pmol/μL (6.02×10^{13} oligos/μL) according to the given instructions using nuclease-free water (Qiagen). To further reduce the concentration to 10^9 oligos/μL, we diluted the mixture with a factor of 60,000. The mixture was diluted serially with the first step being a 1:299 dilution (2 μL:598 μL) and the second step being a 1:199 dilution (5 μL:995 μL). We confirmed the final concentrations using a Qubit Fluorometer (Qiagen) and determined the concentration of each sample as follows; S1: 0.056, S2: 0.044, S3: 0.062, S4: 0.067, S5: 0.054, S6: 0.036, S7: 0.043, S8: 0.053, S9: 0.047, and S10: 0.047 ng/μL. Based on the calculation method of the approximate molecular weight of single-stranded DNA provided by ThermoFisher Scientific (https://www.thermofisher.com/kr/ko/home/references/ambion-tech-support/rna-tools-and-calculators/dna-and-rna-molecular-weights-and-conversions.html), we calculated that the initial 100 pmol/μL mixture had a nucleic acid concentration of 3956 ng/μL, and so the 60,000 times diluted mixture should have a concentration of 0.066 ng/μL. Our previous measurements aligned reasonably with this theoretical value.

Adding UMI to Oligos and NGS Library Preparation for 10 Shifted Sequences. To add UMI to the oligomers, PCR amplification was conducted. The PCR mixture for each sample included 1 μL of the diluted oligos, 5 μL of 10 μM forward primer solution, 5 μL of 10 μM, reverse primer solution, 25 μL of KAPA HiFi HotStart ReadyMix (2×), and nuclease-free water up to a total volume of 50 μL. The thermocycling protocol comprised the 95 °C stage for 5 min followed by 2 cycles of 98 °C for 30 s, 52 °C for 30 s, 72 °C for 60 s, and a final elongation at 72 °C for 3 min. The PCR mixture was purified using Celemag cleanup beads with a bead:sample ratio of 1.8:1.0 and an elution volume of 20 μL. The forward primer contained the 20 bp UMI, while both primers contained the flanking sequences of the synthesized oligos and the Illumina adapter sequences. Because the total number of forward primers going into the reaction was 3.01×10^{13} with a theoretical diversity of 10^{12}; while the number of oligos was 10^9, we expect that all oligos were paired with unique UMIs during the reaction.

Taking into account the initial dilution at the start of the reaction (1:49), the 2 cycle amplification (X4), purification efficiency, and reduction in elution volume (×2.5), we estimated the concentration of oligos in the final product to be in the range of 10^9 to 10^8 copies/μL. We then diluted the purified mixture 1000 fold via 3-stage 10-fold serial dilution (10 μL:90 μL) and used 5 μL of the resulting mixture for an additional PCR stage with 1 μL of 40 μM forward adapter solution, 1 μL of 40 μM, reverse adapter solution, 10 μL of KAPA HiFi HotStart ReadyMix (2×), and nuclease-free water up to a total volume of 20 μL. Illumina index adapters were used in this stage as primers. The thermocycling protocol is similar to the one stated previously with the initial 95 °C stage reduced by 2 min and the total number of cycles increased to 25. The resulting product was purified using Celemag beads as previously described and was sequenced using the Illumina NovaSeq platform with 100 PE kit.

Adding UMI to Oligos and NGS Library Preparation for the Degenerate Base Sequences. For the preparation of the front and end degenerate base sequences, we took the initial 100 pmol/μL oligo solution synthesized by Macrogen and diluted it 200 fold. The PCR mixture for each sample for the first PCR stage included 1 μL of the diluted oligos, 5 μL of 10 μM forward primer solution, 5 μL of 10 μM reverse primer solution, 25 μL of KAPA HiFi HotStart ReadyMix (2×), and nuclease-free water up to a total volume of 50 μL. The thermocycling protocol comprised the 95 °C stage for 5 min followed by 3 cycles of 98 °C for 30 s, 52 °C for 30 s, 72 °C for 60 s, and a final elongation at 72 °C for 3 min. The resulting solution was purified using Celemag beads, as described previously. The second PCR stage was conducted by using Illumina index adapters. The PCR mixture included 5 μL of the purified product, 5 μL of 10 μM forward adapter solution, 5 μL of 10 μM reverse adapter solution, 25 μL of KAPA HiFi HotStart ReadyMix (2×), and nuclease-free water up to a total volume of 50 μL. The thermocycling protocol comprised of 95 °C stage for 5 min followed by 3 cycles of 98 °C for 30 s, 52 °C for 30 s, 72 °C for 60 s, and a final elongation at 72 °C for 3 min. The resulting product was purified using Celemag beads as previously described and was sequenced using an Illumina iSeq platform 150PE kit (ATG Lifetech Inc.).

Preprocessing Raw Sequencing Data. NGS was conducted by Illumina NovaSeq SE sequencing with 100,000 reads per sample. For total positional error profiling, each sequencing read was aligned to the design sequence for Shift 1−10 using Burrows-Wheeler Aligner (BWA) mem aligner

https://doi.org/10.1021/acssynbio.3c00308
ACS Synth. Biol. 2023, 12, 3567−3577

followed by processing with SAMtools; view, sort, and mpileup. For calling variants, we used VasrScan; pileup2cns. To analyze synthesis and sequencing error, the sequencing reads were split by the respective UMIs into family reads and were stored in separate files. Then, each file was aligned to the original design sequence for Shift 1−10 using BWA mem aligner followed by processing with SAMtools; view, sort, and mpileup.

Removing Noise from the Preprocessed Data. To determine the threshold to remove noise reads, the indel error frequency of 10 shift sequences was plotted according to their distribution using R (ggplot). Given that the indel error frequency is the positive control of oligo synthesis error, the noise reads were randomly distributed according to frequency while the signal has a certain pattern of normal distribution (Figure S3). When the number of family reads with the same UMI was under 100, they were discarded before further analysis (Figure S4). Then, each reference and variant bases with under 100 reads allocated were discarded using the cns file from preprocessed data through VasrScan; pileup2cns with the version of VarScan.v2.3.9.

Determining Variant Frequency Threshold To Delineate Synthesis Error and Sequencing Error. From the filtered data set, we extracted indel errors from each position and calculated the error rate along frequencies using a histogram in R. In the plotted data, there were two normal distributions corresponding to sequencing error and synthesis error, respectively, and the variant frequency was determined when the two peaks were ideally separated. The variant frequency was applied to the 10 shifted sequences and validated (Figure S4).

Counting Syn/SeqError and Data Visualization with Heatmap. SynError is defined as the amount of error accumulated in each oligo strand, and its unit is the number of errors per oligo. SeqError is defined as the amount of error accumulated in each sequencing read, and its unit is errors per read. Therefore, SynError was calculated by counting all errors with over 75% variant frequency from the filtered cns files when normalized by the number of oligo strands corresponding to each shift, respectively. SeqError was calculated by summing all errors with their variant frequency under 75% from the filtered cns files when normalized by the number of oligo strands corresponding to each shift. All heatmaps were visualized by R and, in the case of SeqError, an outlier with a 0.6% error rate was excluded (P4 of Shift 2).

Degenerate Base Bias Analysis. NGS was conducted by iSeq PE Illumina sequencing with 1 M reads per sample. Because the degenerate bases were in the middle of the sequences, each sequencing reads were aligned to the original design sequence first, then BWA mem aligner was used and followed by processing with SAMtools; view, sort, and mpileup. For calling variants, we used VasrScan; pileup2cns. Then, the barcodes that had the full length of 8 bases were extracted from the cns files.

■ ASSOCIATED CONTENT

Data Availability Statement

All data are available from the authors upon request.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acssynbio.3c00308.

Further modeling to simultaneously reduce SynError and SeqError; read count distribution; read counts for the 10 sequences; indel error frequency distribution; base and homopolymer sequence transitional error profiling; mutation spectrum of SeqError (Figure S6); variant counts along sequence position; Indel length distribution; sequence logo of the degenerate bases; shift 1−10 of 'dataset 2'; accumulated positional error rate; positional SynError and SeqError; single base transitional SynError and SeqError; homopolymer base transitional SynError and SeqError; and deletion and insertion in SynError and SeqError (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Huiran Yeom − *Division of Data Science, College of Information and Communication Technology, The University of Suwon, Hwaseong 18323, Republic of Korea*; Email: hyeom@suwon.ac.kr

Yeongjae Choi − *School of Materials Science and Engineering, Gwangju Institute of Science and Technology (GIST), Gwangju 61105, Republic of Korea*; Email: yeongjae@gist.ac.kr

Authors

Namphil Kim − *Department of Electrical and Computer Engineering, Seoul National University, Seoul 08826, South Korea*

Amos Chungwon Lee − *Meteor Biotech, Co. Ltd., Seoul 08813, Republic of Korea*; orcid.org/0000-0002-0350-7080

Jinhyun Kim − *Department of Electrical and Computer Engineering, Seoul National University, Seoul 08826, South Korea*

Hamin Kim − *Department of Interdisciplinary Program for Bioengineering, Seoul National University, Seoul 08826, South Korea*

Hansol Choi − *Bio-MAX Institute, Seoul National University, Seoul 08826, Republic of Korea*

Seo Woo Song − *Basic Science and Engineering Initiative, Children's Heart Center, Stanford University, Stanford, California 94304, United States*

Sunghoon Kwon − *Department of Electrical and Computer Engineering and Department of Interdisciplinary Program for Bioengineering, Seoul National University, Seoul 08826, South Korea; Bio-MAX Institute, Seoul National University, Seoul 08826, Republic of Korea*; orcid.org/0000-0003-3514-1738

Complete contact information is available at: https://pubs.acs.org/10.1021/acssynbio.3c00308

Author Contributions

H.Y. designed the study and performed all experiments. N.K. established protocol of adding UMI to oligos and NGS library preparation. A.C.L. and S.W.S. provided conceptual idea which is possible to apply to DNA data storage. J.K., H.K., and H.C. interpreted all experimental data and sequencing result, and wrote the manuscript with input from all authors.

Notes

The authors declare no competing financial interest.

https://doi.org/10.1021/acssynbio.3c00308
ACS Synth. Biol. 2023, 12, 3567−3577

■ ACKNOWLEDGMENTS

This work was supported by the Ministry of Science and ICT (MSIT) of the Republic of Korea and the National Research Foundation of Korea (NRF-2020R1A3B3079653, 2021R1C1C2010079, 2022M3C1A3081366, 2022R1C1C1010938, 2022R1C1C2002904, and 2020R1C1C1007665); the BK21 FOUR program of the Education and Research Program for Future ICT Pioneers, Seoul National University in 2023; and the Technology development Program(RS-2023-00222659) funded by the Ministry of SMEs and Startups(MSS, Korea).

■ REFERENCES

(1) Zhirnov, V.; Zadegan, R. M.; Sandhu, G. S.; Church, G. M.; Hughes, W. L. Nucleic Acid Memory. *Nat. Mater.* **2016**, *15* (4), 366−370.
(2) 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* (3), 242−248.
(3) 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.* **2015**, *54* (8), 2552−2555.
(4) 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. 2206201.
(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 Stat. Mech. Appl.* **2013**, *392* (5), 1072−1079.
(6) Erlich, Y.; Zielinski, D. DNA Fountain Enables a Robust and Efficient Storage Architecture. *Science* **2017**, *355* (6328), 950−954.
(7) Goldman, N.; Bertone, P.; Chen, S.; Dessimoz, C.; Leproust, E. M.; Sipos, B.; Birney, E. Towards Practical, High-Capacity, Low-Maintenance Information Storage in Synthesized. *DNA* **2013**, *494*, 77.
(8) Wang, Y.; Noor-a-rahim, M.; Gunawan, E.; Liang guan, Y.; Poh, C. L. Modelling, Characterization of Data-Dependent and Process-Dependent Errors in DNA Data Storage, 2021; pp 1−8.
(9) 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* (12), 6687−6701.
(10) 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* (8), 2522−2540.
(11) 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* (48), 6714.
(12) Sidore, A. M.; Plesa, C.; Samson, J. A.; Lubock, N. B.; Kosuri, S. DropSynth 2.0: High-Fidelity Multiplexed Gene Synthesis in Emulsions. *Nucleic Acids Res.* **2020**, *48* (16), E95.
(13) Palluk, S.; Arlow, D. H.; De rond, T.; Barthel, S.; Kang, J. S.; Bector, R.; Baghdassarian, H. M.; Truong, A. N.; Kim, P. W.; Singh, A. K.; Hillson, N. J.; Keasling, J. D. De Novo DNA Synthesis Using Polymerasenucleotide Conjugates. *Nat. Biotechnol.* **2018**, *36* (7), 645−650.
(14) Lee, H.; Wiegand, D. J.; Griswold, K.; Punthambaker, S.; Chun, H.; Kohman, R. E.; Church, G. M. Photon-Directed Multiplexed Enzymatic DNA Synthesis for Molecular Digital Data Storage. *Nat. Commun.* **2020**, *11* (1), 5246.
(15) Flamme, M.; Hanlon, S.; Marzuoli, I.; Püntener, K.; Sladojevich, F.; Hollenstein, M. Evaluation of 3′-Phosphate as a Transient Protecting Group for Controlled Enzymatic Synthesis of DNA and XNA Oligonucleotides. *Commun. Chem.* **2022**, *5* (1), 1−12.
(16) 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* (1), 1−15.
(17) Yeom, H.; Lee, Y.; Ryu, T.; Noh, J.; Lee, A. C.; Lee, H.-B.; Kang, E.; Song, S. W.; Kwon, S. Barcode-Free next-Generation Sequencing Error Validation for Ultra-Rare Variant Detection. *Nat. Commun.* **2019**, *10* (1), 977.
(18) Heckel, R.; Mikutis, G.; Grass, R. N. A Characterization of the DNA Data Storage Channel *Sci. Rep.* 2019, *9*, 9963, DOI: 10.1038/s41598-019-45832-6.
(19) Yeom, H.; Ryu, T.; Lee, A. C.; Noh, J.; Lee, H.; Choi, Y.; Kim, N.; Kwon, S.; Kwon, S.; Kwon, S.; Kwon, S.; Kwon, S. Cell-Free Bacteriophage Genome Synthesis Using Low-Cost Sequence-Verified Array-Synthesized Oligonucleotides. *ACS Synth. Biol.* **2020**, *9* (6), 1376−1384.
(20) Press, W. H.; Hawkins, J. A.; Jones, S. K.; Schaub, J. M.; Finkelstein, I. J.; Plotkin, J. B.; Vincent, H. HEDGES Error-Correcting Code for DNA Storage Corrects Indels and Allows Sequence Constraints. *Biophys. Comput. Biol.* DOI: 10.1073/pnas.2004821117/-/DCSupplemental.y.
(21) Choi, H.; Choi, Y.; Choi, J.; Lee, A. C.; Yeom, H.; Hyun, J.; Ryu, T.; Kwon, S. Purification of Multiplex Oligonucleotide Libraries by Synthesis and Selection. *Nat. Biotechnol.* **2022**, *40* (1), 47−53.
(22) Zhou, G.; Huang, X.; Ping, Z.; Chen, S.; Joezhu, S.; Zhang, H.; Lee, H. H.; Lan, Z.; Cui, J.; Chen, T.; Zhang, W.; Yang, H.; Xu, X.; Church, G. M.; Shen, Y. Towards Practical and Robust DNA-Based Data Archiving Using the Yin-Yang Codec System *Nat. Comput. Sci.* 2022, *2*, 234−242, DOI: 10.1038/s43588-022-00231-2.
(23) Filges, S.; Mouhanna, P.; Ståhlberg, A. Digital Quantification of Chemical Oligonucleotide Synthesis Errors *Clin. Chem.* 2021, *67*, 1384−1394, DOI: 10.1093/clinchem/hvab136.
(24) Gao, Y.; Chen, X.; Qiao, H.; Ke, Y.; Qi, H. Low-Bias Manipulation of DNA Oligo Pool for Robust Data Storage. *ACS Synth. Biol.* **2020**, *9* (12), 3344−3352.
(25) Hawkins, J. A.; Jones, S. K.; Finkelstein, I. J.; Press, W. H. Indel-Correcting DNA Barcodes for High-Throughput Sequencing. *Proc. Natl. Acad. Sci. U. S. A.* **2018**, *115* (27), E6217−E6226.
(26) Meiser, L. C.; Antkowiak, P. L.; Koch, J.; Chen, W. D.; Kohll, A. X.; Stark, W. J.; Heckel, R.; Grass, R. N. Reading and Writing Digital Data in DNA. *Nat. Protoc.* **2020**, *15* (1), 86−101.
(27) Xu, C.; Zhao, C.; Ma, B.; Liu, H. SURVEY AND SUMMARY Uncertainties in Synthetic DNA-Based Data Storage. *Nucleic Acids Res.* **2021**, *49* (10), 5451−5469.
(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* (1), 1−9.
(29) Schirmer, M.; Ijaz, U. Z.; D'amore, R.; Hall, N.; Sloan, W. T.; Quince, C. Insight into Biases and Sequencing Errors for Amplicon Sequencing with the Illumina MiSeq Platform. *Nucleic Acids Res.* **2015**, *43* (6), No. e37.
(30) Press, W. H.; Hawkins, J. A.; Jones, S. K.; Schaub, J. M.; Finkelstein, I. J. HEDGES Error-Correcting Code for DNA Storage Corrects Indels and Allows Sequence Constraints *Proc. Natl. Acad. Sci. U. S. A. D117*, 18489−18496, DOI: 10.1073/pnas.2004821117/-/DCSupplemental.y.
(31) Wang, Y.; Noor-a-rahim, M.; Zhang, J.; Gunawan, E.; Guan, Y. L.; Poh, C. L. High Capacity DNA Data Storage with Variable-Length Oligonucleotides Using Repeat Accumulate Code and Hybrid Mapping. *J. Biol. Eng.* **2019**, *13* (1), 1−11.
(32) Smith, T.; Heger, A.; Sudbery, I. UMI-Tools: Modeling Sequencing Errors in Unique Molecular Identifiers to Improve Quantification Accuracy. *Genome Res.* **2017**, *27* (3), 491−499.

(33) Pfeiffer, F.; Gröber, C.; Blank, M.; Händler, K.; Beyer, M.; Schultze, J. L.; Mayer, G. Systematic Evaluation of Error Rates and Causes in Short Samples in Next-Generation Sequencing. *Sci. Rep.* **2018**, *8* (1), 1−14.
(34) Stothard, P. Internet On-Ramp Internet On-Ramp. *Biotechniques* **2000**, *28* (6), 1102−1104.
(35) 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* (1), 12095.
(36) Choi, Y.; Ryu, T.; Lee, A. C.; Choi, H.; Lee, H.; Park, J.; Song, S. H.; Kim, S.; Kim, H.; Park, W.; Kwon, S. High Information Capacity DNA-Based Data Storage with Augmented Encoding Characters Using Degenerate Bases. *Sci. Rep.* **2019**, *9* (1), 6582.
(37) Hwang, B.; Bang, D. Toward a New Paradigm of DNA Writing Using a Massively Parallel Sequencing Platform and Degenerate Oligonucleotide. *Sci. Rep.* **2016**, *6*, 37176.
(38) Blawat, M.; Gaedke, K.; Hütter, I.; Chen, X. M.; Turczyk, B.; Inverso, S.; Pruitt, B. W.; Church, G. M. Forward Error Correction for DNA Data Storage. *Procedia Comput. Sci.* **2016**, *80*, 1011−1022.

https://doi.org/10.1021/acssynbio.3c00308
ACS Synth. Biol. 2023, 12, 3567−3577