Modelling, Characterization of Data-Dependent and Process-Dependent Errors in DNA Data Storage

Yixin Wang, Md. Noor-A-Rahim, Erry Gunawan, Yong L. Guan, and Chueh L. Poh

Abstract—Using DNA as the medium to store information has recently been recognized as a promising solution for long-term data storage. While several system prototypes have been demonstrated, the error characteristics in DNA data storage are discussed with limited content. Due to the data and process variations from experiment to experiment, the error variation and its effect on data recovery remain to be uncovered. To close the gap, we systematically investigate the storage channel, i.e., error characteristics in the storage process. In this work, we first propose a new concept named sequence corruption to unify the error characteristics into the sequence level, easing the channel analysis. Then we derived the formulations of the data imperfection at the decoder including both sequence loss and sequence corruption, revealing the decoding demand and monitoring the data recovery. Furthermore, we extensively explored several data-dependent unevenness observed in the base error patterns and studied a few potential factors and their impacts on the data imperfection at the decoder both theoretically and experimentally. The results presented here introduce a more comprehensive channel model and offer a new angle towards the data recovery issue in DNA data storage by further elucidating the error characteristics of the storage process.

Index Terms—Channel modelling, DNA data storage, error characterization, long-term storage

1 Introduction

The explosion of data has driven scientists to explore new technologies to store information. In recent years, owing to the superior properties like extremely high physical density and preservation duration, using DNA molecules as the data storage medium has drawn a rising attention [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11]. In a typical DNA data storage system, the basic data unit is a DNA strand that represents a string of nucleotide bases consisting of Adenine (A), Thymine (T), Cytosine (C), and Guanine (G). Data writing in DNA data storage is performed by encoding the digital information into an assembly of DNA sequences. Taking the encoded DNA sequences as the reference, corresponding DNA molecules are synthesized and very often the number of molecule copies (e.g., copies of oligo) of each reference sequence varies. The synthesized DNA can then be stored and sequenced during which several random processes are involved, leading to *sequence loss* at the decoder [8], [12]. Here, sequence loss refers to the loss of all copies/reads of the reference sequence. In other words, if all sequence data of one reference sequence could not be found at the decoder, this reference sequence is considered lost.

Besides the sequence loss, *base error* is the other type of error in DNA data storage. While the general base error statistics have been reported [9], [13], [14] and the sequence loss has been studied [12], [13], the overall impact of these two types of errors, i.e., base level and sequence level, on the decoder, remains unclear. To unify the error characteristics and ease the analysis at the decoder, we introduce the concept of *sequence corruption*, transmitting the base type error to sequence type error by incorporating the effects of physical redundancy (i.e., multiple sequence copies at the receiver/sequencer) and post-processing method into the channel model. Different from sequence loss, sequence corruption refers to the failure of reconstructing the reference sequence from its erroneous copies (i.e., with base errors) at the receiver, which is the direct consequence of the base error. Theoretically, the sequence corruption rate collectively depends on the base error rate in the received copies of the reference sequence, the received copy counts of the reference sequence, and the post-processing methods of reconstructing the reference sequence. From another perspective, this new concept leverages the multi-count physical redundancy feature of DNA data storage, enabling the anticipation of the required logical redundancy in code design with the presence of specific physical redundancy in

- *Yixin Wang, Erry Gunawan, and Yong L. Guan are with the School of Electrical and Electronic Engineering (EEE), Nanyang Technological University (NTU), Singapore 639798. E-mail: {ywang065, egunawan, eylguan}@ntu.edu.sg.*
- *Md. Noor-A-Rahim is with the School of Computer Science and IT, University College Cork, T12 K8AF Cork, Ireland. E-mail: m.rahim@cs.ucc.ie.*
- *Chueh L. Poh is with the Department of Biomedical Engineering and the Synthetic Biology for Clinical & Technological Innovation, National University of Singapore (NUS), Singapore 119077. E-mail: poh.chuehloo@nus.edu.sg.*

Manuscript received 24 October 2021; revised 13 December 2022; accepted 30 December 2022. Date of publication 4 January 2023; date of current version 5 June 2023.
This work was supported in part by the Synthetic Biology Initiative of the National University of Singapore under Grant DPRT/943/09/14, in part by the Summit Research Program of the National University Health System under Grant NUHSRO/2016/053/SRP/05, and in part by the Research Scholarship of Nanyang Technological University, and NUS startup grant.
(Corresponding author: Yixin Wang.)
This article has supplementary downloadable material available at https://doi.org/10.1109/TCBB.2023.3233914, provided by the authors.
Digital Object Identifier no. 10.1109/TCBB.2022.3233914

the experiment. As a result, we define the data imperfection at the decoder of DNA data storage channel consisting of sequence loss and sequence corruption. Investigating the characteristic of base errors where some variance might exist is also important since it essentially relates to sequence corruption and can guide the sequence (codeword) design. Meanwhile, many factors, including data structure design, experiment design, and computational processing, could affect the degree of data imperfection at the decoder, leading to process-dependent errors. Understanding how the original data (reference sequences) and these factors affect the overall error rate at the decoder could provide insights into several aspects of designing advanced DNA data storage systems, including codec designs, sequence structure designs, experiment designs, and data processing methods.

In this paper, we theoretically formulated the sequence corruption which is cooperatively dictated by the base error statistics, copy counts of the reference sequence, and downstream processing methods. Combining the sequence corruption with the sequence loss, we then quantified the data imperfection by deriving the overall sequence error rate of the DNA data storage channel. The derivation explicitly takes the unevenness in both the count distribution and the error patterns into consideration, revealing distinct data recovery demands in DNA data storage using different sequencing techniques. Furthermore, we investigated the base error properties by analyzing the data from our previous work [11], [15]. Specifically, we first looked into the single base error and then analyzed the 2/3/4-mer patterns with different types of errors, i.e., substitutions, insertions, and deletions. We observed that there are profound biases in transition errors among DNA bases, and certain k-mer patterns (not only homopolymers) are prone to a certain type of errors. Lastly, with data collected from two independent experiments and theoretical analysis, we broadly studied the factors that might affect the data integrity in aspects spanning from structure design to biological and analytical handling methods. By conducting the most comprehensive study on the imperfect and uneven data in DNA data storage so far, the results in this work could offer insights and instructions to the design and processing pipeline of more effective and efficient DNA data storage systems.

2 Method

2.1 Data Flow and Errors in DNA Data Storage

Data are represented in different forms at different stages in the DNA data storage, such as binary stream, DNA sequences, and physical DNA molecules (see Fig. 1A). Binary data are encoded and converted into DNA sequences before sending them to DNA synthesis. At the synthesis stage, the count of the oligos may vary, and the count distribution can be approximated by gamma or normal distribution based on the different synthesis techniques [12]. Following that, the sample might be stored in a distributed fashion to increase data accessibility, where a random process happens. To illustrate, Fig. 1A describes one scenario where physical copies of certain (i.e., purple-colored) reference sequences are all lost, rendering *physical sequence loss* (i.e., 0 physical copy of the reference sequence).

To prepare the stored sample for DNA sequencing, the sample is usually PCR amplified to meet the sequencing requirements. The PCR amplification is another random process, where the count of newly generated molecules follows a binomial distribution with a probability of successful amplification. This process is likely to exacerbate the bias on the count distribution. This biased count distribution usually leads to additional data (sequence) loss at the DNA sequencing stage since the next-generation sequencing process is another round of random sampling [16], i.e., the chip only reads certain amounts of molecules from a molecule population. With a population of highly biased distribution, each element may have a different probability of being sampled. Hence, if the sampling size is inadequate (i.e., low sequencing coverage), the Poisson sampling effect would cause another round of data loss. We categorize data loss at this stage as *sequencing loss* (see Fig. 1A).

Apart from sequence loss, the base error is the other type of error in the sequencing data at the receiver. In Figs. 1A and 1B, hypothesized random base errors are black-colored for illustration. Before sending to the decoder, the received raw sequencing data is usually post-processed for a preliminary data reconstruction as shown in Figs. 1A and 1C. Note that no standard has been set yet for processing the sequencing data while the processing results given by different processing methods directly affect how many remained errors that the decoder needs to handle. Fig. 1C shows two potential processing results, i.e., successfully reconstructed sequence and sequence corruption, of which the sequence corruption remains to be resolved by the decoder.

2.2 Pair-End Sequencing and Sequence Alignment

Next-generation sequencing technologies provide protocols to generate reads from two ends of the DNA strand. These protocols enable the sequencer to recover long DNA sequences given that the sequence length is no longer than twice the read length. Besides, pair-end reading is also recognized to improve the sequencing accuracy due to the overlapping between the pair of reads. In our two previous works, Pair-end 150 (PE150) protocols were used to read DNA oligos with lengths from 190 to 199 [11], [15] which makes full use of the current synthesis and sequencing technologies. To merge the pair-end short reads into the long reads, several prevalent tools were designed [17], [18], [19], [20], among which we used FLASH [19] to merge PE150 reads (see Supplementary S1, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2022.3233914). To estimate the base error statistics, the merged reads from PE reads are aligned to their corresponding reference/original sequences using sequence alignment tools. Several tools were devised for sequence alignment [21], [22], [23], among which we used Bowtie 2 [21].

3 Results

3.1 Deriving the Overall Sequence Error Rate Consisting of Sequence Loss and Sequence Corruption

Sequence loss in DNA data storage channel might be due to the physical sequence loss in the sample preparation and storage and/or the sequencing loss in sequencing. We used

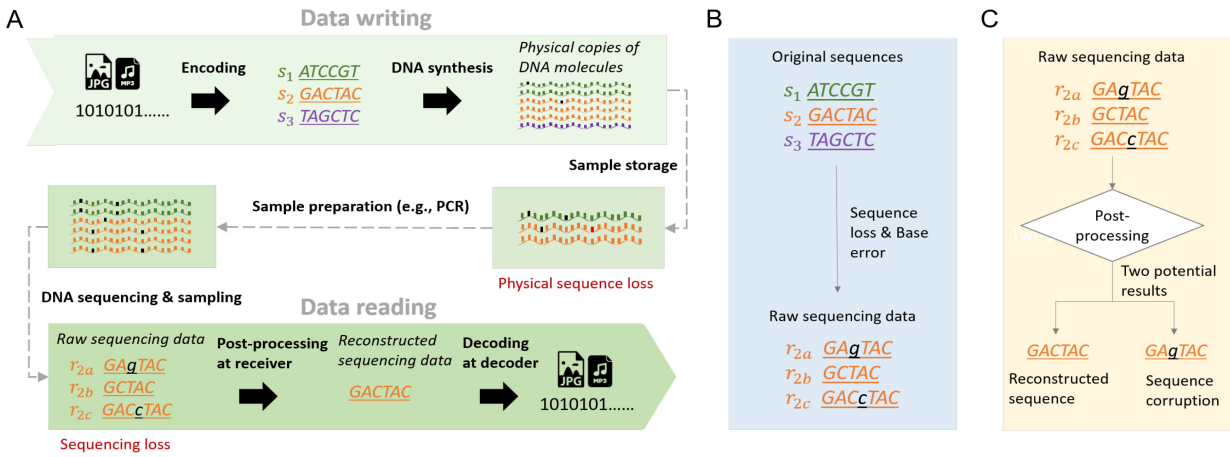

Fig. 1. Data flow and error characterization in DNA data storage. (A) Data writing starts from binary data encoded as DNA sequences and then synthesized. The physical (synthesized) copies of sequences are stored with a certain probability of physical sequence loss due to the random sampling process, and then the sample is prepared before sending for data reading. DNA sequencing and sampling initiate data reading where sequences might be lost due to another round of random sampling process. The sequenced raw data can be post-processed first before being sent to the decoder where binary data are recovered. (B) The general error characterization in DNA data storage where original encoded sequences might be lost due to random sampling processes or erroneous due to base errors such as insertion, deletion, and substitution. (C) Two outcomes with post-processing applied to raw sequencing data. The processed data can succeed or fail in reconstructing an encoded sequence back, and we name the failed result sequence corruption.

a model which computationally simulates the whole process of DNA data storage in [12] to study the sequence loss at the decoder. By defining *channel coverage* as the average number of reads per reference sequence the decoder receives, we found that when the channel coverage is sufficient, the overall sequence loss rate is lower bounded by the physical sequence loss (see Supplementary S2 Fig. 3, available online). Without loss of generality, the sequence loss is found to be higher when data are sampled from a population with more severe over-dispersion, i.e., smaller coefficient of variation (C.V.) (Supplementary S2 Fig. 4, available online). For insufficient channel coverage scenario (i.e., less than 10x), it was found that the overall sequence loss rate no longer changes linearly with the physical sequence, implying that the sequencing loss dominates the overall dropout rate (Supplementary S2 Fig. 5, available online). We also evaluated the model by fitting it with data from our previous work [15], where the correlation coefficient (i.e., $R^2 = 0.96$) shows that the sequencing sampling effect is well-simulated (Supplementary S2 Fig. 6, available online). Moreover, by comparing the experimental dilution effect in [8] with the simulated dilution effect in a modified version of the computational model in [12], we found that there is still a notable gap between the experiment and the simulation (see Supplementary S2 and S2 Fig. 7, available online). To further understand the gap, we used three different types of amplification efficiency p, i.e., constant, random, and strand-specific random, in the computational model to probe the association between PCR amplification and the source of the gap (see Supplementary S2 Fig. 8, available online). Setting the amplification efficiency as a strand-specific random variable in the computational model gives the closest approximation to the experimental results in [8].

Sequence corruption which is the undesired result after employing clustering algorithms but before decoding was not the focus of most existing works [8], [9], [12], [13]. However, the impact of sequence corruption on the decoding is not trivial when establishing more cost-effective and large-scale DNA data storage with less accurate synthesis and sequencing technologies where base error rates are higher and the copy counts of the reference sequences are limited at the receiver. In the following, we extensively study the overall sequence error rate by incorporating both sequence loss and sequence corruption.

3.1.1 Simplified Derivation of Sequence Error Rate

We start from the simplest formulation in which the copy count is assumed to be even and the base error rate is assumed to be constant, i.e., each base has the same error probability. With these assumptions, there is no sequence loss but only sequence corruption that is stemmed from the base error; and it highly depends on the available copy counts at the receiver which is denoted as channel coverage η. Besides, the sequence corruption rates may vary if different post-processing methods are used before decoding. Here, we formulate the sequence error rate for two commonly adopted methods, i.e., non-consensus (or trial-and-error) and consensus (i.e., majority selection at each position). The trial-and-error means one reference sequence is regarded as correctly recovered if at least one read copy of it at the receiver is error-free. This can be achieved by incorporating an error detection mechanism within each sequence, e.g., cyclic redundancy check (CRC) [11]. The majority selection is a well-known consensus algorithm for generating representative data of clustered data. One reference sequence is regarded as correctly recovered if the representative sequence is error-free. For simplicity, the formulation temporarily assumes binary majority selection at each position.

Illumina and Nanopore sequencing are the two commonly used sequencing techniques in the existing DNA data storage where Nanopore sequencing can sequence longer sequences but provides lower sequencing accuracy. To show how the channel coverage affects the sequence error rate in Illumina- and Nanopore-based DNA data storage differently, we applied correspondingly different values to parameters

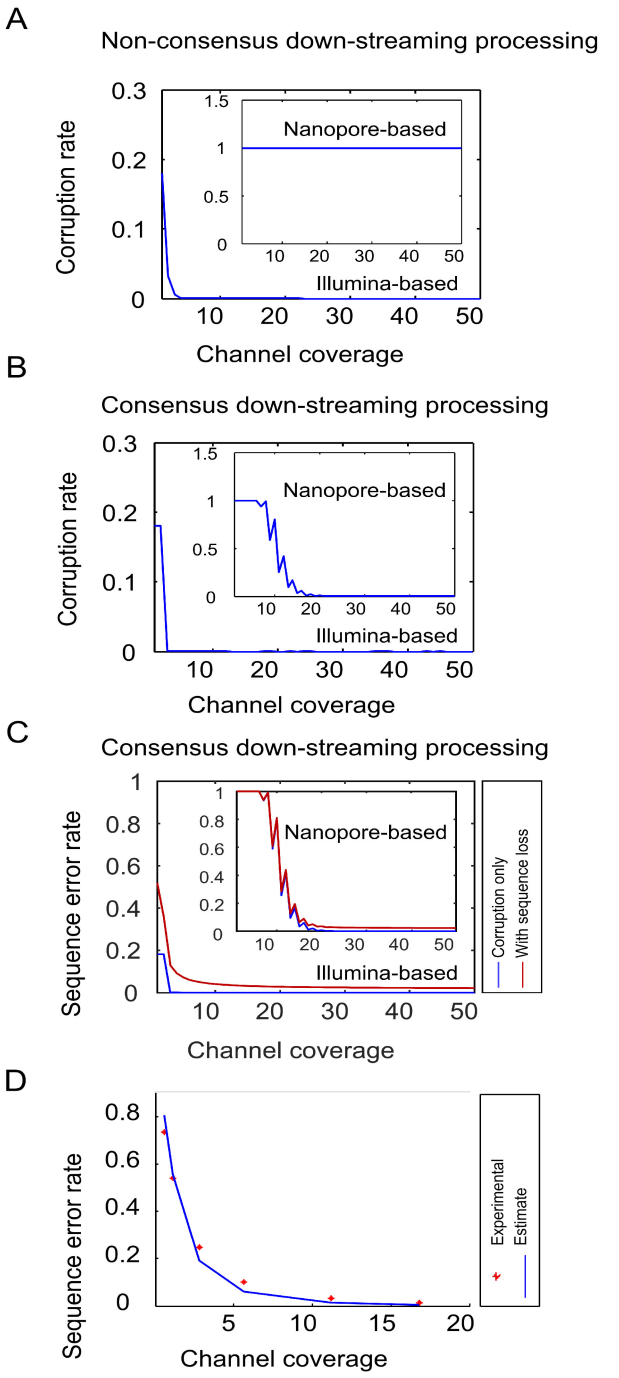

Fig. 2. Theoretical analysis and experimental result of the sequence corruption and sequence error rate against the channel coverage. The major graph in each sub-figure refers to the Illumina-based systems while the top right embedded figure in each sub-figure refers to the Nanopore-based systems. (A) The sequence corruption rate with the assumption of non-consensus method at the receiver. (B) The sequence corruption rate with the assumption of consensus method (i.e., majority selection at each position) at the receiver. (C) The overall sequence error rate that consists of sequence loss and sequence corruption. (D) Theoretical and experimental sequence error rate against channel coverage with the assumption of uneven copy distribution and using a non-consensus method. The overall sequence error rate decreases with the increase of coverage.

including base error rate ϵ, sequence length M, to the formulation $\Omega = (1-(1-\epsilon)^M)^\eta$ and depicted the sequence error rate against the channel coverage. Based on the practical measures, the base error rate ϵ is set to 0.1%, and the sequence length M is set to 200 to simulate the sequence corruption of Illumina-based DNA storage, i.e., $\Omega_{\text{illu}}(\eta;\epsilon,M)$. For simulating the Nanopore-based DNA storage, i.e., $\Omega_{\text{nano}}(\eta;\epsilon,M)$, the base error rate ϵ is set to 10% and the sequence length M is set to 1000. For the trial and error (non-consensus) case, i.e., the sequence is considered as corrupted only if all copies of the sequence are erroneous, the curves representing the association between sequencing coverage and sequence corruption are shown in Fig. 2A). The same parameter values corresponding to Illumina- and Nanopore-based are used to represent the consensus case with simplified majority selection at each position as shown in Fig. 2B.

In Fig. 2, all embedded figures in the sub-figures refer to the Nanopore-based while the rest refers to the Illumina-based. It was found that using Illumina sequencing, the sequence corruption rate decreases drastically with the increase of channel coverage for both non-consensus (Fig. 2A) and consensus cases (Fig. 2B). Specifically, with only ~ 5 channel coverage, the corruption rate could be reduced nearly to 0; and with the addition of a consensus processing, the minimum channel coverage is decreased by half, i.e., ~ 2.5. However, for channels using Nanopore sequencing, the corruption rate maintains a high plateau with a non-consensus method (embedded figure in Fig. 2A) and decreases much more gradually with the consensus method (embedded figure in Fig. 2B). With the consensus algorithm, the corruption rate approaches 0 for a minimum of ~ 20 coverage. Overall, the observation indicates that in Illumina-based storage, increasing channel coverage (read copy redundancy) could effectively reduce the error rate even without any consensus algorithm while in Nanopore-based storage, only with an appropriate consensus algorithm and sufficient coverage, the error rate can be reduced to an acceptable level.

Next, we generalize the formulation of sequence error rate with uneven copy count distribution. In this case, the sequence error is composed of sequence loss and sequence corruption. The average sequence loss rate against the average copy count, i.e., the channel coverage (η), can be well described by an exponentially decreasing curve $e^{-\lambda}$ in which λ is a random variable (RV) following an uneven sequence count distribution Λ. The overall sequence error rate against the channel coverage is shown in Fig. 2C, in which the blue and red curves represent the error rate before and after including sequence loss, respectively. Comparing the blue and red curves, we observe that the sequence loss has a more significant impact on the sequence error rate in Illumina-based DNA data storage. On the contrary, sequence corruption affects the sequence error rate more in Nanopore-based DNA data storage (embedded figure in Fig. 2C).

3.1.2 *Elaborated Derivation of Sequence Error Rate*

We adjust the simplified majority mechanism from binary to quaternary and extend the copy count (the channel coverage) from the constant value η to RV η_i subject to certain distribution H. Thus, for the non-consensus approach, the expected sequence error rate of Ω_1 with uneven copy distribution becomes

$$\Omega_1 = \sum_{\eta_i=0} \Pr(\eta_i|H)(1-(1-\epsilon)^M)^{\eta_i}, \tag{1}$$

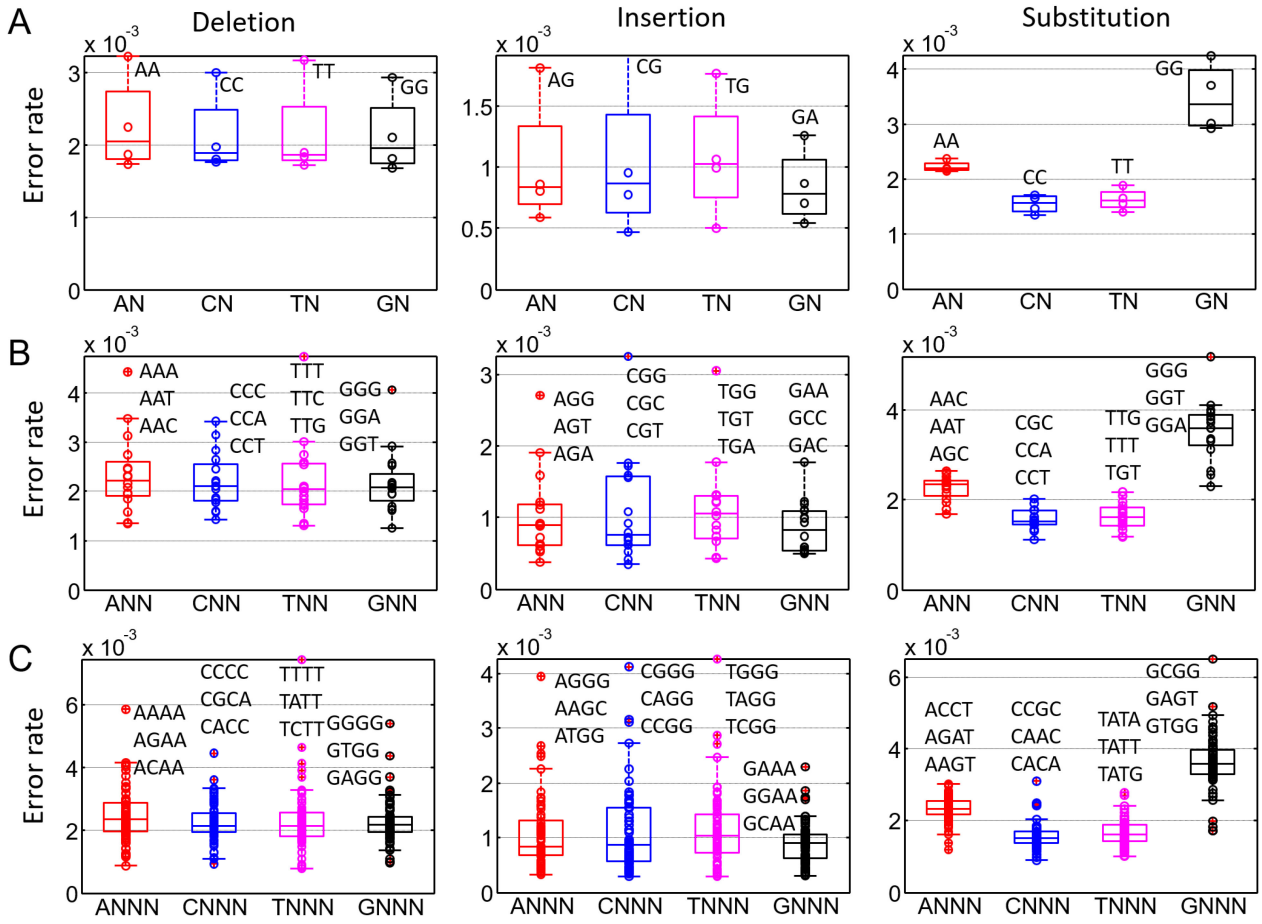

Fig. 3. Experimental results of uneven 2/3/4-mer error patterns for deletions, insertions, and substitutions in DNA data storage, in which the most erroneous patterns are marked correspondingly. With the first base being deleted, following an insertion, and being substituted, the error rates of (A) 2-mer patterns; (B) 3-mer patterns; (C) 4-mer patterns. The homopolymer is observed to have an impact on deletion errors. There are specific patterns that are prone to have an insertion in between. No significant discrimination among patterns is observed for substitution errors.

where η_i represents a copy count subject to a distribution H ($\eta_i \sim H$); ϵ is the base error rate; and M is the sequence length. Fig. 2D compares the sequence error rates of experimental data from [15] (by comparing the sequencing data with original encoded ata) with the estimates derived by (1) with distribution H set as negative binomial fitted from experimental data and other parameters including ϵ and M extracted from experimental data.

With majority selection as the consensus approach, we have Ω_2

$$\Omega_2 = 1 - \sum_{\eta_i=0} \Pr(\eta_i|H) \Bigg\{ \sum_{k_i=\lfloor \frac{\eta_i}{2} \rfloor+1}^{\eta_i} \Bigg\{ \binom{\eta_i}{k_i} (1-\epsilon)^{k_i} \epsilon^{\eta_i-k_i} \Bigg\} + \sum_{k_i=\lfloor \frac{\eta_i}{4} \rfloor+1}^{\frac{\eta_i}{2}} \Bigg\{ \binom{\eta_i}{k_i} (1-\epsilon)^{k_i} \epsilon^{\eta_i-k_i} \Bigg\{ 1 - \sum_{j_i=k_i+1}^{\eta_i-k_i} \Bigg\{ \binom{\eta_i-k_i}{j_i} \cdot \frac{2^{\eta_i-k_i-j_i}}{3^{\eta_i-k_i-1}} \Bigg\} - \frac{1}{2} \binom{\eta_i-k_i}{k_i} \cdot \frac{2^{\eta_i-2k_i}}{3^{\eta_i-k_i-1}} \Bigg\} \Bigg\} + \frac{3}{2} \text{sign}(\eta_i) \cdot \binom{\eta_i}{\frac{\eta_i}{4}} \binom{\frac{3\eta_i}{4}}{\frac{\eta_i}{4}} \binom{\frac{\eta_i}{2}}{\frac{\eta_i}{4}} \epsilon^{\frac{\eta_i}{4}} \left(\frac{1-\epsilon}{3}\right)^{\frac{\eta_i}{4}} \Bigg\}^M, \quad (2)$$

where $\text{sign}(x)$ is a sign function which equals 1 when $x \pmod 4 = 0$ while equals 0 when $x \pmod 4 \neq 0$; other notations are same as (1). The formulation implies that the biased copy count distribution is not only the origin of the sequence loss but also affects the sequence corruption rate after reconstruction. Specifically, the skewed count distribution of data at

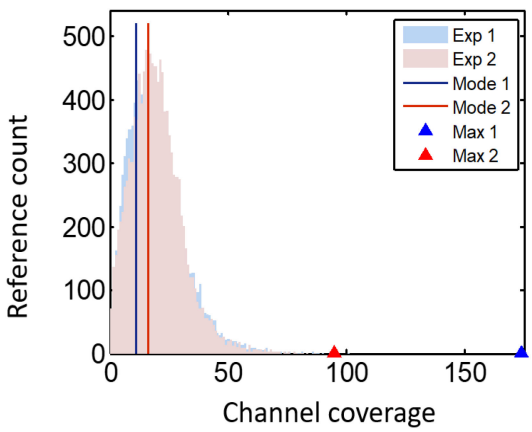

Fig. 4. Experimental copy count distributions for varied sample preparations. Based on data sets with 20x channel coverage, the blue-colored copy count distribution is from the single primer binding site (PBS) experimental set; while the red-colored copy count distribution is from the double PBS experimental set. The blue and red solid lines represent the modes of two distributions, respectively. The maximum counts observed in the two sets are triangle marked, and there is one reference sequence with 174 copy counts in the single PBS set (i.e., blue triangle). Both distributions approximate the negative binomial distribution, the single PBS distribution is with higher bias where the size parameter r is smaller. i.e., 2.7 versus 3.3.

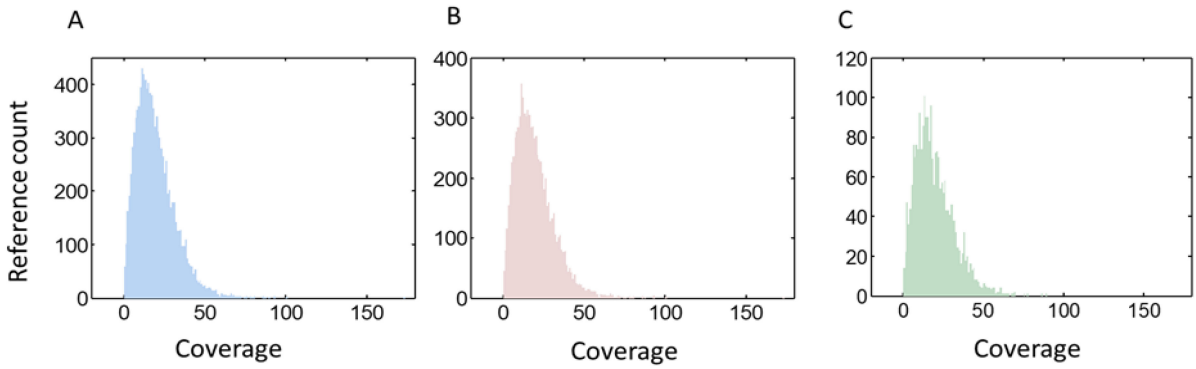

Fig. 5. The comparison of copy count (or channel coverage) distributions of A. all sequences, B. sequences including 4nt homopolymer, and C. sequences without 4nt homopolymer with 20x coverage of single PBS data set.

the receiver jeopardizes the overall performance of consensus methods which are usually designed under simplified assumptions, e.g., the distribution of raw data is even or normal. Acknowledging the bias existing in the distribution helps better design the consensus algorithms and better estimate the data reconstruction performance from the sequencing data.

3.1.3 Customizing the Derivation of Sequence Error Rate With Data-Dependent Errors

In the aforementioned formulations, a constant base error rate ϵ is used to represent the probability of random base errors. However, in systems like Nanopore-based systems, some errors occur in a sequence-dependent way. These systematic errors are quantitatively significant and relevant to the features of the reference sequences (or synthesized DNA molecules). For instance, it was found that around 44% of reads of homopolymer runs no less than 5 were observed to contain a deletion error [24]. This high error rate and the systematic fashion of the error occurrence aggravate the data recovery difficulty at the receiver. We summarize these errors as data-dependent errors and specify two virtual channels to differentiate the consequences of random and systematic errors (see Supplementary S3, available online). We specifically derive the formulation of the sequence error rate of channels that are prone to data-dependent systematic errors. For the non-consensus approach, we have the expected error rate Ω_3

$$\Omega_3 = \sum_{\eta_i=0} \Pr(\eta_i|H)\Bigg\{(1-P(M,l))\Big(1-(1-\epsilon)^M\Big) + P(M,l)\Bigg(1-(1-\epsilon)^M \sum_{\upsilon=1}^{\frac{M}{l+1}} \Pr(\Upsilon=\upsilon)(1-\alpha_h)^{\upsilon}\Bigg)\Bigg\}^{\eta_i}, \quad (3)$$

where $P(M,l) = 1 - q^{\lfloor M \log_q \lambda \rfloor - M}$ is the probability of a M-length q-ary sequence having at least one homopolymer larger than length l where λ is determined by the maximum homopolymer length l (see Supplementary S4, available online); $\Pr(\Upsilon = \upsilon)$ is the probability of a sequence having υ substrings with homopolymer longer than l ; α_h is the specific data-dependent systematic error rate; and other notations are the same as (1). For majority selection approach, we have Ω_4

$$\Omega_4 = 1 - (1 - P(M,l))P_C(\epsilon) - P(M,l)P'_C(\epsilon, \alpha_h), \quad (4)$$

where

$$P_C(x) = \sum_{\eta_i=0} \Pr(\eta_i|H)\Big\{\sum_{k_i=\lfloor \frac{\eta_i}{2} \rfloor+1}^{\eta_i} \Big\{\binom{\eta_i}{k_i}(1-x)^{k_i}x^{\eta_i-k_i}\Big\} + \sum_{k_i=\lfloor \frac{\eta_i}{4} \rfloor+1}^{\frac{\eta_i}{2}} \Big\{\binom{\eta_i}{k_i}(1-x)^{k_i}x^{\eta_i-k_i}\Big(1-\sum_{j_i=k_i+1}^{\eta_i-k_i}\binom{\eta_i-k_i}{j_i}\frac{2^{\eta_i-k_i-j_i}}{3^{\eta_i-k_i-1}} - \frac{1}{2}\binom{\eta_i-k_i}{k_i}\frac{2^{\eta_i;-2k_i}}{3^{\eta_i-k_i-1}}\Big) + \text{sign}(\eta_i)\frac{3}{2}\binom{\eta_i}{\frac{\eta_i}{4}}\binom{\frac{3\eta_i}{4}}{\frac{\eta_i}{4}}\binom{\frac{\eta_i}{2}}{\frac{\eta_i}{4}}\epsilon^{\frac{\eta_i}{4}}\big(\frac{1-\epsilon}{3}\big)^{\frac{\eta_i}{4}}\Big\}^M;$$

$$P'_C(a,b) = P_C(a) \cdot \sum_{y=1}^{\frac{M}{l+1}} \Pr(\Upsilon=\upsilon)\{\sum_{[k_i=\frac{\eta_i}{2}]}^{\eta_i}\binom{\eta_i}{k_i}(1-b)^{k_i}b^{\eta_i-k_i} + \frac{1}{2}$$

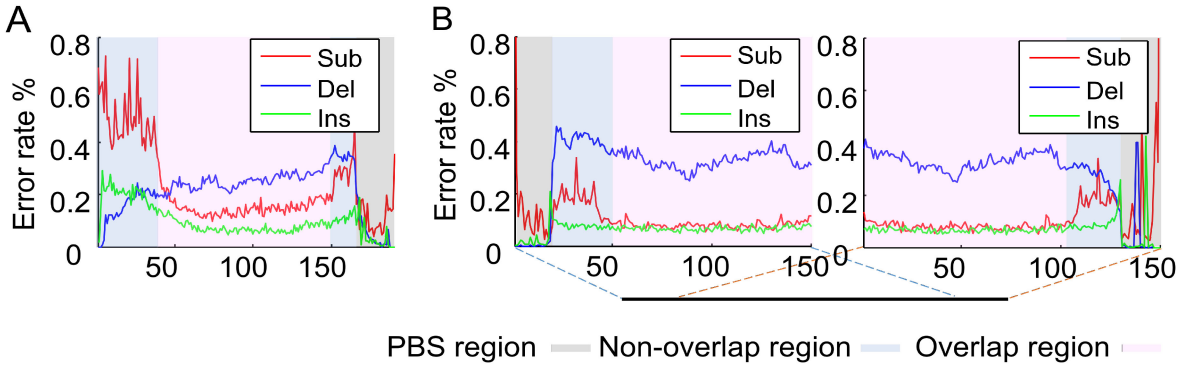

Fig. 6. Experimental positional error profile of reference sequence in DNA data storage using PE150 sequencing protocol and merging processing, in which three regions are highlighted along the coordinate. The positional error rate profile of (A) reference sequences with a length of 190nt. (B) reference sequences with lengths ranging from 190nt to 199nt. Two 150-length positional error profiles starting from two ends of the reference sequences are presented simultaneously to accommodate the variable-length feature of the reference sequences.

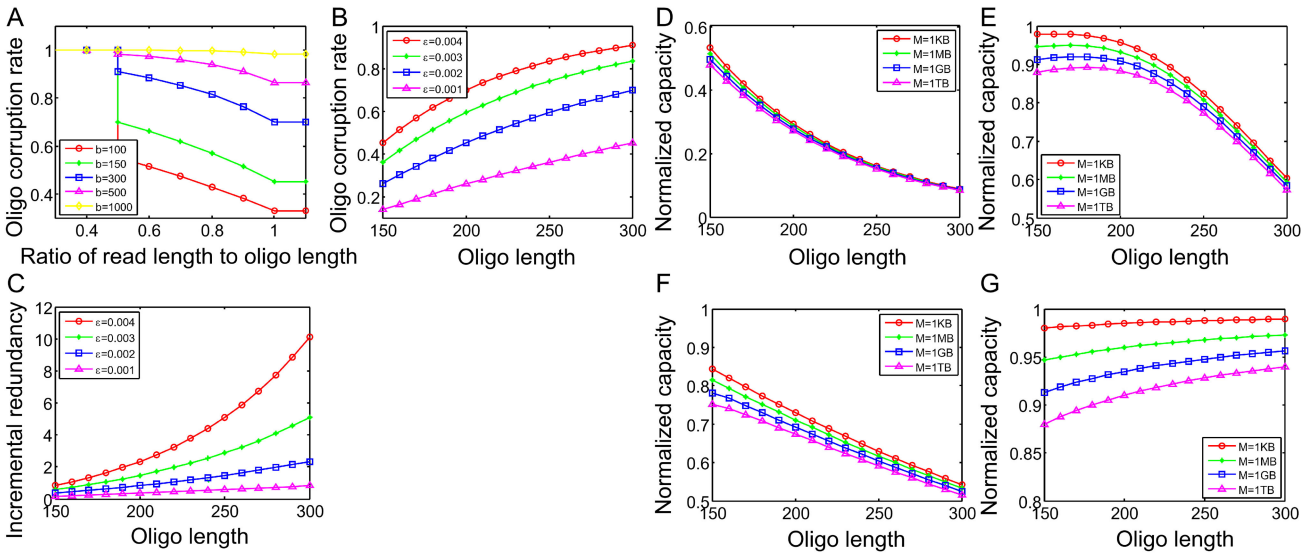

Fig. 7. Theoretical analysis of the impact of sequence length. (A) Sequence corruption rate against the ratio of PE read length to sequence length. Curves with different colors represent different sequence lengths b. With PE read length 150 and curves with different colors representing different base error rates ϵ, (B) sequence corruption rate against sequence length; (C) incremental redundancy for addressing sequence corruption against sequence length. The achieved capacity (normalized) against the sequence length, (D) when channel coverage $\eta = 1$ and raw base error rate $\epsilon = 0.8\%$; (E) when $\eta = 10$ and $\epsilon = 0.8\%$; (F) when $\eta = 1$ and $\epsilon = 0.2\%$; (G) when $\eta = 10$ and $\epsilon = 0.2\%$. Curves with different colors represent different sizes of stored data M.

$\text{sign}'(\eta_i)(1-b)^{k_i} b^{\eta_i - k_i} \upsilon$ and $\text{sign}'(x)$ equals 0 when $x \pmod 2 = 1$ and equals 1 when $x \pmod 2 = 0$. To alleviate the complexities of the formulations, the approximations of (3) and (4) could be found in Supplementary S5, available online. Note that the above formulation is built upon the assumption of storing non-constrained-formatted data where significant increases of systematic errors are present. If data has been constrained-formatted (avoiding the presence of long homopolymer) at the encoding step [25], the severe systematic errors could be avoided in the channel; and the overall sequence error rate of the channel could be measured by (1) and (2) again. However, this constrained formatting/encoding is exploited at the cost of reducing the code rate, i.e., less information stored per nucleotide.

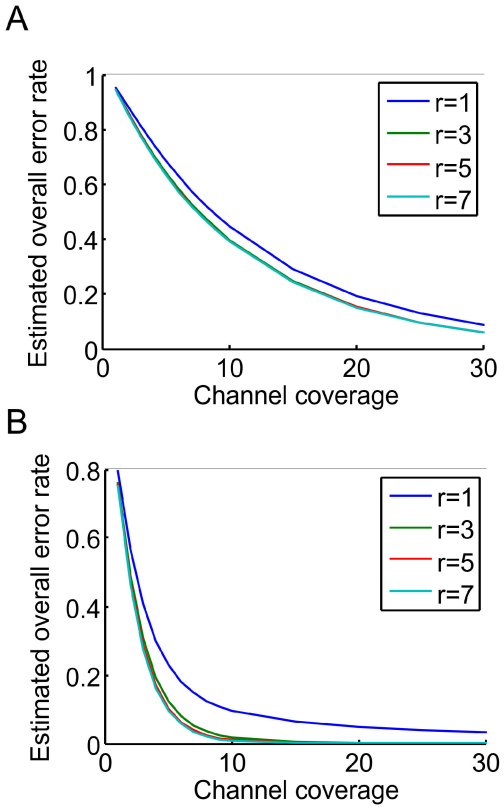

Fig. 8. With the assumption that the sequence length is twice the read length where the merged reads are with the full non-overlapped region, the theoretical overall sequence error rate in Illumina-based storage against the channel coverage, where the raw base error rate is (A) 0.8% similar to Miseq ; (B) 0.3% similar to Hiseq [30]. In each figure, different colors represent different size parameters of the copy count distribution that are ascribed to sequence loss.

3.2 Uneven Base-Level Errors in DNA Data Storage

3.2.1 *Uneven Transition Errors Among Nucleotide Bases*

We first examined the single base error profile. For fair comparison among independent works [11], [13], [15], we used the filtered reads (i.e., reads with the same length as the encoded sequences) to conduct the analysis. Note that with filtered data, many indel errors might be excluded from the analysis. Fortunately, though deletions are the most significant errors in DNA synthesis, its much lower error rates compared to sequencing errors and the prevalence and feasibility of filtering in downstream processing alleviate the impact of using filtered data on the effectiveness of the analysis. It was found that substitution is the dominant error regardless of the choice of sequencing platforms and experimental sets (see Supplementary S6, available online). Furthermore, we analyzed the base transition errors since it could help design more efficient codes as well as facilitate decoding [26]. The normalized transition probabilities are shown in Table 1, where the row labels denote the transmitting/reference bases and the column labels denote the transmitted bases. The bold values represent the transitions with the highest probability for four reference nucleotides. The Italic values represent the overall top 3 erroneous transitions. From the table, we could find that A to C, C to T, T

TABLE 1
The Base Transition Error Rates

	A	C	T	G	N
A	-	**0.157**	0.049	0.037	0.003
C	0.032	-	**0.078**	0.02	0.003
T	0.021	0.062	-	**0.072**	0.004
G	*0.201*	0.073	*0.182*	-	0.007

to G, and G to A are the most potential transitions for each reference base, i.e., A, C, T, and G. Regardless of the minor transitions from any of four bases to the N base, the A to C transition is over 3-fold and 4-fold to other two transitions accordingly, i.e., A to T and A to G. Likely, C is almost 3-fold and 4-fold possible to be recognized as T rather than A and G. For T base, the transitions to C and G are both significantly high with approximately 3-fold over T to A transition. Base G has the most considerable transition rates, in which G to A and G to T are about 10-fold than C to G. Overall, G to A, G to T, and A to C are the top three discernible transitions. Several results in the literature are generally consistent with our observations [27], [28], [29].

3.2.2 Uneven k-mer Error Patterns

In addition to single base error, we estimated the k-mer error patterns in DNA data storage channel. We first analyzed the 2-mer error patterns with respect to deletions, insertions, and substitutions (Fig. 3A). For deletions, we found that the most erroneous 2-mer patterns of four reference bases are their corresponding 2-mer repetitions i.e., AA, CC, TT, and GG. For insertions, the 2-mer patterns that are prone to having an in-between insertion are AG, CG, TG, and GA for A, C, T, and G, respectively. For substitution error rates of patterns where the first base is substituted to other bases, there is no significant discrimination albeit with the most erroneous 2-mer patterns for each base are the corresponding 2-mer repetitions. The G-oriented patterns show higher substitution rates than the error rates of the other patterns, which concurs with the single base error statistics.

Likewise, we analyzed the 3-mer error patterns (Fig. 3B). For deletions, the most error-prone 3-mer patterns for each nucleotide base are the corresponding 3-mer repetitions, i.e., AAA, CCC, TTT, and GGG, which are in line with 2-mer deletion patterns. Moreover, the 3-mer repetitions present higher error rates than the 2-mer repetitions, implying that longer homopolymer might have higher deletion tendency. Again, there are much higher insertion error rates for specific 3-mer patterns, including AGG, CGG, TGG, and GAA. Interestingly, except GCC, all other top 3 erroneous 3-mer patterns (that are shown in Fig. 3B) are with prefixes that are the most error-prone 2-mer patterns, i.e., AG, CG, TG, and GA. This infers that the insertions observed in the received data might highly relate to the neighboring bases. Similarly, most of the top 3 erroneous 3-mer patterns with the first nucleotide being substituted are consistent with the most error-prone 2-mer patterns. However, again, there is no significant discrimination between repetitive patterns and other non-repetitive patterns. Equivalently, the 4-mer repetitions are observed to have the highest tendency toward deletions. And the 4-mer repetitions are with higher deletion rates than the 2-mer and 3-mer repetitions, which further proves that the longer the homopolymer, the higher the probability to encounter deletions (Fig. 3C). For insertions, the most discernible 4-mer patterns for each base are AGGG, CGGG, TGGG, and GAAA. All of them are with 3-mer prefixes that are the most error-prone 3-mer patterns, i.e., AGG, CGG, TGG, and GAA. For most of the 4-mer patterns, we observe similar substitution rates as the 2-mer and 3-mer cases. This suggests that homopolymer does not have a significant impact on substitution rates. We also examined errors occurring at the second position in 2/3/4-mer and the last position in 3-mer and 4-mer (see Supplementary S7, available online) and we find that the result is either with no significant erroneous patterns or in line with the patterns observed in the first position.

3.3 Factors Impacting Overall Sequence Error Rates in DNA Data Storage

3.3.1 Sample Preparations Affect the Copy Count Distribution

To start with, in Fig. 4, we compared the copy count distributions of the reference sequences in our two experiments with different sample preparations [11], [15]. The scheme in [15] was designed with a single primer binding site (PBS) without conducting PCR amplification before proceeding to DNA sequencing, and the other one [11] was designed with double PBSs and the sample was amplified with 9 cycles of PCR before sequencing. We used the data sets with channel coverage 20x, which means that ideally 20 copy counts of each reference sequence could be found at the receiver. The two observed distributions both approximate negative binomial distribution which is the consequence of randomly downsampling from a large population with gamma distribution. Biases in the count distributions are observed for both experiments; and the bias in the single PBS set is larger with size parameter r of smaller value, i.e., 2.7 versus 3.3. Besides, in the single PBS set, one reference sequence has 174 copy counts far away from the mean coverage of 20x. The different degrees of PCR bias and PCR stochasticity (see Supplementary S8, available online) might ascribe to the bias difference between the two experiment sets. To further confirm the major source of the bias differentiation between two distributions, we compared the copy count distributions of all sequences, sequences including 4nt homopolymer, and sequences without 4nt homopolymer as shown in Fig. 5. It is observed that the sequence with maximum copy counts, i.e., 174, is with 4nt homopolymer. And there is no obvious discrepancy in distributions among the three sets. This implies that rather than distinct PCR biases caused by sequence-specific randomness (due to homopolymer differentiation), distinct degrees of PCR stochasticity caused by different sample preparations majorly explain the bias difference.

3.3.2 Sequence Structure and Downstream Processing Affect the Base Error, Sequence Corruption, and System Capacity

The basic data unit in DNA data storage, i.e., DNA sequence, is usually designed with length ~ 200 to tailor the current

synthesis and sequencing techniques. For Illumina-based systems, a PE150 protocol could be used to increase the sequencing accuracy by stitching two paired-end (PE) reads. We aligned the merged reads back to the reference sequences to study how the reference sequence length and the general merging processing cooperatively affect the error profile in DNA data storage channel. The average positional error rates along the coordinate of the reference sequence are shown in Fig. 6. The average base error rate along the coordinate is uneven where the overlapped region (pink-colored in Fig. 6) has a lower error rate than the non-overlapped region (blue-colored) does. And the substitution errors are reduced most notably in the overlapped region. The overlapped region is a region that corresponds to the universal regions shared by two PE reads. The unevenness between overlapped and non-overlapped regions is due to the gap between the length of reference sequences (i.e., 190nt and 190∼199nt) and the length of PE reads (i.e., 150nt). Moreover, by comparing the blue-colored (i.e., non-overlapped) regions in Fig. 6A, we could find that the non-overlapped region with a PBS region (gray-colored) as the adjacent has a lower error rate than the other non-overlapped region.

Additionally, we analyzed the data sets that have been filtered by lengths. This further proves structural design of the sequence, i.e, appended with single PBS or double PBS and fixed length or variable lengths, affects the data integrity at the decoder (see Supplementary S9, available online). It is found that despite that filtering alleviates errors especially indels, filtering might aggravate sequence loss especially when the number of reads provided at the sequencer is limited. This is because after filtering, fewer data are left as if sequencing with small coverage has been performed. Hence, we compared the overall sequence rates before and after filtering with a given number of reads (see Supplementary S10, available online). By incorporating the sequence loss and sequence corruption (under the trial-and-error assumption), the filtered data set shows the same sequence error rates as those of non-filtered data set up to coverage ~ 30x (Supplementary S10 Fig. 15C, available online). This is because the increase rate of sequence loss is the same as the decrease rate of sequence corruption after filtering when trial-and-error is used. Nevertheless, it is worth mentioning that the overall error rate with the assumption of using a consensus algorithm might not be the same case (see Supplementary S10, available online).

Inspired by the unevenness of base error rates along the sequence coordinate caused by the length difference between reference sequences and PE reads, we theoretically analyzed the impact of the sequence length on the sequence corruption rate. The relationship between the ratio of PE read length to the sequence length and the sequence corruption rate Ω follows:

$$\alpha(S_i) = \begin{cases} 1 & \frac{a}{b} < 0.5 \\ 1-(1-\epsilon)^{2(b-a)}(1-\frac{\epsilon}{2})^{(2a-b)} & 0.5 \le \frac{a}{b} < 1, \\ 1-(1-\frac{\epsilon}{2})^{b} & \frac{a}{b} \ge 1 \end{cases} \quad (5)$$

where a is the PE read length; b is the sequence length; ϵ is the raw base error rate.

First, Fig. 7A depicts how the ratio of read length to the sequence length affects the sequence corruption rate. Specifically, when the ratio is below 0.5, stitching two PE reads is unable to recover the reference sequence (see Supplementary S1 Fig. 1C, available online), leading to 100% corruption. When the ratio is from 0.5 to 1, the corruption rate decreases when the ratio increases. When the ratio is no less than 1, each PE read could ideally cover the whole sequence (see Supplementary S1 Fig. 1A, available online), rendering the merged read with a fully overlapped region and consequently reducing the corruption rate to a low floor. The raw error rate used to draw Fig. 7A is set to 0.8% approximating the Illumina Miseq sequencing. Comparing the curves with different colors, the impact of the ratio on the sequence corruption rate was found more noticeable for shorter sequence lengths. With PE read length 150, Figs. 7B and 7C show that the sequence length b directly affects the sequence corruption rate and incremental redundancy for addressing the corruption. Particularly, the increased corruption rate caused by the increased sequence length requires an increased redundancy to recover data. The incremental redundancy follows $\frac{1}{1-\Omega} - 1$, where $1-(1-\epsilon)^{2(b-a)}(1-\frac{\epsilon}{2})^{(2a-b)}$, as $b \in (150, 300)$ which gives the ratio $\frac{a}{b} \in (0.5, 1)$.

We continue to explore the impact of the sequence length on the achieved capacity with an assumption of regarding DNA data storage as an erasure channel where error-correcting code is applied at the sequence level to address the erasure, i.e., sequence corruption. Here, the capacity is determined by the redundancy required for correcting the sequence corruption and the redundancy required for indexing. These two redundancies are highly related to the sequence length, and the impacts of sequence length on them are opposite. Specifically, given a fixed PE read length, the longer the sequence length, the higher the sequence corruption rate and error correction redundancy, and the lower the achieved capacity of the system. In contrast, given fixed data size, the longer the sequence length, the lower number of sequences required to store the data, the lower redundancy required to index the data, and thus the higher the achieved capacity of the system. The achieved capacity with considerations of both error control correct redundancy and indexing redundancy could be derived as $C = (1-\Omega)(1-\frac{b_i}{b})$ where $\Omega = (1-(1-\epsilon)^{2(b-a)}(1-\frac{\epsilon}{2})^{(2a-b)})^{\eta}$ is the sequence corruption rate counting in the effect of coverage η with the assumption of the use of trial-and-error; other parameters are the same as the above mentioned; b_i is the redundancy required to index all encoded sequences including information sequences and redundancy sequences. b_i is resolved by equation $4^{b_i} = \frac{M}{(b-b_i)(1-\Omega)}$ where M is the size of data that is to be stored.

With different channel coverage η and base error rate ϵ, Figs. 7D, 7E, 7F, and 7G shows how the achieved capacity changes with the sequence length in which different colors represent different stored data sizes ranging from 1 Kilobyte to 1 Terabyte. We found that for higher base error rate systems (Figs. 7D and 7E), the achieved capacity decreases with the increase of sequence length, presenting that the impact of increased redundancy for error correction on the capacity plays the prime role. This trend also appears in lower base error rate systems when the coverage is 1x (Fig. 7F), i.e, ideally only one read for each reference sequence could be used for data reconstruction. Interestingly, the trend reverses in the lower base error rate system

when the coverage is 10x (Fig. 7G). The much lower corruption rates ascribed to the high coverage could be the reason for the reversed trend. The error correction redundancy required by the much lower corruption rates no longer weighs higher than the indexing redundancy in regard to affecting the capacity. Therefore, in Fig. 7G, the increased capacity with the increased sequence length is mainly due to the decreased indexing redundancy. This indicates that in most cases, designing sequences with short lengths could improve the achieved capacity (Figs. 7D, 7E, and 7F). However, for storage systems with very low raw base errors and sufficient coverage at the receiver (Fig. 7G), sequence length could be designed as long as possible (up to twice of PE read length) to improve the achieved capacity.

3.3.3 Theoretical Estimation of the Overall Sequence Error Rate

Using different synthesis techniques and experimental settings gives different count distributions at the receiver, leading to distinct sequence loss rates. Meanwhile, using different sequencing techniques and data processing methods gives different base error rates at the receiver, leading to distinct sequence corruption rates. We thus theoretically estimate the overall sequence error rate consisting of sequence loss and sequence corruption by setting a range of practical values to several important impacting factors. With the assumption of using a non-consensus post-processing method after merging PE reads, the formulation of the overall sequence error rate is an extended version of (1) where the base error is no longer constant along the coordinate but varies with the region, i.e., overlapped or non-overlapped.

We separately consider two scenarios, i.e., the merged reads with fully non-overlapped (Fig. 8) and fully overlapped regions (Supplementary S11 Fig. 16, available online). Build upon these considerations and with the assumption of using trial and error before decoding, we formulate the estimation of overall sequence error rate as follows:

$$\left(\frac{r}{\eta+r}\right)^{r}+\left(1-\left(\frac{r}{\eta+r}\right)^{r}\right)\Omega, \tag{6}$$

where r is the size parameter of the copy count distribution (which approximates to negative binomial (NB) distribution at the receiver; η is the mean coverage of reference sequence at the receiver; $\Omega=(1-(1-\epsilon)^{2(b-a)}(1-\frac{\epsilon}{2})^{(2a-b)})^{\eta}$ where ϵ is the raw base error rate (i.e., the error rate in the non-overlapped region); b is sequence length with a range a to $2a$ where a is the PE read length. Note that the formulation separately considers sequence loss which is deduced by size parameter r and mean coverage η and sequence corruption which is deduced by raw base error rate ϵ, sequence length b, PE read length a, and mean coverage η. Another more precise way to construct estimation formulation is using random variable coverage η_i rather than mean coverage η where η_i follows NB distribution $NB(r,p=\frac{r}{\eta+r})$. In this way, the sequence loss ($\eta_i=0$) and sequence corruption ($\eta_i\neq 0$) can be denoted simultaneously.

First, the merged reads are with the fully overlapped region. To comply with the assumption, the sequence length is set equal to the PE read length (i.e., 150nt). The sequence error rate against the channel coverage is illustrated (see Supplementary S11, available online). In general, the higher the base error rate is, the higher the channel coverage is required to achieve a similar sequence error rate at the decoder. The variations among different base error rates, i.e., among sub-figures, are not notable. In the aspect of copy count distribution, the smaller the size parameter is, i.e., the more over-dispersion of the distribution, the higher the sequence error rate. Next, we analyzed the non-overlapped read case by setting the sequence length twice of the PE read length. Similarly, we draw two sub-figures, i.e., Figs. 8A and 8B, corresponding to two different raw base error rates, i.e., 0.8%, 0.3%, corresponding to Miseq and Hiseq Illumina sequencing, respectively [30]. The trend of the sequence error rate against the channel coverage, in this case, is the same as the overlapped case, but the discrimination among different base error rates, i.e., among sub-figures, is more notable. Comparing these two groups of figures (Fig. 8 versus Supplementary S11 Fig. 16, available online), it could be found that to obtain a similar sequence error rate with the same base error rate and size parameter, the required coverage of the non-overlapped case is no less than the overlapped case, i.e., $\sim$ 4-fold for 0.8% base error rate, $\sim$ 1.5-fold for 0.3%. To this end, we conclude that if the raw base error rate could be kept around 0.3% or less, the sequences could be designed with long lengths (i.e., from 150 to 300) where merged reads are all non-overlapped. However, for systems with higher base error rates, short sequence design (i.e., no longer than 150) which leads to all merged reads overlapped or semi-overlapped is a better choice.

4 Discussion

Distinct from other traditional storage systems, DNA data storage systems exhibit a few unique characteristics. Specifically, there are generally amounts of redundant data copies of original data albeit these copies might be corrupted by base errors; the number of redundant data copies for each original data unit is uneven. Some existing works [8], [13] substantially discussed the adverse consequence of the uneven copy count distribution, i.e., sequence loss, while did not discuss the multi copies' benefit to data reconstruction. Also, the physical redundancy of data copies at the decoder was excluded from the channel and discussed separately from the logical redundancy of the error control code. In fact, the multi-count data feature of DNA data storage enables a pre-decoder data reconstruction from multi (erroneous) copies where we term the failure of the reconstruction as sequence corruption. With the preliminary reconstruction before decoding, the data imperfection at the decoder that consists of sequence loss and sequence corruption offers a unified error profile for DNA data storage channel, easing the channel analysis and giving new insights for future code design. Note that this work did not try to differentiate errors presenting at different stages of DNA data storage, e.g., DNA synthesis, sample preparation, and PCR amplification, but characterized the errors from the aspect of data reconstruction at the receiver with centre to errors observed and to be addressed by the decoder. This

simplification of characterization is due to the fact that the decoder essentially disregards where the observed errors come from. However, future works aiming to differentiate the errors could be interesting as they could shed light on the design of biological experiments provided that the challenge of individually investigating those processes with high complexity and uncertainty is overcome.

Diving into the data imperfection observed from the experiments, biases have been found both in copy counts and base error patterns. The existence of these biases further distinguishes DNA data storage from other conventional storage systems, suggesting that facilitating higher performance gains in terms of capacity, reliability, and robustness in DNA data storage are possible. Moreover, the unevenness in the error rates of base patterns, including uneven error rates of single base transition and k-mer deletion/insertion, could be used as prior knowledge for decoding and optimizing encoding. For instance, using the transition tendency as the additional information to the decoder increases the error correction performance [26]. Also, the uneven transition feature could be leveraged to design unequal codes with higher efficiency. In addition, the deletion-prone characteristic in the long homopolymer patterns especially in Nanopore-based systems suggests that coding techniques that restrict the homopolymer length, i.e., constrained coding [25], [31], [32], might be a promising solution provided that the subsequent reduction on code rate/capacity is tolerable. Additionally, with the prevalent PE sequencing protocol and merging processing as the premises, the uneven error rate along the sequence coordinate (between non-overlapped and overlapped regions) is another unevenness that could be used for code design, e.g., unequal encoding, to increase the achieved capacity.

In this work, the impact of sample preparation on the data at the receiver was investigated based on two experiments with different rounds of PCR amplifications before DNA sequencing. We have shown that samples with more rounds of PCR attribute to less biased sequenced data since it provides a more sufficient initial amount of molecules avoiding severe PCR stochasticity. Besides PCR, other sample preparation steps should also affect the data integrity at the receiver and could be investigated in future work. The impact of sequence length on the data imperfection at the decoder and channel capacity was also theoretically studied. It was observed that only with sufficient coverage and a low base error rate, the achieved capacity could increase with the increase of sequence length. However, in other cases, the achieved capacity decreases with the increase of the length because of the increased redundancy for addressing the increased corruption rate. Hence, the system should be designed through comprehensive consideration of the involved impacting factors and the trade-offs among them.

5 Conclusion

We have conducted a comprehensive investigation of errors in DNA data storage channel. Quantitatively, the data imperfection including sequence loss and sequence corruption at the decoder has been presented. Besides deriving the sequence error rate to monitor the data reconstruction demand, we also further studied the imperfect data and found out that unevenness exists in several aspects and it could in turn help design systems with better performance. Additionally, we experimentally and theoretically analyzed the sequence error rates under different experiment settings and various but realistic parameter settings, including sequence lengths, base error rates, and over-dispersion degrees of distribution. From the perspective of data reconstruction, the results reported provide new perspectives for the development of advanced future DNA data storage.

Acknowledgments

We would like to thank Sergio Peisajovich and Saurabh Nirantar from Illumina, and Luo Lei from NovogeneAIT. We thank Ng Yi Mei for assisting on the initial analysis.

References

[1] G. M. Church, Y. Gao, and S. Kosuri, "Next-generation digital information storage in DNA," *Science*, vol. 337, no. 6102, pp. 1628–1628, 2012.
[2] N. Goldman et al., "Towards practical, high-capacity, low-maintenance information storage in synthesized DNA," *Nature*, vol. 494, no. 7435, pp. 77–80, 2013.
[3] R. N. Grass, R. Heckel, M. Puddu, D. Paunescu, and W. J. Stark, "Robust chemical preservation of digital information on DNA in silica with error-correcting codes," *Angewandte Chemie Int. Ed.*, vol. 54, no. 8, pp. 2552–2555, 2015.
[4] S. H. T. Yazdi, Y. Yuan, J. Ma, H. Zhao, and O. Milenkovic, "A rewritable, random-access DNA-based storage system," *Sci. Rep.*, vol. 5, no. 1, pp. 1–10, 2015.
[5] J. Bornholt, R. Lopez, D. M. Carmean, L. Ceze, G. Seelig, and K. Strauss, "A DNA-based archival storage system," in *Proc. 21st Int. Conf. Archit. Support Program. Lang. Oper. Syst.*, 2016, pp. 637–649.
[6] M. Blawat et al., "Forward error correction for DNA data storage," *Procedia Comput. Sci.*, vol. 80, pp. 1011–1022, 2016.
[7] S. H. T. Yazdi, R. Gabrys, and O. Milenkovic, "Portable and error-free DNA-based data storage," *Sci. Rep.*, vol. 7, no. 1, pp. 1–6, 2017.
[8] Y. Erlich and D. Zielinski, "DNA fountain enables a robust and efficient storage architecture," *Science*, vol. 355, no. 6328, pp. 950–954, 2017.
[9] L. Organick et al., "Random access in large-scale DNA data storage," *Nature Biotechnol.*, vol. 36, no. 3, 2018, Art. no. 242.
[10] Y. Choi et al., "High information capacity DNA-based data storage with augmented encoding characters using degenerate bases," *Sci. Rep.*, vol. 9, no. 1, pp. 1–7, 2019.
[11] Y. Wang, M. Noor-A-Rahim, J. Zhang, E. Gunawan, Y. L. Guan, and C. L. Poh, "High capacity DNA data storage with variable-length oligonucleotides using repeat accumulate code and hybrid mapping," *J. Biol. Eng.*, vol. 13, no. 1, pp. 1–11, 2019.
[12] Y.-J. Chen et al., "Quantifying molecular bias in DNA data storage," *Nature Commun.*, vol. 11, no. 1, pp. 1–9, 2020.
[13] R. Heckel, G. Mikutis, and R. N. Grass, "A characterization of the DNA data storage channel," *Sci. Rep.*, vol. 9, no. 1, pp. 1–12, 2019.
[14] O. Sabary, Y. Orlev, R. Shafir, L. Anavy, E. Yaakobi, and Z. Yakhini, "SOLQC: Synthetic Oligo library quality control tool," *Bioinformatics*, vol. 37, no. 5, pp. 720–722, 2021.
[15] Y. Wang, M. Noor-A-Rahim, J. Zhang, E. Gunawan, Y. L. Guan, and C. L. Poh, "Oligo design with single primer binding site for high capacity DNA-based data storage," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 17, no. 6, pp. 2176–2182, Nov./Dec. 2020.
[16] D. Aird et al., "Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries," *Genome Biol.*, vol. 12, no. 2, pp. 1–14, 2011.
[17] J. Zhang, K. Kobert, T. Flouri, and A. Stamatakis, "PEAR: A fast and accurate Illumina paired-end read merger," *Bioinformatics*, vol. 30, no. 5, pp. 614–620, 2014.
[18] A. P. Masella, A. K. Bartram, J. M. Truszkowski, D. G. Brown, and J. D. Neufeld, "PANDAseq: Paired-end assembler for Illumina sequences," *BMC Bioinf.*, vol. 13, no. 1, pp. 1–7, 2012.
[19] T. Magoč and S. L. Salzberg, "FLASH: Fast length adjustment of short reads to improve genome assemblies," *Bioinformatics*, vol. 27, no. 21, pp. 2957–2963, 2011.

[20] B. Liu et al., "COPE: An accurate k-mer-based pair-end reads connection tool to facilitate genome assembly," *Bioinformatics*, vol. 28, no. 22, pp. 2870–2874, 2012.
[21] B. Langmead and S. L. Salzberg, "Fast gapped-read alignment with Bowtie 2," *Nature Methods*, vol. 9, no. 4, 2012, Art. no. 357.
[22] R. Li et al., "SOAP2: An improved ultrafast tool for short read alignment," *Bioinformatics*, vol. 25, no. 15, pp. 1966–1967, 2009.
[23] H. Li and R. Durbin, "Fast and accurate short read alignment with burrows–wheeler transform," *Bioinformatics*, vol. 25, no. 14, pp. 1754–1760, 2009.
[24] R. Lopez et al., "DNA assembly for nanopore data storage readout," *Nature Commun.*, vol. 10, no. 1, pp. 1–9, 2019.
[25] Y. Wang, M. Noor-A-Rahim, E. Gunawan, Y. L. Guan, and C. L. Poh, "Construction of bio-constrained code for DNA data storage," *IEEE Commun. Lett.*, vol. 23, no. 6, pp. 963–966, Jun. 2019.
[26] L. Deng et al., "Optimized code design for constrained DNA data storage with asymmetric errors," *IEEE Access*, vol. 7, pp. 84107–84121, 2019.
[27] F. Pfeiffer et al., "Systematic evaluation of error rates and causes in short samples in next-generation sequencing," *Sci. Rep.*, vol. 8, no. 1, pp. 1–14, 2018.
[28] G. Chen, S. Mosier, C. D. Gocke, M.-T. Lin, and J. R. Eshleman, "Cytosine deamination is a major cause of baseline noise in next-generation sequencing," *Mol. Diagnosis Ther.*, vol. 18, no. 5, pp. 587–593, 2014.
[29] X. Ma et al., "Analysis of error profiles in deep next-generation sequencing data," *Genome Biol.*, vol. 20, no. 1, pp. 1–15, 2019.
[30] M. A. Quail et al., "A tale of three next generation sequencing platforms: Comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers," *BMC Genomic.*, vol. 13, no. 1, pp. 1–13, 2012.
[31] K. A. S. Immink and K. Cai, "Design of capacity-approaching constrained codes for DNA-based storage systems," *IEEE Commun. Lett.*, vol. 22, no. 2, pp. 224–227, Feb. 2018.
[32] W. Song, K. Cai, M. Zhang, and C. Yuen, "Codes with run-length and GC-content constraints for DNA-based data storage," *IEEE Commun. Lett.*, vol. 22, no. 10, pp. 2004–2007, Oct. 2018.

Yixin Wang received the BSc degree from the Harbin Institute of Technology, Weihai, China, in 2016, and the MSc degree from Nanyang Technological University (NTU), Singapore, in 2017, where she is currently working toward the PhD degree with the School of Electrical Electronic Engineering. Her research interests include coding for deoxyribonucleic acid data storage, constrained codes, and error control codes.

Md. Noor-A-Rahim received the PhD degree from the Institute for Telecommunications Research, University of South Australia, Australia, in 2015. He was a postdoctoral research fellow with the Centre for Infocomm Technology (INFINITUS), Nanyang Technological University (NTU), Singapore. He is currently a senior postdoctoral researcher and a Marie-Curie research fellow with the School of Computer Science and IT, University College Cork, Ireland. His research interests include information theory, wireless communications, and vehicular communications. He was a recipient of the Michael Miller Medal from the Institute for Telecommunications Research (ITR), University of South Australia, for the most outstanding Ph.D. thesis, in 2015.

Erry Gunawan received the BSc degree in electrical and electronic engineering from the University of Leeds, and the MBA and PhD degrees from Bradford University. From 1984 to 1988, he was a satellite communication system engineer with Communication Systems Research Ltd., Ilkley, U.K. In 1988, he moved to Space Communication (SAT-TEL) Ltd., Northampton, U.K. In 1989, he joined the School of Electrical and Electronic Engineering, Nanyang Technological University, where he is currently an associate professor. He has been a consultant with Sytek Technical Associates, Singapore, on the development of a device to enhance the security of data transmitted through facsimile machines, Addvalue Communications Pte Ltd., on DECT and Bluetooth systems, and also RFNet Technologies Pte Ltd., Singapore, for IDA project on New Generation Wireless LAN (IEEE 802.11a). He conducted courses for MINDEF and NTUs MBA program. He is appointed as an external examiner by Multimedia University for a MEngSc candidate. He has published more than 80 papers in international journals and more than 70 international conference papers on error correction codings, modeling of cellular communications systems, power control for CDMA cellular systems, MAC protocols, multicarrier modulations, multiuser detections, space-time coding, radio-location systems, MIMO interference channel, and the applications of UWB radar for vital sign sensing and medical imaging. He is a technical reviewer of various international journals, such as the *IEEE Transactions on Vehicular Technology*, *IEEE Journal on Selected Areas in Communications*, *IEEE Transactions on Signal Processing*, and *IEEE Communications Letters*.

Yong L. Guan is currently a tenured associate professor with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. He has led 13 past and present externally funded research projects on advanced wireless communication techniques, coding for 10-Tb/in2 magnetic recording, acoustic telemetry for drilling application, and so on with the total funding of over SGD 9 million. His research interests broadly include coding, signal design, and signal processing for communication systems, storage systems, and information security systems. He has published an invited monograph, three book chapters, and more than 300 journal and conference papers. He was an associate editor of the *IEEE Signal Processing Letters*. He is also an associate editor of the *IEEE Transactions on Vehicular Technology* and the chair of the IEEE ComSoc Singapore Chapter.

Chueh L. Poh received the BEng degree in electrical and electronic engineering from Nanyang Technological University (NTU), Singapore, and the PhD degree in bioengineering from Imperial College London, U.K. He is currently an associate professor with the Department of Biomedical Engineering, National University of Singapore (NUS), Singapore. He is also a principal investigator with NUS Synthetic Biology for Clinical and Technological Innovation (SynCTI) and leads the NUS Biofoundry. He is also the assistant dean (Outreach-External Relations and Outreach) of the Faculty of Engineering, NUS. He is also the co-founder of a Singapore start-up company, AdvanceSyn Pte Ltd., which specializes on providing model-assisted design tools and services for Synthetic Biology. His research group has been reprogramming microbes for medical and industrial applications. His current research interests include microbial biosensors, optogenetics, synthetic gene circuits design and automation, deoxyribonucleic acid data storage, modeling of biological systems for design, and computer-aided design (CAD) tools for SynBio. He has received a number of awards, including the Tan Chin Tuan Fellowship, in 2012, and the NTU Excellence in Teaching Award, in 2010. He is also the co-editor-in-chief of the *IET Engineering Biology Journal*.