RESEARCH ARTICLE

National Science Review
8: nwaa227, 2021
doi: 10.1093/nsr/nwaa227
Advance access publication 3 September 2020

[1]Biomedical Pioneering Innovation Center (BIOPIC), School of Life Sciences, Beijing Advanced Innovation Center for Genomics (ICG), and Peking-Tsinghua Center for Life Sciences, Peking University, Beijing 100871, China; [2]Center for Bioinformatics, State Key Laboratory of Protein Engineering and Plant Genetic Engineering, Peking University, Beijing 100871, China; [3]Center for Quantitative Biology, Peking University, Beijing 100871, China; [4]College of Engineering, Peking University, Beijing 100871, China; [5]College of Chemistry and Molecular Engineering, Peking University, Beijing 100871, China; [6]Institute for Cell Analysis, Shenzhen Bay Laboratory, Guangdong 518132, China and [7]Chinese Institute for Brain Research (CIBR), Beijing 102206, China

*Corresponding authors. E-mails: czitian@gmail.com; yanyi@pku.edu.cn

Received 11 May 2020; **Revised** 16 August 2020; **Accepted** 16 August 2020

CHEMISTRY

A virtual sequencer reveals the dephasing patterns in error-correction code DNA sequencing

Wenxiong Zhou[1], Li Kang[1], Haifeng Duan[1], Shuo Qiao[1], Louis Tao[2,3], Zitian Chen[1,4,*] and Yanyi Huang[1,4,5,6,7,*]

ABSTRACT

An error-correction code (ECC) sequencing approach has recently been reported to effectively reduce sequencing errors by interrogating a DNA fragment with three orthogonal degenerate sequencing-by-synthesis (SBS) reactions. However, similar to other non-single-molecule SBS methods, the reaction will gradually lose its synchronization within a molecular colony in ECC sequencing. This phenomenon, called dephasing, causes sequencing error, and in ECC sequencing, induces distinctive dephasing patterns. To understand the characteristic dephasing patterns of the dual-base flowgram in ECC sequencing and to generate a correction algorithm, we built a virtual sequencer *in silico*. Starting from first principles and based on sequencing chemical reactions, we simulated ECC sequencing results, identified the key factors of dephasing in ECC sequencing chemistry and designed an effective dephasing algorithm. The results show that our dephasing algorithm is applicable to sequencing signals with at least 500 cycles, or 1000-bp average read length, with acceptably low error rate for further parity checks and ECC deduction. Our virtual sequencer with our dephasing algorithm can further be extended to a dichromatic form of ECC sequencing, allowing for a potentially much more accurate sequencing approach.

Keywords: DNA sequencing, error-correction code, dephasing, computer simulation, sequencing-by-synthesis

INTRODUCTION

Next-generation sequencing (NGS) technology has transformed biological and medical research dramatically [1–7]. However, mainstream NGS methods require a DNA amplification step to amplify the signal, causing a phenomenon called 'dephasing': within a clone of identical DNA molecules not every molecule is reacting at the same pace, and such asynchronization leads to a mixture of the measurable signals used for sequencing identification. Therefore, dephasing intrinsically limits the accuracy and read length of DNA sequencing. Although single-molecule sequencing may avoid dephasing and can reach an ultra-high read length, the low signal-to-noise ratio and natural stochasticity make it challenging to achieve satisfactory accuracy [8–13]. To date, NGS methods relying on clonal amplification and sequencing are still the major technology in fundamental biology studies and clinical applications. Inventing new chemistry for sequencing DNA is always intriguing but challenging, and dephasing is one of the major problems most new chemistry needs to overcome before reliable sequencing results can be provided.

To reduce DNA sequencing errors originating from the imperfect chemical reactions, we introduced the error-correction code (ECC) concept into sequencing-by-synthesis (SBS) reactions to deduce an unambiguous DNA sequence using three degenerate sequences obtained by a novel dual-base flowgram [14] (Fig. 1a and b). In this approach, we use three orthogonally generated degenerate sequences, which can be considered as binary strings, to perform a parity check in between, and correct the errors through Bayesian probability calculations. The lengths of consecutive identical degenerate bases, the ideal sequencing signals, are called the 'degenerate polymer length' (DPL). Although this ECC sequencing technology requires specific unnatural nucleotides as substrates

© The Author(s) 2020. Published by Oxford University Press on behalf of China Science Publishing & Media Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from https://academic.oup.com/nsr/article/8/5/nwaa227/5900999 by guest on 10 August 2026

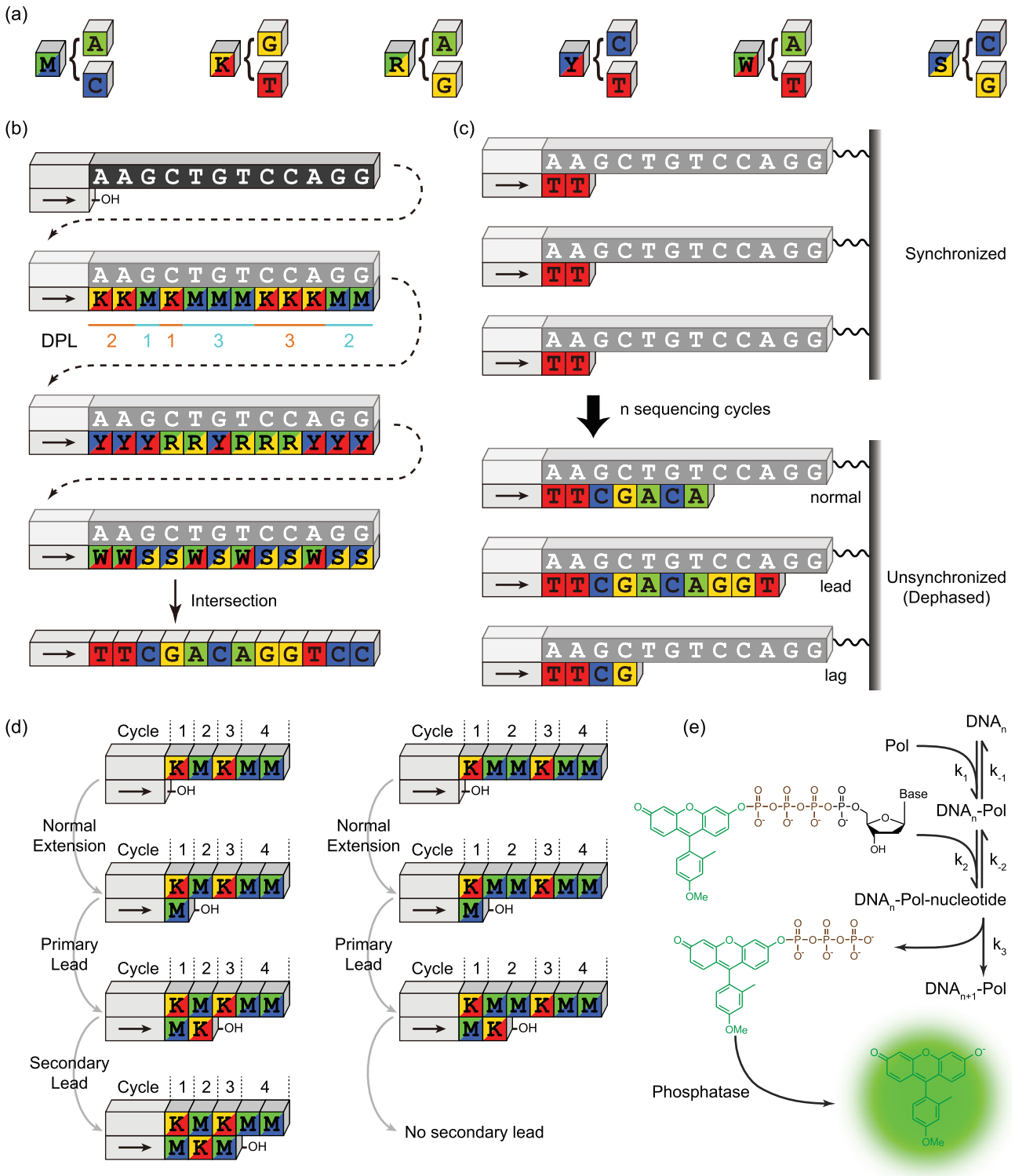

Figure 1. Schematic. (a) The six degenerate bases. (b) Schematic of error-correction code (ECC) sequencing. The ideal sequencing signals, or the lengths of the consecutive identical degenerate bases, is called the 'degenerate polymer length' (DPL). (c) The phenomenon that the nascent DNA strands progressively lose their synchronicity, also called 'dephasing,' is common in sequencing-by-synthesis-based DNA sequencing technologies. (d) Illustration of 'one-base slippage' (OBS) scheme. (e) Chemical reactions simulated in the virtual sequencer.

Downloaded from https://academic.oup.com/nsr/article/8/5/nwaa227/5900999 by guest on 10 August 2026

for polymerase to incorporate into the newly synthesized strand, it does hold great potential for providing long sequencing reads while maintaining high accuracy. We noted a unique dephasing pattern exhibited by the specific dual-base flowgram used in ECC sequencing. As an inevitable phenomenon that describes the loss of synchronization (phase) between the DNA extension reactions in a clone of DNA molecules being sequenced (Fig. 1c), dephasing has two components, lead and lag, meaning that the reaction happens in advance or is delayed, respectively. Hence, the true number of DNA nucleotides incorporated in each reaction cycle becomes difficult to determine.

With the exception of single-molecule based approaches, all sequencing chemistry has characteristic dephasing patterns, through which the asynchronous signal can be reconstructed back to DNA sequences [15–20]. The dephasing pattern is a consequence of the nature of sequencing chemistry, reflecting multiple parameters of the reaction, including the yield, the side products, the kinetics and the impurity of the reactional system.

In this article, we present an ordinary differential equation-based model to simulate the clonal reactions of dual-base flowgram of ECC sequencing. This virtual sequencer can faithfully simulate the experimental results of ECC sequencing and identify the major factors that determine the performance of sequencing results, which is majorly a result of the extent of dephasing. Through such simulation we are able to build an effective dephasing algorithm to correct the phases between molecules in a clone, and eventually improve the read length and the raw accuracy before ECC deduction. We also prove that such an algorithm can work well on a dichromatic degenerate sequencing scheme, to enable more accurate sequencing through elevated information provided in each reaction cycle.

RESULTS AND DISCUSSION

We found that in ECC sequencing the dephasing pattern follows a 'one-base slippage' (OBS) scheme, meaning that the signal leading, primarily because of impurity reactions, will easily cause a secondary lead under specific conditions, and exhibits a characteristic dephasing pattern that is fundamentally different from other SBS sequencing chemistries [14]. For example (Fig. 1d), in a typical case where the DNA template to be sequenced is KM_nKMM, the main reactant in the reaction solution is M (A/C) and the impurity is K (G/T). After the first nucleotide K is extended by the main reactant M, the successive M is partially extended by the impurity K (defined as the 'primary lead'). If $n = 1$, then the second K will be further extended by the excessive main reactant M (defined as the 'secondary lead'). However, the secondary lead is negligible if $n > 1$, because the impurity is a trace amount and will be depleted after the primary lead. Although it has been shown that a practical algorithm for dephasing parameter estimation and signal correction built upon the OBS proposition can be established, it remains unknown when the dephasing correction algorithm is applicable and which factors determine OBS.

To answer these questions, we built a virtual sequencer from first-principles using ordinary differential equations (ODEs) to model the chemical reactions in ECC sequencing (Fig. 1e). The fluorogenic SBS chemistry comprises two major reactions [14,21–23]: the DNA synthesis catalyzed by DNA polymerase (Pol), and the dephosphorylation catalyzed by the alkaline phosphatase. In our model the dephosphorylation can be omitted because it is a much faster reaction compared with DNA synthesis.

By setting the concentrations of the four fluorogenic nucleotides in each cycle, we are able to numerically simulate the sequencing process with any desired flowgram, including the dual-base flowgram used in ECC sequencing and, for comparison, the conventional single-nucleotide addition flowgram. Taking the MK flowgram in ECC sequencing as an example, we add excessive M as the main species and trace amounts of K as the impurity in every odd cycle, and excessive K and trace amounts of M in every even cycle. The final values of the fluorophore F are regarded as the fluorescent intensities detected in each cycle, and $d[F]/dt = k_3[P \cdot D_k - 1 \cdot N_k]$, where D_k stands for the k-bp primed DNA strand, and N_k stands for the nucleotide complementary to the k-th base of the template.

Using our virtual sequencer and by tuning the impurity amount and reaction time, we simulated four typical sequencing conditions, with and without lead or lag, in all combinations. If the SBS reaction for all of the DNA molecules is perfectly synchronous, there will be no lead or lag, the fluorescent signals produced in each cycle will be proportional to the length of each copolymer, and all primed DNA strands will have exactly the same length (Fig. S1a and b). When the reaction loses its synchronicity, the fluorescent signals become aberrant, and there will be lead and/or lag, resulting in an increasing amount of primed DNA strands leading forward or lagging behind the main primed DNA (Fig. S1c–h). It is clear that the lead is mainly caused by impurities in the reaction buffer, and the lag is primarily a result of insufficient reaction time within which some of the molecules do not finish the reaction.

Downloaded from https://academic.oup.com/nsr/article/8/5/nwaa227/5900999 by guest on 10 August 2026

We first simulated the single-cycle sequencing using dual-base flowgram on the sequence $A(G)_nAAA(n = 1,2,3)$. The main nucleotide species is T and the impurity is C. In the simulation, we sampled six parameters 10 000 times in the range given in Column 2, Table S1. Among the six sampling parameters, the concentrations of the impurity and Pol, and the reaction time are sampled uniformly, while the reaction rate constants k_1, k_2 and k_3 are sampled logarithm uniformly (Table S1). We also fixed $k_{-1} = 0.01k_1$ and $k_{-2} = 10k_2$ according to reported chemical equilibrium constants [24–26].

After sequencing, the primed DNA is extended to different lengths, and we use an all lowercase sequence to denote the unextended primed DNA (e.g. agaaa), an uppercase letter in the sequence to denote the primed DNA extending to the position of the uppercase letter (e.g. aGaaa) and multiple upper case letters in the sequence to denote the sum of primed DNA denoted by a single upper case (e.g. agAAA = agAaa + agaAa + agaaA). For $A(G)_nAAA$, the dephasing parameters lag (λ), primary lead (ε^1) and secondary lead (ε^2) can be defined as $[a(g)_naaa]$, $[a(G)_naaa]$ and $[a(g)_nAAA]$, respectively. We also define the total lead $\varepsilon = \varepsilon^1 + \varepsilon^2$ (Fig. 2a).

The λ and ε measured in the three sequences, AGAAA, AGGAAA and AGGGAAA, are very close to each other, especially when the values are small (Figs S2 and S3). Not surprisingly, except for k_2 and λ, other sampling parameters have significant correlations with the dephasing parameters in terms of Spearman's correlation coefficient (Figs S4–S7). However, empirically, the most notable correlations are: 1) ε is equivalent to the impurity concentration when λ is small; 2) insufficient Pol will cause severe lag effect, but it becomes less important after saturating the DNA template (Fig. 2b). Additionally, the ε^1 in AGGAAA and the ε^2 in AGAAA are tightly correlated (Fig. S8).

If OBS holds, then the major contribution of ε in AGGAAA and AGAAA should be ε^1 and ε^2, respectively, as verified in Fig. 2c. To quantitatively assess how OBS approximates the simulation results of the virtual sequencer, we define the OBS index $\omega = (\varepsilon^1/\varepsilon)_{AGGAAA} \cdot (\varepsilon^2/\varepsilon)_{AGAAA}$, a parameter that shows significant correlation to all sample parameters (Figs S9 and S10). Notably, a greater ε or λ causes a lower ω, indicating that an optimized sequencing protocol with less dephasing is desired (Fig. S11).

We selected $\omega \geq 0.99$ as the criterion when OBS holds. Among the 10 000 sets of sampling parameters, 138 sets satisfy $\omega \geq 0.99$ with impurity concentration limited within 0.02 and $k_3 > 0.1$ (Table S1 and Fig. S12). Highly correlated with impurity concentration, the lead is also limited within 0.02. Lag < 0.02 is preferred, but lag as great as 0.78 is also possible to satisfy $\omega \geq 0.99$ (Fig. S13).

To validate the parameter range such that $\omega \geq 0.99$, we performed second round single-cycle sequencing simulation. The second-round simulation was the same as the first round except for the parameter sampling range (Column 3, Table S1) being narrower. Similar distributions of lead and lag and parameter correlations were observed (Figs S14–S20). However, only 5242 out of 10 000 sets of parameters satisfy $\omega \geq 0.99$ with no significant shrink in range (Figs S21 and S22) or correlation in between (Fig. S23). Two extreme parameter subsets exist that help us further clarify the desynchronization mechanism of the dual-base flowgram (Figs S24 and S25). The first subset has a large λ (> 0.1), but OBS still holds. In this subset, a lack of Pol and a small k_1 account for the large lag, but even less impurity and a relatively large k_3 ensure $\omega > 0.99$. The second subset is the opposite: $\lambda < 0.01$ but $\omega < 0.97$. This is mainly a result of the low percentage of ε^2 in AGAAA, which can be explained by the limited reaction time and k_3. Overall, the impurities and k_3 seem to be the main determinants of ω, while the value of λ, mainly determined by the Pol concentration, the reaction time and k_1, is preferred to be small, but is not determinant for ω.

We simulated a 100-cycle MK flowgram with our virtual sequencer, and found that our OBS scheme could approximate the DNA length distribution fairly well. The sequences were picked from the *E. coli* genome with a length containing exactly 110 DPLs (Table S2). All four DNA templates were sequenced under the same 1000 sets of parameters, among which the impurity concentration and reaction time are uniformly sampled in the range of [0, 0.1] and [10, 100], respectively, while the rest parameters are set to default. The dephasing parameter estimation and signal correction of the simulated sequencing signals were performed using our previous algorithm [14]. To test whether the lead and lag were estimated accurately, we also simulated the single-cycle sequencing of AGAAA using the same parameters in the 100-cycle simulation. The lead ε and lag λ are measured as [aGAAA] and [agaaa]. Fig. S26 shows that the estimated lead and lag in the 100-cycle simulation are very close to their measured values in the single-cycle simulation, especially when they are small.

The overall performance of our dephasing correction can be simply indicated using an error number, which is defined as the number of different DPLs between the result c (the DPL after OBS-based dephasing correction) and the true DPL h (the known

Downloaded from https://academic.oup.com/nsr/article/8/5/nwaa227/5900999 by guest on 10 August 2026

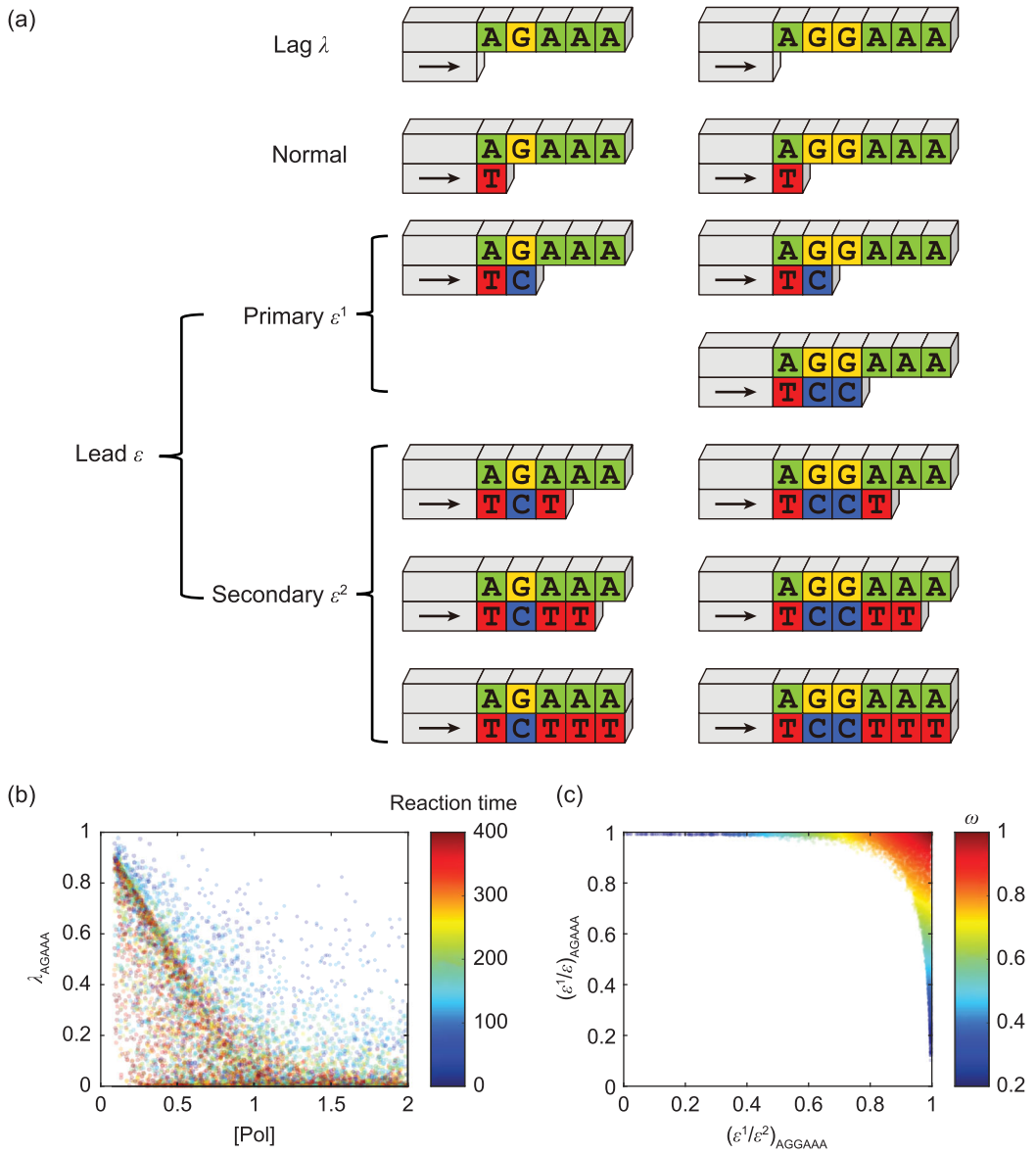

Figure 2. Key factors of dephasing patterns revealed by single-cycle simulation. (a) Dephasing parameters and their corresponding nascent DNA strands. (b) Insufficient Pol will cause severe lag, which can be remedied by prolonged reaction time. (c) Distribution of ε^1 percentage in AGGAAA and ε^2 percentage in AGAAA of each set of parameters, which gives the definition of the OBS index.

input for virtual sequencing). We confirmed that as ω increases, all lead, lag and error numbers of dephasing corrections decrease (Figs S26–S29). When $\omega > 0.99$, the error number is limited to 5, which is sufficient for accurate ECC sequencing. Moreover, as expected, a large ε or λ causes a high error rate (Fig. 3a). As we observed in the virtual sequencing of AGAAA, dephasing correction simulation shows many identical trends; for example many correlations exist between parameters, ε is determined by impurity concentration, λ is determined by reaction time, and a large ε or λ leads to the decrease of ω (Fig. S26).

We further simulated 250-cycle and 500-cycle sequencing scenarios, using sequences from the *E. coli* genome, at the same genomic location as one of the previous 100-cycle simulations (Table S2). Only one error remained after dephasing correction in the 250-cycle simulation, which occurred in the 249th cycle (Figs S30 and S31). Five errors occurred in the 500-cycle simulation, initially at the 335th cycle (Fig. S32). The DPL distribution matrix, which illustrates the loss of reaction synchronization by dephasing, can also be obtained by either fitting the dephasing correction algorithm, or by direct calculations of the ODE-based virtual sequencer result

Downloaded from https://academic.oup.com/nsr/article/8/5/nwaa227/5900999 by guest on 10 August 2026

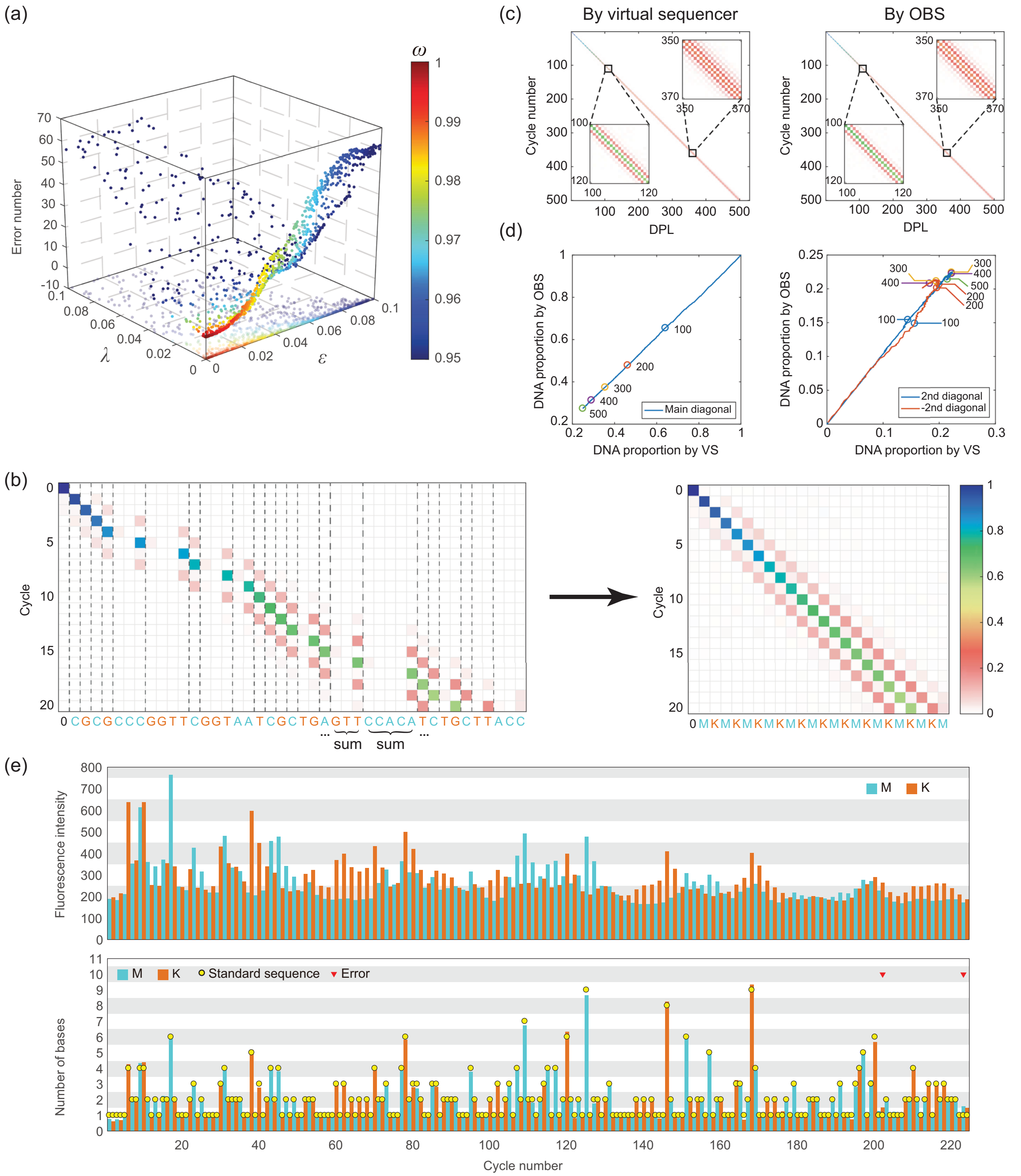

Figure 3. The DPL distribution matrix and dephasing correction. (a) Relationship between ε, λ, ω and error number in the dephasing correction of 100-cycle simulated sequencing signals. (b) The base distribution matrix is transformed into the DPL distribution matrix through summing up columns in the same DPL. (c) The DPL distribution matrix per cycle obtained through two methods using the same colormap in (b). (d) Comparison of the main diagonal (left) and $\pm$2nd diagonal (right) of the DPL distribution matrix per cycle obtained through two methods. Numbered dots indicate cycle numbers. VS, virtual sequencer. (e) Fluorescent signals in an ECC sequencing experiment (top) and their dephasing-corrected signals (bottom).

Downloaded from https://academic.oup.com/nsr/article/8/5/nwaa227/5900999 by guest on 10 August 2026

(Fig. 3b). We found that these two methods yield highly consistent DPL distribution matrices (Fig. 3c and d, Fig. S33), showing that the OBS principle can support accurate dephasing correction for long sequencing signals of at least 500 cycles (1000 bp on average).

We then used our experimental prototype to sequence a DNA template for 224 cycles (read length 451 bp) using the MK flowgram, and tested the OBS-based dephasing correction. We found only two errors, one in the 202nd cycle and the other in the 223rd cycle, indicating that the first 410 bp of sequencing is error-free (Fig. 3e).

So far, our discussion has been based on a monochromatic dual-base flowgram, meaning that the two nucleotides in one degenerate reaction cycle are labeled with the same fluorophore, whose intensity reveals the total number of nucleotides incorporated in a cycle. In fact, the virtual sequencer simulator can be further extended to a dichromatic mode, where the two nucleotides in one reaction cycle are labeled with different fluorophores. For example, in the MK flowgram, A and C are added in every odd cycle and G and T are added in every even cycle, while A and G are labeled with a green dye and C and T are labeled with a red dye (Fig. 4a). Thus, we would be able to separately measure the number of each nucleotide that is extended in any given cycle. Dichromatic ECC sequencing provides 3.37 bit/cycle of information, which is much higher than the 2 bit/cycle afforded by monochromatic ECC sequencing and therefore has the potential to be even more accurate. Because there are two different dyes labeled for the four nucleotides, the sizes of DPL $\mathbf{h}$ and fluorescent intensities $\mathbf{f}$ change from $m \times 1$ and $n \times 1$ to $m \times 2$ and $n \times 2$, respectively. To correct the dephased dichromatic signals, we hypothesize that the basic equation $\mathbf{f} = T\mathbf{h}$ in monochromatic dephasing still holds. Specifically, $\mathbf{f}\bullet_1 = T\mathbf{h}\bullet_1$, and $\mathbf{f}\bullet_2 = T\mathbf{h}\bullet_2$.

To validate this hypothesis, we simulated the dichromatic ECC sequencing for 250 cycles, using the same parameters as in the 250-cycle monochromatic simulation (Fig. S34). The DPL distribution matrix D is given directly by the virtual sequencer, while the flux matrix T is deduced from D by:

$$T_{i,j} = \begin{cases} D_{i,j} - D_{i-1,j} & j = m \\ T_{i,j+1} + D_{i,j} - D_{i-1,j} & j < m \end{cases}$$

We compared the simulated fluorescent signals $\mathbf{f}$ and the signals fitted by $T\mathbf{h}$ and found the signals to be almost identical (Fig. 4b), thus confirming our hypothesis, and suggesting that our OBS-based dephasing correction algorithm can be directly applied to the dichromatic mode. To experimentally validate this suggestion, we sequenced a DNA template for 79 cycles (read length 165 bp) using the dichromatic MK flowgram and corrected the fluorescent signals (Fig. 4c). The first error appeared at the 75th cycle, indicating that the first 152 bp were error-free.

CONCLUSION

In summary, we constructed a virtual sequencer to simulate fluorogenic SBS sequencing reactions, and found that the characteristic 'one-base slippage' dephasing pattern can be applied to correct for dephasing. Using our virtual sequencer, we clarified the dephasing mechanism of ECC sequencing, and extended our understanding to the dichromatic form of ECC sequencing. Our virtual sequencer also revealed that, when applying to dichromatic instead of monochromatic ECC sequencing, the relationship (involving matrix multiplication) linking the DPL and sequencing signals remains unchanged.

The OBS pattern is common in ECC sequencing. However, it also exists in traditional single-nucleotide addition (SNA) sequencing methods such as pyrosequencing [27] and semiconductor sequencing [28], although at a lower frequency. The impurities in SNA sequencing are mainly the nucleotides left over from the previous cycle, thus OBS may occur in DNA motifs $X_mY_nXY_k$. Allowing for OBS in dephasing algorithms of SNA sequencing may also improve their accuracy.

Stochastic models are widely used in systems biology and noise plays an indispensable role in many biochemical processes [29–34]. However, the stochastic version of the virtual sequencer showed negligible difference to the ODE-based deterministic version (data not shown). That is because there are typically $>5 \times 10^3$ DNA molecules and even more enzymes and nucleotides in the sequencing reaction, which greatly reduce the reaction noise. Besides, ECC sequencing chemistry comprises only two consecutive enzymatic reactions and noise is unlikely to play a role in such a simple reaction topology.

Understanding dephasing patterns in ECC sequencing not only provides insights into SBS reactions, but also suggests that our approach can lead to the design and further optimization of dephasing algorithms in general. For example, the fluorescent signals of some DNA reads are incomplete because of occasional chip contaminants or sight offset. The virtual sequencer may be used to simulate abnormal sequencing signals and hence aid the development of a special dephasing algorithm.

Downloaded from https://academic.oup.com/nsr/article/8/5/nwaa227/5900999 by guest on 10 August 2026

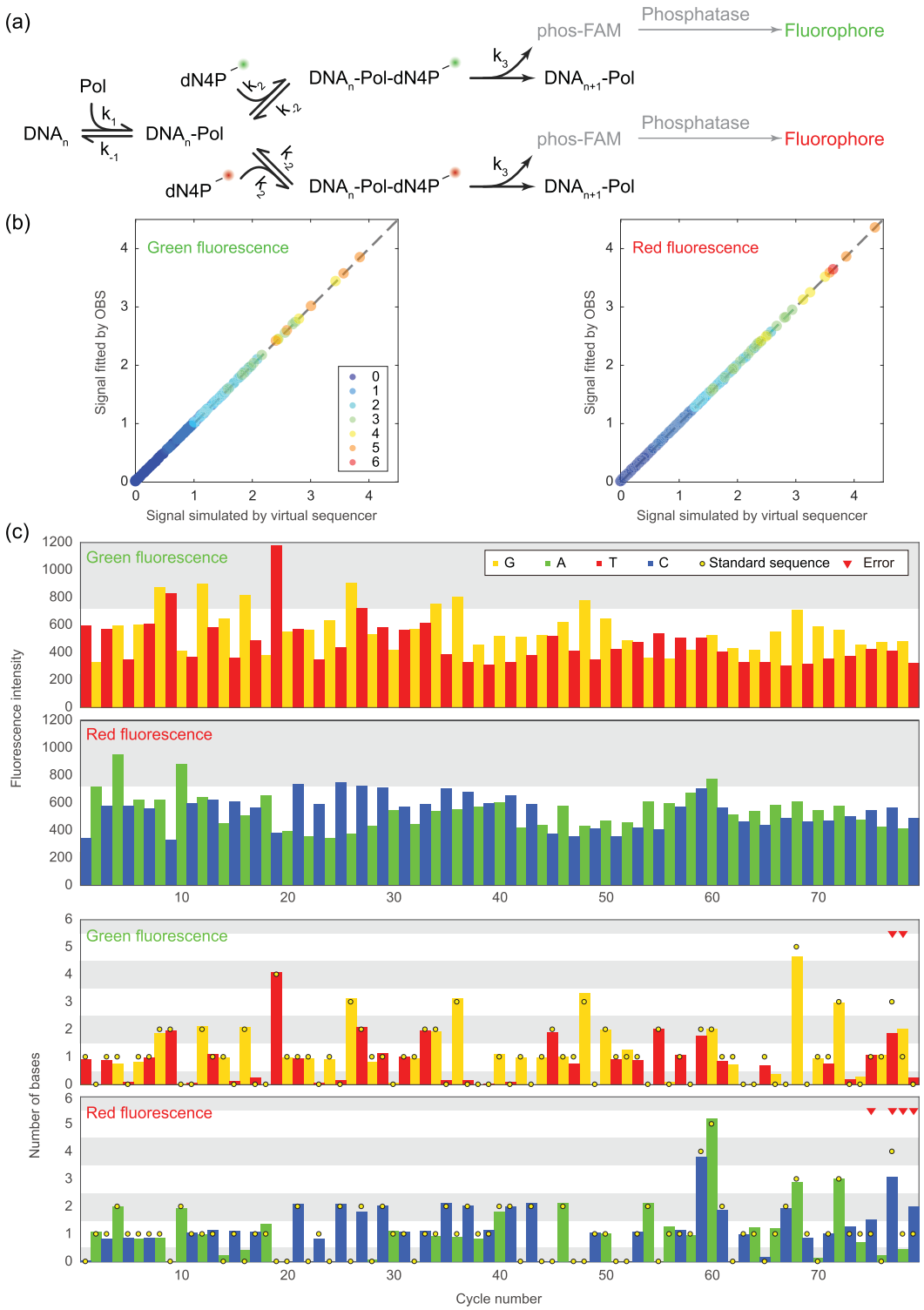

Figure 4. Dichromatic dual-base flowgram simulated by the virtual sequencer. (a) Dichromatic fluorogenic SBS reactions. (b) The signals by the flux matrix fit well with the simulated signals by the dichromatic virtual sequencer. Color denotes the DPL. (c) The fluorescence intensities (top) and their dephasing corrected signals (bottom) of a dichromatic ECC sequencing experiment.

Downloaded from https://academic.oup.com/nsr/article/8/5/nwaa227/5900999 by guest on 10 August 2026

METHODS

Based on mass-action, the ODEs in the virtual sequencer are

$$\begin{cases} \frac{d[P]}{dt} = k_{-1}[P \cdot D_{k-1}] - k_1[P][D_{k-1}] \\ \frac{d[D_{k-1}]}{dt} = k_{-1}[P \cdot D_{k-1}] - k_1[P][D_{k-1}] \\ \frac{d[P \cdot D_0]}{dt} = k_1[P][D_0] - k_{-1}[P \cdot D_0] + k_{-2}[P \cdot D_0 \cdot N_1] - k_2[P \cdot D_0][N_1] \\ \frac{d[P \cdot D_k]}{dt} = k_1[P][D_k] - k_{-1}[P \cdot D_k] + k_{-2}[P \cdot D_k \cdot N_{k+1}] - k_2[P \cdot D_k][N_{k+1}] + k_3[P \cdot D_{k-1} \cdot N_k] \\ \frac{d[N_k]}{dt} = k_{-2}[P \cdot D_{k-1} \cdot N_k] - k_2[P \cdot D_{k-1}][N_k] \\ \frac{d[P \cdot D_{k-1} \cdot N_k]}{dt} = k_2[P \cdot D_{k-1}][N_k] - k_{-2}[P \cdot D_{k-1} \cdot N_k] - k_3[P \cdot D_{k-1} \cdot N_k] \\ \frac{d[F]}{dt} = k_3[P \cdot D_{k-1} \cdot N_k] \end{cases}$$

where P stands for Polymerase (Pol), D_k stands for the k-bp primed DNA strand, N_k stands for the nucleotide complementary to the k-th base of the template, and F for the fluorophore. In one run of the virtual sequencer, the ODE is serially solved numerically for many cycles. In Cycle 1, the virtual sequencer sets the initial value of D_0 to 1 and D_k to 0. And in Cycle $i+1$, the initial values of D_0 and D_k are set to their final values in Cycle i. The final values of F in each cycle are regarded as the detected fluorescent intensities (signal). These settings are followed in this article if there is no additional explanation: the concentration of Pol is 1, the concentration of the main species is 30, $k_1 = k_2 = k_3 = 1$, $k_{-1} = 0.01$, $k_{-2} = 10$.

The dephasing parameter estimation and signal correction of the simulated sequencing signals are done using a simplified version of the algorithm described in Ref. [14]. Specifically, for DPL $\mathbf{h}_{m\times 1}$ ($m = 110$ denotes the DPL number) and fluorescent intensities $\mathbf{f}_{n\times 1}$ ($n = 100$ denotes the cycle number) simulated by the virtual sequencer, we calculate the DPL distribution matrix D and flux matrix T as functions of lead and lag:

$$D_{i,j} = \begin{cases} 1 & i = j = 1 \\ 0 & i = 1,\ j > 1 \\ D_{i-1,j} - T_{i-1,j} & j = 1 \\ D_{i-1,j} - T_{i-1,j} + T_{i-1,j-1} & j > 1 \end{cases}$$

$$T_{i,j} = \begin{cases} (1-\lambda) D_{i,j} & j = 1,\ (i+j) \equiv 0\ (\mathrm{mod}2) \\ 0 & j = 1,\ (i+j) \equiv 1\ (\mathrm{mod}2) \\ (1-\lambda)\, D_{i,j} + h' T_{i,j-1} & j > 1,\ (i+j) \equiv 0\ (\mathrm{mod}2) \\ \varepsilon\,(1-\lambda)\, D_{i,j-1} & j > 1,\ (i+j) \equiv 1\ (\mathrm{mod}2) \end{cases}$$

where λ and ε are lag and lead coefficients to be estimated, respectively, and

$$h'_i = \begin{cases} 0 & h_i > 1 \\ 1 & h_i \le 1 \end{cases}$$

The DPL distribution matrix D is used to describe how the primed DNA is dephased during the sequencing, and the flux matrix T links the DPL $\mathbf{h}$ and fluorescent intensities $\mathbf{f}$ through the basic equation:

$$\mathbf{f} = T\mathbf{h}$$

Hence, we estimate λ and ε from $\mathbf{h}$ and $\mathbf{f}$ by:

$$\min_{\lambda,\varepsilon} ||T\mathbf{h} - \mathbf{f}||_2.$$

And the corrected signals are calculated as:

$$\mathbf{c} = T^+\,\mathbf{f}$$

where T^+ is the Moore-Penrose pseudoinverse of T.

SUPPLEMENTARY DATA

Supplementary data are available at *NSR* online.

FUNDING

This work was supported by the National Natural Science Foundation of China (21927802, 21525521), the Beijing Brain Initiative (Z181100001518004) and the Beijing Advanced Innovation Center for Genomics.

AUTHOR CONTRIBUTIONS

Y.H., Z.C. and W.Z. proposed the project. Y.H. supervised the project. W.Z. performed the simulation. W.Z., Z.C. and L.T. analyzed data. L.K., H.D. and S.Q. performed the experiments. W.Z., L.T. and Y.H. wrote the manuscript. All authors discussed the results and commented on the manuscript.

Conflict of interest statement. None declared.

Downloaded from https://academic.oup.com/nsr/article/8/5/nwaa227/5900999 by guest on 10 August 2026

REFERENCES

1. Levy SE and Myers RM. Advancements in next-generation sequencing. *Annu Rev Genomics Hum Genet* 2016; **17**: 95–115.
2. Manolio TA. Genomewide association studies and assessment of the risk of disease. *N Engl J Med* 2010; **363**: 166–76.
3. Mortazavi A, Williams BA and McCue K *et al.* Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* 2008; **5**: 621–8.
4. Johnson DS, Mortazavi A and Myers RM *et al.* Genome-wide mapping of in vivo protein-DNA interactions. *Science* 2007; **316**: 1497–502.
5. Ku CS, Naidoo N and Wu M *et al.* Studying the epigenome using next generation sequencing. *J Med Genet* 2011; **48**: 721–30.
6. Fullwood MJ, Liu MH and Pan YF *et al.* An oestrogen-receptor-α-bound human chromatin interactome. *Nature* 2009; **462**: 58–64.
7. Wu AR, Wang J and Streets AM *et al.* Single-cell transcriptional analysis. *Annu Rev Anal Chem* 2017; **10**: 439–62.
8. Pushkarev D, Neff NF and Quake SR. Single-molecule sequencing of an individual human genome. *Nat Biotechnol* 2009; **27**: 847–50.
9. Eid J, Fehr A and Gray J *et al.* Real-time DNA sequencing from single polymerase molecules. *Science* 2009; **323**: 133–8.
10. Chaisson MJ and Tesler G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinform* 2012; **13**: 238.
11. Clarke J, Wu HC and Jayasinghe L *et al.* Continuous base identification for single-molecule nanopore DNA sequencing. *Nat Nanotechnol* 2009; **4**: 265–70.
12. Goodwin S, Gurtowski J and Ethe-Sayers S *et al.* Oxford nanopore sequencing, hybrid error correction, and *de novo* assembly of a eukaryotic genome. *Genome Res* 2015; **25**: 1750–6.
13. Jain M, Fiddes IT and Miga KH *et al.* Improved data analysis for the MinION nanopore sequencer. *Nat Methods* 2015; **12**: 351–6.
14. Chen Z, Zhou W and Qiao S *et al.* Highly accurate fluorogenic DNA sequencing with information theory-based error correction. *Nat Biotechnol* 2017; **35**: 1170–8.
15. Metzker ML. Sequencing technologies—the next generation. *Nat Rev Genet* 2010; **11**: 31–46.
16. Shendure J and Ji H. Next-generation DNA sequencing. *Nat Biotechnol* 2008; **26**: 1135–45.
17. Mardis ER. Next-generation DNA sequencing methods. *Annu Rev Genom Hum Genet* 2008; **9**: 387–402.
18. Mardis ER. A decade's perspective on DNA sequencing technology. *Nature* 2011; **470**: 198–203.
19. Goodwin S, McPherson JD and McCombie WR. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* 2016; **17**: 333–51.
20. Shendure J, Balasubramanian S and Church GM *et al.* DNA sequencing at 40: past, present and future. *Nature* 2017; **550**: 345–53.
21. Sood A, Kumar S and Nampalli S *et al.* Terminal phosphate-labeled nucleotides with improved substrate properties for homogeneous nucleic acid assays. *J Am Chem Soc* 2005; **127**: 2394–5.
22. Sims PA, Greenleaf WJ and Duan H *et al.* Fluorogenic DNA sequencing in PDMS microreactors. *Nat Methods* 2011; **8**: 575–80.
23. Chen Z, Duan H and Qiao S *et al.* Fluorogenic sequencing using halogen-fluorescein-labeled nucleotides. *ChemBioChem* 2015; **16**: 1153–7.
24. Englund PT, Huberman JA and Jovin TM *et al.* Enzymatic synthesis of deoxyribonucleic acid XXX. Binding of triphosphates to deoxyribonucleic acid polymerase. *J Biol Chem* 1969; **244**: 3038–44.
25. Muise O and Holler E. Interaction of DNA polymerase I of *Escherichia coli* with nucleotides. Antagonistic effects of single-stranded polynucleotide homopolymers. *Biochemistry* 1985; **24**: 3618–22.
26. Patel SS, Wong I and Johnson KA. Pre-steady-state kinetic analysis of processive DNA replication including complete characterization of an exonuclease-deficient mutant. *Biochemistry* 1991; **30**: 511–25.
27. Margulies M, Egholm M and Altman WE *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 2005; **437**: 376–80.
28. Rothberg JM, Hinz W and Rearick TM *et al.* An integrated semiconductor device enabling non-optical genome sequencing. *Nature* 2011; **475**: 348–52.
29. Wilkinson DJ. Stochastic modelling for quantitative description of heterogeneous biological systems. *Nat Rev Genet* 2009; **10**: 122–33.
30. Norman TM, Lord ND and Paulsson J *et al.* Stochastic switching of cell fate in microbes. *Annu Rev Microbiol* 2015; **69**: 381–403.
31. Silva-Rocha R and de Lorenzo V. Noise and robustness in prokaryotic regulatory networks. *Annu Rev Microbiol* 2010; **64**: 257–75.
32. Balázsi G, van Oudenaarden A and Collins JJ. Cellular decision making and biological noise: from microbes to mammals. *Cell* 2011; **144**: 910–25.
33. Zhang L, Radtke K and Zheng L *et al.* Noise drives sharpening of gene expression boundaries in the zebrafish hindbrain. *Mol Syst Biol* 2012; **8**: 613.
34. Ge H, Qian H and Xie XS. Stochastic phenotype transition of a single cell in an intermediate region of gene state switching. *Phys Rev Lett* 2015; **114**: 078101.

Downloaded from https://academic.oup.com/nsr/article/8/5/nwaa227/5900999 by guest on 10 August 2026