


Highly accurate fluorogenic DNA sequencing with information theory–based error correction

Zitian Chen^{1–4}, Wenxiong Zhou^{1,2,4}, Shuo Qiao^{1,2,4}, Li Kang^{1,2,4}, Haifeng Duan^{1,2,4}, X Sunney Xie^{1,2,4} & Yanyi Huang^{1–5} 

Eliminating errors in next-generation DNA sequencing has proved challenging. Here we present error-correction code (ECC) sequencing, a method to greatly improve sequencing accuracy by combining fluorogenic sequencing-by-synthesis (SBS) with an information theory–based error-correction algorithm. ECC embeds redundancy in sequencing reads by creating three orthogonal degenerate sequences, generated by alternate dual-base reactions. This is similar to encoding and decoding strategies that have proved effective in detecting and correcting errors in information communication and storage. We show that, when combined with a fluorogenic SBS chemistry with raw accuracy of 98.1%, ECC sequencing provides single-end, error-free sequences up to 200 bp. ECC approaches should enable accurate identification of extremely rare genomic variations in various applications in biology and medicine.

In the last decade, the advent of next-generation sequencing (NGS) has dramatically extended our capability to investigate questions in fundamental biology and medicine by reducing sequencing cost by about four orders of magnitude compared to Sanger sequencing^{1–5}. Many sequencing chemistries have been developed and commercialized^{6–10}. Currently the prevailing NGS approaches are SBS-based methods that use DNA polymerase to extend a new DNA strand and deduce the template sequence by detecting the incorporated oligonucleotides during the strand synthesis^{6,8,9,11}.

With the exception of the technology used by Pacific Biosciences single-molecule sequencer, which monitors the *in situ* strand synthesis continuously¹², all other SBS approaches detect the synthesis process with cyclic reactions⁶. These approaches can be further divided into two major categories based on the flowgrams used in reaction cycles⁶: cyclic reversible termination (CRT; e.g., Illumina/Solexa¹³, Direct/Helicos^{14–16}, and others^{17–19}) and single-nucleotide addition (SNA; e.g., Roche/454²⁰, Ion Torrent²¹, and others^{22,23}). CRT approaches use terminating nucleotides on which the 3′-OH groups have been blocked and, hence, after the nucleotides have been incorporated onto the complementary synthetic strand, the strand will not be further extended. In each of these reaction cycles, four nucleotides covalently linked with differentiable labels, are introduced simultaneously and one extended base can be deduced in each cycle. By contrast, SNA approaches introduce only one specific nucleotide, with an open 3′-OH group, in each reaction cycle. Elongation may occur when the introduced nucleotide is complementary, and until a mismatch occurs.

CRT and SNA approaches exhibit different strengths and weaknesses. CRT approaches use fluorescence, which is stable and highly efficient, to signal nucleotide incorporation. Although the strength

of the signal is an advantage of these approaches, the complexity and imperfection of reaction chemistries, such as scars left on the nascent strand after cleaving the fluorescent tags, cause sequencing errors and limit the read length²⁴. A major intrinsic advantage of SNA approaches is that the product of each incorporation reaction is natural DNA, without terminating groups or scars left from the cleaved labels, providing a potential to generate longer sequencing reads²⁵. However, the detectable signals in current SNA instruments (both the chemiluminescence of Roche/454 and the hydrogen ions of Ion Torrent) are transient and have poor linearity between the signal intensity and the number of dNTPs incorporated in a single reaction cycle in homopolymer regions^{26–28}. In our previous reports we have demonstrated that a fluorogenic SNA approach²², which also produces natural DNA product but releases stable fluorescent molecules for ratiometric deduction of the number of extended nucleotides, can reduce the difficulty of homopolymer sequencing²³.

Here we present a strategy for DNA sequencing, ECC sequencing, that can greatly improve sequencing accuracy and read length using a dual-base flowgram combined with fluorogenic SBS chemistry. The ECC sequencing approach allows a mixture of two types of nucleotide substrates to be introduced into each reaction cycle. The synthetic strands expose free 3′-OH groups that can be continuously extended until no nucleotides in the mixture can be further incorporated. Although each of such reactions provides only one degenerate sequence with partially defined base composition, one DNA template can be sequenced three times with three orthogonal combinations of dual-base mixes to provide three degenerate sequences, from which an unambiguous sequence can be accurately deduced. Sequencing errors in any degenerate sequence can be further identified and corrected in this approach. The ECC sequencing approach does not require more

¹Beijing Advanced Innovation Center for Genomics (ICG), Peking University, Beijing, China. ²Biodynamic Optical Imaging Center (BIOPIIC), Peking University, Beijing, China. ³College of Engineering, Peking University, Beijing, China. ⁴School of Life Sciences, Peking University, Beijing, China. ⁵Peking-Tsinghua Center for Life Sciences, Peking University, Beijing, China. Correspondence should be addressed to Y.H. (yanyi@pku.edu.cn).

Received 10 November 2016; accepted 30 August 2017; published online 6 November 2017; doi:10.1038/nbt.3982

sequencing reaction time than SNA, but provides higher confidence of the sequence accuracy through the extra information received in the orthogonal flowgrams. In principle, ECC sequencing strategy can be applied to any SNA-based sequencing chemistries. We have built a laboratory prototype DNA sequencer to demonstrate the complete ECC process using fluorogenic SBS chemistry, and obtained single-end reads up to 250 bp with the first 200 bp free of error.

RESULTS

The principle of degenerate-base fluorogenic sequencing

We have developed a family of fluorogenic sequencing substrates using Tokyo Green (TG)²⁹, a high-performance fluorophore, to terminally label tetraphosphate nucleotides (dN4P or dN, see **Fig. 1a**, and **Supplementary Note**, Section 1). TG offers higher fluorescence quantum yield (0.82), higher absorption coefficient (8×10^4), higher on-off ratio (2.8×10^2), and better photostability than previously reported fluorogenic dyes (**Supplementary Note**, Section 2). During the fluorogenic SBS process, the single-strand DNA templates are grafted onto the surface of a glass flowcell using solid-phase PCR (**Supplementary Note**, Section 3). Each template is then annealed to a sequencing primer with its 3'-end serving as the starting point for SBS reactions. In each cycle of the sequencing, a reaction mix (*Bst* polymerase, alkaline phosphatase, and fluorogenic nucleotides) is brought to react with those immobilized primed DNA templates. When the polymerase incorporates a correct nucleotide onto the primer terminus, a non-fluorescent dark state dye-triphosphate will be released, and then immediately switched to a highly-fluorescent bright state through dephosphorylation^{22,23,30}. This fluorogenic SBS reaction produces native DNA duplex, leaving the 3'-end of the synthesized strand not terminated. The substrates that can form correct Watson-Crick pairs at the primer terminus will continuously extend until the first mismatch encounters.

In our previous work this feature has been used to sequence 30–40 bases through a single-base flowgram, in which one of the four substrates was introduced into the reaction in each cycle. In this work we employ a dual-base flowgram. For example, in the first cycle of the sequencing (**Fig. 1b**), a K reaction mix (i.e., containing dG and dT, **Supplementary Note**, Section 4.2) is added to the primed DNA template with the starting sequence ACTTGAAA. DNA polymerase will incorporate one dT and one dG to pair the first two bases (AC) and yield two fluorophores, then stop upon the third base T because of the base mismatch. In the following M (dA & dC) cycle, two dA and one dC are paired with the next three bases (TTG) and yield three fluorophores. Conjugated mixes M and K are alternately introduced to react with the primed DNA template (**Fig. 1c**). The amount of fluorophores produced in each cycle is equivalent to the number of extended bases.

Fluorescence signal is measured upon the completion of polymerase elongation in each reaction cycle. Normalized fluorescence signal, representing the number, not the actual composition and sequence, of bases extended in each cycle, is named degenerate polymer length (DPL). In **Figure 1c**, the DPL array (0, 2, 3, 3, 1, ...) can be transformed to a degenerate sequence (KKMMMKKKM...). Besides this M-K dual-base flowgram, there are two additional dual-base flowgrams R (dA & dG)-Y (dC & dT) and W (dA & dT)-S (dC & dG), through which the same template can be expressed as different degenerate sequences (YRRYYYYRRYY...) and (WSWWSWWWW...). To acquire these three orthogonal degenerate sequences, a reset operation is needed between sequencing rounds to denature the nascent strand and reanneal the sequencing primer. Each actual base can be deduced from three sequences by calculating the intersection of degenerate bases.

Base-calling for degenerate sequences

We have built a lab prototype to perform fluorogenic sequencing using dual-base flowgrams (**Supplementary Note**, Section 4). Similar to other SBS sequencing approaches, some fluorescence intensity decay is inevitable. This decay, mainly due to the reaction imperfection and the loss of template or primer in the flowcell, has caused severe challenges in base-calling (**Fig. 2a**). In a typical fluorogenic degenerate sequencing run, the fluorescence intensity decline could be normalized by an exponential decay function with ~1% of signal drop between reaction cycles (**Supplementary Note**, Section 6.5).

Ideally, normalized fluorescence signal in each cycle should be equal to the DPL (**Fig. 2c**). However, the correspondence between intensity and DPL could only be preserved in about the first 30 cycles, after which dephasing could not be neglected; that is, the signal of each cycle became significantly affected by the neighboring cycles (**Fig. 2c**). Dephasing, the asynchronization of primer ensemble, has two major components, lag and lead. Lagging strands are caused mainly by incomplete extension, while leading strands are attributed mainly to unexpected extension caused by contaminating bases (**Supplementary Note**, Section 2.2). Therefore, in a given cycle, the fluorescence signal, contributed from the asynchronized primer ensemble, is different from the corresponding DPL. The accumulation of dephasing will gradually reduce the correlation between the sequencing signals and the DPL array.

Nevertheless, we have discovered that the accumulation effect of signal dephasing and decay could be well estimated assuming first-order reaction kinetics (**Supplementary Note**, Section 6), with the residues between estimated and measured value <0.3 (**Fig. 2b**). Furthermore, a sequence-independent, iterative, dephasing-rectification algorithm was developed to deduce the DPL array of each sequencing round. With dephasing rectification, we can substantially extend the low-error span of DPL array length from the first 50 cycles (~100 nt) to more than 125 cycles (~250 nt), beyond which the crowded errors could not be correctly rectified with our dephasing algorithm (**Fig. 2d**). For the same template, such a rectification method can also be applied to the other two orthogonal degenerate sequences (**Fig. 2d**). Each of the three degenerate sequences harbors infrequent errors that are unlikely to be located on the same base position.

Information communication model for error correction

We analyzed the information redundancy in dual-base degenerate sequencing using the principles of information theory. Obviously, a DPL array acquired from one dual-base sequencing round cannot provide an explicit DNA sequence. When there is no sequencing error, the information entropy of an L -nt-long random DNA sequence is $2L$ bits, while that of its DPL array is only L bits. The orthogonal nature guarantees that the mutual information entropy of two DPL arrays acquired from different flowgrams is 0 bits, and the joint information entropy is $2L$ bits (**Supplementary Note**, Section 5.3). Therefore, two degenerate sequences provide both sufficient and necessary information of an explicit DNA sequence ($L + L - 0 = 2L$). The explicit DNA sequence can be deduced by taking the intersection of the degenerate bases in two DPL arrays from different flowgrams. For example, if a base in the MK-DPL array is sequenced as M(A/C), and in the RY DPL array as R(A/G), then it can be deduced as base A ($\{A, C\} \cap \{A, G\} = \{A\}$).

However, due to experimental sequencing errors, the entropy of a DPL array (denoted as l) is lower than L bits. Two of such error-containing DPL arrays provide insufficient joint information to deduce the DNA sequence ($l + l - 0 < 2L$). With our current experimental error rate, an extra DPL array is introduced to provide the

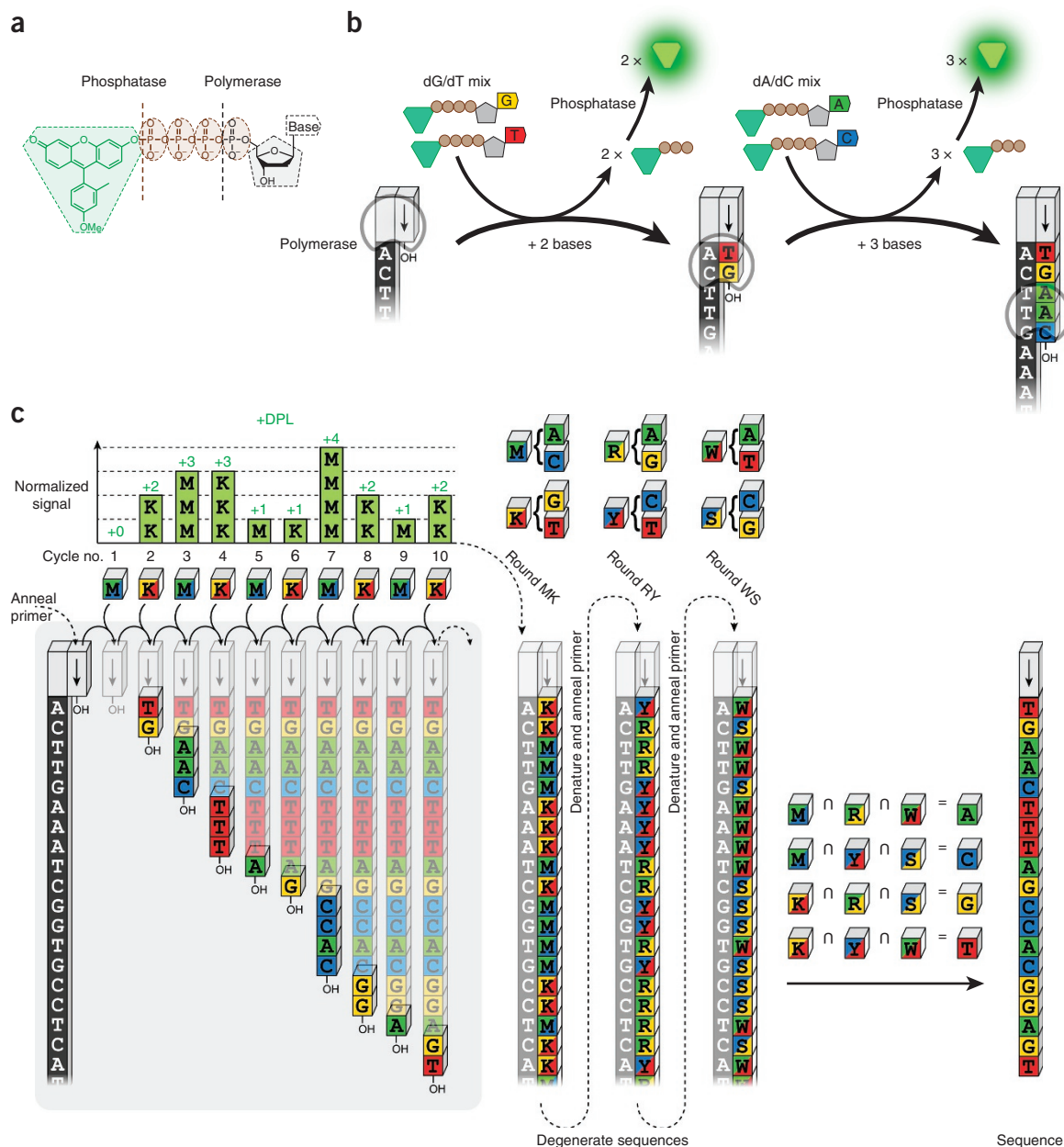


Figure 1 The schematic of the ECC sequencing approach. **(a)** The chemical structure of a Tokyo Green (TG)-terminated fluorogenic nucleotide substrate. During the SBS reaction, the polymerase will incorporate the substrate into the nascent strand of DNA and release TG-triphosphate that can be further dephosphorylated by phosphatase co-existing in the reaction buffer. **(b)** In the fluorogenic SBS reaction, two bases are mixed and introduced in a single reaction cycle. A polymerase binds to a DNA template (black) annealed with sequencing primer (gray), leaving the 3'-OH exposed. The dual-base mix allows any complementary bases to be added to the 3'-end of the synthetic strand (color), and releases an equivalent amount of the fluorophore molecules with bases extended. **(c)** The DNA sequence can be deduced from three orthogonal degenerate sequences obtained through dual-base flowgrams. Four nucleotides can be split as three orthogonal pairs of dual-base mixes, M-K, R-Y, and W-S. In each reaction cycle the degenerate polymer length (DPL) is experimentally determined through fluorescent intensity measurement, and then the DPL reaction is rewritten into a degenerate sequence. After one round of reaction, the synthesized strand is melted away and another round of SBS reaction, with a different set of dual-base mixes, is carried out using a newly annealed sequencing primer. The explicit base information can be extracted by taking the intersection between three degenerate bases.

redundant information ($2L < 3l < 3L$), which can be used to both detect errors and deduce the explicit sequence.

We have established an information communication model, which contains an encoder, a decoder, and a communication channel, to depict dual-base sequencing with the intrinsic characteristic of error detection and correction (Fig. 3a). Three orthogonal dual-base flowgrams encode a DNA sequence, the information source, into three

original DPL arrays (n). We analyzed the DPL distributions in human, yeast, and *Escherichia coli* genomes and found they are close to $P(n) = 1/2^n$, the theoretical distribution of DPL from a random DNA sequence (Fig. 3b). We also found that only 1.15% of DPL is >8 in the human genome.

The sequencing reaction is regarded as a communication channel, through which sequencing errors are inevitably introduced

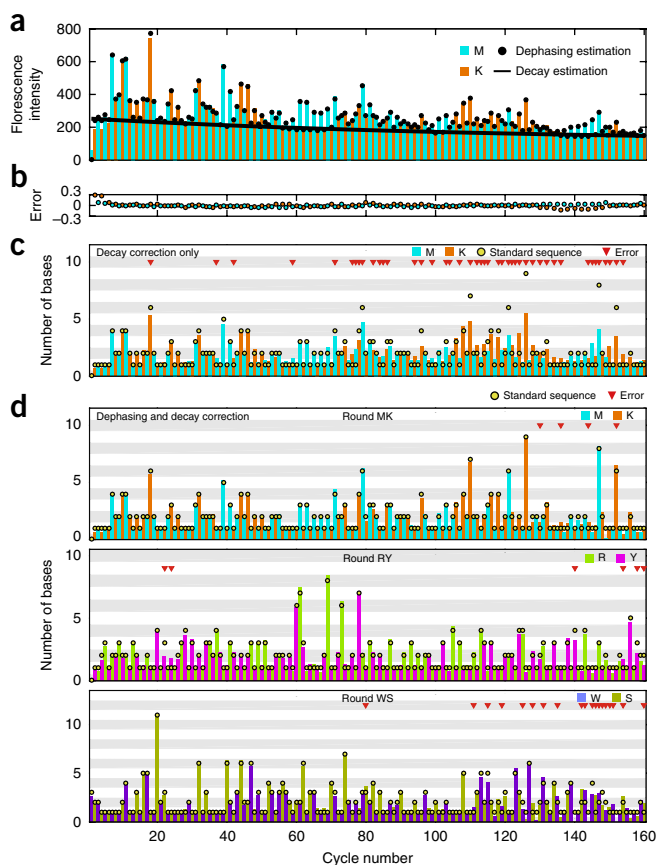


Figure 2 Sequencing signal in a typical fluorogenic degenerate sequencing run. (a) Raw fluorogenic sequencing signals from Round MK. Bars: experimentally measured fluorescence intensity increment F during each sequencing reaction cycle. Black dots: estimated intensity increment P according to the estimated sequencing parameters. (b) Raw sequencing signal errors, calculated by $(P - F)/F$. (c) Normalized fluorescence signal with decay correction only. The discordance between the measurement and the estimation has been marked as errors, which are caused mainly by signal dephasing. Yellow dots: DPL inferred from the DNA sequence. Red triangle: sequencing errors where the normalized signal cannot be rounded into the corresponding DPL. (d) Normalized fluorescence signal from Round MK, RY and WS with both dephasing and decay corrected. Note the numbers of errors have decreased greatly after dephasing correction.

into the received message. For instance, in cycle 3 of the R-Y round, original DPL $n = 3$ is mistakenly measured as $m = 4$ (a 3-to-4 insertion error, Fig. 3a). We carefully analyzed the concordance of original and measured DPLs in our 45 rounds of dual-base sequencing data. 5,503 out of 5,609 (98.1%) original DPL ($n \leq 8$) were faithfully transmitted (Fig. 3c).

We rewrote the measured DPL arrays into degenerate base sequences and defined a codeword as the 3-tuple of degenerate bases in the same position from these degenerate sequences in the order of MK, RY and WS. In the case of Figure 3a, the first few codewords are (KYW), (KRS), and (MRW). Such codewords can be further compiled into a binary format, with M, R, and W assigned as logical 1, and K, Y, and S as logical 0. Each degenerate sequence in any single flowgram became a bit string. We defined the parity of a codeword as the result of XOR (exclusive or) operation of its three bits (Fig. 3d). The degenerate bases in a codeword have only one common base if and only if the parity is logical 1, and this common base is regarded as the decoded result. Specifically, 111 (MRW) is decoded as a base A,

100 (MYS) as a C, 010 (KRS) as a G, and 001 (KYW) as a T. These four legitimate codewords have Hamming distances of 2 in between. On the other hand, the remaining four illegitimate codewords with parity logical 0 (no common base) indicate sequencing errors. As is the case in Figure 3a, the DNA sequence is decoded from the bit string and a 3-to-4 error at the fifth codeword (MRS/110) was caught by decoder through a parity check.

Conventionally, memoryless codewords with Hamming distance 2 are only error-detectable but not correctable. However, we discovered that the dual-base sequencing results in bit string format are not memoryless but context-dependent, providing extra information for error correction besides error detection.

Sequence decoding through dynamic programming

The error correction decoding is performed through an algorithm based on dynamic programming, which has been used in short-read sequence alignment softwares³¹. Those dual-base sequencing errors resulting from mistakenly measured DPLs are unique errors which contain only bit insertions or deletions, but not bit alterations, in a bit string. When an error is found, it is possible to be rectified by changing the corresponding DPLs based on bit string context. Errors must be rectified sequentially from the first error, because the changes of DPL, corresponding to bit string-shift operations, will affect the downstream codewords.

In Figure 4a the first illegitimate codeword is detected at codeword 5, which has three highly possible error sources with a one-base error: (1) an insertion error in cycle 2 of round MK, original DPL ($n = 2$) is erroneously measured as 3; (2) an insertion error in cycle 2 of round RY, original DPL ($n = 3$) is measured as 4; (3) a deletion error in cycle 3 of round WS, original DPL ($n = 3$) is measured as 2. In this case, the insertion error in cycle 2 of round RY is true, and it can be corrected by left-shifting bit string (BS) 2 after the 5th bit. With this shift operation, the following illegitimate codewords pass parity check concomitantly until a second error is detected at base 14. This deletion error, together with the remaining illegitimate codewords, is rectified by right-shifting bit string 1 after the 13th bit. In this case nine codeword illegitimacies are legitimized by only two correction operations, resulting in an error-free decoded DNA sequence.

In fact, there are numerous possible operation combinations to decode the sequence. Moreover, the number of combinations increases exponentially with the read length, making it practically impossible to obtain the optimal sequence by enumerating all possible combinations. Therefore, we used dynamic programming to determine the global optimal decoded sequence. We constructed a codeword space as a three-dimensional (3D) matrix with the three bit strings as its axis. Each node (i, j, k) represents the codeword consisting of the i -th bit of bit string 1, the j -th bit of bit string 2, and the k -th bit of bit string 3, and it can be classified as Pass or Error according to the parity checking (Fig. 4b). Any path starting from the node $(1,1,1)$ and only passing through the Pass nodes in this 3D matrix represents a possible decoded DNA sequence.

The probability of a given path in the codeword space can be calculated by the Bayesian formula. The prior probability of the occurrence of DPL with length n is $1/2^n$ (Fig. 3b), and the probability of DPL with length n to be sequenced as length m , $P(m|n)$, can be obtained from reference sequences and the data compared to theoretical values (Fig. 3c). Then for round r (r is MK, RY, or WS), the posterior probability $P_r(n_i|m_i)$ that its i -th measured DPL of length m_i is produced from a DPL of length n_i can be given below:

$$P_r(n_i|m_i) = \frac{P(m_i|n_i)/2^{n_i}}{\sum_{k=1}^{\infty} P(m_i|k)/2^k}, r \in \{\text{MK, RY, WS}\}$$

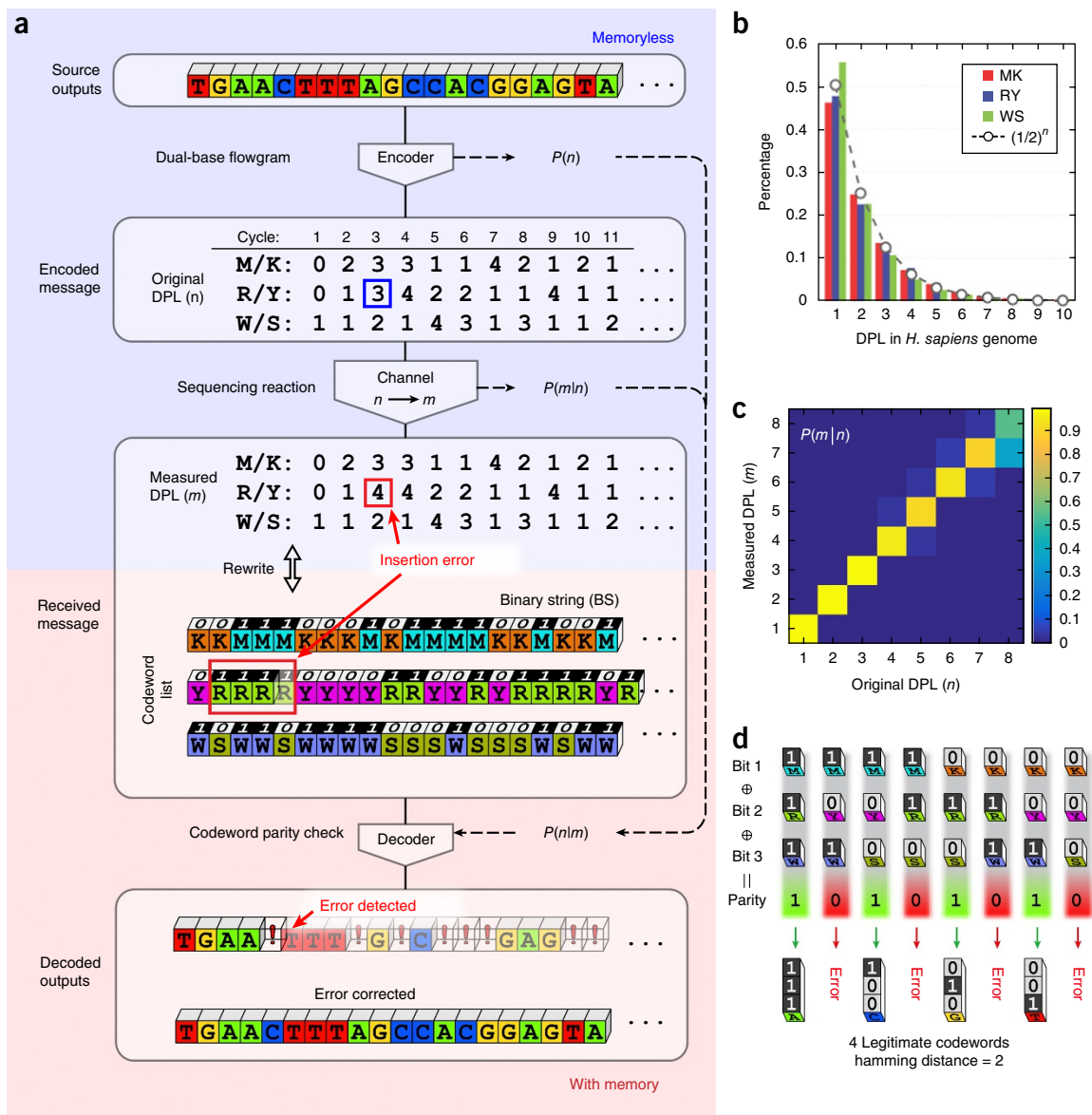


Figure 3 Information communication model for ECC sequencing. (a) Schematic of the information communication model. The DNA template serves as the information source, and is encoded into three original DPL arrays (n). The sequencing reaction is regarded as a communication channel, through which sequencing errors are inevitably introduced. In this case, in Cycle 3 of Round RY, the original DPL $n = 3$ is mistakenly measured as $m = 4$ and causes an insertion error. Then the measured DPL arrays are rewritten into degenerate base sequences and further compiled into binary strings. By parity check of each 3-tuple codeword, the decoder will detect the sequencing errors. The measured DPL arrays (m) are memoryless but the rewritten binary strings are context-dependent, offering the possibility of error correction. (b) DPL distribution in the human genome, close to $P(n) = 1/2^n$, the theoretical distribution of DPL from a random DNA sequence. Only about 1.15% of the DPLs are greater than 8. (c) The probabilities of DPL with length n to be sequenced as length m . (d) The eight possible codewords and their parity. Only four codewords among them are legitimate (with parity 1).

Then the probability P_r that a measured DPL array is produced from a certain DNA is the cumulative product of $P_r(n_i|m_i)$. And under the hypothesis that the three rounds of ECC sequencing are independent of each other, the probability of a given path is $P_{\text{path}} = P_{\text{MK}} \cdot P_{\text{RY}} \cdot P_{\text{WS}}$. And the probability of every path in the codeword space can be calculated in the same way (Fig. 4b). A dynamic programming approach was adopted to obtain the path with the maximum probability (Supplementary Note, Section 7.1).

ECC decoding improves sequencing accuracy

ECC decoding can efficiently rectify errors for long sequencing reads. We performed 15 long-length ECC sequencing experiments

to sequence three different templates of lambda phage DNA. Before ECC decoding, there were minor occasional errors in the sequencing signals. After decoding, these errors were completely eliminated up to 200 bp, and also substantially reduced in the 200–250 bp range (Fig. 5a, and Supplementary Note, Section 7.2). In the case illustrated in Figure 5a, although the first sequencing error occurred in base 39 of round RY, it was successfully corrected after ECC decoding along with other several sequencing errors in round WS. In this case the first error only occurred after 250 bps.

The ECC decoding algorithm has the power to accurately identify complex error forms. Compared to scattered sequencing errors, neighboring errors in the same or different rounds are more challenging

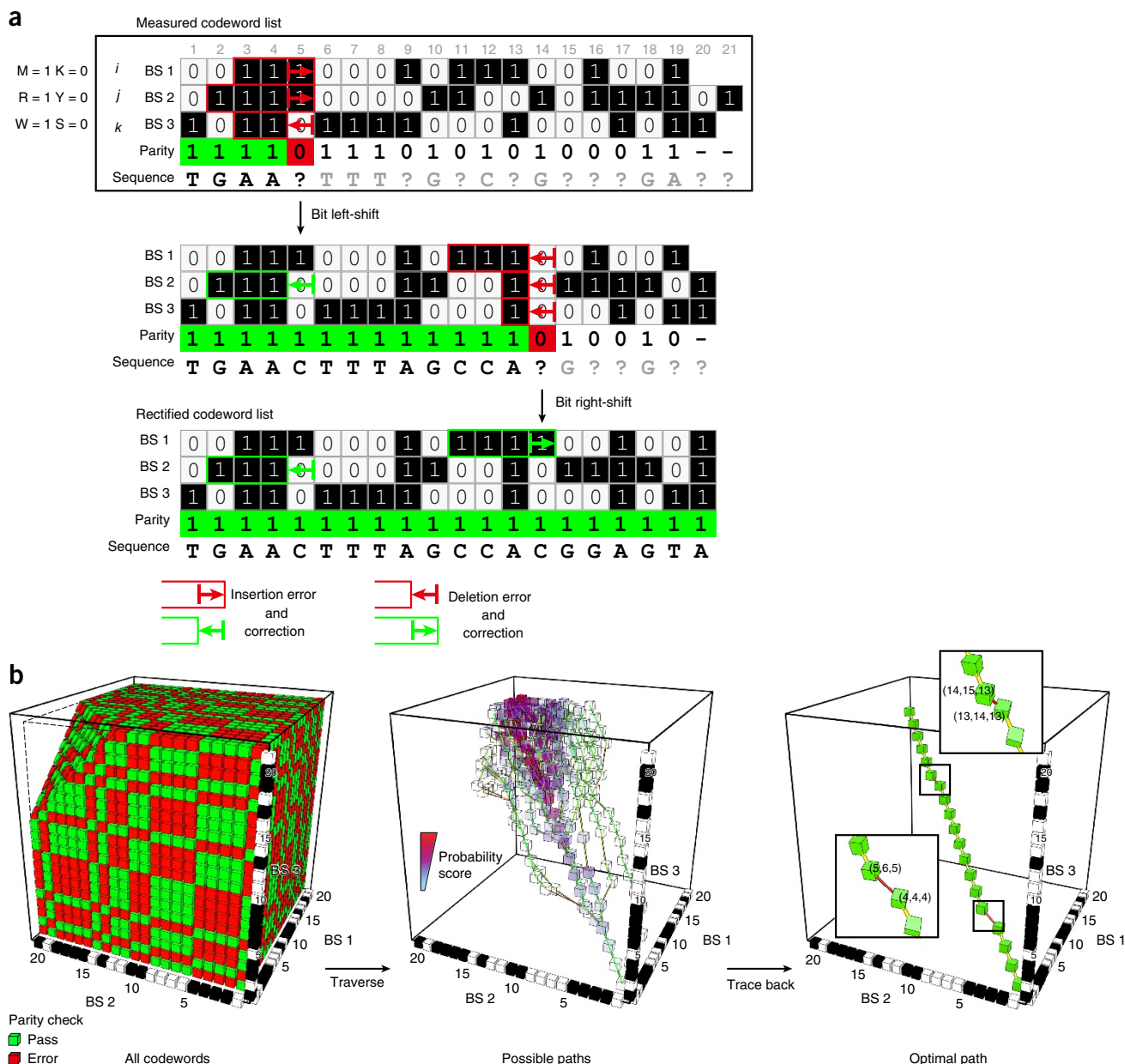


Figure 4 Dynamic programming-based decoding algorithm. **(a)** A simplified example of sequencing error correction. The first illegitimate codeword is detected at codeword 5, and there are three possible error sources. In this case, the insertion error in cycle 2 of Round RY is identified and corrected by left-shifting the bit string (BS) 2 since the 6th bit. Likewise, the second error at codeword 14 (a deletion) is identified and corrected by right-shifting the bit string 1 since the 14th bit. After these two corrections, all codewords are legitimate and an error-free sequence is obtained. **(b)** The decoding algorithm. A 3D codeword space is constructed with the 3-bit string as axes. Each node represents a 3-bit codeword that can be classified as Pass or Error according to the parity checking. Any path starting from the node (1,1,1) and only passing through the Pass nodes represents a possible decoded DNA sequence. We traverse the codeword space using dynamic programming and assign each path a probability by the Bayesian formula. The path with the maximum probability is traced back to produce the decoded DNA sequence.

to correct since more and sophisticated correction operations are required in the decoding algorithm. When parity check failed between the three-round sequencing signals, the algorithm would calculate the probabilities of different operations.

In one case, two sequencing errors occurred within three cycles in Round RY (a 1-base deletion at cycle 22 and a 1-base insertion at cycle 24). At least two alternative correction approaches, each of which contains two correction operations, can fix these errors (Fig. 5b). The first approach operates a 1-to-2 insertion correction and a 2-to-1

deletion correction ($P(1|2) \cdot P(2|1) = 0.00015$), whereas the second approach contains an 1-to-2 insertion correction and a 3-to-2 deletion correction ($P(3|2) \cdot P(2|1) = 0.00022$). Therefore, the second approach is preferred because of the higher probability.

Fluorogenic degenerate sequencing has intrinsically high accuracy. We analyzed the error frequencies of different DPL along the sequencing read every 50 nt (Fig. 5c,d). Without ECC correction, we found 106 errors in 11,062 bases. These errors were more likely to happen on longer DPLs and on posterior positions, similar to other

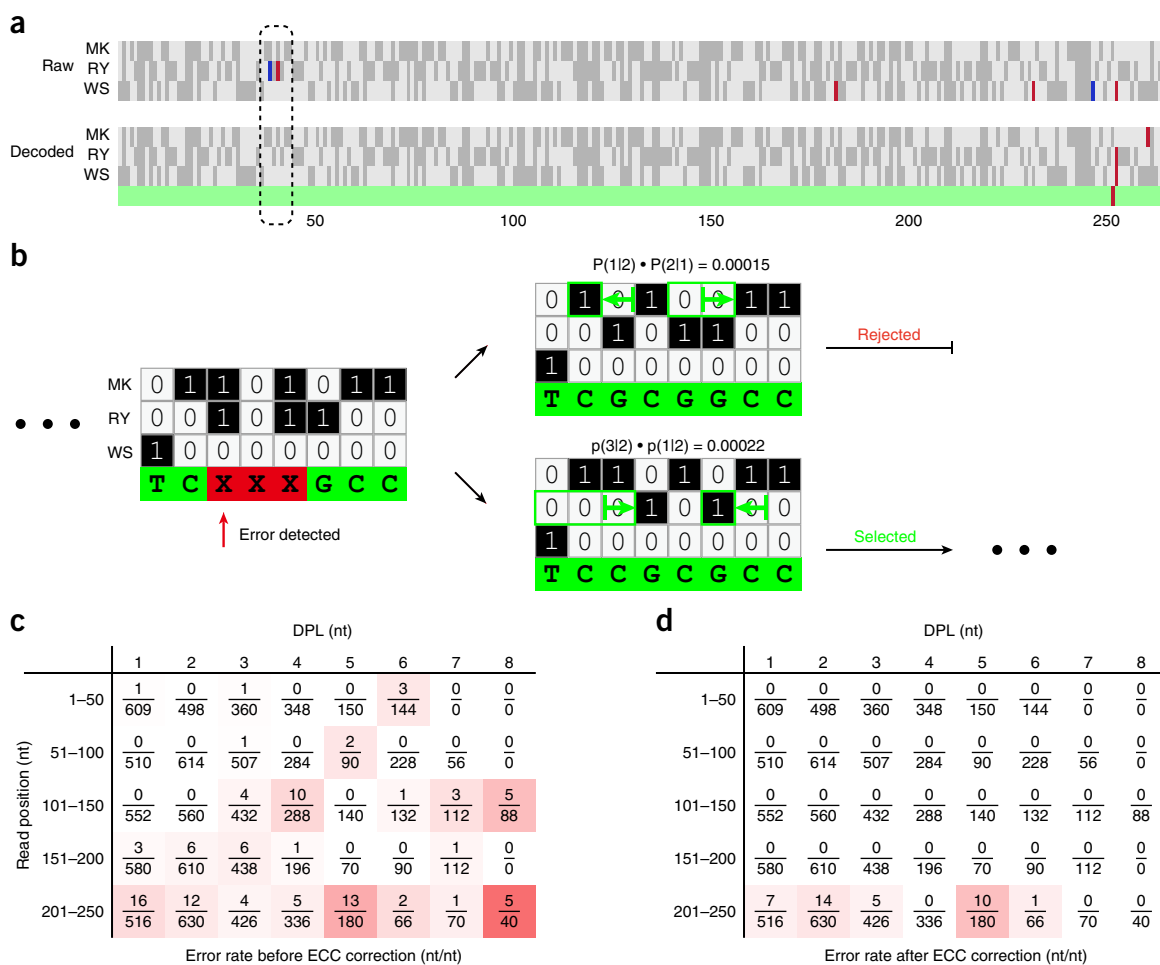


Figure 5 High accuracy of ECC sequencing. (a) The ECC decoding performance. The neighboring errors are more challenging to correct since more sophisticated correction operations are required. Dark gray: bit 1, light gray: bit 0, blue: deletion error, red: insertion error, green: error-free region. (b) Probability assessment of alternative correction approaches for a case of neighboring errors. (c) Error rate before ECC correction. (d) Error rate after ECC correction. The denominator and numerator are the number of bases in total DPL and erroneously measured DPL in this read position range, respectively. Note that errors in 1–200 nt are completely eliminated, and those in 201–250 nt are also mostly corrected.

sequencing methods^{26,27}. The raw accuracy is 99.82% in the first 100 nt, and 99.45% in the first 200 nt. With 99% accuracy cut-off, the read length of more than 250 nt can be achieved (**Supplementary Note**, Section 7).

ECC decoding eliminates the majority of raw sequencing errors. The high raw accuracy of the fluorogenic degenerate sequencing approach provides a foundation for ECC correction to completely eliminate all errors in the first 200 nt (**Fig. 5d**), including the errors in DPL up to 8 nt (for estimation of the upper boundary of the error rate, see **Supplementary Note**, Section 7.4). In addition, ECC decoding effectively reduced the cumulative error rate of 250 nt, from 0.96% to 0.33%.

DISCUSSION

We have built a complete system for fluorogenic degenerate DNA sequencing as well as the corresponding ECC theory based on the communication model. Even in our laboratory prototype we obtained a read length of 250 bp and were able to eliminate all errors in the first 200 bp by ECC decoding.

A highly accurate sequencing method would be beneficial in various applications, including fetal genetic mutation detection in

maternal blood and rare mutation identification in circulating tumor DNA or in highly heterogeneous tumor tissues. However, the accuracy currently available in prevalent sequencing platforms is still not ideal for these applications, thus sophisticated sample preparation and bioinformatics strategies are indispensable^{32–34}. The ECC technology could fundamentally improve the data quality and fit the accuracy requirement of precision medicine.

Compared to conventional SNA approaches, a significant change in our ECC approach is that the dark cycles, in which the read length does not increase, are eliminated using the dual-base flowgram since the four bases are split into two complementary sets (**Supplementary Note**, Section 5). As a result, the average primer extension length is increased from 0.67 bp/cycle in the conventional flowgram to 2 bp/cycle in the dual-base flowgram (**Supplementary Note**, Section 5.3). On average, a dual-base flowgram uses threefold fewer reaction cycles to sequence the same length of a DNA template as the single-base flowgram does. This fast-extension is advantageous because it substantially reduces the effect of cycle-wise side reactions, including the primer loss and the misincorporation of bases, that cause signal-to-noise ratio decline along the sequencing reaction. It is also worthy to point out that using TG as the fluorophore has greatly improved

the signal intensity and consequently increased the signal-to-noise ratio for measurement. The detailed analysis and modeling of the enzymatic kinetics (**Supplementary Note**, Section 2) have also helped determine the best experimental conditions to perform fluorogenic SBS reactions, ensuring the completion of reactions with minimal substrate hydrolysis-induced fluorescence background, and substantially improving the sequencing quality by providing both longer read length and fewer raw errors. The quantitative understanding of the reaction kinetics also provides basic parameters to build a complete model of intensity fitting for deciphering the lead and lag components in each reaction cycle (**Supplementary Note**, Section 6.5).

Every SBS chemistry has its own coding strategy. To our knowledge, we are the first to introduce error-correction encoding into SBS technologies, although other SBS technologies adopted either non-redundant encoding or error-detection-only encoding strategies. Using conventional SNA pyrosequencing flowgram, the transient analog signals have not provided redundant information of the sequence, hence the errors are neither detectable nor correctable. As for CRT strategies, when labeling the four nucleotides using four different dyes, the bases are encoded in a memoryless code with Hamming distance 2, capable of error detection but not error correction. And when labeling the nucleotides using only two different dyes, even error detection is unachievable. Uniquely, there are two codes used in ECC sequencing, the DPL code for channel description, and the bit string code for decoding codewords. The DPL code and the DNA sequence itself are memoryless. However, when coded in bit string, the sequencing errors will accumulate and be transmitted to the latter sequencing readouts, and this can be well-described using the memory-encoding framework.

It is beneficial to view DNA sequencing chemistry using a communication model. The whole process of ECC sequencing can be divided as a few key modules including encoding, decoding, and the generation and elimination of errors. In the field of communication, mature models have been developed for these operations. However, unlike the cases in communication applications, in ECC sequencing the encoding is applied specifically to DPL but the codeword decoding is applied specifically to the bit string. Besides, memory-encoding allows us to use error probability to decode the original message. Further analysis shows that, in the communication model of the dual-base flowgram, each homopolymer in the DNA template will be encoded into three DPLs (one DPL in each round). Every homopolymer will be extended exclusively in at least one of the three related cycles (**Supplementary Note**, Section 5.2). The reaction in these exclusive cycles preserves the accuracy of conventional flowgram because its DPL is equal to the homopolymer length and the additional mismatching substrate causes negligible kinetic delay (**Supplementary Note**, Section 2.2). Therefore undetected errors are extremely rare (**Supplementary Note**, Section 7.4) because they only happen when a base is mistakenly measured in all three rounds. The problem of accuracy declining with long homopolymers has been solved to a large extent owing to the three-round sequencing in the ECC approach.

Despite the simplicity of modular interpretation using a communication model, decoding the memory-containing signals is not straightforward in ECC sequencing. The probabilities of sequencing errors are readily available in DPL, but it is difficult to describe the mutual correlation between DPLs in three rounds. On the other hand, errors can be detected in bit string, but it is not feasible to obtain the error probabilities from bit string. Besides, the number of possible decoding operations increases exponentially with the read length. Our algorithm tackled these challenges by representing all codewords

in the codeword space, enabling the description of both parity check and all possible decoding solutions. The optimal solution has a natural Bayesian probability background and can be efficiently obtained via dynamic programming (**Supplementary Note**, Section 7.1), which utilizes the relatively easily obtained $P(n|m)$ to calculate both local and global optima with reduced complexity.

Currently, we have only acquired a limited amount of ECC sequencing data, which is not sufficient for accurately revealing the error profiles of more complicated genomic contexts. With the accumulation of ECC sequencing data and the optimization of the correction algorithm, we will be able to better understand and to further improve the ECC performance.

The ECC approach is theoretically compatible with any SBS chemistry using nucleotides with the 3'-OH unblocked such as SMRT of PacBio, semiconductor sequencing of Ion Torrent, or pyrosequencing. Our analysis reveals that when it is above 95%, the higher raw accuracy leads to more effective accuracy enhancement by ECC (**Supplementary Note**, Section 7.3). If the raw accuracy, especially in quantifying homopolymer lengths, is close to or above that of the fluorogenic sequencing chemistry, a similar accuracy improvement can be reasonably anticipated.

To become a truly practical sequencing technology, ECC fluorogenic sequencing may be further developed to a high-throughput sequencer, which may be implemented by conducting reactions in chips with multiple microreactors. Another promising improvement will be to label the two nucleotides in the same reaction cycle with two different dyes that can be differentiated by color filters. This dichromatic ECC sequencing will provide more information per round (~1.7 bit/base, **Supplementary Note**, Section 5.3) than the monochromatic method herein (1 bit/base), thus possessing further enhancing performance. In addition, when applied to resequencing applications, the ECC sequencing can also be combined with the newly emerging graph genome mapping algorithm³⁵, greatly reduce the computing cost.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the [online version of the paper](#).

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

The authors thank Y. Men, Z. Yu, H. Qiu, C. Zheng, Y. Fu, X. Zhang, T. Chen, L. Wu, S. Zhang, X. Jiang, J. Bu, P.A. Sims, L.L. Tao, and H. Ge for experimental assistance and discussion. This work was supported by the Ministry of Science and Technology of China (863 Program 2012AA02A101), National Natural Science Foundation of China (21327808 and 21525521), Beijing Municipal Commission of Science and Technology (Z111100059111002), and Beijing Advanced Innovation Center for Genomics.

AUTHOR CONTRIBUTIONS

Y.H. and X.S.X. conceived the project. Y.H. and Z.C. designed the experiment. Z.C., W.Z., S.Q., L.K., and H.D. performed the experiments. All authors analyzed the data and wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the [online version of the paper](#).

Reprints and permissions information is available online at <http://www.nature.com/reprints/index.html>. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

- Shendure, J., Mitra, R.D., Varma, C. & Church, G.M. Advanced sequencing technologies: methods and goals. *Nat. Rev. Genet.* **5**, 335–344 (2004).

2. Koboldt, D.C., Steinberg, K.M., Larson, D.E., Wilson, R.K. & Mardis, E.R. The next-generation sequencing revolution and its impact on genomics. *Cell* **155**, 27–38 (2013).
3. Drmanac, R. The advent of personal genome sequencing. *Genet. Med.* **13**, 188–190 (2011).
4. Mardis, E.R. & Wilson, R.K. Cancer genome sequencing: a review. *Hum. Mol. Genet.* **18**, R2, R163–R168 (2009).
5. Schrijver, I. *et al.* Opportunities and challenges associated with clinical diagnostic genome sequencing: a report of the Association for Molecular Pathology. *J. Mol. Diagn.* **14**, 525–540 (2012).
6. Goodwin, S., McPherson, J.D. & McCombie, W.R. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **17**, 333–351 (2016).
7. Mardis, E.R. A decade's perspective on DNA sequencing technology. *Nature* **470**, 198–203 (2011).
8. Mardis, E.R. Next-generation sequencing platforms. *Annu. Rev. Anal. Chem. (Palo Alto, Calif.)* **6**, 287–303 (2013).
9. Metzker, M.L. Sequencing technologies - the next generation. *Nat. Rev. Genet.* **11**, 31–46 (2010).
10. Shendure, J. & Ji, H. Next-generation DNA sequencing. *Nat. Biotechnol.* **26**, 1135–1145 (2008).
11. Fuller, C.W. *et al.* The challenges of sequencing by synthesis. *Nat. Biotechnol.* **27**, 1013–1023 (2009).
12. Eid, J. *et al.* Real-time DNA sequencing from single polymerase molecules. *Science* **323**, 133–138 (2009).
13. Bentley, D.R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).
14. Braslavsky, I., Hebert, B., Kartalov, E. & Quake, S.R. Sequence information can be obtained from single DNA molecules. *Proc. Natl. Acad. Sci. USA* **100**, 3960–3964 (2003).
15. Pushkarev, D., Neff, N.F. & Quake, S.R. Single-molecule sequencing of an individual human genome. *Nat. Biotechnol.* **27**, 847–850 (2009).
16. Gao, Y. *et al.* Single molecule targeted sequencing for cancer gene mutation detection. *Sci. Rep.* **6**, 26110 (2016).
17. Ju, J. *et al.* Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators. *Proc. Natl. Acad. Sci. USA* **103**, 19635–19640 (2006).
18. Guo, J., Yu, L., Turro, N.J. & Ju, J. An integrated system for DNA sequencing by synthesis using novel nucleotide analogues. *Acc. Chem. Res.* **43**, 551–563 (2010).
19. Stupi, B.P. *et al.* Stereochemistry of benzylic carbon substitution coupled with ring modification of 2-nitrobenzyl groups as key determinants for fast-cleaving reversible terminators. *Angew. Chem. Int. Ed.* **51**, 1724–1727 (2012).
20. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380 (2005).
21. Rothberg, J.M. *et al.* An integrated semiconductor device enabling non-optical genome sequencing. *Nature* **475**, 348–352 (2011).
22. Sims, P.A., Greenleaf, W.J., Duan, H. & Xie, X.S. Fluorogenic DNA sequencing in PDMS microreactors. *Nat. Methods* **8**, 575–580 (2011).
23. Chen, Z. *et al.* Fluorogenic sequencing using halogen-fluorescein-labeled nucleotides. *ChemBioChem* **16**, 1153–1157 (2015).
24. Wu, W. *et al.* Termination of DNA synthesis by N6-alkylated, not 3'-O-alkylated, photocleavable 2'-deoxyadenosine triphosphates. *Nucleic Acids Res.* **35**, 6339–6349 (2007).
25. Rothberg, J.M. & Leamon, J.H. The development and impact of 454 sequencing. *Nat. Biotechnol.* **26**, 1117–1124 (2008).
26. Forgetta, V. *et al.* Sequencing of the Dutch elm disease fungus genome using the Roche/454 GS-FLX Titanium System in a comparison of multiple genomics core facilities. *J. Biomol. Tech.* **24**, 39–49 (2013).
27. Loman, N.J. *et al.* Performance comparison of benchtop high-throughput sequencing platforms. *Nat. Biotechnol.* **30**, 434–439 (2012).
28. Liu, L. *et al.* Comparison of next-generation sequencing systems. *J. Biomed. Biotechnol.* **2012**, 251364 (2012).
29. Urano, Y. *et al.* Evolution of fluorescein as a platform for finely tunable fluorescence probes. *J. Am. Chem. Soc.* **127**, 4888–4894 (2005).
30. Sood, A. *et al.* Terminal phosphate-labeled nucleotides with improved substrate properties for homogeneous nucleic acid assays. *J. Am. Chem. Soc.* **127**, 2394–2395 (2005).
31. Rumble, S.M. *et al.* SHRiMP: accurate mapping of short color-space reads. *PLoS Comput. Biol.* **5**, e1000386 (2009).
32. Kinde, I., Wu, J., Papadopoulos, N., Kinzler, K.W. & Vogelstein, B. Detection and quantification of rare mutations with massively parallel sequencing. *Proc. Natl. Acad. Sci. USA* **108**, 9530–9535 (2011).
33. Hoang, M.L. *et al.* Genome-wide quantification of rare somatic mutations in normal human tissues using massively parallel sequencing. *Proc. Natl. Acad. Sci. USA* **113**, 9846–9851 (2016).
34. Schmitt, M.W. *et al.* Detection of ultra-rare mutations by next-generation sequencing. *Proc. Natl. Acad. Sci. USA* **109**, 14508–14513 (2012).
35. Paten, B., Novak, A. & Haussler, D. Mapping to a reference genome structure. Preprint available at <https://arxiv.org/abs/1404.5010v1> (2014).

ONLINE METHODS

Synthesis of terminal phosphate-labeled fluorogenic nucleotides (TPLFNs).

Here we take the Tokyo Green (TG) terminal phosphate-labeled deoxyadenosine fluorogenic substrate, TG-dA4P, as an example. The TG dye was synthesized based on the reported procedure²⁹. Briefly, 332 mg of TG (1 mmol) was suspended into 15 mL anhydrous CH₂Cl₂ in a flame-dried flask under Ar. To this solution Proton Sponge (759 mg, 3.50 mmol) was added with stirring. After 10 min the mixture was cooled to -10 °C and phosphorous(V) oxychloride (275 µL, 3.00 mmol) was added. The reaction was kept on at the same temperature for 30 min. Then TEAA buffer (20 mL of 1 M solution) was added to quench the reaction and to hydrolyze the phosphoryl chloride intermediate for 1 h at 0 °C. After that, the two phases were separated and the aqueous solution was filtered and concentrated in vacuum for further purification by reversed phase flash LC system. Conditions: AQ C-18 column (Agela 40 g) using 0–50% acetonitrile in 50 mM triethylammonium acetate buffer (PH 7.4), flow rate 20 mL/min. The purified TG-monophosphate triethylammonium (100 mM DMF solution) was kept in a -20 °C freezer for further usage. 2'-deoxyadenosine-5'-triphosphate (dATP) disodium salt (12.5 µL of 100 mM solution) was converted to the tributylammonium salt by treatment with ion-exchange resin (Bio-Rad AG-50W-XB) and tributylamine. The obtained tributylammonium salt was evaporated with anhydrous DMF (1 mL) twice and then dissolved in 0.5 mL anhydrous DMF under Ar. To the solution, carbonyldiimidazole (10.1 mg, 63 µmol) was added, and the mixture was stirred at room temperature for 12 h. After that, MeOH (3.2 µL) was added, and the solution was stirred for 0.5 h. Then TG-monophosphate triethylammonium salt (25 µmol) DMF solution (0.25 mL) from the previous step was transferred into the reaction by syringe, and MgBr₂ (25 mg) in DMF (0.5 mL) was added subsequently. The mixture was stirred for 30 h at room temperature. Then, the reaction mixture was concentrated, diluted with water, and purified on a C18 reversed phase high-performance liquid chromatography (HPLC) system (Shimadzu LC-20A) using preparative sephax Amethyst C18-H column (21.2 × 150 mm) at 3 mL/min flow rate, with a gradient of B (CH₃CN) in A (50 mM TEAA pH 7.3) (0–20% of B over 13 min, 20–30% of B over 10 min, 30–30% of B over 10 min). The desired fraction was collected and concentrated using a Hi-Trap Q-HP 1 mL anion exchange column (GE Healthcare). The collected solution containing the desired product can be purified again by HPLC using the same eluting conditions and concentrated by Hi-Trap Q-HP column. The product solution was stored at -20 °C for further usage. Mass spectral analyses were carried out with Bruker APEX IV Mass Spectrometer and AB Sciex MALDI-TOF 5800 Spectrometer. Other TG terminal phosphate-labeled fluorogenic substrates, TG-dC4P, TG-dT4P, and TG-dG4P, were synthesized following the same procedure. For the details of the synthesis and characterization, please check **Supplementary Note**, SI Section 1.1.

Surface modification and primer grafting of flowcells. We collected the used Illumina HiSeq flowcells, and stripped the original surface coating with chromic acid and then rinsed with DI water. The detailed coating process of the surface modification is described in **Supplementary Note**, SI Section 3.1. Briefly, *N*-(5-(2-bromoacetamido)pentyl)acrylamide (BRAPA) (70 mg in 700 µL DMF) was added into 10 mL 2% acrylamide in DI water. Then the mixture was well mixed, filtered by 0.22 µm filter, bubbled with argon for 15 min, then added sequentially with 11.5 µL tetramethylethylenediamine (TEMED) and 10 µL 50 mg/mL potassium persulfate in DI water. The solution was immediately vortexed and injected into the flowcell and the polymerization took place under humid argon atmosphere for 35 min. Then the hydrogel-coated flowcell was washed thoroughly with 200 mL DI water. The hydrogel-coated flowcell was then injected with 10 µM 5'-phosphorothioate oligonucleotide PS-T10-P7 (5'-T*⁺T*⁺T*⁺TTTTTTTCAAGCAGAAGACGGCATAAC-3', * = phosphorothioate) solution in PBS buffer (pH 8.0), reacted for 1 h at 50 °C, blocked by 10 mM 2-mercaptoethanol solution in PBS buffer (pH 8.0), and washed thoroughly with DI water.

DNA template preparation and immobilization. Briefly, the DNA templates were prepared from lambda phage genomic DNA (New England BioLabs) by nested PCR and flanked with two common adaptors, including the sequencing primer region. Then the template was mixed with PCR reagents and injected into the primer-grafted flowcell. The PCR mix contained the prepared DNA

template (1 nM), primer P5 (500 nM), primer P7 (62.5 nM), MgCl₂ (6 mM), dNTP (0.5 mM), Platinum Taq polymerase (0.5 U/ µL, Life Tech), BSA (0.2 mg/mL), and PCR buffer (200 mM Tris HCl, 500 mM KCl). To amplify the DNA onto the surface of the flowcell: the solid-phase PCR comprised two stages; the first stage was (1) hot start 95 °C for 90 s; (2) 15 thermal cycles, each composed of 30 s at 95 °C, 15 s at 65–60 °C gradually, and 30 s at 72 °C. The second stage contained 30 thermal cycles, each composed of 30 s at 95 °C and 300 s at 65 °C. After the solid-phase PCR, the products were denatured using formamide, removed using pipette, and washed by wash buffer (20 mM Tris-HCl buffer, pH = 8.0, 50 mM KCl). Details are in **Supplementary Note**, SI Section 3.2.

Instrumentation. We have built a laboratory prototype to perform the ECC sequencing experiments. Briefly, the sequencing flowcell was set on a temperature controller, under which was a 3D translation stage used to move the sequencing flowcell in three dimensions. Above the flowcell was a highly sensitive camera (Hamamatsu Flash4.0 sCMOS) and 10× microscope objective (Nikon CFI PlanAPO Lambda 10×, NA 0.45). When the blue light irradiated on the flowcell during reaction, the emitted green light was captured by the camera through microscope. On one end of the flowcell, there was a slim tube connected with valve and pump, to import the reaction buffer and wash buffer, while on the other end, the flowcell was mounted to a tube to export waste liquid.

Sequencing. For details, see **Supplementary Note**, SI Section 4.2. Briefly, there were six sequencing reagents, each containing TPLFN M, K, R, Y, W, and S, respectively. All six reagents contained *Bst* DNA polymerase (1 U/µL, NEB), calf intestinal alkaline phosphatase (CIP, 0.5 U/mL, NEB), MnCl₂ (1 mM), DTT (10 mM) in reaction buffer (40 mM trisbase, 40 mM NH₄Cl, 100 mM KCl). The concentrations of the four TPLFN were: 3 µM TG-dA4P, 2.5 µM TG-dC4P, 3 µM TG-dG4P, and 5 µM TG-dT4P. Each sequencing reagent contained only two of the four TPLFNs. For example, reagent M contained TG-dA4P and TG-dC4P since M stands for A and/or C. After annealing the sequencing primer, the sequencing instrument conducted the sequencing process automatically under the control of a LabVIEW program. Each sequencing cycle contained the following steps: (1) rinsed the flowcell with wash buffer; (2) cooled the flowcell to 4 °C, selected one of the sequencing reagents and loaded the reagent by a syringe pump; (3) set the temperature to 15 °C and took background fluorescence image; (4) raised the temperature to 65 °C for 1 min to perform the elongation reaction; (5) cooled the flowcell to 15 °C and acquired the fluorescence image. The sequencing instrument took a dual-base flowgram in each round. Taking round MK as the example, the sequencing reagents in each cycle were in the order M,K,M,K,M,K,M,K,... After one round of sequencing, the nascent DNA strand was denatured by formamide, and the sequencing primer was re-annealed to the DNA template to initiate a new sequencing round. Every sequencing run contained three rounds, including MK, RY, and WS.

Data analysis. Details and necessary codes can be found in **Supplementary Note**, SI Section 6.3, 6.5 and 7.1. In the dephasing part, a flux matrix *T* is constructed according to the “One Pass, More Stop” principle (see **Supplementary Note**, SI Section 6.2.2). Let **h** be the DPL array of the DNA sequence, **f** be the fluorescence intensity array, *a* the unit signal, *b* the decay coefficient, *c* and *d* the signal offsets for amending the background fluorescence intensities of different substrates, ξ the noise, and $\mathbf{t} = (1, 2, \dots, n)$, $\mathbf{s}^{(1)} = (s_1^{(1)}, s_2^{(1)}, \dots, s_n^{(1)})$, $\mathbf{s}^{(2)} = (s_1^{(2)}, s_2^{(2)}, \dots, s_n^{(2)})$, where

$$s_j^{(1)} = \begin{cases} 1 & \text{if } j \text{ is odd} \\ 0 & \text{if } j \text{ is even} \end{cases}$$

$$s_j^{(2)} = \begin{cases} 0 & \text{if } j \text{ is odd} \\ 1 & \text{if } j \text{ is even} \end{cases}$$

then the following formula holds:

$$\mathbf{f} = a \cdot b^{\mathbf{t}} \mathbf{Th} + c \mathbf{s}^{(1)} + d \mathbf{s}^{(2)} + \xi$$

When sequencing DNA templates with known sequences, the parameters (*T*, *a*, *b*, *c*, and *d*) were estimated by fitting the formula above using gradient descent. And when sequencing new DNA templates, these estimated parameters

were used to deduce the DPL array in an iterative manner. In the ECC part, a codeword space was constructed to represent every legitimate 3-tuple of the degenerate signals. A graph was constructed for each of the three degenerate signals (MK, RY and WS). Every common path of the three graphs represented a possible decoded DNA sequence and could be assigned with a probability according to the Bayesian formula. The maximum common path of the three graphs, namely, the common path with the maximum probability, was found

by traversing the codeword space using dynamic programming and then traced back to obtain the DNA sequence.

Data availability. The data that support the findings of this study are available from the corresponding author upon reasonable request.

A **Life Sciences Reporting Summary** is available.

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see [Reporting Life Sciences Research](#). For further information on Nature Research policies, including our [data availability policy](#), see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

▶ Experimental design

1. Sample size

Describe how sample size was determined.

not applicable to our work.

2. Data exclusions

Describe any data exclusions.

No data is excluded for analyses.

3. Replication

Describe whether the experimental findings were reliably reproduced.

not exactly applicable. but if we consider sequences are samples. Three different sequences are sequenced, each has more than four replicates.

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

not applicable to our work.

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

not applicable to our work.

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a Confirmed

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
- A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- A statement indicating how many times each experiment was replicated
- The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
- A description of any assumptions or corrections, such as an adjustment for multiple comparisons
- The test results (e.g. P values) given as exact values whenever possible and with confidence intervals noted
- A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
- Clearly defined error bars

See the web collection on [statistics for biologists](#) for further resources and guidance.

► Software

Policy information about [availability of computer code](#)

7. Software

Describe the software used to analyze the data in this study.

Matlab

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* [guidance for providing algorithms and software for publication](#) provides further information on this topic.

► Materials and reagents

Policy information about [availability of materials](#)

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.

Yes

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

not applicable to our work.

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.

not applicable to our work.

b. Describe the method of cell line authentication used.

not applicable to our work.

c. Report whether the cell lines were tested for mycoplasma contamination.

not applicable to our work.

d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by [ICLAC](#), provide a scientific rationale for their use.

not applicable to our work.

► Animals and human research participants

Policy information about [studies involving animals](#); when reporting animal research, follow the [ARRIVE guidelines](#)

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.

not applicable to our work.

Policy information about [studies involving human research participants](#)

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.

not applicable to our work.