Published online 16 May 2016

Nucleic Acids Research, 2016, Vol. 44, No. 13 ***e118***
doi: 10.1093/nar/gkw436

Mapping DNA polymerase errors by single-molecule sequencing

David F. Lee[1,†], Jenny Lu[1,†], Seungwoo Chang[2], Joseph J. Loparo[2] and Xiaoliang S. Xie[1,*]

[1]Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA 02138, USA and [2]Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, MA 02115, USA

Received February 12, 2016; Revised April 27, 2016; Accepted May 07, 2016

ABSTRACT

Genomic integrity is compromised by DNA polymerase replication errors, which occur in a sequence-dependent manner across the genome. Accurate and complete quantification of a DNA polymerase's error spectrum is challenging because errors are rare and difficult to detect. We report a high-throughput sequencing assay to map *in vitro* DNA replication errors at the single-molecule level. Unlike previous methods, our assay is able to rapidly detect a large number of polymerase errors at base resolution over any template substrate without quantification bias. To overcome the high error rate of high-throughput sequencing, our assay uses a barcoding strategy in which each replication product is tagged with a unique nucleotide sequence before amplification. This allows multiple sequencing reads of the same product to be compared so that sequencing errors can be found and removed. We demonstrate the ability of our assay to characterize the average error rate, error hotspots and lesion bypass fidelity of several DNA polymerases.

INTRODUCTION

DNA polymerases act during DNA replication and repair to catalyze the synthesis of a complementary DNA strand from a DNA template. Errors made during this replication process are rare but can drive disease (1,2) or evolution (3,4). The impact of DNA polymerase errors depends on the type of error and its location and frequency, but these are difficult to predict. This is because each DNA polymerase has a unique error spectrum, and each organism contains a diverse mix of DNA polymerases that are recruited by different pathways (5). Error rates also vary with the template sequence, and bases where a DNA polymerase is particularly error prone ('error hotspots') undergo accelerated mutagenesis (6,7). In addition, DNA bases are subject to chemical modifications *in vivo* which can compromise fidelity to different degrees depending on the replicating polymerase (8,9). Altogether, DNA polymerase fidelity and its impact on genome stability have been challenging to understand.

DNA polymerase fidelity can be measured by quantifying the errors made during *in vitro* DNA replication, but errors are rare and existing methods of quantification have significant limitations. Early methods involved transfecting the replication products into bacteria for clonal amplification and sequencing (10). This method allows the average error rate of a polymerase to be determined, but error rates cannot be quantified at base resolution because very few errors can be collected. Mutation assays, which follow a similar methodology, select a target gene that causes a phenotypic change in the transfected bacteria if incorrectly replicated, allowing colonies with error-containing products to be selected (11–14). This modification improves throughput but can only be used to detect errors at phenotypically detectable sites on a limited number of template sequences, and remains relatively low-throughput. Non-phenotypically detectable errors can only be scored if multiple errors are made during each round of replication, which only occurs frequently for highly inaccurate polymerases (15). As an alternative to phenotypic selection, denaturing electrophoresis or thin layer chromatography can be used to separate error-containing products (16,17). Separation using these techniques is simple when only a few products are present, such as when a DNA lesion causes most errors to be made at a single position. However, when the error diversity is greater, achieving good error resolution is challenging because multiple cycles of separation, purification and sequencing identification are required. In light of these limitations, our understanding of DNA polymerase fidelity would benefit from a new technique that has greater throughput and fewer practical restrictions.

We have developed a more powerful approach to quantify DNA polymerase fidelity using high-throughput sequencing. With high-throughput sequencing, a large number of replication products can be sequenced at the single-molecule level, allowing direct quantification of rare errors without intermediate error detection and product separation steps. A significant obstacle to this approach is

*To whom correspondence should be addressed. Tel: +1 617 496 9925; Fax: +1 617 496 8709; Email: xie@chemistry.harvard.edu
† These authors contributed equally to the work as first authors.

the high rate of sequencing error in high-throughput sequencing instruments. However, this obstacle has been overcome with a strategy known as barcoding (18–22), which allows sequencing errors to be identified and separated. Barcoded high-throughput sequencing techniques have previously been used to quantify DNA polymerase fidelity, but these previous approaches had limitations. In one approach, the error rate of the engineered Phusion DNA polymerase was determined by quantifying the proportion of PCR products that contained errors (18). However, errors can affect PCR efficiency and cause amplification bias (23). Another approach had a high background error rate, making DNA polymerase errors difficult to distinguish (24).

In this report, we present a new approach to quantify DNA polymerase fidelity using barcoded high-throughput sequencing on the Illumina platform. Our method avoids PCR quantification bias by quantifying error rates from a single round of DNA synthesis. We demonstrate that our barcoding approach can remove sequencing errors, resulting in a low error background. We then evaluate its ability to quantify overall DNA polymerase error rate, obtain reproducible error spectra, identify mutation hotspots and assess the impact of a single-base DNA lesion on fidelity.

MATERIALS AND METHODS

Primer design

Twelve forward primers (F0X, F00–F10) and 12 reverse primers (R0X, R00–R10) were used in these experiments (Supplementary Table S1). From 5′ to 3′, these primers contained a 33 nucleotide partial Illumina adapter sequence, a 'condition' barcode sequence (bold) that varies between each primer, a 12–15 nucleotide random 'product' barcode region and a priming region for the target sequence. A 12 bp random nucleotide region represents 4×10^{12} unique combinations, which is much greater than the 1×10^{4}–1×10^{5} products we collect for each reaction condition. Combinations of forward and reverse primers allows for a maximum of 144 reaction conditions or replicates to be separated during sequencing (12×12 condition barcodes). The condition barcodes were variable in length to ensure cluster diversity during sequencing, and designed for near-equal base representation. Primers were also analyzed *in silico* for primer homodimers and heterodimers (Multiple primer analyzer, Thermo Scientific). The sequences of library-preparation primers PE 1.0 and PE 2.0 were obtained from Illumina. The 3′ end of PE 1.0 and PE 2.0 are complementary to the 5′ end of the forward and reverse primers. Ultramer synthesis was used for all primers (Integrated DNA Technologies).

Templates and proteins

Escherichia coli DNA Polymerase IV was purified as previously described (25). All other polymerases were obtained from New England Biolabs (NEB).

The plasmids pBeloBAC11 (NEB) and pOPINP (gift from Ray Owens, Addgene plasmid # 41139) were clonally amplified and isolated from *E. coli K12 ER2420* and *DH5alpha* strains, respectively. Cells were streaked on Luria Broth (LB) agar plates with 25 μg/ml chloramphenicol or 50 μg/ml carbenicillin, respectively, and incubated for 12 h at 37°C. Single colonies were inoculated in liquid LB media with 25 μg/ml chloramphenicol or 50 μg/ml carbenicillin at 37°C in a shaking incubator. OD 600 measurements were taken to determine the start of stationary phase, upon which the ZR Plasmid Miniprep kit (Zymo Research) was used to isolate the plasmid. Purified plasmids were quantified using a Nanodrop 2000 UV-Vis spectrophotometer (Thermo Scientific).

Single-stranded DNA with a site-specific DNA lesion was constructed using M13mp7(L2), a mutant phage that contains an EcoRI site within a stable hairpin in its genome, as previously described (16). Briefly, a 20-mer oligonucleotide (Chemgenes: 5′-CTA CCT XTG GAC GGC TGC GA-3′) containing a fluoro substituent at the N2 position of guanine (X) was treated by furfurylamine and purified by HPLC and MALDI-TOF MS. The oligonucleotide was verified to 99.9% purity by mass spectromoetry. The M13mp7(L2) phage genome was purified and the lesion-containing oligonucleotide ligated into the digested EcoRI site using annealed scaffold oligonucleotides. Scaffold oligonucleotides, unligated linear M13 DNA and excess insert were removed by treatment at 37°C for 4 h with 18 U T4 DNA polymerase and 80 U exonuclease I (New England Biolabs). A control 20-mer oligonucleotide (IDT: 5′-CTA CCT GTG GAC GGC TGC GA-3, 99.9% purity by mass spectrometry) was ligated into the M13mp7(L2) plasmid in a similar manner to create the control substrate.

Generation of replication product

For each reaction involving double-stranded template DNA, 1–10 ng of template was digested using 1U of either BsmI (pBeloBac11), BseRI and BspHI (pOPINP) or PvuII-HF (M13mp7(L2)) (New England Biolabs) with incubation for 10 min at 37°C. This left a shorter double-stranded region containing the target locus. The resulting DNA was purified using the AMPure XP bead system (Beckman Coulter).

The extension step with the polymerase of interest was conducted under variable buffer and temperature conditions depending on the polymerase and DNA template. For 3′→5′ exonuclease deficient Klenow Fragment (New England Biolabs) and Taq (New England Biolabs), 1X Taq Reaction Buffer (New England Biolabs) was supplemented up to 3 mM Mg^{2+} with $MgCl_2$. For *E. coli* DNA Polymerase IV and DNA Polymerase I (New England Biolabs), the buffer consisted of 50 mM pH 7.9 HEPES-NaOH, 12 mM $Mg(OAc)_2$, 80 mM KCl, 0.1 mg/ml BSA and 5 mM DTT. An alternative buffer formulation for DNA Polymerase IV consisted of 20 mM pH 7.5 Tris, 8 mM $MgCl_2$, 5 mM DTT, 0.1 mM EDTA, 25 mM sodium glutamate, 40 μg/ml BSA and 4% glycerol. For Q5 DNA Polymerase (New England Biolabs), extension was conducted with 1X Q5 Buffer at 72°C. Each reaction mixture also contained 0.025 μM of forward primer (Integrated DNA Technologies) and 200 uM dNTP. For polymerases from New England Biolabs, 1 unit was used for extension. For *E. coli* DNA polymerase IV, a 5:1 ratio of polymerase to primer-template junction was used. To begin the extension reaction, a 10 μl solution consisting of the reaction buffer, forward primer and template sequence was prepared. If the template was

double-stranded, the mixture was subjected to denaturation at 95°C for 30 s, primer annealing for 2 min at 52°C, followed by ramping of the mixture to extension temperature. For single-stranded template DNA, the mixture was held at 65°C for 3 min then ramped to the extension temperature at 0.1°C/s. The extension temperature was 37°C for Klenow Fragment (exo-), Taq, DNA Pol IV and DNA Pol I and 72°C for Q5. Once the extension temperature was reached, a 10 μl solution containing buffer, dNTPs and the polymerase was added. After extension, the reactions were quenched with 5 μl of 50 mM EDTA and then purified using AMPure XP beads. We used quantitative polymerase chain reaction (qPCR) to determine the extension time required for saturation, which was 5–20 min depending on the polymerase and template used.

The synthesis of the complementary strand was conducted with the purified extension product as template, 0.025 μM reverse primer, 1U Q5 High Fidelity DNA polymerase, 200 μM of each dNTP and 1X Q5 buffer. The product was denatured at 98°C for 30 s, the reverse primer annealed at 52°C for 2 min and extension performed at 72°C for 5 min. Extensions were quenched with 5 μl of 50 mM EDTA and then purified using AMPure XP beads.

The concentration of the complementary strand was quantified using qPCR with primers complementary to the partial Illumina adapters (Forward: 5′- TAC ACG ACG CTC TTC CGA TCT -3′, Reverse: 5′- CAT TCC TGC TGA ACC GCT CT -3′). Standards were generated by serial dilution of 10 mM PhiX v3 Control Template (Illumina) to concentrations of 1×10^8, 1×10^7, 1×10^6, 1×10^5 and 1×10^4 copies/μl. qPCR reactions were performed using the DyNAmo SYBR Green qPCR kit (Thermo-scientific) on the 7500 Real-Time PCR system (Applied Biosystems).

All experiments were conducted with at least two replicates. Replicates were performed in parallel with aliquots taken from the same starting template pool.

Library preparation and sequencing

The qPCR quantified products were amplified using primers PE 1.0 and PE 2.0 (Illumina, Integrated DNA Technologies). The volume of product and the number of cycles used for amplification depended on the number of unique products and the desired depth of sequencing coverage. As an example, we amplified 20 000 unique products in a 20 μl reaction mix using two rounds of 13 cycles of denaturation at 98°C for 8 s and annealing/extension at 72°C for 30s. The PCR products were purified after each round using the AMPure XP Bead system. Two separate rounds were used because a single round of 26 cycles produced PCR side-products. After amplification, the library was quantified using a Qubit 2.0 Fluorometer (Life Technologies) and sized using the 2200 Tapestation (Agilent). Sequencing was performed on the MiSeq Desktop Sequencer using the MiSeq V2 300 cycle kit (Illumina). Base calling was conducted under the standard Illumina pipeline.

Criteria for quantifying DNA polymerase error rate

The paired-end FASTQ sequencing reads were separated into groups that shared the same condition barcodes. Any reads which had a base with quality score (Q) < 20 was removed. The first 5 bases of the read in the target region were compared to the reference sequence. If these did not match, the read was removed since this mismatch could have been the result of primer truncations or improper annealing. The filtered forward and reverse reads were then aligned to the corresponding section of the reference sequence using the multiple sequence alignment by log-expectation (MUSCLE) algorithm (26). These two aligned sequences were then compared to call errors. The total length of forward and reverse reads was greater than the target region so that insertion errors (which add bases into the sequence) are also covered. Errors that occurred in adjacent base positions were grouped together as 'multi-base' errors. After errors were called, reads with the same product barcode were grouped together since these were copies of the same original product. Errors were accepted as polymerase errors only if the product had 3 or more copies and the error was present in all of them. When calculating the average error rate over a template, the error rate of Q5 (used for complementary strand synthesis) for that sequence was subtracted from the overall substitution rate. Base-resolution error spectra are presented without subtraction. Our error spectra represent errors made by DNA polymerase during replication of the template. This is different from the convention used in reports that utilize the LacZα mutation assay (11), where the spectra represent errors made in the transcript that is synthesized from the replicated locus.

Sequencing analysis software for the identification of mutations was written in Python and run on the Harvard Odyssey Computing Cluster. Data analysis and figure construction were carried out using numpy, pandas and matplotlib.

Accession codes and code availability

Sequencing data are available from the sequencing read archive (SRA) with accession number SRX1559518. Python scripts for splitting of condition barcodes and filtering of sequencing errors are available in the Supplementary Material.

RESULTS

Description of barcoding assay

Our approach to quantify DNA polymerase fidelity is illustrated in Figure 1. A pool of templates with identical sequences undergoes one round of extension with the polymerase of interest (Figure 1A). The primers (Figure 1B) contain a randomized 12 bp barcode sequence to tag each product with a unique 'product barcode'. The primers also contain a 'condition barcode' unique to each reaction condition, allowing multiple reactions to be pooled and sequenced simultaneously. After extension by the polymerase of interest, the complementary strand is synthesized by a high-fidelity polymerase using a primer of the same structure. This complementary strand is then PCR amplified using primers complementary to the partial Illumina adapters on both ends of the product, generating a library with multiple barcoded copies of each original product. After paired-end sequencing, the reads are grouped according

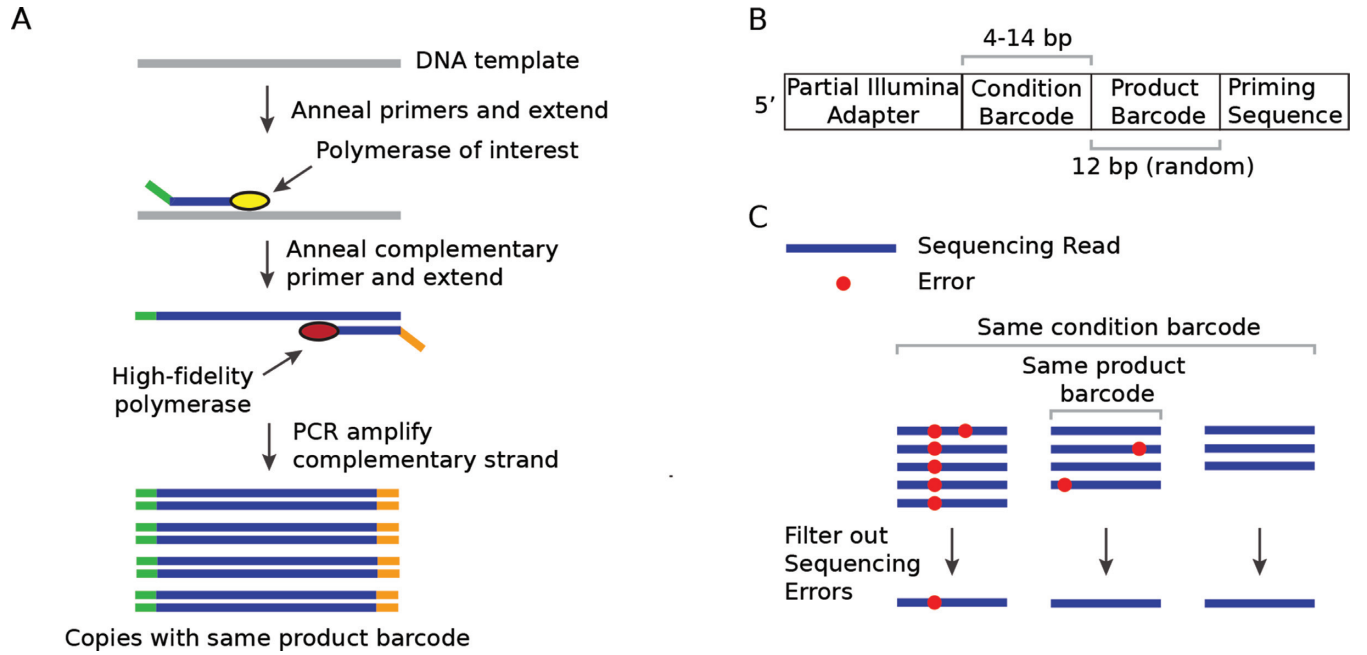

Figure 1. Schematic of barcoding strategy. (**A**) Workflow to generate products for paired-end sequencing. The pool of templates is replicated using the polymerase of interest. The complementary strands are then synthesized using a high-fidelity polymerase. In both cases, a special primer (green and orange) containing a partial Illumina Adapter, a random product barcode and a condition barcode is used. Primers complementary to the partial Illumina adapter are used to PCR amplify the complementary strands, forming the sequencing library. Each amplification product is tagged with a unique set of product barcodes that indicates its origin. (**B**) The special primer contains a part of the Illumina sequencing adapter, a 'condition barcode' that is unique to each reaction, a 12 bp randomized 'product barcode' that uniquely tags each product and the priming sequencing for the region of interest. (**C**) After sequencing, reads are grouped according to condition barcode and product barcode. Sequences are aligned to the correct sequence and errors are called. Errors are only kept if they are present in all copies, otherwise they are discarded as sequencing error.

to the product barcodes on both ends (Figure 1C). Errors generated in the initial extension by the polymerase of interest should be present in all copies of the product. Sequencing errors can be recognized because they are most likely present in only a fraction of copies and can therefore be eliminated. After sequencing errors are filtered out, the DNA polymerase errors are obtained for each product. This approach is not subject to PCR quantification biases because error rates are quantified using the number of unique products and not their final amplified amount.

In addition to removing sequencing errors, we took measures to minimize other sources of false positives throughout the protocol. We generated the starting templates by clonally amplifying a plasmid containing the template sequence in *E. coli*. *E. coli* replication has a low error rate of about 1×10^{-9} errors per base pair per replication (27) and generates a homogenous starting template pool. We also minimized errors during synthesis and PCR amplification of the complementary strand by using Q5 DNA polymerase (Q5), the highest fidelity DNA polymerase available.

Sequencing error removal and DNA polymerase error rate quantification

To test if the barcoding strategy could reduce sequencing errors, we determined the error rate of Q5 DNA polymerase (Q5) when different numbers of product copies were used to filter sequencing errors. To do this, we grouped products according to the number of copies captured by sequencing and determined the error rate as a function of copy number. Since Q5 was used for both initial extension and complementary strand synthesis, the true error rate was calculated as half the recorded value. These error rates were averaged over two template sequences: a 188 base sequence within the chloramphenicol acetyltransferase (Cm^R) gene of the *pBeloBac11* plasmid vector and a 281 base sequence within the *LacZα* gene of the *pOPINP* plasmid vector (Supplementary Tables S2 and S3). One replicate over the *LacZα* locus was excluded because the template showed evidence of DNA damage (Supplementary Figure S1).

For products with only one copy, sequencing errors and polymerase errors could not be separated and the recorded error rate was 1.3×10^{-4} substitutions/bp (Figure 2). For products with 2 copies, the error rate decreased to 5.6×10^{-6} substitutions/bp because sequencing errors were removed. When more copies were present, the substitution error rate decreased further to 4.4×10^{-6} substitutions/bp for 5 copies. Sequencing generated deletions and insertions were also removed, with these error rates decreasing from 0.99×10^{-5} deletions/bp and 2.2×10^{-7} insertions/bp at 1 copy to no detected deletions or insertions at 5 copies (Figure 2). This shows that our barcoding method successfully allows the separation and removal of sequencing errors.

We measured the average error rates of 3′→5′ exonuclease deficient Klenow Fragment (Klenow (exo-)), Taq, *E. coli* Y-family DNA Polymerase IV (Pol IV) and Q5 over the Cm^R, *LacZα* (-) strand, and *LacZα* (+) strand loci (Supplementary Tables S4 and S5) for comparison with published values. To minimize sequencing errors while maxi-

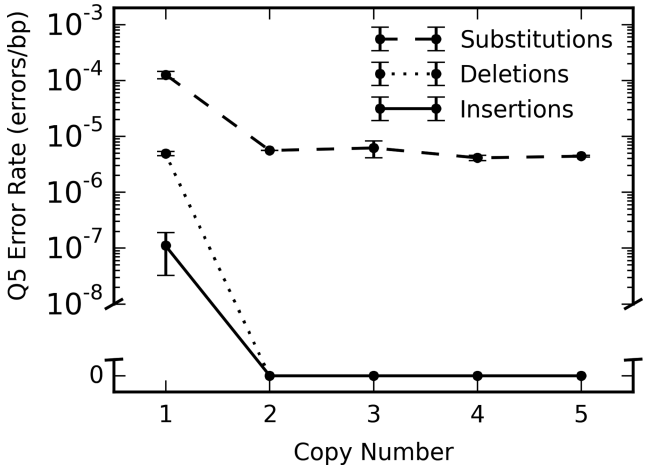

Figure 2. Q5 error rate as a function of product copy number. As the number of product copies used for comparison is increased, sequencing errors are increasingly eliminated. Error bars indicate standard error.

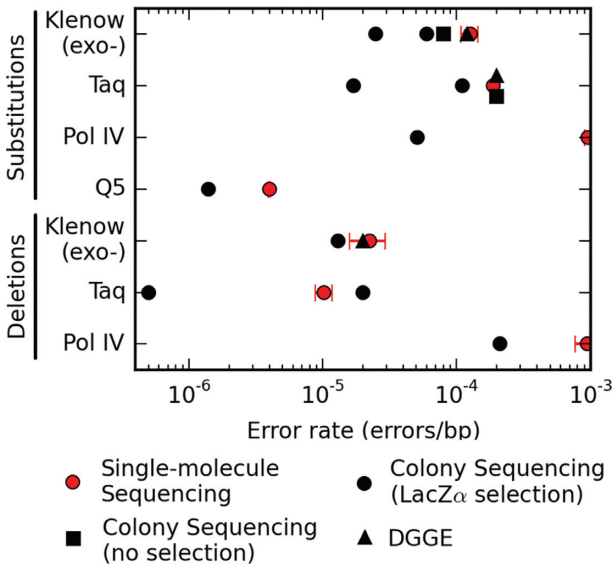

Figure 3. Comparison of error rates from our barcoded sequencing assay (red circle) with results from denaturing gradient gel electrophoresis (DGGE) (black triangles), sequencing of mutant bacterial colonies with *LacZα* forward mutation selection (black circles), or direct sequencing of bacterial colonies without phenotypic selection (black squares) (see Supplementary Table S6). For our assay values, the red points are means and the red bars indicate the standard error.

mizing product number, we analyzed products with 3 or more copies. The Q5 error rate was subtracted from our measurements to account for errors made during complementary strand synthesis. We measured the error rates over two technical replicates, where replicates were done by taking aliquots from the template pool and conducting the assay on each in parallel. We compared our results with those previously collected using denaturing gradient gel electrophoresis (DGGE) (23), sequencing of mutant bacterial colonies with *LacZα* forward mutation selection (11,28–31), or direct sequencing of bacterial colonies without phenotypic selection (10,32) (Supplementary Table S6). Our error rates for Klenow (exo-) and Taq were similar to both DGGE and direct sequencing values (Figure 3). In contrast, our error rates for Klenow (exo-), Taq, Pol IV and Q5 were on average 7 times higher for substitutions and 3 times higher for deletions than the *LacZα* forward mutation assay values. To investigate the cause of this difference, we considered in greater detail the case of Pol IV replication over the *LacZα* (+)-strand. We first modified the buffer conditions for Pol IV replication over the *LacZα* (+)-strand to match that in the corresponding forward mutation assay study (see Materials and Methods for buffer composition). This reduced the error rates from 1.06×10^{-3} sub/bp and 1.3×10^{-3} del/bp to 4.6×10^{-4} sub/bp and 6.3×10^{-4} del/bp. We then removed the errors which are not phenotypically detectable (11), further reducing the error rate to 3.2×10^{-4} sub/bp and 4.6×10^{-4} del/bp (Supplementary Methods). This overall 3-fold reduction shows that extension conditions and phenotypic detectability have a significant influence on error rates and may partially account for the discrepancy. However, since some *LacZα* forward mutation assay results were up to 20 times lower than our measurements, a significant deviation still remains. Thus, our measurements corresponded with results from other non-phenotypic techniques but appeared to deviate from those of the mutation assay.

Variations in error rate and identification of error hotspots

DNA polymerase fidelity varies across a template. To investigate this variation and its reproducibility, we mapped the frequency of single-base substitutions for Pol IV (39 641 substitutions in 163 949 products) and Klenow (exo-) (4046 substitutions in 122 846 products) replication over the *LacZα* locus (−) strand for two technical replicates each (Supplementary Table S4). The error spectra (Figure 4A and Supplementary Figure S2) illustrate that error rate varied substantially from base to base along the template. The variation ranged over two orders of magnitude, and the error rates at each base were reproducible between replicates for Pol IV (Pearson $\rho = 0.97$, $P < 0.01$) and Klenow (exo-) (Pearson $\rho = 0.74$, $P < 0.01$) (Figure 4B), indicating that the variation was not due to sampling noise. Pol IV and Klenow (exo-) were strikingly different in their error spectra (Spearman $\rho = 0.13$, $P < 0.01$) (Figure 4C), suggesting that these variations were polymerase specific. To demonstrate how sampling noise would make these variations difficult to accurately characterize, we randomly under-sampled the error spectra to between 50 and 2000 errors and compared it to the original. We repeated this 100 times to obtain the average similarity at each error sampling number. The error rates averaged over the entire template remained similar (Supplementary Figure S3), but the correlation between error rates at each base improved from $\rho \sim 0.30$ at 50 errors to $\rho \sim 0.95$ at 2000 errors (Figure 4D). Next, to identify error hotspots, we fit the distribution of substitution errors per base position to a combination of Poisson distributions using the omputer-assisted analysis of mixture distributions (C.A.MAN) package (33) (Figure 5A and Supplementary Figure S4). Hotspots were first defined as those

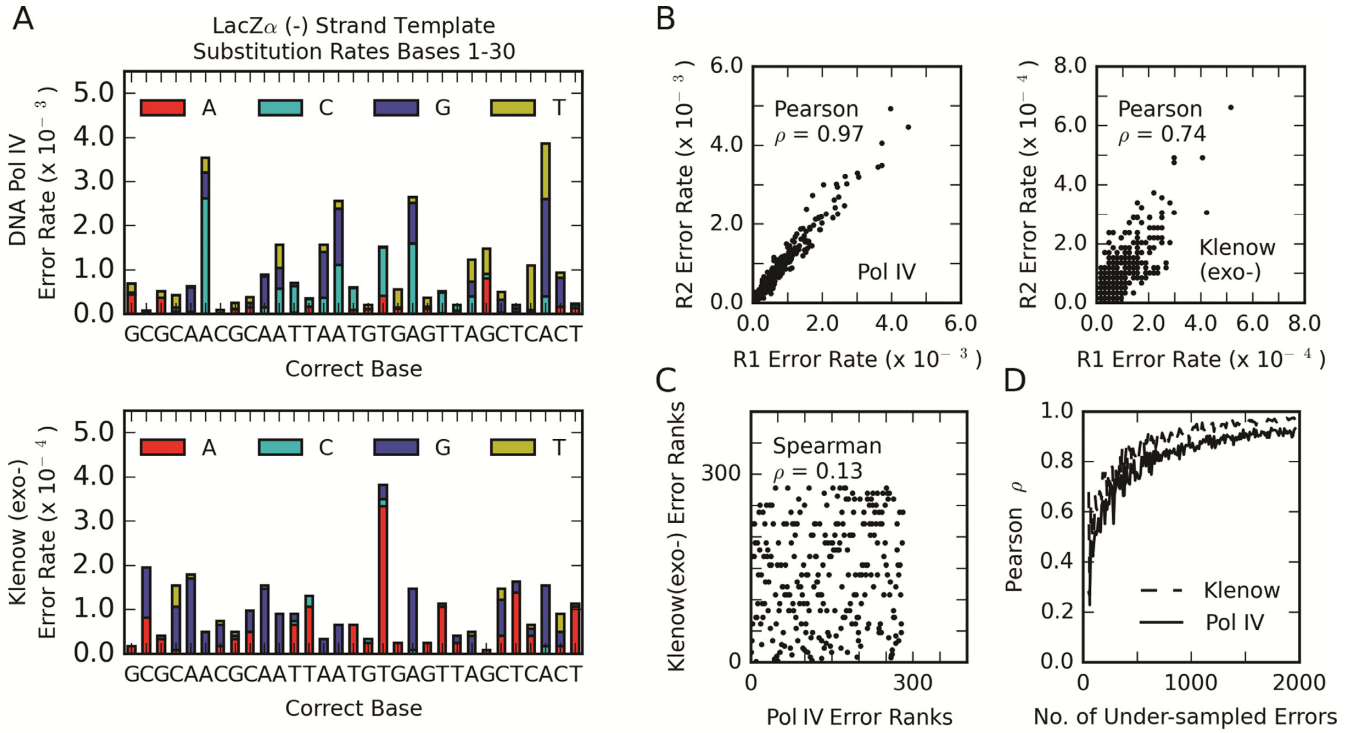

Figure 4. Variations in substitution error rate when replicating the *LacZα* (−) strand template. (**A**) A snapshot of the first 30 bases of the replication product for Pol IV and Klenow (exo-) illustrates variations in substitution error rate across the template. The length of the bar indicates the error rate. (**B**) Correlation plots of error rates at each base position between technical replicates for Pol IV replicates (21 212 and 18 429 mutations) and Klenow (exo-) (1973 and 2073 mutations) show that the variations were reproducible and hence not due to sampling noise. R1 and R2 designate the first and second replicates respectively. An outlier for Klenow (exo-) at 1.66×10^{-3} (replicate 1) and 2.04×10^{-3} errors/bp (replicate 2) was excluded from analysis. (**C**) Correlation plots of error ranks for Pol IV versus Klenow (exo-) show that error spectra are strikingly different depending on the polymerase. (**D**) Pearson correlation coefficient between the original error spectrum and an under-sampled copy as the number of errors per under-sampled copy is changed. Average of 100 repeats at each error number.

positions that deviated significantly from the fitted distribution ($P < 0.05$ with Benjamini–Hochberg correction assuming independence). This yielded 1 error hotspot in the Klenow (exo-) spectrum at position 260 which was 13 times more error prone than average (Supplementary Figure S2). Under-sampling made this hotspot difficult to distinguish, as even with 500 errors the hotspot was only called in 50% of under-sampling replicates (Figure 5B). If the definition of 'hotspot' was relaxed to include all positions that belong to the Poisson distribution with the highest mean error parameter, we could identify 2 hotspots at positions 106 and 132 in the Pol IV spectrum that were 5 times more error prone than average, and 2 hotspots at positions 80 and 260 in the Klenow (exo-) spectra that were 5 and 13 times more error prone than average (Supplementary Figure S2). To rule out the possibility that the error spectra of Pol IV and Klenow (exo-) were different because their extension buffers were different, we repeated the analysis after using Pol IV buffer for both polymerases. As before, there was a poor correlation (Spearman $\rho = 0.31$, $P < 0.01$) and no overlap in hotspots. In summary, we were able to identify variations in error rate that are both reproducible and polymerase specific.

To demonstrate the advantages of our assay in characterizing error spectra, we mapped the single-base substitution spectrum of Pol IV replicating over the *LacZα* locus (+) strand (837 substitutions in 6578 products) and compared our results to those mapped using the *LacZα* forward mutation assay (66 substitutions) (31) (Supplementary Figure S5). Our extension was performed in the same buffer as that reported in the forward mutation assay. The correlation between the error profiles was quite poor even after limiting our spectrum to phenotypically detectable errors (Spearman $\rho = 0.30$, $P < 0.01$), probably due to the small sample size of the forward mutation assay spectrum. Furthermore, whereas three error hotspots (under the relaxed definition) could be identified at positions 23, 95 and 270 from our data, only 1 error hotspot at base position 95 could be identified from the mutation assay data (Supplementary Figure S4) because the other two hotspots are not phenotypically detectable. This comparison illustrates the importance of a high-throughput and non-phenotypic method in error profiling and hotspot identification.

Impact of a DNA lesion on fidelity

DNA lesions compromise the fidelity and replication kinetics of DNA polymerase, but cells contain translesion synthesis polymerases that are specially adapted to synthesize across lesions. To test if our assay can correctly measure the effects of a lesion on DNA polymerase fidelity, we investigated the error rate of lesion bypass by Pol IV, a translesion synthesis polymerase, over an N^2-furfuryl-dG

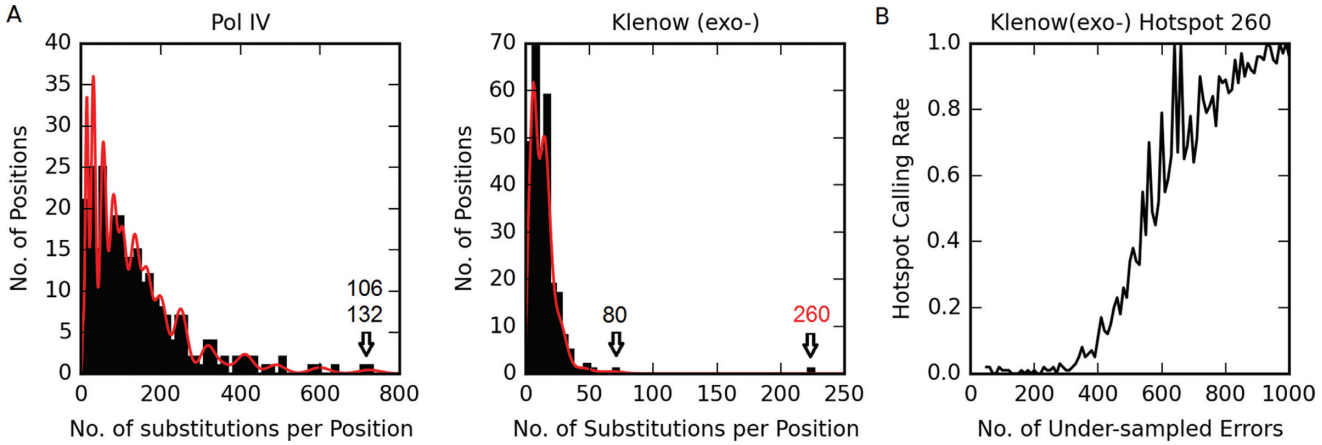

Figure 5. Identification of substitution error hotspots. (**A**) Histograms showing the distribution of substitution errors per base position for Pol IV and Klenow (exo-) replication across the *LacZα* (−) strand template. The histograms are fit to a combination of Poisson distributions (red) using the C.A.MAN package. Hotspots and their positions are indicated. The red-lettered hotspot was identified as being exceptional to the fitted distribution ($\alpha < 0.05$, Benjamini–Hochberg corrected). The black-lettered hotspots were identified under a more relaxed definition, which included all positions that belonged to the Poisson distribution with highest mean error parameter. (**B**) Frequency with which the Klenow (exo-) position 260 hotspot is identified when the original spectrum is under-sampled. Average of 100 repeats at each error sampling number.

lesion. The N^2-furfuryl-dG lesion is a structural analogue of the main lesion formed in cells treated with nitrafurazone, and Pol IV has been shown to bypass this lesion effectively (34). We compared the error spectrum of Pol IV replicating across the N^2-furfuryl-dG lesion to that of Pol I, an accurate 3′→5′ exonuclease-proficient replicative polymerase, as well as Klenow (exo-). The lesion-containing ('damaged') substrate and the lesion-free (control) substrate were made by ligating 20 base oligonucleotides into the M13mp7(L2) plasmid (see Materials and Methods). To account for false positives due to oligonucleotide synthesis errors, we subtracted the control substrate error spectra from the damaged substrate error spectra.

All three polymerases had increased substitution error rate when synthesizing DNA across the lesion. As expected, Pol IV had the lowest error rate of 1.27×10^{-2} substitutions/bp (Table 1), while Pol I and Klenow (exo-) were much more error prone and made errors at rates of 1.25×10^{-1} substitutions/bp and 1.93×10^{-1} substitutions/bp, respectively. The relatively low error rate of Pol IV is consistent with previous kinetic measurements for DNA replication across the N^2-furfuryl-dG lesion, which showed that Pol IV incorporates the correct base at a greater kinetic rate than Pol I (34). The dominant error type was different among the DNA polymerases. The G*·dTTP mismatch (C→T transition) was most common for Pol IV at 1.13×10^{-2} occurrences/bp and also for Pol I at 1.17×10^{-1} occurrences/bp. This result matches with previous kinetic measurements for Pol IV, which report that thymine was incorporated opposite the lesion at the greatest rate (34) (Supplementary Table S7). In contrast, the Klenow (exo-) substitution spectrum was dominated by the G*·dGTP mismatch, which occurred at a rate of 1.42×10^{-1} errors/bp. We also characterized the fidelity of replication at bases adjacent to the lesion site (Supplementary Figure S6). Although it appears that replication fidelity is reduced in the proximity of the lesion, there are error hotspots in both the lesion and control spectra that seem to be caused by inconsistencies in oligonucleotide synthesis or damage from template preparation. Overall, the error rate at the lesion can be measured with confidence, but characterization of fidelity around the lesion requires additional investigation.

DISCUSSION

We have taken advantage of barcoded high-throughput sequencing to develop a more powerful approach to map DNA polymerase errors. Errors are detected by high-throughput sequencing of replication products at the single-molecule level, so there are no restrictions on template sequence or error resolution imposed by phenotypic selection, colony picking or gradient separation. The majority of sequencing errors can be removed with only 2 barcoded copies of a product, which minimizes the additional sequencing cost of barcoding. Our assay is minimally affected by PCR biases because error quantification is based on the number of unique products before amplification. With these features, our assay can obtain accurate and reproducible single base-resolution maps of error rates for any type of DNA polymerase and template substrate.

We cross-validated our measurements for average DNA polymerase error rates with published values and found that correspondence was technique dependent. Our results matched with those from DGGE and clonal colony sequencing, which are both non-phenotypic methods, but were on average 7 times higher for substitution rates and 3 times higher for deletion rates than measurements made using the phenotypic selection and sequencing of mutant *LacZα* colonies. We found that this discrepancy could be partially explained after accounting for phenotypically detectable mutations and differences in replication buffer conditions. Despite this, the *LacZα* forward mutation assay still appears to underestimate the error rates relative to non-phenotypic methods. Furthermore, measurements made using the mutation assay can vary by an order of magnitude. There is no perfect reference for the 'true' polymerase error rate because template DNA damage, which influences error

Table 1. Error rates for each DNA polymerase when replicating across the N^2-furfuryl-dG lesion

	Error Rate at N^2-furfuryl-dG lesion (errors/bp)		
Type of Mutation	DNA Polymerase IV	Klenow (exo-)	DNA Polymerase I
G*·dTTP	1.13×10^{-2}	4.47×10^{-2}	1.17×10^{-1}
G*·dATP	1.30×10^{-3}	6.17×10^{-3}	1.11×10^{-3}
G*·dGTP	8.58×10^{-5}	1.42×10^{-1}	7.33×10^{-3}
Total	1.27×10^{-3}	1.93×10^{-1}	1.25×10^{-1}

rates, can be difficult to detect. Nonetheless, our comparison suggests that there could be a source of bias and variability in phenotypic detection that is not yet accounted for. Our technique should, in principle, produce the most accurate values since error-containing products are directly sequenced without intermediate steps or PCR quantification bias.

One important application of sequencing-based assays is to quantify variations in DNA polymerase error rate across a template and identify mutagenic hotspots. We were able to reproducibly characterize substitution error spectra and hotspots by collecting substitution densities of 4000–40 000 over a 281 base pair region. The distribution of substitution errors per site could only be fit by a mixture of Poisson distributions, which implies that the error rate of DNA polymerase replication varied along the template. There are numerous factors which could cause this heterogeneity to arise: template base identity, surrounding DNA sequence, DNA damage or variations in DNA structure. Our technique will facilitate a more detailed dissection of these influences, although in future investigations a more rigorous assessment of the fitting algorithm may be needed. We also found that the substitution spectra and hotspot locations were very different between Klenow (exo-) and Pol IV, emphasizing that error spectra are both polymerase and template specific. We showed that accurate characterization of the error spectrum would be challenging with low-throughput methods. To produce an error spectrum with at least $\rho = 0.7$ ($r^2 \approx 0.50$) correspondence with the high-density spectrum, at least 300–400 substitutions would need to be collected. Since most products contain 0 or 1 errors, low-throughput methods would require hundreds of products to be individually sequenced. Furthermore, by comparing our substitution spectrum for Pol IV with a published *LacZα* forward mutation assay spectrum, we found that the mutation assay missed the location of two hotspots which did not contain phenotypically detectable mutations. This demonstrates that our high-throughput and non-phenotypic technique is capable of producing a more accurate and complete characterization of DNA polymerase errors than existing methods.

Sequencing-based assays are also used to investigate the mutagenic effect of DNA lesions by quantifying their impact on DNA polymerase fidelity. We confirmed that Pol IV, a translesion synthesis polymerase, was best adapted to replicate across the N^2-furfuryl-dG DNA lesion in comparison to Pol I or Klenow (exo-). Furthermore, we observed that the most common error type varied depending on the polymerase. These results are in correspondence with previous kinetic measurements for Pol IV and Pol I replication fidelity in the presence of a lesion. We were also able to measure the impact of a lesion on the replication fidelity of surrounding bases, but our spectra were likely complicated by errors in oligonucleotide synthesis and DNA damage from template preparation. With improved template purification and preparation, this bypass mutagenicity could be accurately characterized. Given the large variety of DNA lesions and the polymerase specificity of their mutation signatures, our high-throughput assay would facilitate the characterization of their mutagenicity and shed light on their roles in disease.

The fidelity of DNA polymerase replication is a complex phenomenon and has been difficult to study in spite of the many techniques that have been developed to probe this process. Sequencing assays provide an important complementary approach to kinetic and structural studies, but current methods have significant technical limitations. With our high-throughput *in vitro* approach, DNA polymerase fidelity can be studied with greater ease and accuracy.

NOTE IN THE PROOF

It has been brought to our attention that a report on the use of a PacBio sequencer to map DNA polymerase errors was recently published (35).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors would like to thank Dr Graham Walker, Dr Igor Rogozin and Dr Patricia Purcell for many useful discussions and feedback on the manuscript.

FUNDING

National Institutes of Health (NIH) [TR01 5R01EB010244 to X.S.X., R01 GM114065 to J.J.L.]; Department of Energy [DE-SC0012411 to X.S.X.]; Medical Scientist Training Program from the National Institute of General Medical Sciences [T32GM007753 to J.L.]; Richard and Susan Smith Family Foundation [to D.F.L.]. Funding for open access charge: Department of Energy [DE-SC0012411 to X.S.X.].
Conflict of interest statement. None declared.

REFERENCES

1. Lange,S.S., Takata,K. and Wood,R.D. (2011) DNA polymerases and cancer. *Nat. Rev. Cancer*, **11**, 96–110.
2. Henninger,E.E. and Pursell,Z.F. (2014) DNA polymerase epsilon and its roles in genome stability. *IUBMB Life*, **66**, 339–351.

3. Svarovskaia,E.S., Cheslock,S.R., Zhang,W.H., Hu,W.S. and Pathak,V.K. (2003) Retroviral mutation rates and reverse transcriptase fidelity. *Front Biosci.*, **8**, d117–134.
4. Tompkins,J.D., Nelson,J.L., Hazel,J.C., Leugers,S.L., Stumpf,J.D. and Foster,P.L. (2003) Error-prone polymerase, DNA polymerase IV, is responsible for transient hypermutation during adaptive mutation in Escherichia coli. *J. Bacteriol.*, **185**, 3469–3472.
5. Kunkel,T.A. (2009) Evolving views of DNA replication (in)fidelity. *Cold Spring Harb. Symp. Quant. Biol.*, **74**, 91–101.
6. Zheng,W., Khrapko,K., Coller,H.A., Thilly,W.G. and Copeland,W.C. (2006) Origins of human mitochondrial point mutations as DNA polymerase gamma-mediated errors. *Mutat. Res.*, **599**, 11–20.
7. Bebenek,K. and Kunkel,T.A. (2000) Streisinger revisited: DNA synthesis errors mediated by substrate misalignments. *Cold Spring Harb. Symp. Quant. Biol*, **65**, 81–91.
8. Chang,D.J. and Cimprich,K.A. (2009) DNA damage tolerance: when it's OK to make mistakes. *Nat. Chem. Biol.*, **5**, 82–90.
9. Sale,J.E., Lehmann,A.R. and Woodgate,R. (2012) Y-family DNA polymerases and their role in tolerance of cellular DNA damage. *Nat. Rev. Mol. Cell Biol.*, **13**, 141–152.
10. Scharf,S.J., Horn,G.T. and Erlich,H.A. (1986) Direct cloning and sequence analysis of enzymatically amplified genomic sequences. *Science*, **233**, 1076–1078.
11. Bebenek,K. and Kunkel,T.A. (1995) Analyzing fidelity of DNA polymerases. *Methods Enzymol.*, **262**, 217–232.
12. Garibyan,L., Huang,T., Kim,M., Wolff,E., Nguyen,A., Nguyen,T., Diep,A., Hu,K., Iverson,A., Yang,H. *et al.* (2003) Use of the rpoB gene to determine the specificity of base substitution mutations on the Escherichia coli chromosome. *DNA Repair (Amst)*, **2**, 593–608.
13. Barnes,W.M. (1992) The fidelity of Taq polymerase catalyzing PCR is improved by an N-terminal deletion. *Gene*, **112**, 29–35.
14. Kunkel,T.A. (1985) The Mutational Specificity of DNA Polymerase-beta during in vitro DNA synthesis. Production of frameshift, base substitution, and deletion mutations. *J. Biol. Chem.*, **260**, 5787–5796.
15. Arana,M.E., Takata,K., Garcia-Diaz,M., Wood,R.D. and Kunkel,T.A. (2007) A unique error signature for human DNA polymerase nu. *DNA Repair (Amst)*, **6**, 213–223.
16. Delaney,J.C. and Essigmann,J.M. (2006) Assays for determining lesion bypass efficiency and mutagenicity of site-specific DNA lesions in vivo. *Methods Enzymol.*, **408**, 1–15.
17. Khrapko,K., Coller,H., Andre,P., Li,X.C., Foret,F., Belenky,A., Karger,B.L. and Thilly,W.G. (1997) Mutational spectrometry without phenotypic selection: human mitochondrial DNA. *Nucleic Acids Res.*, **25**, 685–693.
18. Kinde,I., Wu,J., Papadopoulos,N., Kinzler,K.W. and Vogelstein,B. (2011) Detection and quantification of rare mutations with massively parallel sequencing. *Proc. Natl. Acad. Sci. U.S.A.*, **108**, 9530–9535.
19. Schmitt,M.W., Kennedy,S.R., Salk,J.J., Fox,E.J., Hiatt,J.B. and Loeb,L.A. (2012) Detection of ultra-rare mutations by next-generation sequencing. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, 14508–14513.
20. Lou,D.I., Hussmann,J.A., McBee,R.M., Acevedo,A., Andino,R., Press,W.H. and Sawyer,S.L. (2013) High-throughput DNA sequencing errors are reduced by orders of magnitude using circle sequencing. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 19872–19877.
21. Gregory,M.T., Bertout,J.A., Ericson,N.G., Taylor,S.D., Mukherjee,R., Robins,H.S., Drescher,C.W. and Bielas,J.H. (2015) Targeted single molecule mutation detection with massively parallel sequencing. *Nucleic Acids Res.*, **44**, e22.
22. Guo,X., Lehner,K., O'Connell,K., Zhang,J., Dave,S.S. and Jinks-Robertson,S. (2015) SMRT sequencing for parallel analysis of multiple targets and accurate SNP phasing. *G3 (Bethesda)*, **5**, 2801–2808.
23. Keohavong,P. and Thilly,W.G. (1989) Fidelity of DNA polymerases in DNA amplification. *Proc. Natl. Acad. Sci. U.S.A.*, **86**, 9253–9257.
24. Zamft,B.M., Marblestone,A.H., Kording,K., Schmidt,D., Martin-Alarcon,D., Tyo,K., Boyden,E.S. and Church,G. (2012) Measuring cation dependent DNA polymerase fidelity landscapes by deep sequencing. *PLoS One*, **7**, e43876.
25. Beuning,P.J., Simon,S.M., Godoy,V.G., Jarosz,D.F. and Walker,G.C. (2006) Characterization of Escherichia coli translesion synthesis polymerases and their accessory factors. *Methods Enzymol.*, **408**, 318–340.
26. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
27. Lee,H., Popodi,E., Tang,H. and Foster,P.L. (2012) Rate and molecular spectrum of spontaneous mutations in the bacterium Escherichia coli as determined by whole-genome sequencing. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, E2774–2783.
28. Tindall,K.R. and Kunkel,T.A. (1988) Fidelity of DNA synthesis by the Thermus aquaticus DNA polymerase. *Biochemistry*, **27**, 6008–6013.
29. Bebenek,K., Joyce,C.M., Fitzgerald,M.P. and Kunkel,T.A. (1990) The fidelity of DNA synthesis catalyzed by derivatives of Escherichia coli DNA polymerase I. *J. Biol. Chem.*, **265**, 13878–13887.
30. Eckert,K.A. and Kunkel,T.A. (1990) High fidelity DNA synthesis by the Thermus aquaticus DNA polymerase. *Nucleic Acids Res.*, **18**, 3739–3744.
31. Kobayashi,S., Valentine,M.R., Pham,P., O'Donnell,M. and Goodman,M.F. (2002) Fidelity of Escherichia coli DNA polymerase IV. Preferential generation of small deletion mutations by dNTP-stabilized misalignment. *J. Biol. Chem.*, **277**, 34198–34207.
32. Saiki,R.K., Gelfand,D.H., Stoffel,S., Scharf,S.J., Higuchi,R., Horn,G.T., Mullis,K.B. and Erlich,H.A. (1988) Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science*, **239**, 487–491.
33. Bohning,D., Ekkehart,D. and Schlattmann,P. (1998) Recent developments in computer-assisted analysis of mixtures. *Biometrics*, **54**, 525–536.
34. Jarosz,D.F., Godoy,V.G., Delaney,J.C., Essigmann,J.M. and Walker,G.C. (2006) A single amino acid governs enhanced activity of DinB DNA polymerases on damaged templates. *Nature*, **439**, 225–228.
35. Hestand,M.S., Houdt,J.V., Cristofoli,F. and Vermeesch,J.R. (2016) Polymerase specific error rates and profiles identified by single molecule sequencing. *Mutat. Res.*, **784–785**, 39–45.