Mutation Research 784-785 (2016) 39–45

Contents lists available at ScienceDirect

Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis

journal homepage: www.elsevier.com/locate/molmut
Community address: www.elsevier.com/locate/mutres

Polymerase specific error rates and profiles identified by single molecule sequencing

Matthew S. Hestand, Jeroen Van Houdt, Francesca Cristofoli, Joris R. Vermeesch*

Department of Human Genetics, KU Leuven, O&N I Herestraat 49—box 602, 3000 Leuven, Belgium

ARTICLE INFO

Article history:
Received 23 July 2015
Received in revised form
16 December 2015
Accepted 14 January 2016
Available online 19 January 2016

Keywords:
Single molecule sequencing
Polymerase fidelity
Heteroduplex

ABSTRACT

DNA polymerases have an innate error rate which is polymerase and DNA context specific. Historically the mutational rate and profiles have been measured using a variety of methods, each with their own technical limitations. Here we used the unique properties of single molecule sequencing to evaluate the mutational rate and profiles of six DNA polymerases at the sequence level. In addition to accurately determining mutations in double strands, single molecule sequencing also captures direction specific transversions and transitions through the analysis of heteroduplexes. Not only did the error rates vary, but also the direction specific transitions differed among polymerases.

1. Introduction

Low grade mosaicism detection is increasingly important to unravel the causes of both constitutional and acquired human disorders. Mosaic mutations underlie an increasing number of human genetic diseases (reviewed in Ref. [6,11]). Cancers, for instance, arise out of mixtures of cells with various parallel or cumulative mutations that drive proliferation and metastatic potential. Therefore, detection of low grade mosaicism is becoming important in cancer characterization and monitoring its progression, response, and remission [34,8,30,5]. However, detection of low frequency mutations is hampered by the innate mutational errors introduced by DNA polymerases. These enzymes are at the heart of many core genomic technologies, including the polymerase chain reaction (PCR) and most massive parallel sequencing methods [13]. In cases where the tumor/normal-cell ratio is very low, to avoid expensive high-depth genome wide sequencing it will become essential to use polymerases with low error rates.

Several methods exist to measure polymerase error rates, but each has technology specific limitations. The M13mp2 forward mutation assay uses single-stranded M13mp2 DNA containing the α-complementation region of the *Escherichia coli* lacZ gene as a template for a single cycle of DNA synthesis. This construct is then transfected into an appropriate *E. coli* strain that shows dark blue spots when there is no mutation (i.e. synthesis error), but lighter blue or no plaques upon synthesis errors [22,10]. This feature of the assay requires additional sequencing steps to identify the precise error(s) at the sequence level and is limited to evaluating coding (i.e. reporter) DNA. The strategy of denaturing gradient gel electrophoresis (DGGE) denatures DNA at ever increasing concentrations of a chemical denaturant, before applying the DNA material to a gel where sequences containing a heteroduplex (i.e. an error) will migrate at a different speed compared to properly paired strands [12,19,10]. Again, this method requires additional sequencing steps to identify errors at the nucleotide sequence level. BEAMing (Beads, Emulsion, Amplification, and Magnetics) is a method used for quantifying rare variants in which a population of amplicons is amplified and converted to a population of beads [9,26]. These beads are then assessed by sequence specific probes, which are bound by fluorescently labeled antibodies. These are then counted fluorescently via flow cytometry to determine the exact nature of the nucleotide sequence [9,26]. This technique is limited to a small number of targets per experiment, though it has been suggested the BEAMing method creates an ideal template for high throughput sequencing to detect polymerase errors [26].

Sequencing based methods do already exist, including cloning of PCR products and traditional sequencing to evaluate mutations over multiple target sequences [28]. However, this suffers from small data size and therefore high-fidelity polymerases may not identify enough mutations to reliably call error rates [28]. For increased data size, high throughput sequencing using input amplicon molecules tagged with unique identifiers (UIDs) has been used to discriminate sequencing errors from errors present in amplicon sequences [20]. Taking this to the next level, Duplex Sequencing

* Corresponding author. Fax: +32 16 3 46060.
E-mail address: joris.vermeesch@uzleuven.be (J.R. Vermeesch).

http://dx.doi.org/10.1016/j.mrfmmm.2016.01.003
0027-5107/

and CypherSeq utilize UIDs on both DNA strands to further reduce introduced errors [33,18,14]. These high-throughput sequencing approaches are powerful, though each still require a critical PCR based step which is subject to the method's polymerase fidelity [20,33,18,14].

The above methods have been used to estimate error rates across different DNA polymerases. Several methods exist for reporting error rates, but we have used observed nucleotide errors per total nucleotides sequenced per PCR cycle. See Supplemental Table S1 for conversion of reference error rates (when needed). Initial error rates for Taq and (modified) T7 polymerases were 2.0×10^{-4} and 5.4×10^{-5}, respectively, but these numbers could be improved by optimizing PCR conditions (e.g. pH, dNTP concentration, and magnesium ion concentration) to 7.2×10^{-5} and 4.4×10^{-5}, respectively [27]. In addition, it has been estimated that sequence context and conditions can create up to a 10 fold differences in error rates (9.2×10^{-5} to less than 6.2×10^{-6} when using a Taq polymerase) [10]. Hence, accurately determining polymerase error rates remains challenging. In addition to different error rates, polymerases generate different error profiles [19]. For example, T4 and modified T7 polymerases show primarily transitions of G·C > A·T, while Taq polymerase preferential shows A·T > G·C [19].

We hypothesized that single molecule sequencing would enable the determination of polymerase error rates and profiles directly. This assumption may at first be counter-intuitive, since single molecule sequencing is error prone with accuracies of only about 85% [7]. However, a single double stranded DNA molecule is circularized and both strands are sequenced multiple times (each sequence a subread) to form a long linear read (Fig. 1). Considering errors are randomly distributed across reads, the consensus of the subreads (termed the "read-of-insert") is increasingly accurate with an increasing number of passes [7,17]. Hence, with multiple passes on the same molecule, sequencing errors are eliminated and all variants are molecule specific. In addition, we hypothesized that PacBio sequencing could provide the unique capability of determining when a base in the 5'–3' strand is not complementary to the base in the other strand, termed a heteroduplex. In a double-stranded heteroduplex molecule, the subreads from one direction of a read-of-insert should match the read-of-insert sequence, and the subreads from the other direction would identify the sequence mismatch (Fig. 1). Here we demonstrate that the unique features of PacBio circular sequencing allow accurate detection and characterization of mutations introduced by six commonly used polymerases during PCR.

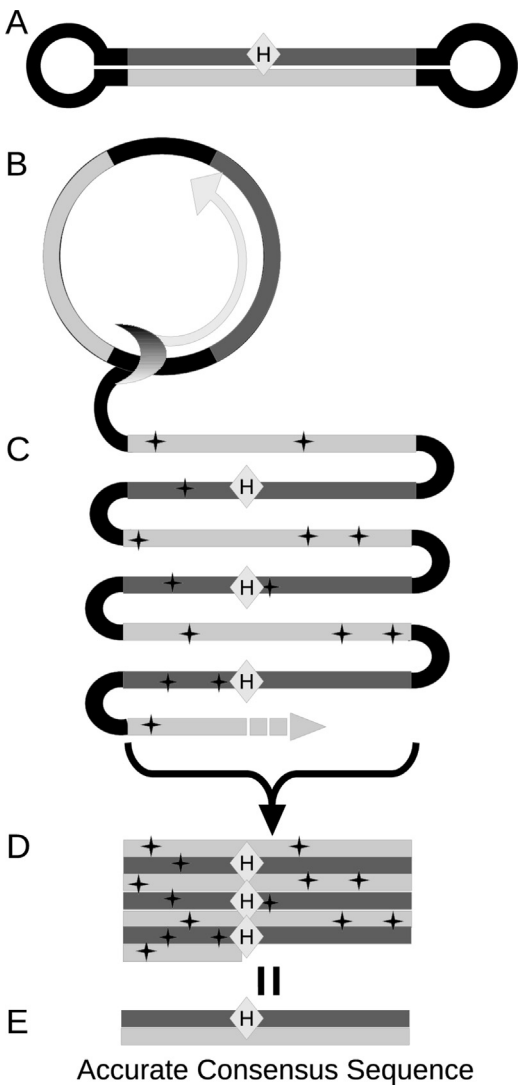

Fig. 1. During PacBio sequencing a construct made up of a double stranded DNA molecule with ligated adaptors (black loops) (A) is circularized (B), and a polymerase (moon shape) repeatedly sequences the first strand, an adaptor, the second strand, an adaptor, and repeats generating many subreads (C). Though the subreads have a high error rate, errors (indicated by stars) are random. The error prone subreads can be assembled (D) and since errors are random, a high quality consensus sequence can be generated (E). In addition, if a variant is found only on one strand (i.e. a heteroduplex, as indicated by H diamonds), it should be found back only in the subreads matching that strand.

2. Results

2.1. Polymerase mutation rates and profiles

To test the assumption that single molecule sequencing of amplicons would permit the determination of mutational profiles, we sequenced a single PCR fragment generated with Platinum Taq polymerase using a single PacBio SMRTcell. This SMRTcell provided 57,556 read-of-inserts with a minimum of two passes. On average, each read-of-insert was a consensus of eight passes (Supplemental Fig. S1A).

When plotting the errors per base per cycle as a function of the number of passes of the same molecule, error rates become asymptotic (Fig. 2A). Hence, when a consensus read is made up of ten or more read passes the variants are not due to mutational errors of the PacBio polymerase, but rather due to variation present in the input molecule. Using ten or more passes, we identified an error rate per base per cycle of 3.28×10^{-5}. Watson-Crick base pair errors were 4.44×10^{-5} for A·T and 1.29×10^{-5} for G·C. This minimum number of ten passes provided ~9.6 k times coverage of the amplicon. At this depth, transitions were more dominant than transversions (Fig. 2B). Similar to previous Taq polymerase findings [19], A·T > G·C transitions were more common than G·C > A·T transitions.

Encouraged by this result, six different polymerases were selected that are commonly used (Table 1). To evaluate the mutational profile of those polymerases two amplicons were generated by each polymerase in duplicate. The size of the PCR products was verified by agarose gel electrophoresis (Supplemental Fig. S2), amplicons pooled, PacBio libraries generated per replicate pool, and sequencing performed on a PacBio single molecule sequencing instrument. The newer PacBio instrument, chemistries, and longer movie time provided significantly more subreads per linear read (Supplemental Fig. S1B), with 18 mean passes when analyzing read-of-inserts with a minimum of two passes. Viewing initial alignments in IGV [32] (data not shown) revealed no SNPs within these amplicons; therefore all reads with a sequence variant should be due to a mutation introduced by the polymerase.