Polymerase Specific Error Rates and Profiles Identified by Single Molecule Sequencing

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

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

*Tel: +32 16 3 45941; Fax: +32 16 3 46060; Email: joris.vermeesch@uzleuven.be

This is a preprint: for the final published article please see: http://dx.doi.org/10.1016/j.mrfmmm.2016.01.003

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.

KEYWORDS

single molecule sequencing, polymerase fidelity, heteroduplex

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 Biesecker and Spinner 2013; Erickson 2014). 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 (Shah et al. 2010; Ding et al. 2012; Nik-Zainal et al, 2012; Bedard et al. 2013). 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 (Fuller et al. 2009). 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 E.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 (Kunkel 1985; Eckert and Kunkel 1991). 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 (Fischer and Lerman 1983; Keohavong and Thilly 1989; Eckert and Kunkel 1991). 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 (Dressman et al. 2003; Li et al. 2006). 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 (Dressman et al. 2003; Li et al. 2006). 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 (Li et al. 2006).

Sequencing based methods do already exist, including cloning of PCR products and traditional sequencing to evaluate mutations over multiple target sequences (McInerney et al. 2014). However, this suffers from small data size and therefore high-fidelity polymerases may not identify enough mutations to reliably call error rates (McInerney et al. 2014). 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 (Kinde et al. 2011). Taking this to the next level, Duplex Sequencing and CypherSeq utilize UIDs on both DNA strands to further reduce introduced errors (Schmitt et al. 2012; Kennedy et al. 2014, Gregory et al. 2015). 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 (Kinde et al. 2011; Schmitt et al. 2012; Kennedy et al. 2014, Gregory et al. 2015).

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 $2.0x10^{-4}$ and $5.4x10^{-5}$, respectively, but these numbers could be improved by optimizing PCR conditions (e.g. pH, dNTP concentration, and magnesium ion concentration) to $7.2x10^{-5}$ and $4.4x10^{-5}$, respectively (Ling et al. 1991). In addition, it has been estimated that sequence context and conditions can create up to a 10 fold differences in error rates ($9.2x10^{-5}$ to less than $6.2x10^{-6}$ when using a Taq polymerase) (Eckert and Kunkel 1991). Hence, accurately determining polymerase error rates remains challenging. In addition to different error rates, polymerases generate different error profiles (Keohavong and Thilly 1989). For example, T4 and modified T7 polymerases show primarily transitions of G•C>A•T, while Taq polymerase preferential shows A•T>G•C (Keohavong and Thilly 1989).

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% (Carneiro et al. 2012). 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 (Carneiro et al. 2012, Jiao et al. 2013). 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' to 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.

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 $3.28x10^{-5}$. Watson-Crick base pair errors were $4.44x10^{-5}$ for A•T and $1.29x10^{-5}$ for G•C. This minimum number of ten passes provided ~9.6k times coverage of the amplicon. At this depth, transitions were more dominant than transversions (Fig. 2B). Similar to previous Taq polymerase findings (Keohavong and Thilly 1989), A•T>G•C

transitions were more common than G•C>A•T transitions.

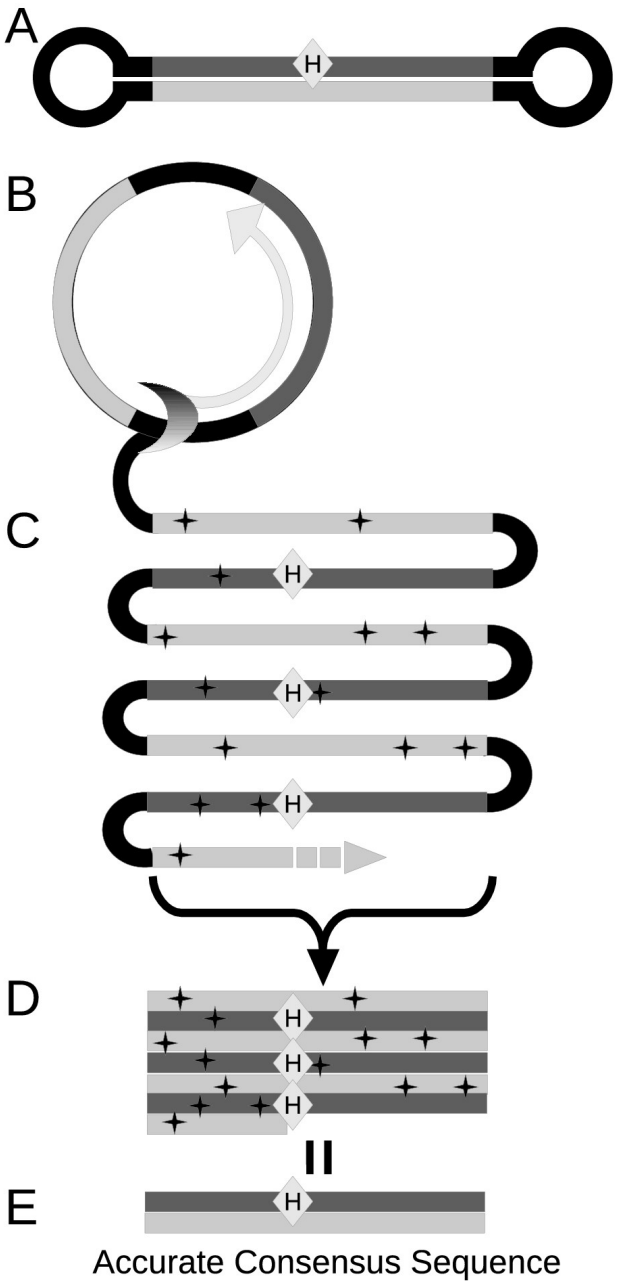

Figure 1: During PacBio sequencing a construct made up of a double stranded DNA molecule with ligated adaptors (black loops) (A) is circularized (B), and a polymerase (moon shape) repeatedly sequences the first strand, an adaptor, the second strand, an adaptor, and repeats generating many subreads (C). 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.

This is a preprint: for the final published article please see: http://dx.doi.org/10.1016/j.mrfmmm.2016.01.003

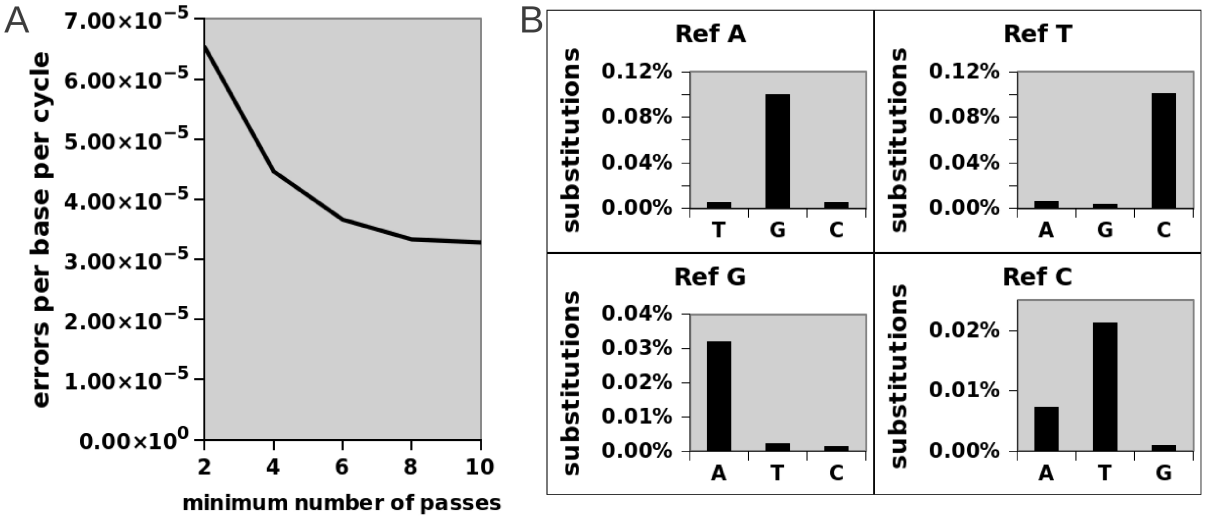

Figure 2: (A) Across the PacBio deep sequenced CASK amplicon is the errors per base per cycle when using different cutoffs for the number of passes to call a PacBio read-of-insert. (B) The percent of non-reference calls for each reference (Ref) nucleotide when using a minimum of 10 passes per PacBio read-of-insert.

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 (Robinson et al. 2011) (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.

The observed error rate per base per cycle as a function of the number of read passes is plotted for all six polymerases and, as in the pilot experiment, flattens at ten passes (Fig. 3A). Hence, to measure the mutational profile of the polymerases used in the PCR reaction we proceeded with minimal ten pass read-of-inserts. The number of ten pass molecules assessed is indicated in Table 2. Platinum Taq and TaKaRa Taq showed higher error rates per base per cycle compared to the other polymerases (Table 2). To assess reproducibility, error rates were calculated per polymerase for each SMRTcell (Supplemental Table S2). Single factor Anovas (alpha=0.05) showed no statistical difference between SMRTcells (p-value 0.99996), but a statistical difference between polymerases (p-value 4.46109×10^{-39}). To evaluate the influence of PCR-pool/library replicates, the mean SMRTcell error per polymerase was calculated. A single factor Anova (alpha=0.05) showed no statistical difference between PCR-pools/libraries (p-value 0.97670).

When separating the error analysis by Watson-Crick base pairs, Platinum Taq and TaKaRa Taq showed more errors on A•T than G•C base-pairs. The other polymerases showed the opposite trend, with less errors on A•T than G•C base-pairs (Table 2). Looking at the percent of non-reference bases sequenced per polymerase revealed these to be primarily transitions (Fig. 3B and Supplemental Fig. S3). Hence Platinum Taq and TaKaRa Taq preferentially introduce errors of A•T> G•C, while the other polymerases preferentially introduce errors of G•C>A•T .

This is a preprint: for the final published article please see: http://dx.doi.org/10.1016/j.mrfmmm.2016.01.003

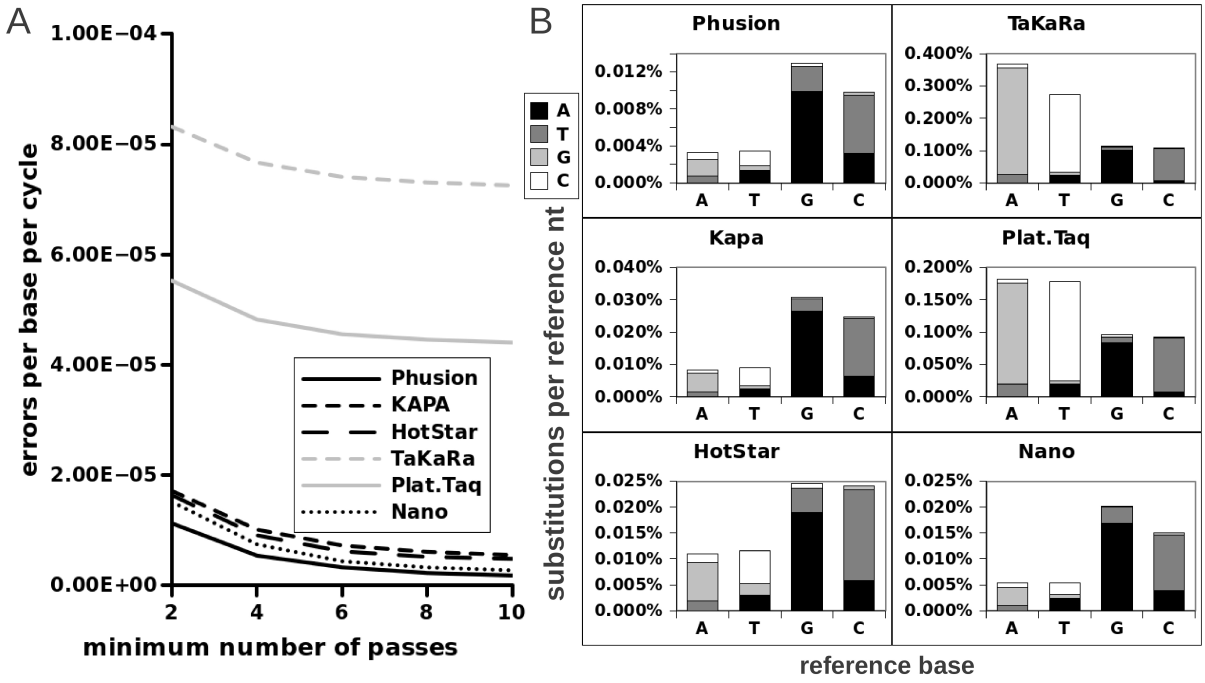

Figure 3: (A) Indicated is the errors per base per cycle when using different cutoffs for the number of passes to call a PacBio read-of-insert. (B) Across the PacBio sequenced CASK amplicons is indicated the percent of non-reference calls for each reference nucleotide per polymerase.

2.2. Heteroduplex Evaluation by Polymerase

A heteroduplex is the result of either: 1) a polymerase incorporating the wrong nucleotide during extension or 2) two strands annealing with a mismatched base. Since we found few multiple identical heteroduplex positions (<30%) we believe assumption 1 to be the most common in our datasets. This conclusion is also supported by estimated errors per base-pair sequenced (Table 3) in the range with the non-directional errors per base per cycle (10^{-5} to 10^{-6} error range, Table 2). This finding indicates that heteroduplexes are from a single round of extension and not an accumulation of errors. Hence, if we assume the wild-type strand of a heteroduplex molecule was the original and the mutant strand was altered during extension we can pinpoint which nucleotide is in error.

Only read-of-inserts containing primer pairs (68-77% of reads retained) and aligning to the amplicon targets (loss of 0.2-4.1%) were analyzed (Table 3). This eliminates the potential of detecting a heteroduplex initiated by a primer synthesized with a mismatch. Since heteroduplex identification requires reliably identifying each strand's consensus base separately, while giving some margin for indels due to reference bias when aligning the mismatch strand to the consensus molecule, we performed a first filter for read-of-inserts with more than 30 subreads. 28-31% of the aligned and primer containing reads were retained by applying this filter. The percent of reads containing a heteroduplex (Fig. 4 as an example) was higher in TaKaRa Taq and Platinum Taq relative to the other polymerases (Table 3). This was a difference of approximately ten fold, similar to the nucleotide error analysis fold difference (ten pass minimum).

This is a preprint: for the final published article please see: http://dx.doi.org/10.1016/j.mrfmmm.2016.01.003

Using the heteroduplex analysis with the MD and CIGAR strings, strand specific mutations were identified (Table 4). TaKaRa Taq and Platinum Taq showed direction specific A>G and T>C transitions to be the most prominent, supporting the observed non-directional nucleotide error profiles (A•T>G•C). Phusion, KAPA, and Nano heteroduplexes also supported the observed non-directional nucleotide error profiles with prominent direction specific G>A and C>T transitions. HotStar did not produce enough heteroduplexes to confirm or refute the nucleotide error analysis.

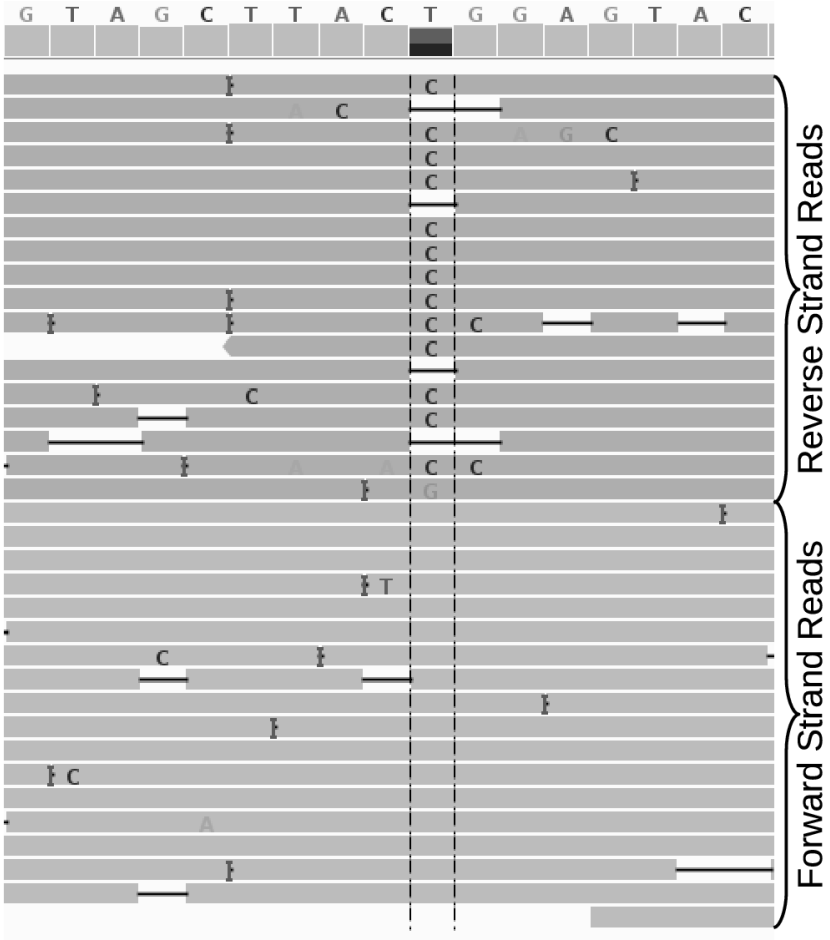

Figure 4: Example of a Heteroduplex. This shows all subreads aligning to the corresponding read-of-insert identified in a Phusion FAM120C amplicon. Reads are sorted by which strand they align to.

3. DISCUSSION

Here we have used the advantages of PacBio single molecule sequencing to identify errors per base per cycle and directly observed at the sequence level on which strand the errors are generated. Two general classes of polymerases, low-fidelity and high fidelity (McInerney et al. 2014), demonstrate approximately ten fold difference in error rates (10^{-5} and 10^{-6} error ranges, respectively) and can also be discriminated by error profiles (primarily A•T>G•C transitions and more G•C>A•T transitions, respectively) (Tables 2 and 4). The unique capability of sequencing heteroduplexes enabled the direct observation that Watson-Crick base pairing errors are not equally distributed, but that there is a bias for C>T over G>A across most polymerases and T>C over A>G for KAPA and Platinum Taq polymerases (Table 4). Hence, within traditional DNA base-pairing, transitions more often originate on the pyrimidine nucleotide than on the purine nucleotide.

It is not evident why pyrimidines are favored in the same direction transition over purines. The observed T>C & G>A and C>T & A>G result in heteroduplexes composed of C•A and T•G. Thermodynamics studies show T•G pairings to be much more stable than C•A parings (Aboul-ela et al. 1985). This would support the observed C>T over G>A frequencies. However, this would not explain the T>C over A>G observations for KAPA and Platinum Taq polymerases. It has been suggested that substitution mutagenesis occurs through incorporation of minor tautomeric forms of bases (Watson and Crick 1953; Harris et al. 2003). Heteroduplexes composed of C•A or T•G preserve the geometry of a correct Watson-Crick base pair when using the enol form of G/T and the imino form of A/C (Watson and Crick 1953; Bebenek et al. 2011). The unequal strand-error distribution could be due to either polymerase bias towards incorporating different minor tautomeric base forms or differing concentrations of minor tautomeric base forms in the PCR reaction. An alternative theory to the polymerases introducing specific errors could be specific types of DNA damage occurring in specific PCRs.

Though the template DNA, PCR conditions, and sequencing conditions differed, the error rates using Platinum Taq in the single amplicon and pooled experiments showed similar trends (Fig. 2A and 3A). In particular, final errors per base per cycle were in a similar range at $3.28x10^{-5}$ for the initial CASK amplicon and $4.41x10^{-5}$ for the pooled amplicons, when using a minimum of ten passes. This value is also close to that previously reported for Platinum Taq ($2.3x10^{-5}$) (Li et al. 2006) and within the range of many other reported Taq error ranges ($9.2x10^{-5}$ to less than $6.2x10^{-6}$) (Eckert and Kunkel 1991; Ling et al. 1991; McInerney et al. 2014). Reproducibility is further supported in the pooled experiment, where both CASK and FAM120C amplicons gave similar error profiles (Fig. 3B and Supplemental Fig. S3). The observation of A•T>G•C transitions being the most prominent (0.10% A>G, 0.10% T>C, 0.03% G>A, and 0.02% C>T for the single CASK amplicon and 0.16% A>G, 0.15% T>C, 0.08% G>A, and 0.08% C>T for the pooled CASK amplicon) also supports previous Taq polymerase findings (Keohavong and Thilly 1989; McInerney et al. 2014).

The Phusion error rate ($1.81x10^{-6}$) was worse than previously reported ($4.2x10^{-7}$) (Li et al. 2006), but almost exactly that reported using Phusion with an alternate buffer ($1.81x10^{-6}$) (McInerney et al. 2014). With the alternate buffer, equal levels of G•C>A•T and A•T>G•C transitions were reported (McInerney et al. 2014), whereas we identify 2.7 times more G•C>A•T than A•T>G•C transitions (Fig. 3B). Considering our considerably higher counts (495 G•C>A•T and 186 A•T>G•C, compared to 4 G•C>A•T and 6 A•T>G•C in McInerney et al. 2014), we believe our sampling to more realistically estimate the error profile. This demonstrates the advantage of using high-throughput sequencing over traditional sequencing, even with PacBio's limited throughput. Compared to Illumina's, PacBio's throughput does limit the number of targets that can be evaluated at a time. However, compared to all Illumina based approaches reviewed in this manuscript (Kinde et al. 2011, Schmitt et al. 2012; Kennedy et al. 2014, Gregory et al. 2015), this PacBio approach is the only library preparation that is PCR free.

For future applications to analyze pairs of mutations (i.e. phasing of SNPs directly from sequence (Mensah et al. 2014, Guo et al. 2015)), PacBio also permits the generation of longer inserts to evaluate more distant variant pairs. With an average linear read length of 15kb, a minimum of 10 passes still approaches utilizing inserts of ~1.5kb. This is almost double what a normal Illumina library insert can be to achieve good read quality (Kircher et al. 2011). For additional future applications, we also propose that input PCR molecules could be tagged with UIDs. This would then provide an additional level of technical error removal.

As low grade mosaicism detection increasingly becomes important in studying and treating human genetic disorders, it becomes ever important to understand the error rates and profiles of utilized technologies. It is also known that cancer genomes can be characterized by specific mutational profiles (Alexandrov et al. 2013), such as chronic lymphocytic leukemias with or without mutated immunoglobin genes and relapsed versus primary acute myeloid leukemias showing different substitution profiles (Puente et al. 2011; Ding et al. 2012). Higher proportions of C>T transitions are also found in early than late developing breast cancers (Nik-Zainal et al. 2012). Such differences could be important in making treatment decisions in the personal genomics era. Confidence in low-depth variant calling can be improved knowing the characterization of these polymerase

enzymes and including them in error modeling to improve bioinformatic resources. Even at high coverage PCR validation, the PCR will introduce errors that can be accounted for if we better know the technological properties. In addition, this single molecule sequencing approach could be an important tool for scientists studying the enzymology of polymerases.

4. METHODS

To evaluate polymerase errors during PCR we followed a general method, as described in detail in the following subsections. This consists of the following steps: 1) perform PCR to create an amplicon, 2) create an amplification-free PacBio library from the amplicon(s), 3) PacBio sequence the library, and 4) analyze the library for errors generated during PCR.

4.1. Amplicon Generation and Sequencing

The preliminary experiment was a single 488bp amplicon targeting the CASK gene from a healthy male family member, approved for research purposes under an institutional review board protocol nr. S-52853 (see Supplemental Methods). The second experiment, performed as full replicates, were amplicon pools using primers targeting two genic positions, barcoded per polymerase type. Primer sequences were the previous CASK primers and previously published primers targeting exon 10 of the FAM120C gene (De Wolf et al. 2014) (resulting in a 410bp amplicon), including the 5' addition of PacBio specific barcodes 1-6 (of PacBio's initial 48 barcode design) with padding sequence. For the PCRs, six commonly used polymerases used for a variety of methods were selected (Table 1). Input DNA was from a healthy male, though not the same individual as used for the initial experiment. We aimed to stay as close as possible to the manufacturers' guidelines for PCR cocktails and thermocycler profiles (see Supplemental Methods).

The individual CASK amplicon and pools were purified on Qiagen QIAquick PCR Purification Kit columns and fragmentation checked on a DNA 12000 chip analyzed on a Bioanalyzer 2100 (Agilent). The single CASK amplicon was prepared for sequencing according to Pacific Biosciences Standard Seq v2 protocol using PacBio's DNA Template Prep Kit 2.0 (250bp-3kb). This was sequenced on a PacBio RS using PacBio's DNA/Polymerase Binding Kit 2.0 on a single SMRT cell for 2x55 minute movies. The two pools were prepared for sequencing according to Pacific Biosciences Standard Seq v3 protocol using PacBio's DNA Template Prep Kit 2.0 (3-10kb). The two libraries of pools were sequenced on a PacBio RSII using a DNA/Polymerase Binding Kit P4, each on four SMRT cells for 180 minute movies. All runs used PacBio DNA Sequencing Kit 2.0 sequencing reagents.

4.2. Error Rate Analysis

The individual CASK SMRTcell was run through PacBio's SMRT Portal (v2.2.0) pipeline RS_ReadsOfInsert.1 for minimum full passes of 2, 4, 6, 8, or 10. For the pools, all SMRTcells were run together through the same pipeline (v2.3.0) for minimum full passes of 2, 4, 6, 8, or 10, but also included the barcode setting for “Different on each end (paired)” and minimum barcode score 30. Fastq files (per barcode in the pools) were aligned to the 1000 Genome reference file (Abecasis et al. 2012) using BWA-SW (Li and Durbin 2010) v0.7.5a. Samtools (Li et al. 2009) v0.1.18 was used to convert SAM to BAM (view), sort, index, and mpileup (-A -d “60000 for CASK, 40000 for pool” -l “bed files of amplicon coordinates, excluding primer positions”). Custom scripts were then used to count the number of reference and substitution calls in the mpileup file. The errors per base per cycle were then defined as the number of substitution calls divided by the total number of A, T , G , and C bases sequenced, divided by the number of PCR cycles.

4.3. Heteroduplex Analysis

To identify heteroduplexes, the fastq file for minimum 10 full passes was filtered for sequences containing a forward and reverse primer pair. Only sequence and corresponding quality scores between the primer pairs

(excluding barcodes) were retained. Read-of-inserts were aligned with BWA-SW to the reference genome and filtered (Samtools view) for only reads aligning to the primer locations. In the error rate analysis 10 full passes was the optimal minimum coverage needed, but for heteroduplex analysis each strand must be evaluated separately. Therefore, we aimed to have more than ten reads per strand instead of per molecule. In addition, to give some margin for indels due to reference bias when aligning the mismatch strand to the consensus molecule, we performed a first filter for read-of-inserts with more than 30 subreads. This was done with a custom perl script which also made a reference of each read-of-insert and aligned the subreads to the read-of-insert with BWA-SW using the settings of Carneiro et al. 2012 to account for the high error rate per subread. The script then generated an mpileup (Samtools: view to convert SAM to BAM>sort>index>mpileup) and identified heteroduplex variant positions defined as:

- To achieve overall sequence coverage similar to the circular consensus sequence (consensus sequence is now per strand instead of per molecule), non-indel nucleotides had to come from over 20 subreads.
- Similar to calling a heterozygous SNP, excluding indel positions, the subread nucleotides not matching the read-of-insert had to account for between 25% and 75% of the non-indel subreads.
- To discriminate that the heteroduplex variant occurs only on one strand and not the other, >90% of the subread nucleotides supporting the read-of-insert had to come from one strand and >90% of the subread nucleotides not matching the read-of-insert had to come from the other strand.

To evaluate read-of-inserts for which strand the error was on, the heteroduplex read-of-inserts were aligned with BWA-SW and MD fields generated (Samtools: view to convert SAM to BAM>sort>index>view to select amplicon alignment locations>view to convert SAM to BAM>calmd).

ACCESSION NUMBERS

The raw data is available in the European Nucleotide Archive (ENA) (Leinonen et al. 2011) under study accession numbers PRJEB12157 and PRJEB12175.

CONFLICT OF INTEREST

The authors declare that there are no conflicts of interest.

FUNDING

This work was supported by research grants from the KU Leuven (SymBioSys [PFV/10/016], GOA/12/015 to J.R.V), the Hercules foundation (ZW11-14) and from the Belgian Science Policy Office Interuniversity Attraction Poles (BELSPO-IAP) programme through the project IAP P7/43-BeMGI. M.S.H. is a postdoctoral fellow at KU Leuven supported by F+/12/037. F.C. is PhD aspirant of the Research Foundation - Flanders (FWO).

ACKNOWLEDGMENTS

We wish to thank Greet Peeters for preparing amplicons, Matthias Declercq and Wim Meert for PacBio library preparations and runs, Veerle De Wolf for initial FAM120C primer design, and Carlo Sala Frigerio for providing materials. We also wish to thank Carlo Sala Frigerio and Thierry Voet for providing useful discussions and comments. In addition, we thank David Wilson (National Institute on Aging) for critical reading and editing of the manuscript.

This is a preprint: for the final published article please see: http://dx.doi.org/10.1016/j.mrfmmm.2016.01.003

TABLES

Table 1. Polymerases of This Study

Polymerase	Provider	Example Usages
Phusion High Fidelity PCR Master Mix w/HF Buffer	New England BioLabs	exome amplifications
KAPA HiFi HotStart ReadyMix	KAPA Biosystems	Illumina library preps
HotStar HiFidelity Polymerase Kit	QIAGEN	targeted amplicon generation for sequencing
TaKaRa LA Taq DNA Polymerase	TaKaRa/Clontech	long range PCRs
Platinum Taq DNA Polymerase	Invitrogen	general usage
TruSeq Nano DNA Sample Prep Kit	Illumina	Non-Invasive Prenatal Testing (NIPT)

These polymerases are commonly used, as demonstrated by examples of usage.

Table 2. Polymerase Error Rates

Polymerase	Number Molecules	Overall Error Rate	A•T Error Rate	G•C Error Rate
Phusion	51,732	1.81×10^{-6}	1.05×10^{-6}	2.83×10^{-6}
KAPA	48,450	5.52×10^{-6}	2.99×10^{-6}	8.98×10^{-6}
HotStar	37,706	4.85×10^{-6}	3.78×10^{-6}	6.20×10^{-6}
TaKaRa	47,110	7.25×10^{-5}	1.03×10^{-4}	3.11×10^{-5}
Plat.Taq	40,731	4.41×10^{-5}	5.90×10^{-5}	2.42×10^{-5}
Nano	50,161	2.76×10^{-6}	1.89×10^{-6}	4.50×10^{-6}

For the pooled amplicons is indicated the number of molecules with a minimum of ten sequencing passes. Error rates are errors per base per cycle.

This is a preprint: for the final published article please see: http://dx.doi.org/10.1016/j.mrfmmm.2016.01.003

Table 3. Heteroduplex Identification

	Phusion	KAPA	HotStar	TaKaRa	Plat.Taq	Nano
# Reads	51,732	48,450	37,706	47,110	40,731	50,161
w/primers	40,058	35,417	25,797	31,958	28,415	38,037
aligned	39,462	35,330	24,858	31,513	28,233	36,475
>30subreads	11,453	9,938	7,666	9,075	8,253	10,420
# w/Het.D	28	30	8	188	198	36
% w/Het.D	0.2%	0.3%	0.1%	2.1%	2.4%	0.3%
Errors/bp	$5.44x10^{-6}$	$6.72x10^{-6}$	$2.32x10^{-6}$	$4.61x10^{-5}$	$5.10x10^{-5}$	$7.05x10^{-6}$

Indicated is the initial number of read-of-inserts, the number after filtering for primer sequences, followed by the number after filtering for alignment, and finally those containing over 30 subreads and length >250bp. These were analyzed for the number in which a heteroduplexes (Het.D) was identified. Errors per bp are estimated as the number of heteroduplex positions (Table 4) divided by the product of the number of molecules with >30 subreads times the average length of the two amplicons (449bp). Note, some reads may contain more than one mismatch position.

Table 4. Heteroduplex Nucleotide Content

	Phusion	KAPA	HotStar	TaKaRa	Plat.Taq	Nano
T>C	3	3	-	63	62	1
A>G	3	-	1	63	56	2
C>T	11	15	2	22	26	14
G>A	8	8	2	16	22	13
A>C	-	-	-	3	1	-
A>T	1	2	-	9	13	-
C>A	1	1	1	1	2	1
C>G	-	-	-	-	1	-
G>C	-	-	-	1	3	-
G>T	-	1	1	2	1	-
T>A	1	-	1	8	1	1
T>G	-	-	-	-	1	1

REFERENCES

Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA, Altshuler DM et al. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491(7422):** 56–65.

Aboul-ela F, Koh D, Tinoco I, and Martin FH. 1985. Base-base mismatches. Thermodynamics of double helix formation for dCA3XA3G + dCT3YT3G (X, Y = A,C,G,T). *Nucleic Acids Res.* **13(13):** 4811–4824.

Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, Bignell GR, Bolli N, Borg A, Børresen-Dale AL et al. 2013. Signatures of mutational processes in human cancer. *Nature* **500(7463):** 415–421.

Bebenek K, Pedersen LC, and Kunkel TA. 2011. Replication infidelity via a mismatch with Watson-Crick geometry. *Proc. Natl. Acad. Sci. U.S.A.* **108(5):** 1862–1867.

Bedard PL, Hansen AR, Ratain MJ, and Siu LL. 2013. Tumour heterogeneity in the clinic. *Nature* **501(7467):** 355–364.

Biesecker LG and Spinner NB. 2013. A genomic view of mosaicism and human disease. *Nat. Rev. Genet.* **14(5):** 307–320.

Carneiro MO, Russ C, Ross MG, Gabriel SB, Nusbaum C, and DePristo MA. 2012. Pacific biosciences sequencing technology for genotyping and variation discovery in human data. *BMC Genomics* **13:** 375.

Ding L, Ley TJ, Larson DE, Miller CA, Koboldt DC, Welch JS, Ritchey JK, Young MA, Lamprecht T, McLellan MD et al. 2012. Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature* **481(7382):** 506–510.

Dressman D, Yan H, Traverso G, Kinzler KW, and Vogelstein B. 2003. Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proc. Natl. Acad. Sci. U.S.A.* **100(15):** 8817–8822.

Eckert KA and Kunkel TA. 1991. DNA polymerase fidelity and the polymerase chain reaction. *PCR Methods Appl.* **1(1):** 17–24.

Erickson RP. 2014. Recent advances in the study of somatic mosaicism and diseases other than cancer. *Curr. Opin. Genet. Dev.* **26:**73–78.

Fischer SG and Lerman LS. 1983. DNA fragments differing by single base-pair substitutions are separated in denaturing gradient gels: correspondence with melting theory. *Proc. Natl. Acad. Sci. U.S.A.* **80(6):** 1579–1583.

Fuller CW, Middendorf LR, Benner SA, Church GM, Harris T, Huang X, Jovanovich SB, Nelson JR, Schloss JA, Schwartz DC, and Vezenov DV. 2009. The challenges of sequencing by synthesis. *Nat. Biotechnol.* **27(11):** 1013–1023.

Gregory MT, Bertout JA, Ericson NG, Taylor SD, Mukherjee R, Robins HS, Drescher CW, and Bielas JH. 2015. Targeted single molecule mutation detection with massively parallel sequencing. *Nucleic Acids Res.* Gkv915.

Guo X, Lehner K, O'Connell K, Zhang J, Dave SS, and Jinks-Robertson S. 2015. SMRT Sequencing for Parallel Analysis of Multiple Targets and Accurate SNP Phasing. *G3 (Bethesda).* pii: g3.115.023317.

Harris VH, Smith CL, Jonathan Cummins W, Hamilton AL, Adams H, Dickman M, Hornby DP, and Williams DM. 2003. The effect of tautomeric constant on the specificity of nucleotide incorporation during DNA replication: support for the rare tautomer hypothesis of substitution mutagenesis. *J. Mol. Biol.* **326(5):** 1389–1401.

Jiao X, Zheng X, Ma L, Kutty G, Gogineni E, Sun Q, Sherman BT, Hu X, Jones K, Raley C et al. 2013. A Benchmark Study on Error Assessment and Quality Control of CCS Reads Derived from the PacBio RS. *J Data*

Mining Genomics Proteomics **4(3):** pii: 16008.

Kennedy SR, Schmitt MW, Fox EJ, Kohrn BF, Salk JJ, Ahn EH, Prindle MJ, Kuong KJ, Shen JC, Risques RA, and Loeb LA. 2014. Detecting ultralow-frequency mutations by Duplex Sequencing. *Nat Protoc.* **9(11):** 2586–2606.

Keohavong P and Thilly WG. 1989. Fidelity of DNA polymerases in DNA amplification. *Proc. Natl. Acad. Sci. U.S.A.* **86(23):** 9253–9257.

Kinde I, Wu J, Papadopoulos N, Kinzler KW, and Vogelstein B. 2011. Detection and quantification of rare mutations with massively parallel sequencing. *Proc. Natl. Acad. Sci. U.S.A.* **108(23):** 9530–9535.

Kircher M, Heyn P, and Kelso J. 2011. Addressing challenges in the production and analysis of illumina sequencing data. *BMC Genomics*. **12:**382.

Kunkel TA. 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(9):** 5787–5796.

Leinonen R, Akhtar R, Birney E, Bower L, Cerdeno-Tarraga A, Cheng Y, Cleland I, Faruque N, Goodgame N, Gibson R et al. 2011. The European Nucleotide Archive. *Nucleic Acids Res.* **39(Database issue):** 28–31.

Li, H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, and Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25(16):** 2078–2079.

Li H and Durbin R. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26(5):** 589–595.

Li M, Diehl F, Dressman D, Vogelstein B, and Kinzler KW. 2006. BEAMing up for detection and quantification of rare sequence variants. *Nat. Methods* **3(2):** 95–97.

Ling LL, Keohavong P, Dias C, and Thilly WG. 1991. Optimization of the polymerase chain reaction with regard to fidelity: modified T7, Taq, and vent DNA polymerases. *PCR Methods Appl.* **1(1):** 63–69.

McInerney P, Adams P, and Hadi MZ. 2014. Error Rate Comparison during Polymerase Chain Reaction by DNA Polymerase. *Mol Biol Int.* **2014:** 287430.

Mensah MA, Hestand MS, Larmuseau MH, Isrie M, Vanderheyden N, Declercq M, Souche EL, Van Houdt J, Stoeva R, Van Esch H et al. 2014. Pseudoautosomal Region 1 Length Polymorphism in the Human Population. *PLoS Genet.* **10(11):**e1004578.

Nik-Zainal S, Van Loo P, Wedge DC, Alexandrov LB, Greenman CD, Lau KW, Raine K, Jones D, Marshall J, Ramakrishna M. et al. 2012. The life history of 21 breast cancers. *Cell* **149(5):** 994–1007.

Puente XS, Pinyol M, Quesada V, Conde L, Ordonez GR, Villamor N, Escaramis G, Jares P, Bea S, Gonzalez-Diaz M et al. 2011. Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. *Nature* **475(7354):** 101–105.

Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, and Mesirov JP. 2011. Integrative genomics viewer. *Nat. Biotechnol.* **29(1):** 24–26.

Schmitt MW, Kennedy SR, Salk JJ, Fox EJ, Hiatt JB, and Loeb LA. 2012. Detection of ultra-rare mutations by next-generation sequencing. *Proc. Natl. Acad. Sci. U.S.A.* **109(36):** 14508–14513.

Shah SP, Morin RD, Khattra J, Prentice L, Pugh T, Burleigh A, Delaney A, Gelmon K, Guliany R, Senz J et al. 2010. Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature* **461(7265):** 809–813.

Watson JD and Crick FH. 1953. Genetical implications of the structure of deoxyribonucleic acid. *Nature* **171(4361):** 964–967.

De Wolf V, Crepel A, Schuit F, van Lommel L, Ceulemans B, Steyaert J, Seuntjens E, Peeters H, and Devriendt

K. 2014. A complex Xp11.22 deletion in a patient with syndromic autism: Exploration of FAM120C as a positional candidate gene for autism. *Am. J. Med. Genet. A* **164A(12):** 3035-41.