Dna Research, 2016, 23(6), 547–559
doi: 10.1093/dnares/dsw038
Advance Access Publication Date: 31 July 2016
Full Paper

Full Paper

Artifactual mutations resulting from DNA lesions limit detection levels in ultrasensitive sequencing applications

Barbara Arbeithuber[1,2], Kateryna D. Makova[2], and Irene Tiemann-Boege[1,*]

[1]Institute of Biophysics, Johannes Kepler University, Linz 4020, Austria, and [2]Department of Biology, Pennsylvania State University, University Park, PA 16802, USA

*To whom correspondence should be addressed. Tel: +43 732 2468 7620. Fax: +43 732 2468 27620.
Email: irene.tiemann@jku.at

Edited by Prof. Masahira Hattori

Received 22 February 2016; Accepted 25 June 2016

Abstract

The need in cancer research or evolutionary biology to detect rare mutations or variants present at very low frequencies ($<10^{-5}$) poses an increasing demand on lowering the detection limits of available methods. Here we demonstrated that amplifiable DNA lesions introduce important error sources in ultrasensitive technologies such as single molecule PCR (smPCR) applications (e.g. droplet-digital PCR), or next-generation sequencing (NGS) based methods. Using templates with known amplifiable lesions (8-oxoguanine, deaminated 5-methylcytosine, uracil, and DNA heteroduplexes), we assessed with smPCR and duplex sequencing that templates with these lesions were amplified very efficiently by proofreading polymerases (except uracil), leading to G->T, and to a lesser extent, to unreported G->C substitutions at 8-oxoguanine lesions, and C->T transitions in amplified uracil containing templates. Long heat incubations common in many DNA extraction protocols significantly increased the number of G->T substitutions. Moreover, in ~50-80% smPCR reactions we observed the random amplification preference of only one of both DNA strands explaining the known 'PCR jackpot effect', with the result that a lesion became indistinguishable from a true mutation or variant. Finally, we showed that artifactual mutations derived from uracil and 8-oxoguanine could be significantly reduced by DNA repair enzymes.

Key words: artifactual mutations, DNA lesions, sequencing errors, PCR jackpot, ultrasensitive detection

1. Introduction

The last decade has seen an extensive development of ultrasensitive technologies capable of detecting DNA variants present at very low frequencies ($\sim 10^{-4}$–10^{-5}) in a sample driven mainly by cancer research to better understand tumour development. Additionally, evolutionary biologists and geneticists have invested considerable efforts in measuring mutation frequencies to gain further insights into the process of mutagenesis. Directly measuring rare mutations or sequence variants in a large DNA pool is still quite difficult given the vast excess of background, which has been reduced with special tweaks relying on restriction enzyme digests to reduce the wild-type pool[1,2] or the highly selective enzymatic activity of pyrophosphorolysis coupled with polymerization of blocked primers, known as pyrophosphorolysis-activated PCR.[3,4]

Alternative methods to measure ultra-rare sequence variations have been developed using single-molecule amplification approaches that amplify each DNA molecule in its own compartment, which allows to count each mutation since clonal copies of the initial single DNA molecule are produced.[5–7] The amplification of single DNA molecules, often referred as dilution PCR, single-molecule PCR (smPCR), or digital PCR (dPCR), has been considerably improved in its throughput by in-house platforms[8] for rare mutation detection[5,9–13] or DNA methylation analysis,[14] but also by commercial platforms available either as chip-based systems (using nanofluidic chips), or droplet-based systems (e.g. droplet digital PCR).[15] In fact, over the last years, we have seen an explosion of applications using the nanoliter-sized droplet PCR for the highly precise absolute quantification of nucleic acids, especially for DNA methylation analysis,[16] and the highly sensitive detection of cancer mutations[17] or fetal DNA.[18 and references within]

Next-generation sequencing (NGS) can produce billions of reads per sequencing run, and also provides an ideal platform with the appropriate throughput to identify ultra-rare sequence variations. However, a fundamental limitation of this technology has been the high error rates associated with library preparation (involving also PCR), sequencing, and base calling, resulting in error rates of ∼1–0.05% with common sequencing instruments like Illumina (reviewed by Kinde et al.[19]). As a result, different approaches for library preparation have been published to distinguish sequencing or PCR errors from real DNA variants, decreasing error rates to less than 10^{-5} per base sequenced (depending on the method).[19–24] This is achieved by tagging strategies that allow the formation of consensus sequences (also known as read families), each family representing one initial DNA molecule. A true mutation is present in the majority of the reads in a family; whereas PCR and sequencing errors are only present in a subset of reads within a family. Strategies to accomplish this individual tagging include the addition of a random sequence in the amplification primers,[19,23] ligation of tag-containing adapters,[22] or the hybridization of molecular inversion probes (MIPs),[20,24] which contain random tags, followed by extension and ligation, forming a single-stranded circle. Alternatively, in circle sequencing, family members are created by circularization of small DNA fragments followed by rolling circle amplification.[21] This rolling circle amplification has the advantage over other conventional PCR methods that it filters out errors introduced in early PCR cycles that are exponentially amplified resulting in artifactual mutations (known as 'jackpot mutations').

While all these described consensus sequence-based ultrasensitive technologies report extremely low error rates by filtering out PCR and sequencing-associated errors, amplifiable DNA lesions (such as 8-oxoguanine, or deaminated cytosine or 5-methylcytosine) opposite of which the polymerase can insert a wrong base,[25 and references within] cannot be distinguished with most of these technologies, and therefore, have a profound effect in the sensitivity and the detection limits of the methods. So far, 'duplex sequencing' has been shown to be a suitable ultrasensitive NGS method that can filter out amplifiable lesions, in addition to PCR and sequencing errors, by a special tagging strategy.[22] Duplex sequencing independently tags each strand of a double-stranded DNA, such that the resulting forward and reverse consensus reads can be traced back to the same initial double-stranded DNA molecule. By comparing the forward and reverse strand of a single initial double-stranded DNA molecule, errors coming from DNA lesions, present in only one strand, can be identified and corrected. However, even duplex sequencing has its limitations due to the high number of reads necessary to obtain the final consensus sequence of one initial DNA molecule. While smaller genomes/targets are affordable, sequencing the whole human genome is very costly and not feasible yet. Additionally, a high amount of starting material is required, which makes the method unsuitable for the analysis of limited material, e.g. in forensics or specific clinical samples.

Other sequencing-based methods developed over the last years allow direct sequencing of different lesion types, either by the analysis of alterations in the kinetics of DNA polymerases with single-molecule real-time (SMRT) DNA sequencing,[26] or by the introduction and sequencing of an additional unnatural base pair at the lesion sites that can be directly sequenced with the α-hemolysin nanopore system.[27] However, considering the high error rates of both these methods and the technical challenge when measuring higher lesion numbers, to date they provide a tool for the direct study of several DNA lesions, but are not well suited for the analysis of ultra-rare sequence variants.

Amplifiable template lesions occur frequently during DNA extraction and preparation protocols in which increased oxidation and hydrolytic deamination rates have been associated.[28–35] For example, Costello et al.[30] reported an increase of G->T transversions that were directly linked to DNA preparation protocols that induced 8-oxoguanine (8-oxoG) lesions, a common product of oxidative DNA damage. This lesion does not only occur during sample preparation, but was also shown to be present in genomic DNA at levels of 10^{-6}, with increased numbers at meiotic recombination sites and regions with a high SNP density.[36] The oxidized guanine (8-oxoG) pairs either with cytosine, or more frequently, with adenine during amplification.[37,38] Another common form of DNA damage is the deamination of cytosine or 5-methylcytosine (5-meC). The deamination product of cytosine is uracil, which can be easily removed by treating the DNA with a uracil DNA glycosylase (UDG), followed by the cleavage of the abasic site by an AP endonuclease resulting in an unamplifiable template.[39] However, the deamination of 5-meC forms thymine, a base naturally occurring in DNA, which cannot be easily removed, and is, therefore, a common source of artifacts.

This highlights that DNA lesions arising during template preparation are a critical problem, especially for applications requiring ultrasensitive detection, since they often result in false positive mutation calls. Yet, very little is known on how different lesions are amplified and what kind of biases result. In theory, a lesion on one DNA strand would only render about half to a quarter of the PCR or sequencing products with the wrong base. However, this might not be the case and lesions could be an important driver of 'jackpot' or artifactual mutations due to amplification biases happening during PCR. More importantly, only very few studies report on sample preparation procedures that can reduce sequencing artifacts, and improve the sensitivity and detection limits of ultrasensitive methods. For this reason, we evaluated in this work the artifact formation of the most common amplifiable DNA lesions (uracil, 8-oxoG, deaminated 5-methylcytosine, and DNA heteroduplexes) with two different PCR-based approaches used in ultrasensitive technologies, smPCR and duplex sequencing. Specifically, we analyzed synthetically produced inserts with DNA lesions at defined positions. Additionally, we also analyzed artifact formation in plasmids and genomic DNA and also tested different sample preparation procedures to reduce specific DNA lesions. Finally, we characterized the efficiency of different DNA repair enzymes in eliminating specific lesions before PCR.

2. Materials and methods

2.1. DNA sources

Plasmid and human genomic DNA: A plasmid (HSI_vector) containing a 4187-bp region of the human chromosome 21 was prepared as described in Supplementary methods. Human sperm and blood samples from anonymous donors were collected by informed consent approved by the ethics commission of Upper Austria (Approval: F1-11) at the IVF Landes-Frauen- und Kinderklinik Linz, Austria. Blood DNA was extracted using the PAXgene blood DNA kit (Qiagen), sperm DNA was extracted with the Gentra Puregene Cell Kit, as described previously.[40,41] In short, we extracted sperm DNA following instructions of the manufacturer except for the proteinase K digest which was performed overnight at 37 °C. The plasmid (HSI_vector: Supplementary Fig. S1) was extracted with a standard plasmid extraction protocol detailed in Supplementary Methods.

Synthetic DNA inserts: We produced six different double-stranded inserts with known DNA lesions by hybridizing synthetic single-stranded DNA fragments (Supplementary Table S1) in different combinations (Fig. 1A and Supplementary Fig. S2). The sequence of the synthetic DNA was designed to produce 20 bp single-stranded overhangs at the 3′ ends after hybridization to allow the integration of the synthetic fragment in the HSI_vector. Two bases at positions 2540-2541 of the vector were designed to be different from the original HSI_vector sequence. The hybridization was carried out by mixing equal amounts (2–10 µl) of 100 µM of two different single-stranded synthetic DNAs in hybridization buffer (50 mM Tris-HCl pH 7.4, 0.1 M NaCl) and incubating them with the following temperature program: 98 °C for 3 min, with a temperature decrease of 1 °C per min, and storage at 8 °C; the hybridization efficiency was monitored on a 10% polyacrylamide gel. All inserts were synthesized as Ultramers (which provide a lower synthesis error rate compared with standard oligos,[42] and for which, the synthesis chemistry provides an improved coupling efficiency above 99.5%), except for the strand with 8-oxoG, which was only available as standard DNA Oligo (both from Integrated DNA Technologies).

Vector-insert (HSI_insert) DNA constructs: Within the HSI_vector, we substituted a 110 bp fragment with the different synthetic inserts (Fig. 1A and Supplementary Figs S1 and S2). The steps involved in the preparation of the vector-insert constructs were modified from the Gibson assembly[43] since standard cloning by restriction digests and ligation was too inefficient to exchange the majority of the old inserts by a synthetic one. An overall scheme of the preparation of the HSI_insert constructs is shown in Supplementary Fig. S3. The HSI_vector was first linearized by digestion with the restriction enzyme XmaI (NEB) to allow better amplification, and 10^8 molecules (which is the minimal amount of DNA that was amplified successfully) were then amplified with primers (vector linearization) designed such that the PCR product did not contain position 2465-2574 of HSI_vector (site of the synthetic insert) (Supplementary Table S1). Reactions contained 10^8 molecules HSI_vector, 0.33 ng *Escherichia coli* DNA, 0.5 µM of each primer, 0.1× SYBR Green I (Invitrogen), 1× Expand Long Range Buffer with $MgCl_2$, and 0.35 U Expand Long Range Enzyme Mix (Roche). The reactions were carried out with an initial heating step of 92 °C for 2 min, followed by 35 cycles at 92 °C for 10 s, 61 °C for 15 s, and 68 °C for 8 min. The amplified vector was then purified by using the Wizard SV Gel and PCR Clean-Up Kit (Promega) according to the instructions of the manufacturer, followed by resection of ~4 µg vector DNA in 3′–5′ direction with 0.29 U/µl T5 Exonuclease in a 35 µl reaction at 37 °C for 15 min. The product was then purified again with the Wizard SV Gel and PCR Clean-Up Kit.

After hybridization of the inserts, they were phosphorylated at the 5′ end with a T4 Polynucleotide Kinase (PNK) (NEB) as follows: 6.02×10^{13} molecules insert were treated with 10 U PNK and 1× T4 Ligase Buffer (containing ATP) in a 10 µl reaction and incubated at 37 °C for 30 min. The six different inserts were then hybridized to the HSI_vector in separate reactions and extended with Platinum *Taq* DNA Polymerase (Invitrogen) with the following protocol: 0.185 mM dNTPs, 1.5 mM $MgCl_2$, 1× PCR Buffer, 5.36×10^{10} molecules of the vector, and 6.02×10^{11} molecules of each insert, respectively, were mixed in a 12 µl reaction and incubated at 55 °C for 15 min for hybridization, followed by an extension step with 0.6 U Platinum *Taq* DNA Polymerase (activated by incubation at 94 °C for 5 min before addition) at 50 °C for 15 min. After this extension, 2 µl of the reaction were mixed with 20 U *Taq* DNA Ligase (NEB) and 1× *Taq* DNA Ligase Reaction Buffer in a 10 µl reaction and incubated at 55 °C for 15 min. The HSI_insert constructs containing different fragments were then digested with XhoI (linearization of the vector for better downstream amplification efficiencies) and DpnI (to remove the original vectors without the synthetic inserts) (both restriction enzymes from NEB).

2.2. Template treatments with DNA repair enzymes

We tested Fpg (formamidopyrimidine [fapy]-DNA glycosylase or 8-oxoguanine DNA glycosylase) and the USER (Uracil-Specific Excision Reagent) enzyme system to reduce the amplification of different DNA lesions. Treatment with Fpg was performed as follows: 5×10^7 copies linearized HSI_vector or 5×10^7 copies linearized HSI_insert_5 construct were treated with 4 U Fpg (NEB) in 1× NEBuffer 1 and 1× BSA in a reaction volume of 50 µl at 37 °C for 1 h. Control reactions were set up without Fpg.

Treatments with the USER enzyme were performed on HSI_insert_1 by incubating 2×10^7 copies HSI_insert construct with 1 U USER enzyme (NEB) in 1× Phusion HF Buffer in a reaction volume of 20 µl at 37 °C for 30 min prior to amplification.

For duplex sequencing, 10 µl purified, adapter ligated library of insert 3 were incubated with 1 U USER enzyme in 1× NEB CS Buffer in a reaction volume of 20 µl at 37 °C for 1h. As a control, another 10 µl were incubated for the same time without addition of the enzyme. The reactions were then purified with 1.2 volumes Agencourt AMPure XP beads (Beckmann Coulter), eluted in 20 µl TE_{low}.

2.3. Single-molecule PCR

For single-molecule PCR (smPCR) of HSI_insert constructs, PCR was performed with the 'SMA Inserts' primers (Supplementary Table S1) in 20 µL reactions containing vector-insert construct DNA (diluted to 0.2 molecules per reaction that rendered on average ~20% smPCR reactions with an amplification product), 0.33 ng *E. coli* DNA, 0.25 µM of the appropriate forward and reverse primer, 0.16 mM dNTPs, 0.5× EvaGreen (Jena Bioscience), 1× Phusion HF Buffer (ThermoFisher Scientific) and 0.1 U Phusion Hot Start II High-Fidelity DNA Polymerase. The reactions were carried out with an initial heating step of 94 °C for 2 min, followed by 55 cycles at 94 °C for 15 s, 65 °C for 15 s, and 72 °C for 20 s.

Given that none of the steps during the HSI_insert construction were 100% efficient, we used several control passes to ensure that we are analyzing the HSI_insert construct: (1) we placed our primers such that only an insert ligated within the vector was amplified

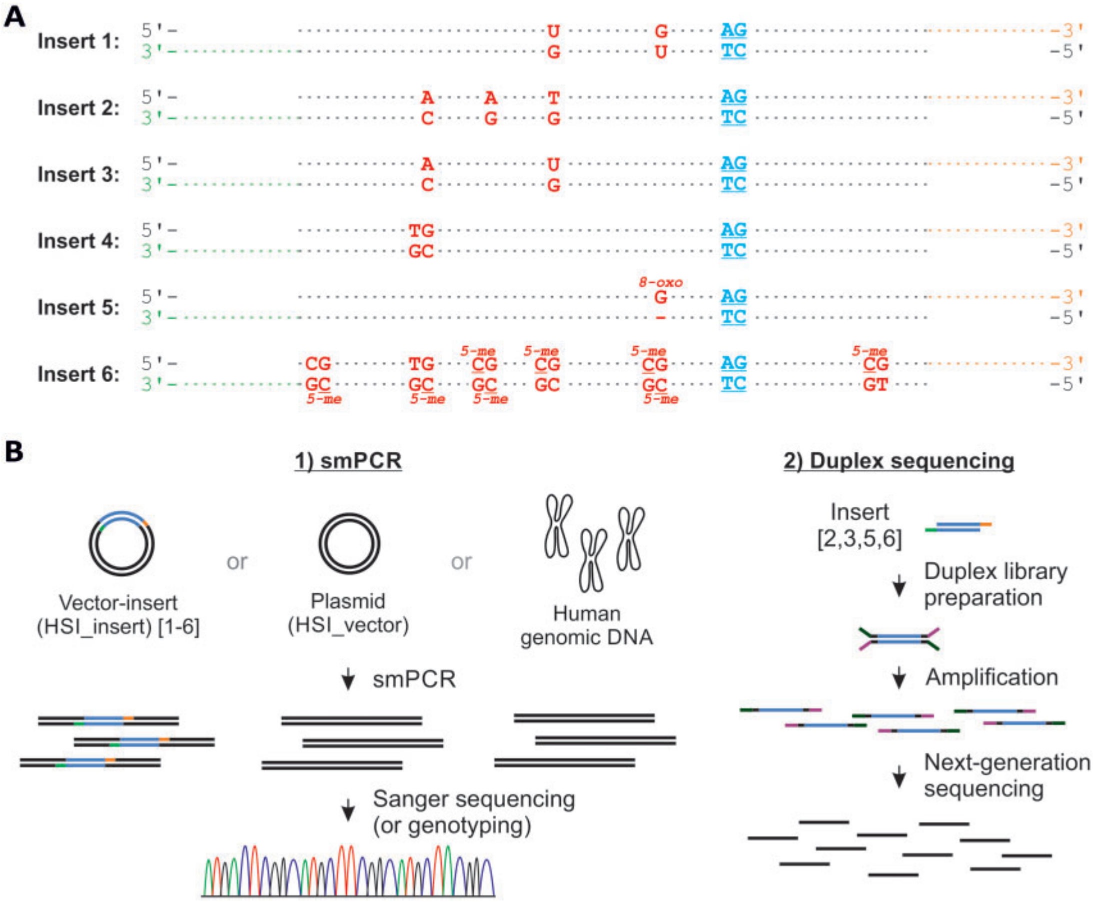

Figure 1. (A) Inserts used in the analysis of different lesions. The inserts were designed with uracils (U) on one or both strands, different mismatches placed randomly in the sequence, an 8-oxoG with a deletion (-) at the opposite position, or methylated cytosines (5-me) within a CpG context. The underlined dinucleotides represent the sequence difference between the plasmid (HSI_vector) and the vector-insert construct (HSI_insert). (B) Strategy used for the analysis of amplifiable DNA lesions. Three different DNA sources were amplified with smPCR: the vector-insert construct (HSI_insert 1-6), a plasmid DNA (HSI_vector), and human genomic DNA, and then analyzed with Sanger sequencing (or in some cases with genotyping). Duplex sequencing was performed directly on the inserts 2, 3, 5, and 6.

during smPCR, and (2) amplifiable vectors without the synthetic insert (present at ~10% as circular vectors with the original sequence) were distinguished by a dinucleotide unique to the insert identified by genotyping before sequencing (Supplementary Methods).

In selected experiments, Phusion U Hot Start DNA Polymerase (ThermoFisher Scientific) was used for the amplification in otherwise identical reactions in order to test amplification efficiencies in uracil containing templates.

For smPCR of human genomic molecules and the HSI_vector DNA, the same PCR conditions were used as described previously for HSI.[41] For the amplification of the HSI_vector, equal primers and PCR conditions were used as for the genomic DNA, with only minor differences in the reaction composition: in short, the HSI_vector DNA was diluted to 0.2 molecules on average per reaction to reduce the number of reactions with more than one molecule to less than 2%, based on the Poisson distribution. As carrier DNA, 0.33 ng *E. coli* or blood DNA per reaction was used when amplifying the HSI_vector. All smPCR reactions were set up in laminar flow hoods to avoid carry-over contamination and no-template controls (NTCs) were included in all experiments.

The smPCR amplicons were either analyzed by sequencing or genotyping (Supplementary Methods). Sequencing was performed by standard capillary Sanger sequencing in a 96-well format (by LGC Genomics GmbH), as described previously.[41] For the HSI_insert constructs, only the sequencing primer HSInt15-Reg3-fwd was used.[41] In some cases, the fidelity of *de novo* mutations was confirmed by sequencing in reverse direction with the sequencing primer HSInt15-Reg3-2rev. Chromatograms were analyzed with the Mutation Surveyor software[44] for new mutations, and outcomes of mismatches and DNA lesions (uracils, 8-oxoG and 5-meC) as described before.[41]

2.4. Temperature treatments of DNA

In order to study the effect of temperature on lesion formation, we performed different heating steps (as often used in DNA extraction protocols) or freeze and thaw cycles on HSI_insert_6 or insert 6 and analyzed them with smPCR or duplex sequencing, respectively. For each analysis (smPCR or duplex sequencing), we compared three general treatments: an untreated control, a frozen, and a heated sample.

smPCR: A freshly prepared HSI_insert_6 construct (control) was frozen at −20 °C for 24 h (frozen sample). We also used a standard DNA extraction protocol (Gentra Puregene Cell Kit) as used previously for sperm DNA extraction[40,41] on 5 × 10^7 molecules unfrozen, freshly prepared HSI_insert_6, but we included a heating step by performing the proteinase K digest at 65 °C for 2 h and the samples/solutions were mixed by vortexing (heated sample).

Duplex sequencing: Prior to the library preparation of insert 6, an aliquot was left untreated (control), a second one was thawed, frozen again at −20 °C for 14 days, and thawed and refrozen two additional times for 24 h freezing steps (frozen sample). The third aliquot was equivalent to the control, but then heated to 65 °C for 3 h in TE_{low} during library preparation after the adapter ligation step to avoid sample denaturation beforehand (heated sample).

2.5. Duplex sequencing of inserts

We evaluated artifact formation at lesions by the independent analysis of the forward and reverse strands using single strand consensus sequences (SSCSs), which have a ∼100-fold reduced error rate compared with conventional NGS methods.[22] Duplex libraries were prepared as described by Kennedy et al.,[45] with minor modifications. An aliquot of 3,630 ng double-stranded synthetic DNA inserts (inserts 2, 3, 5, and 6) containing different DNA lesions were end-repaired with the NEBNext End Repair Module (NEB) according to the instructions of the manufacturer, except that the incubation time was elongated to 1 h (1.5 h for insert 5) instead of the suggested 30 min. This adaptation was necessary to blunt the 20 bp 3′-overhangs of the fragments (one 50 bp overhang in insert 5). Libraries were then A-tailed, and the adapter was ligated with 1800 U T4 ligase (NEB) with 20× molar excess at 16 °C for 30 min.

Duplex adapters were prepared as described by Kennedy et al.,[45] with some minor modifications in the protocol. In brief: T-tailed adapters were prepared by hybridization of the oligos MWS51 and MWS55 (sequences reported by Kennedy et al.[45]), followed by extension with the Klenow Fragment (3'→5' exo-) (NEB) and a restriction digest with TaaI (HypCH4III) at 60 °C for 16 h. Adapters were purified by ethanol precipitation with 2 volumes absolute ethanol and 0.5 volumes 5 M NH_4OAc. The different steps of adapter preparation were monitored on a 3% agarose gel (1.5% normal agarose and 1.5% low-melt agarose). One attomole of adapter-ligated DNA was used for the generation of amplified tag families, and the optimal cycle number for amplification was evaluated by real-time PCR, as suggested by Kennedy et al.[45] Purifications of DNA between the different steps of library preparation were performed with Agencourt AMPure XP beads (Beckmann Coulter), or ethanol precipitation for the shorter (80 bp) insert 5 (as described for the adapters).

The libraries were quantified with the KAPA Library Quantification Kit (Kapa Biosystems). Sequencing was performed on an Illumina MiSeq platform producing 151 bp paired-end reads (v2 chemistry). Reported results represent one sequencing run; however, the analysis of a second independent sequencing run of the same libraries showed very similar results (Supplementary Fig. S4). Duplex sequencing data were analyzed according to the pipeline published by Kennedy et al.[45] with several modifications as described below.

To obtain independent sequencing data for the forward and reverse strands, single strand consensus sequences (SSCSs) were analyzed instead of duplex consensus sequences (DCSs). The 151 base paired-end reads were trimmed to 126 bases (96 bases for insert 5) with the FASTX-Toolkit before analysis. This trimming was necessary since the inserts had only a length of 110 bp (80 bp for insert 5), to which the duplex adapters were ligated. After trimming, the SSCS bam-files were created with published duplex sequencing analysis protocols.[45] An insert-specific reference containing mixed bases at sites differing between forward and reverse strands was used to avoid biases in the preferential strand-specific mapping (as demonstrated for insert 2; Supplementary Table S2). Since the used read-length produced paired-end reads overlapping the whole insert sequence (80 bp for insert 5 and 110 bp for all other inserts), it was possible to merge the SSCSs of both paired-ends during sequence analysis. This allowed to recover sequence information at positions with high error rates (represented with N in one of the paired SSCS reads, when the predominant base was present in less than 70% of the family members) by substituting the N with the correct base derived from the opposite paired-end SSCS read. For this purpose, the reads in the SSCS bam-file were first separated into read 1 and read 2 and further divided into forward- and reverse-mapping reads, allowing the separate analysis of the forward and reverse strands of the initial DNA molecule, using samtools.[46,47] Only the reads with both paired-end reads available were used for further analysis. The bam-files were first converted to fastq format using bamtools,[48] and only reads with both paired-ends were merged using PEAR.[49] After merging, the forward and reverse mapping reads were again separately mapped to the sample-specific reference sequence using BWA[50] for the further independent analysis of forward and reverse strands. Alignments were inspected with the Integrative Genomics Viewer (IGV).[51] The first and the last 15 bases of the inserts (which could represent initially 3′ resected ends that were repaired to blunt ends) were not used for analysis since errors in end-repair is a potential source of sequence biases.

2.6. Statistical analysis

Error bars for mutations frequencies were calculated as 95% Poisson confidence intervals (CIs),[52,53] the significance of differences between obtained mutations rates was tested with Fisher's Exact Test.

3. Results and discussion

3.1. Construction and analysis of inserts with different template lesions

In this work, we evaluated how template lesions can introduce different biases or artifactual mutations with considerable impact on the detection limits in applications measuring extremely rare events. We used two different ultrasensitive methods for this purpose: smPCR and a special application of NGS, known as duplex sequencing.[22] In smPCR, we analyzed amplification products of single molecules by Sanger sequencing. In duplex sequencing, each strand of a DNA duplex is tagged and sequenced by NGS, such that the original complementary DNA can be reconstructed.[22] In this work, we only used the forward and reverse single strand consensus sequences (SSCSs) of the duplex sequences, which was sufficient for filtering out PCR and sequencing errors. Both ultrasensitive methods provide their own advantages: in smPCR, strand-specific amplification biases can be identified via the assessment of one initial molecule at a time; whereas, a higher number of molecules can be inspected with duplex sequencing compared with smPCR.

We tested the effect of lesions on three different template types: a vector with synthetic inserts (HSI_insert) or just the synthetic DNA inserts with known lesions, an oxidized plasmid (the exact location of lesion sites was not predefined), and human genomic DNA (template source in a high variety of biological applications). We produced six different inserts with various DNA lesions or heteroduplexes at predefined positions, representing the most common amplifiable lesions arising within cells and during DNA template preparation (Fig. 1A; Supplementary Fig. S2). The ∼100-bp double-stranded synthetic fragments were ligated either into a plasmid (HSI_insert) for smPCR, or were directly analyzed with duplex sequencing (Fig. 1B). For smPCR, the observed substitution frequency of the HSI_insert (~800 nucleotides that included the ~100 bp insert) was not significantly different between insert and plasmid DNA (3.7×10^{-4} versus 5.9×10^{-4} substitutions/bp, respectively; P value = 0.11, Fisher's exact test). In comparison, error rates measured by smPCR in human genomic DNA or plasmid DNA directly prepared from bacterial extracts (HSI_ vector) rendered mutation frequencies of the order of $\sim 10^{-6}$ and $\sim 10^{-4}$, respectively. However, inserts showed a higher frequency of insertions and deletions (1.5×10^{-3} versus 3.8×10^{-5} indels/bp, respectively; P value = 4.4×10^{-8}, Fisher's exact test). Comparable frequencies of synthesis errors in ultramers were also found previously[42] and could be confirmed with duplex sequencing (calculated for the insert 6 control sample) to be 3.9×10^{-4} substitutions/bp and 1.1×10^{-3} indels/bp. These low synthesis errors did not interfere with our measurements of artifact formation from lesions in the inserts.

3.2. Guanine oxidation is an important source of artifactual mutations

The modified base 8-oxo-7,8-dihydroguanine, more commonly referred as 8-oxoguanine (8-oxoG), is the oxidation product of guanine and one of the most frequently occurring form of DNA damage. Since 8-oxoG tends to pair with adenine instead of cytosine,[37,38] this often leads to sequencing artifacts (G->T or C->A transversions, depending on the strand used as a reference). In this work, we analyzed templates with an 8-oxoG in one strand and a deletion in the complementary strand at the same position (HSI_insert_5 or insert 5; Fig. 1A), which allowed the identification of the amplified strand and to the understanding of how 8-oxoG leads to artifactual mutations.

By screening 56 smPCR reactions with single molecules of HSI_insert_5, we observed that 8-oxoG lesions are amplified with the same efficiency as strands without lesions (35 and 31, respectively). This could also be confirmed with duplex sequencing, resulting in similar numbers of forward- and reverse-mapping reads (57,092 and 56,899, respectively). A closer look at the substitution types resulting at 8-oxoG lesions in smPCR showed that the polymerase added an adenine opposite the 8-oxoG in 42.9% of all reactions, resulting in a G->T homogenous chromatogram peak in 25% of all reactions (these amplified only the forward 8-oxoG containing strand) or a or a G/del->T/del (17.9%) heterogeneous chromatogram in the reactions that amplified both the forward and the reverse strand. The polymerase also inserted in rare occasions (1.8%) a guanine opposite the 8-oxoG resulting in a G->C transversion, and only in 5.4% of all reactions the polymerase incorporated a cytosine opposite the 8-oxoG and no nucleotide change was observed (Table 1). When considering reactions in which only the forward strand (with the 8-oxoG lesion) was amplified, substitution frequencies of 56% and 4% of the total number of amplified 8-oxoGs for the G->T and G->C transversion were observed, respectively (Fig. 2; Supplementary Table S3). A few samples (12.5%) that also amplified exclusively the 8-oxoG strand rendered two bases at the same position at the 8-oxoG lesion, GT or GC (observed as double peak in the sequencing chromatogram), with one peak being predominant in most samples (reaching ratios of 45:55–22:78), which may be the result of biases occurring later in the PCR cycling.

Very similar artifactual mutation frequencies were also measured with duplex sequencing (Fig. 2; Supplementary Table S3) with the G->T substitution being the most predominant type (~80%) as was already reported previously[30]; however, the unreported G->C substitution was also observed (~4%). The 8-oxoG lesion did not result in a substitution only in ~10% of the cases. The few observed G->A substitutions (0.03%) were not significantly different from G->A substitutions at non-oxidized guanines (0.02%; P value = 0.376, Fisher's exact test) (Supplementary Table S4) and, therefore, most likely represent synthesis errors.

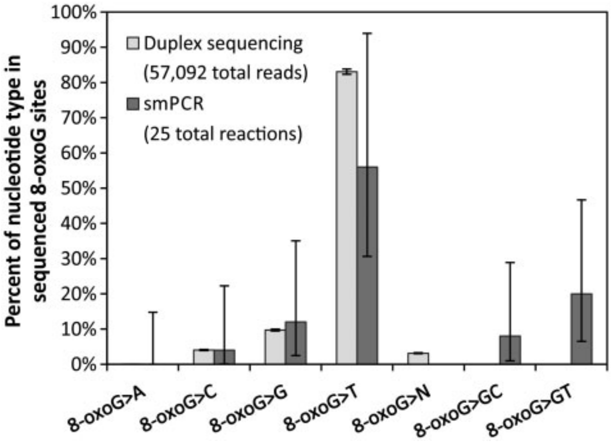

Figure 2. Types of nucleotide substitutions in the amplification of the 8-oxoG lesion. The percentage of different substitutions observed at the 8-oxoG site is shown for SSCSs in duplex sequencing (57,092 total analyzed reads) and Sanger sequencing reads of smPCR products (25 total analyzed smPCR reactions), detailed numbers are shown in Supplementary Table S3. GC and GT in smPCR represent different nucleotides opposite the 8-oxoG in the heterogeneous peak of the sequencing chromatogram. In duplex sequencing, these are represented as an *N*, where *N* represents the presence of more than one nucleotide in the duplex sequencing reads, with the predominant nucleotide being found in less than 70% of the reads. Error bars represent Poisson 95% CIs.

3.3. Fpg treatment reduces artifactual mutations caused by oxidized guanines

After demonstrating that 8-oxoG is an important source of artifactual mutations, we explored the efficiency of the formamidopyrimidine [fapy]-DNA glycosylase or more commonly referred as 8-oxoguanine DNA glycosylase (Fpg), an enzyme that cleaves the glycosylic bond of 8-oxoG resulting in an unamplifiable abasic site. Fpg is an enzyme often used to assess oxidative DNA damage (reviewed by Collins et al.[54]), but to date circle sequencing[21] is the only application for which Fpg is included in the library preparation protocol for the analysis of rare sequence variants. Here, we tested the efficiency of Fpg in reducing artifact formation at 8-oxoG lesions. For this purpose, we first analyzed the HSI_insert_5 construct as a control. Our results demonstrate that Fpg can efficiently reduce artifact formation, since in all of the 50 smPCR reactions treated with Fpg, only the reverse strand (without the 8-oxoG) was amplified (Table 1). These results show that artifacts arising in smPCR at 8-oxoG lesions can be efficiently removed with pre-treatments of the DNA with the repair enzyme Fpg.

We also examined whether rare 8-oxoG lesions present in larger amounts of DNA (e.g. human genomic DNA or oxidized plasmid DNA spiked with 0.33 ng of carrier DNA per smPCR reaction) are also efficiently removed by Fpg, since genomic DNA is prone to this lesion during DNA extraction or purification.[29–32] To address this, we analyzed in total ~9500 smPCRs, each one representing an

Table 1. Artifactual mutations introduced in the amplification of 8-oxoG in smPCR

	Amplified strand									
	Both strands	One strand	Forward (8-oxoG)							Reverse (del)
	(forward + reverse)	(reverse or forward)	G	T	C	A	GT	GC	Total	del
w/o Fpg (#)	10 (17.9%)	46 (82.1%)	3 (5.4%)	14 (25.0%)	1 (1.8%)	0 (0%)	5 (9.0%)	2 (3.6%)	25 (44.6%)	21 (37.5%)
+ Fpg (#)	— —	50 (100%)	— —	— —	— —	— —	— —	— —	— —	50 (100%)

Sanger sequencing was performed on smPCR reactions of the HSI_insert_5 construct. The identity of the amplified strand was determined by the deletion (del) present in the reverse strand opposite the 8-oxoG. The strand amplification preference is indicated by the amplification of either both strands or only one strand (reverse or forward). The identity of the base present in the DNA sequence at the position of the 8-oxoG (2529) is reported. The number of reactions (#) and the percentage of the total are reported. All the reactions in which both strands were amplified showed a heterogeneous chromatogram peak (T/del) resulting from a G->T substitution at the 8-oxoG lesion. Samples without Fpg treatment (w/o Fpg) and samples with Fpg treatment (+ Fpg) are shown separately.

Table 2. Enzymatic removal of 8-oxoG analyzed with smPCR

Sample	G>T or C>A artifacts	G>GT or C>CA artifacts	Effective analyzed sites	Mutation frequency
HSI_vector	21	37	357,712	1.6×10^{-4}
HSI_vector +enzyme	0	0	78,720	$< 10^{-5}$
Genomic DNA	7	8	9,766,956	1.5×10^{-6}
Genomic DNA +enzyme	2	2	10,343,596	3.9×10^{-7}

Sanger sequences of smPCR reactions of the plasmid HSI_vector and human sperm genomic DNA were analyzed for homogeneous G->T or C->A, and heterogeneous G->GT or C->CA mutations. Samples not treated with Fpg and samples treated with Fpg (+enzyme) prior to amplification were analyzed. Mutation frequencies were calculated using the effective sites and both, single (G>T or C>A) and doublet (G>GT or C>CA) sequence chromatogram peaks. Data was obtained from 4,547 smPCR reactions of genomic DNA without and 4,921 with enzyme (Fpg) treatment, and 284 and 96 reactions for plasmid DNA without and with Fpg treatment, respectively. No G > C transversions were observed in this data.

amplifiable human genome and ∼400 smPCR reactions of one oxidized control plasmid (HSI_vector without synthetic insert). An elevated number of G->T and G->GT mutations (observed either as G->T or C->A or G->GT or C->CA, depending on which strand was used as the reference sequence) was observed with an effective mutation frequency of 1.5×10^{-6} and 1.6×10^{-4} in genomic DNA and the oxidized HSI_vector, respectively (Table 2). Fpg treatment completely eliminated G->T/C->A and G->GT/C->CA transversions on the oxidized HSI_vector and the mutation frequency was significantly reduced (P value $= 1.8 \times 10^{-5}$, Fisher's exact test). Fpg treatments of genomic DNA also showed a significant ∼4-fold reduction (P value $= 0.01$, Fisher's exact test) in G->T/C->A and G->GT/C->CA transversions (Table 2), but given already the low frequencies of this substitution, it is unclear if the remaining transversions in genomic DNA are still due to 8-oxoG lesions, other types of artifacts, or represent rare true mutations in the genome.

Although, we showed that Fpg can effectively eliminate rare 8-oxoG products even in larger amounts of DNA like genomic DNA, it is very important to consider that after Fpg treatment only the complementary strand without the 8-oxoG gets amplified, which could introduce other types of biases at nearby amplifiable lesions on the non 8-oxoG containing strand. In order to overcome this issue, instead of Fpg, DNA repair mixes could be used, such as available for formalin fixed, paraffin embedded (FFPE) samples (e.g. NEBNext FFPE DNA Repair Mix from NEB) or forensic samples (e.g. PreCR Repair Mix[55]), which include Fpg and other enzymes for the repair of generated abasic sites generating an amplifiable DNA template. However, this enzymatic repair is not 100% accurate and might also be a source of error.

3.4. Amplification of uracil with proofreading polymerases

Deamination of cytosine is another common source of DNA damage often resulting in artifactual mutations, since it produces uracil, a base that pairs with adenine resulting in a C->T substitution. Different proofreading polymerases used in ultrasensitive protocols (e.g. Phusion Hot Start II or KAPA HiFi DNA Polymerase) amplify templates containing uracil very inefficiently. For this reason, we analyzed in this work if and how uracil lesions produce artifactual mutations when using proofreading polymerases. Specifically, we evaluated the amplification of (1) templates containing uracil in both strands, (2) templates containing uracil only in one strand, and (3) uracil containing templates treated with the repair enzyme mix uracil-specific excision reagent (USER) (a mixture of UDG and the DNA glycosylase-lyase Endonuclease VIII).

First, we assessed the amplification efficiency of uracil-containing templates (HSI_insert_1, containing uracil in both strands in close proximity) by counting the number of positive smPCR reactions that amplified the HSI_insert_1. The percentage of positive smPCR reactions obtained with Phusion Hot Start II (inhibited by uracils) was compared with the positives obtained with Phusion U, a modified version of the Phusion polymerase designed to efficiently amplify uracils. The numbers of positive reactions obtained with Phusion U were close to the predicted 30.5%, based on the amount of input DNA and Poisson distribution. In comparison, with Phusion Hot Start II, we obtained a ∼37.7-fold lower number of positive smPCR reactions, reflecting the lower amplification efficiency of this polymerase for uracil-containing templates (Fig. 3). An analysis of the sequence of the templates that got amplified with Phusion Hot Start II

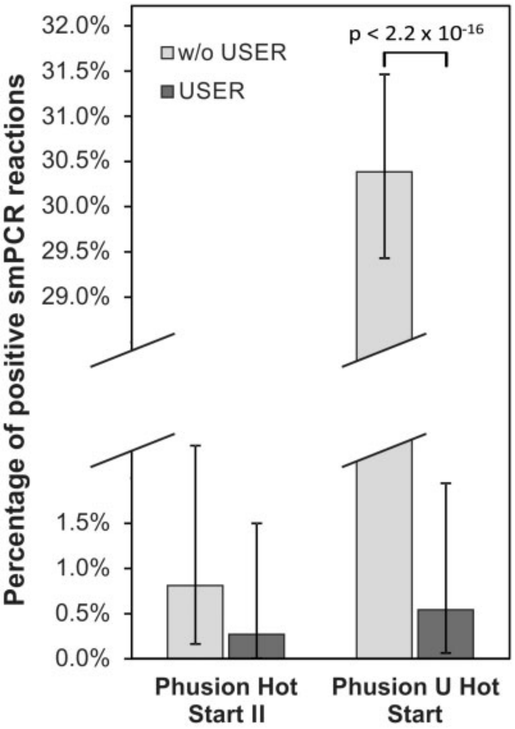

Figure 3. Amplification of uracils with different Phusion polymerases with smPCR. The amplification efficiency of Phusion Hot Start II and Phusion U was compared for samples that contain uracil in both strands (forward and reverse; HSI_insert_1 construct). Efficiency was measured as the percentage of positive smPCR reactions. In total, 372 smPCR reactions were analyzed for each condition (without USER treatment, and USER treatment before amplification). Error bars represent Poisson 95% CIs.

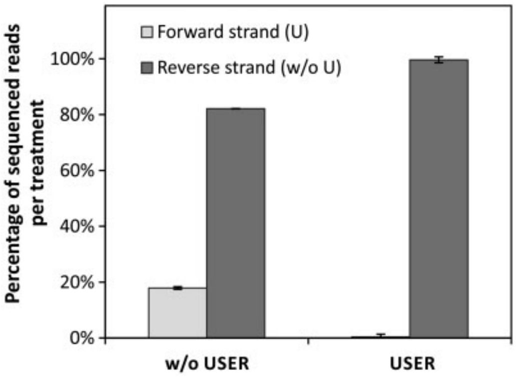

Figure 4. Strand-amplification bias of insert_3 with and without USER treatment measured by duplex sequencing. Based on a total of 34,072 reads for insert_3 (one uracil in the forward strand), 6,096 SSCSs were formed with the strand containing uracil (forward strand) and 27,976 for the strand without the uracil (reverse strand). After USER treatment, a total of 34,636 reads were obtained, of which, 138 and 34,498 formed SSCSs for the forward and reverse strand, respectively. Detailed numbers can be found in Supplementary Table S6. The proportion of amplifiable forward strands could be significantly decreased with USER treatment ($P < 2.2 \times 10^{-16}$, Fisher's exact test). Error bars represent Poisson 95% CIs.

indicated that the uracil was deleted in the majority of these reads may be due to synthesis errors or alternatively due to PCR chimera formation between truncation products leading up to the uracil. One chimera was indeed observed (Supplementary Table S5). We observed a C->T mutation (or G->A) only in one read (11.1% of all successfully sequenced reads), most likely from the amplification of a uracil. This insert was not analyzed with duplex sequencing.

Second, we analyzed the amplification of templates with only one strand containing a uracil lesion (HSI_insert_3 and insert 3), which could be a more likely scenario found in genomic DNA. In addition to a uracil in the forward strand, HSI_insert_3 had a mismatch at position 2487 used to identify the amplified strand. The overall amplification efficiency of this insert in smPCR was approximately the same as for inserts without a uracil. As expected, the majority (89.7%) of the smPCR reactions (29 in total) showed a preferential amplification of the strand without uracil (reverse strand) when using Phusion Hot Start II (Supplementary Table S5). This percentage is comparable to the measured 82.1% obtained with the KAPA HiFi DNA Polymerase for the 34,072 analyzed SSCSs reads in duplex sequencing (Fig. 4; Supplementary Table S6). In rare cases, the forward strand containing the uracil was amplified in smPCR and duplex sequencing (~3% and ~18%, respectively). Uracil was sequenced as expected as a T in 50% and in 94.0% of these cases by smPCR or duplex sequencing, respectively. The frequency of this substitution was significantly different from the C->T frequency observed at Cs on the forward strand of insert 3 (Supplementary Table S4). Note that a fraction of amplified forward strands had other nucleotides instead of the uracil present at very low levels (5.9×10^{-5} to 1.5×10^{-3}) representing inherent synthesis errors (Supplementary Table S6), as was also reported for the other HSI_inserts.

Third, we tested the effect of the enzymatic USER treatment, a DNA enzyme mix commonly used to prevent sequencing errors due to cytosine deamination.[56] The USER treatment significantly reduced the amplification of uracil containing templates of HSI_insert_1 by ~56-fold ($P < 2.2 \times 10^{-16}$, Fisher's exact test) when using Phusion U for the smPCR (Fig. 3). Similar results were obtained for USER treatments of insert 3 with duplex sequencing (Fig. 4; Supplementary Table S6). USER reduced the percentage of amplified uracil containing strands by more than two orders of magnitude of the total sequences. Artifactual mutations (measured as the percentage of reads with a T at the position of the uracil) were reduced from 94.0% to 10.9% (16.82% to 0.04% of the total reads considering both strands), down to a frequency of 4.35×10^{-4} equivalent to the inherent synthesis error frequency. For genomic DNA that was amplified with Phusion Hot Start II, mutation frequencies were already so low that significant differences with USER treatments could not be measured (Supplementary Table S7). No C->T transitions were observed for the plasmid (HSI_vector).

These data show that the amplification of templates containing uracil using polymerases like Phusion Hot Start II or KAPA HiFi DNA polymerase resulted in a strong biased amplification of the strand without the uracil. This effectively reduced the number of mutational artifacts in smPCR and duplex sequencing resulting from this type of lesion. However, these polymerases alone are not sufficient to completely eliminate artifactual mutations arising from uracil lesions present in one DNA strand. Indeed, approximately one-fifth of all reads were amplified from templates with an initial uracil, leading to a C->T substitution. USER treatments considerably reduced this artifact. It is possible that USER can eliminate additional uracils, but we do not have the power to distinguish artifacts from synthesis errors at these low levels.

3.5. Heteroduplexes in double-stranded DNA lead to artifactual mutations due to strand amplification bias

A more difficult lesion to detect and remove is the deamination product of 5-methylcytosine. While most DNA lesions result in bases not

Table 3. Amplification behavior of different mismatches in double-stranded DNA

Amplified strand	HSI_insert_2 construct				HSI_insert_4 construct	
	n (%)	2487(mismatch)	2498 (mismatch)	2509 (mismatch)	*n* (%)	2486 (mismatch)
Forward	7 (22.6)	A	A	T	30 (38.5)	T
Reverse	8 (25.8)	G	C	C	28 (35.9)	C
Forward+ reverse	15 (48.4)	A/G	A/C	T/C	20 (25.6)	T/C
	1 (3.2)	A/G	del/C	T/C		

The number of smPCR reactions (n) in which only the forward strand, only the reverse strand, or both strands were amplified, is shown for the HSI_insert_2 construct (with three mismatches) and the HSI_insert_4 construct (with one mismatch). The identity of the bases in the sequencing results of the forward strand is inferred from positions 2487 (A/C mismatch), 2498 (A/G mismatch), and 2509 (T/G mismatch) in the HSI_insert_2 construct, and for position 2486 (T/G mismatch) in the HSI_insert_4 construct. Nucleotides in the table represent the sequences after mapping to the forward strand. In one smPCR sequence, a heterozygous deletion of one of the analyzed bases was observed, which most likely represents a synthesis error. Note that HSI_insert_2 was analyzed by Sanger sequencing and HSI_insert_4 by genotyping.

naturally occurring in the DNA that can be removed by DNA repair enzymes, some lesions cannot be distinguished from native DNA. One example is the deamination product of 5-methylcytosine resulting in thymine which forms a T/G mismatch in double-stranded DNA. Thus, in this work, we also analyzed whether these lesions forming mismatches can be distinguished and filtered out, e.g. using smPCR. For this purpose, we first created the HSI_insert_2 construct that contains three different randomly placed DNA mismatches (A/C, A/G and T/G). If both strands are amplified, then the mismatches should be identifiable downstream in the resulting data (e.g. as a doublet/heterogenous chromatogram peak).

Our results show that more than half of all smPCR reactions of HSI_insert_2 (31 reactions in total) show a strand amplification bias, or 'PCR jackpot', in which only the forward or the reverse strand was amplified in ~23% and ~26% of the reactions, respectively (Table 3). In all these reactions, the sequencing chromatogram showed a peak of just one nucleotide (homogeneous peak) instead of the expected doublet chromatogram. This strand amplification bias was confirmed by three independent bases in the sequencing read (AAT or GCC for the forward or reverse strand, respectively), making this observation rather unlikely the result of a sequencing error. For those reactions in which both strands were amplified, the amount of rendered smPCR product varied for the different strands, ranging from ratios of 80/20 to 50/50 in the chromatograms for all three different heteroduplexes (Table 3). Summing up the frequency of the forward or the reverse strand, each strand is represented in almost equal proportions. Similar amplification proportions were also obtained with duplex sequencing, with 21,807 reads (48.2%) and 23,396 (51.8%) for the forward strand and the reverse strand, respectively; however, a specific strand amplification bias cannot be accurately measured with duplex sequencing since additional factors during library preparation can also cause an absence of a strand in this method (e.g. incomplete ligation of the duplex adapters).

When analyzing templates with only one mismatch (HSI_insert_4 construct), which represents the deamination product of a 5-methylcytosine in the context of a CpG site, similar results were obtained suggesting that a strand amplification bias is a common phenomenon in smPCR regardless of the sequence. For this experiment, we screened the amplified strand by genotyping position 2486, which contains the mismatch (Fig. 1A; Supplementary Figs S1 and S2), allowing us to screen a higher number of reactions (78 total reactions). As shown in Table 3, both strands, only the forward, or the reverse strand were amplified in 25.6%, 38.5%, and 35.9% of the cases, respectively. The number of smPCR reactions in which both strands were amplified varies between the HSI_insert_2 and HSI_insert_4 construct (51.6% versus 25.6%); however, this can be due to random sampling since only 31 smPCR reactions were analyzed for the HSI_insert_2 construct ($P = 0.1$, Fisher's exact test).

3.6. Strand amplification biasis an important source of artifactual mutations in templates with lesions

We have shown that templates with lesions that are amplified very efficiently by PCR can lead to artifactual mutations. However, so far it was not obvious how lesions in one strand of a double-stranded template could lead to artifactual mutations, especially in applications analyzing the products of single-molecule amplifications such as droplet digital PCR.[15] Assuming that both strands of a double-stranded DNA template are amplified with the same efficiency, positions with mismatches or specific amplifiable lesions should appear as reads with more than one nucleotide that can be identified and filtered out. In our smPCR data for the 8-oxoG lesions (HSI_insert_5), both strands (forward and reverse) were amplified only in 17.9% of the smPCR reactions (Table 1). However, in the majority of the cases (82.1%), only one of the both strands was used as a template for amplification, with either the forward (8-oxoG) or the reverse (deletion) strand being amplified in 44.6% or 37.5% of the times, respectively. The reactions that only amplified the strand containing the 8-oxoG lesion resulted in substitutions that could not be differentiated from *bona fide* G->T or G->C mutations.

Interestingly, our smPCR data for the highly oxidized plasmid (HSI_vector) also possibly indicated strand amplification bias. The sequencing reads showed single chromatogram peaks or *bona fide* G->T substitutions (observed previously in our vector-insert constructs from strand amplification bias). These were as frequent as doublet chromatogram peaks G->GT or C->CA (likely representing the product of the amplification of both strands). Both types of peaks (single and doublet chromatograms) got eliminated with the Fpg treatment (Table 2), suggesting that the *bona fide* G > T substitution was the result of an 8-oxoG lesion that got likely enriched by a strand amplification bias. Similar results were also observed for smPCR of genomic DNA.

A strand amplification bias was observed in at least a third or even a higher proportion of smPCR reactions with synthetic samples for which an equal efficiency in the amplification of forward and reverse strands was expected (HSI_insert_2, 4, 5, and 6). This strand amplification bias would also explain why amplifiable lesions such as mismatches, deamination products of 5-meC or 8-oxoG can result

Table 4. Effect of sample storage and preparation on the mutation rate

	Control			Frozen			Heated		
	n	Effective sites	μ	*n*	Effective sites	μ	*n*	Effective sites	μ
Transition (A>G; T>C)	3	10,660	2.8×10^{-4}	5	9,840	5.1×10^{-4}	8	12,300	6.5×10^{-4}
Transition (G>A; C>T)	1	10,140	9.9×10^{-5}	2	9,360	2.1×10^{-4}	5	11,700	4.3×10^{-4}
Transversion (G>T; C>A)	1	10,140	9.9×10^{-5}	3	9,360	3.2×10^{-4}	4	11,700	3.4×10^{-4}
Total	5	20,800	2.4×10^{-4}	10	19,200	5.2×10^{-4}	17	24,000	7.1×10^{-4}

smPCR reactions of the HSI_insert_6 construct were either collected right after construct generation (untreated control), after storage of an aliquot at − 20 °C for 24 h (frozen), or after performance of the DNA extraction protocol on an aliquot (heated) with a 2 h heating step at 65 °C (25, 24, and 29 reactions were analyzed for the different treatments, respectively). Different mutations found in the samples are reported in the table, *n* gives the number of found mutations, μ is the mutation frequency (n/effective sites for a mutation type) for the different types of observed mutations.

in artifactual mutations, especially if the strand carrying the lesion is exclusively amplified, propagating artifactual mutations by the known 'PCR jackpot' effect that cannot be filtered out with standard sequencing set-ups.

3.7. Effect of sample storage and preparation on artifactual mutations

DNA extraction and storage can introduce different types of amplifiable DNA lesions leading to false base calls.[29–33,57] To analyze these effects, we treated the HSI_insert_6 construct (containing eight 5-methylcytosines, four on each strand, in the context of CpG sites, and two T/G mismatches) with different storage and heating conditions. For this experiment, 5×10^7 molecules of freshly prepared HSI_insert_6 construct were either (1) never frozen (control); (2) frozen at −20 °C for 24 h (frozen); or (3) treated with a standard genomic DNA extraction protocol with a 2 hour incubation at 65 °C (heated) (Materials and Methods section). In total, we analyzed 320 methylated CpG sites with smPCR, but did not observe a deamination event since we did not have the power (with only 320 analyzed 5-meCs) to accurately determine the deamination rate ($<3.13 \times 10^{-3}$; Supplementary Table S8). We also analyzed the effect of the different sample storage and preparations on all mutations in general. The lowest number of artifactual mutations was obtained in control molecules (control), and was highest in molecules treated with the DNA extraction protocol including a heating step (heated), with a significant difference between control, and extracted and heated samples in the total number of artifactual mutations ($P = 0.03$, Fisher's Exact Test) (Table 4). However, given that the number of analyzed samples by smPCR, and therefore, the number of effective sites per treatment type was low, we performed a similar experiment with duplex sequencing.

For duplex sequencing, we reanalyzed the formation of artifacts in slightly different treated/stored samples of insert 6: without treatment (control); three freeze-thaw cycles (frozen); heating to 65 °C for 3 h (heated). In total, 51,996, 48,302, and 25,157 5-methylcytosines were obtained for the control, frozen, and heated samples, respectively (only the six 5-methylcytosines located in the middle of insert 6 were used for the analysis to avoid a bias from the end-repair introduced during sequencing-library preparation). The observed deamination frequencies were 8.65×10^{-4}, 9.11×10^{-4}, 1.07×10^{-3} for control, frozen, and heated samples, respectively (45, 44, and 27 total events), with a slightly increased deamination frequency at 5-meC sites rendering C->T artifactual mutations (Fig. 5A; Supplementary Table S9) and were significantly higher than the inherent synthesis errors. The C->T transitions were significantly higher for methylated Cs compared with non-methylated Cs (~3.7-fold). This represents a higher difference compared with what is reported for hydrolytic deamination of 5-methylcytosine and non-methylated Cs in double-stranded DNA (2.2-fold difference in the rate constants),[58,59] but could be attributed to the 1,000× faster deamination of 5-meC when heating single-stranded DNA. Considering that a high temperature was used during the hybridization step of our synthetic fragments (when the inserts were single-stranded), the larger difference in deamination products between 5-meC and C is not surprising. Additionally, partial inhibition of amplification by the deamination product of C (uracil) might have an effect on the increased difference. Given that heat plays an important role in the deamination rate, avoidance of extensive heating steps that induce single-stranded DNA during extraction and DNA preparation can decrease the formation of this lesion, especially when the source DNA is highly methylated. Thus, heat should be avoided to reduce the formation of artifactual mutations.

We also analyzed with duplex sequencing the effect of sample storage and heat treatment in insert 6 on the formation of other types of artifactual mutations (Fig. 5B; Supplementary Table S9). Repeated freeze–thaw cycles at −20 °C did not have a significant effect on observed mutation frequencies—neither for the individual substitution types nor for the total number of mutations. Heating the sample to 65 °C (for 3 h), which is a common step during tissue lysis in the extraction of DNA, significantly increased the total frequency of artifactual mutations by ~1.4-fold ($P = 4.87 \times 10^{-7}$, Fisher's exact test). When analyzing the individual mutation types separately, only two types of transversions showed a significant difference from the untreated control, with G->T transversions having the most predominant effect ($P = 5.04 \times 10^{-13}$, Fisher's exact test) and with C->G transversions being on the border of significance ($P = 0.04$, Fisher's exact test). These results show that especially heat (either in combination with DNA extraction or heat treatment alone) has the power to increase artifactual mutations, predominately G->T transversions. Interestingly, the sequence context also played a role in the frequency of the G->T transversion with G in the context of CGG, GGG, or GGT showing a significantly higher transversion frequency in heated samples compared with control or frozen samples (Supplementary Fig. S5).

The main type of increased artifactual mutations with heat were G->T transversions (2.4-fold increase compared with the untreated control), which could have resulted from heat-induced ROS formation, as described previously.[35] This difference between heated and non-heated samples could be even greater, but not measurable at the error rates of our synthetic templates. Within many commercially available genomic DNA extraction kits, the proteinase K digest or

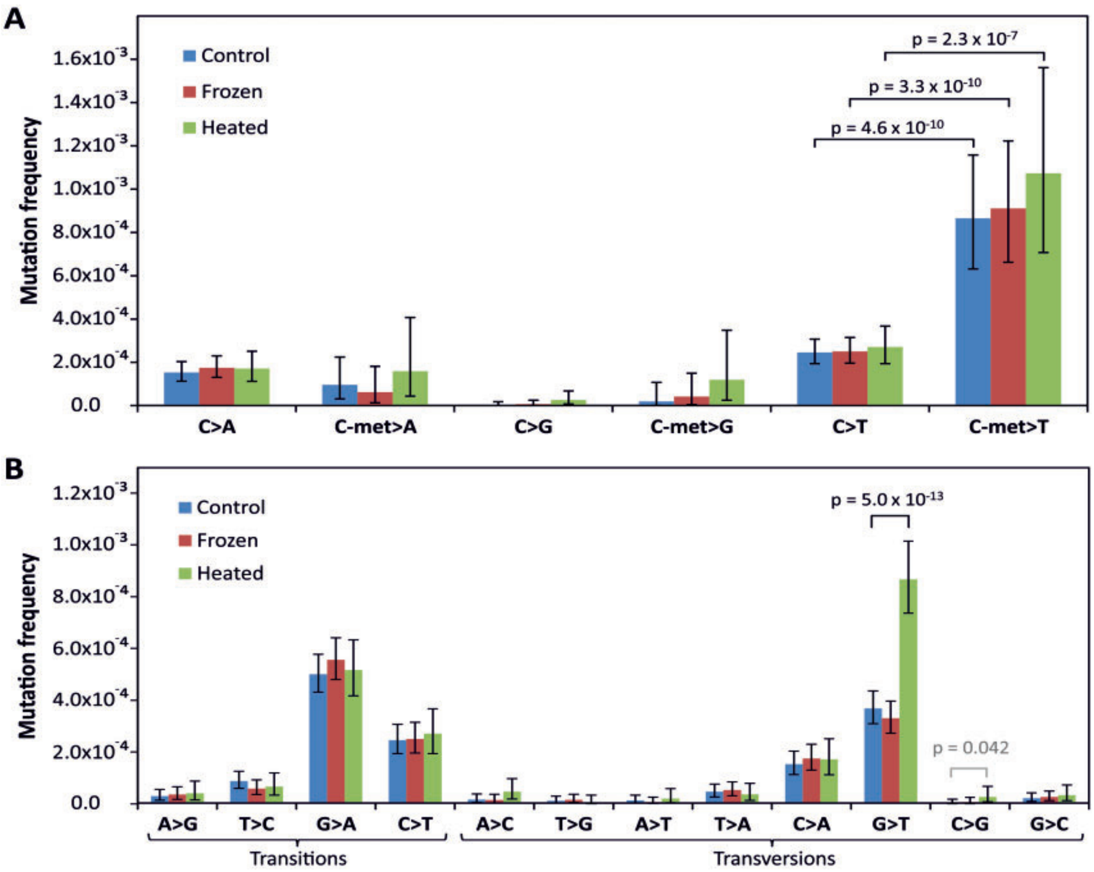

Figure 5. Measured substitutions in duplex sequencing of insert 6 in differently treated samples. Duplex sequencing of insert 6 was performed without any treatment (control), after repeated freeze-thaw cycles (frozen), and after heating to 65 °C for 3 h (heated). (A) Comparison of substitution frequencies at 5-methyl-cytosines (C-met) versus unmethylated C; detailed numbers are shown in Supplementary Table S9. (B) Transition and transversion frequencies in SSCS reads of differently treated/stored samples (repeatedly frozen/thawed or heated to 65 °C). Due to the relatively high number of indels introduced during DNA Ultramer synthesis, this type of mutation was not considered in the analysis. Significant differences were only observed between the control and heated sample for G->T and C->G transversions. An analysis of the nucleotide context of the observed G->T mutations is shown in Supplementary Fig. S5. Error bars represent Poisson 95% CIs. Significance values were estimated using a Fisher's exact test.

DNA rehydration is performed at elevated temperatures (exact temperatures and incubation times depend on the used kit, sample type, and input sample amount). Just some random examples are the Wizard Genomic DNA Purification Kit from Promega, in which DNA rehydration is recommended at 65 °C for 1 h, the DNA Extraction Kit from Stratagene that includes a pronase digestion at 55 °C for 2 h or 60 °C for 1 h depending on the sample type, or the DNeasy Blood & Tissue Kit from QIAGEN, in which for DNA extraction from animal tissue heating to 65 °C for 1–8 h is suggested for the proteinase K digest. Additionally, we have to consider that samples are heated not only within the DNA extraction protocol itself. If, for example, a restriction digest is required in an application, protocols often include a heat-inactivation of the enzyme (an incubation at 80 °C for 20 min is suggested for many commercially available restriction enzymes). Finally, PCR itself requires very high initial temperatures above 90 °C. Not all the heating steps can be avoided, but precautions can be taken in experiments to reduce extensive DNA oxidation during sample preparation, which can include the decrease of temperature during cell lysis (proteinase K digestion) and DNA rehydration to 37 °C, with an increase in the incubation time, or the exclusion of an heat inactivation step after a restriction enzyme digest and the use of DNA repair enzymes.

4. Conclusions

In this work, we demonstrated that amplifiable lesions play an important role in increasing the number of artifactual mutations, and have to be considered in ultrasensitive detection technologies such as smPCR sequencing or NGS applications. While duplex sequencing[22] can effectively discriminate between template lesions and real mutations, most other ultrasensitive detection methods do not have this power. Even in applications using smPCR, lesions can lead to artifactual mutations due to the strand amplification bias of single DNA strands. Therefore, it is very important to reduce lesion formation by diminishing ROS formation and heating steps, as well as by including treatments with DNA repair enzymes. Moreover, it is critical to estimate the lesion induced artifactual mutation rate in control samples and correct for this value when estimating 'true' mutation rates.

Acknowledgements

The authors are grateful to Kristin Eckert and Suzanne Hile for experimental advice regarding establishing duplex sequencing and to Wilfried Guiblet for computational suggestions.

Supplementary data

Supplementary data are available at www.dnaresearch.oxfordjournals.org. The supplementary data file contains: Supplementary Methods, Supplementary Figs S1–S5 and Supplementary Tables S1–S9.

Funding

This work was supported by the 'Austrian Science Fund' (FWF) P25525-B13 to I.T-B., the DOC Fellowship of the Austrian Academy of Sciences 23722 to B.A., the Marshall Plan Scholarship of the Austrian Marshall Plan Foundation 53772424132014 to B.A., and an NIH Grant R01 GM116044 to K.D.M. Funding for open access charge: FWF/P25525-B13.

References

1. Goriely, A., Hansen, R.M., Taylor, I.B., et al. 2009, Activating mutations in FGFR3 and HRAS reveal a shared genetic origin for congenital disorders and testicular tumors, *Nat. Genet.*, **41**, 1247–52.
2. Tiemann-Boege, I., Navidi, W., Grewal, R., et al. 2002, The observed human sperm mutation frequency cannot explain the achondroplasia paternal age effect, *Proc. Natl. Acad. Sci. U S A.*, **99**, 14952–7.
3. Liu, Q. and Sommer, S.S. 2004, PAP: detection of ultra rare mutations depends on P* oligonucleotides: "sleeping beauties" awakened by the kiss of pyrophosphorolysis, *Hum. Mutat.*, **23**, 426–36.
4. Qin, J., Calabrese, P., Tiemann-Boege, I., et al. 2007, The molecular anatomy of spontaneous germline mutations in human testes, *PLoS Biol.*, **5**, e224.
5. Vogelstein, B. and Kinzler, K.W. 1999, Digital PCR, *Proc. Natl. Acad. Sci. U S A.*, **96**, 9236–41.
6. Li, H.H., Gyllensten, U.B., Cui, X.F., Saiki, R.K., Erlich, H.A. and Arnheim, N. 1988, Amplification and analysis of DNA sequences in single human sperm and diploid cells, *Nature*, **335**, 414–7.
7. Li, H., Cui, X. and Arnheim, N. 1990, Direct electrophoretic detection of the allelic state of single DNA molecules in human sperm by using the polymerase chain reaction, *Proc. Natl. Acad. Sci. U S A.*, **87**, 4580–4.
8. Lukyanov, K.A., Matz, M.V., Bogdanova, E.A., Gurskaya, N.G. and Lukyanov, S.A. 1996, Molecule by molecule PCR amplification of complex DNA mixtures for direct sequencing: an approach to in vitro cloning, *Nucleic Acids Res.*, **24**, 2194–5.
9. Diehl, F., Li, M., Dressman, D., et al. 2005, Detection and quantification of mutations in the plasma of patients with colorectal tumors, *Proc. Natl. Acad. Sci. U S A.*, **102**, 16368–73.
10. Dressman, D., Yan, H., Traverso, G., Kinzler, K.W. 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**, 8817–22.
11. Li, M., Diehl, F., Dressman, D., Vogelstein, B. and Kinzler, K.W. 2006, BEAMing up for detection and quantification of rare sequence variants, *Nat. Methods*, **3**, 95–7.
12. Boulanger, J., Muresan, L. and Tiemann-Boege, I. 2012, Massively parallel haplotyping on microscopic beads for the high-throughput phase analysis of single molecules, *PLoS One*, **7**, e36064.
13. Tiemann-Boege, I., Curtis, C., Shinde, D.N., Goodman, D.B., Tavare, S. and Arnheim, N. 2009, Product length, dye choice, and detection chemistry in the bead-emulsion amplification of millions of single DNA molecules in parallel, *Anal. Chem.*, **81**, 5770–6.
14. Chhibber, A. and Schroeder, B.G. 2008, Single-molecule polymerase chain reaction reduces bias: application to DNA methylation analysis by bisulfite sequencing, *Anal. Biochem.*, **377**, 46–54.
15. Hindson, C.M., Chevillet, J.R., Briggs, H.A., et al. 2013, Absolute quantification by droplet digital PCR versus analog real-time PCR, *Nat. Methods*, **10**, 1003–5.
16. Li, M., Chen, W.D., Papadopoulos, N., et al. 2009, Sensitive digital quantification of DNA methylation in clinical samples, *Nat. Biotechnol.*, **27**, 858–63.
17. Holdhoff, M., Schmidt, K., Diehl, F., et al. 2011, Detection of tumor DNA at the margins of colorectal cancer liver metastasis, *Clin. Cancer Res.*, **17**, 3551–7.
18. Chiu, R.W., Cantor, C.R. and Lo, Y.M. 2009, Non-invasive prenatal diagnosis by single molecule counting technologies, *Trends Genet.*, **25**, 324–31.
19. 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**, 9530–5.
20. Hiatt, J.B., Pritchard, C.C., Salipante, S.J., O'Roak, B.J. and Shendure, J. 2013, Single molecule molecular inversion probes for targeted, high-accuracy detection of low-frequency variation, *Genome Res.*, **23**, 843–54.
21. Lou, D.I., Hussmann, J.A., McBee, R.M., et al. 2013, High-throughput DNA sequencing errors are reduced by orders of magnitude using circle sequencing, *Proc. Natl. Acad. Sci. U S A.*, **110**, 19872–7.
22. 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**, 14508–13.
23. Jabara, C.B., Jones, C.D., Roach, J., Anderson, J.A. and Swanstrom, R. 2011, Accurate sampling and deep sequencing of the HIV-1 protease gene using a Primer ID, *Proc. Natl. Acad. Sci. U S A.*, **108**, 20166–71.
24. O'Roak, B.J., Vives, L., Fu, W., et al. 2012, Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders, *Science*, **338**, 1619–22.
25. Stiller, M., Green, R.E., Ronan, M., et al. 2006, Patterns of nucleotide misincorporations during enzymatic amplification and direct large-scale sequencing of ancient DNA, *Proc. Natl. Acad. Sci. U S A.*, **103**, 13578–84.
26. Clark, T.A., Spittle, K.E., Turner, S.W. and Korlach, J. 2011, Direct detection and sequencing of damaged DNA bases, *Genome Integr.*, **2**, 10.
27. Riedl, J., Ding, Y., Fleming, A.M. and Burrows, C.J. 2015, Identification of DNA lesions using a third base pair for amplification and nanopore sequencing, *Nat. Commun.*, **6**, 8807.
28. Lindahl, T. 1993, Instability and decay of the primary structure of DNA, *Nature*, **362**, 709–15.
29. Ravanat, J.L., Douki, T., Duez, P., et al. 2002, Cellular background level of 8-oxo-7,8-dihydro-2'-deoxyguanosine: an isotope based method to evaluate artefactual oxidation of DNA during its extraction and subsequent work-up, *Carcinogenesis*, **23**, 1911–8.
30. Costello, M., Pugh, T.J., Fennell, T.J., et al. 2013, Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation, *Nucleic Acids Res.*, **41**, e67.
31. Claycamp, H.G. 1992, Phenol sensitization of DNA to subsequent oxidative damage in 8-hydroxyguanine assays, *Carcinogenesis*, **13**, 1289–92.
32. Finnegan, M.T., Herbert, K.E., Evans, M.D., Griffiths, H.R. and Lunec, J. 1996, Evidence for sensitisation of DNA to oxidative damage during isolation, *Free Radic. Biol. Med.*, **20**, 93–8.
33. Milowska, K. and Gabryelak, T. 2007, Reactive oxygen species and DNA damage after ultrasound exposure, *Biomol. Eng.*, **24**, 263–7.
34. Chen, G., Mosier, S., Gocke, C.D., Lin, M.T. and Eshleman, J.R. 2014, Cytosine deamination is a major cause of baseline noise in next-generation sequencing, *Mol. Diagn. Ther.*, **18**, 587–93.
35. Bruskov, V.I., Malakhova, L.V., Masalimov, Z.K. and Chernikov, A.V. 2002, Heat-induced formation of reactive oxygen species and 8-oxoguanine, a biomarker of damage to DNA, *Nucleic Acids Res.*, **30**, 1354–63.
36. Ohno, M., Miura, T., Furuichi, M., et al. 2006, A genome-wide distribution of 8-oxoguanine correlates with the preferred regions for recombination and single nucleotide polymorphism in the human genome, *Genome Res.*, **16**, 567–75.
37. McAuley-Hecht, K.E., Leonard, G.A., Gibson, N.J., et al. 1994, Crystal structure of a DNA duplex containing 8-hydroxydeoxyguanine-adenine base pairs, *Biochemistry*, **33**, 10266–70.
38. Beard, W.A., Batra, V.K. and Wilson, S.H. 2010, DNA polymerase structure-based insight on the mutagenic properties of 8-oxoguanine, *Mutat. Res.*, **703**, 18–23.

39. Lindahl, T., Ljungquist, S., Siegert, W., Nyberg, B. and Sperens, B. 1977, DNA N-glycosidases: properties of uracil-DNA glycosidase from Escherichia coli, *J. Biol. Chem.*, **252**, 3286–94.
40. Meyer, W.K., Arbeithuber, B., Ober, C., et al. 2012, Evaluating the evidence for transmission distortion in human pedigrees, *Genetics*, **191**, 215–32.
41. Arbeithuber, B., Betancourt, A.J., Ebner, T. and Tiemann-Boege, I. 2015, Crossovers are associated with mutation and biased gene conversion at recombination hotspots, *Proc. Natl. Acad. Sci. U S A.*, **112**, 2109–14.
42. Gibson, D.G. 2009, Synthesis of DNA fragments in yeast by one-step assembly of overlapping oligonucleotides, *Nucleic Acids Res.*, **37**, 6984–90.
43. Gibson, D.G., Young, L., Chuang, R.Y., Venter, J.C., Hutchison, C.A., 3rd and Smith, H.O. 2009, Enzymatic assembly of DNA molecules up to several hundred kilobases, *Nat. Methods*, **6**, 343–5.
44. Minton, J.A., Flanagan, S.E. and Ellard, S. 2011, Mutation surveyor: software for DNA sequence analysis, *Methods Mol. Biol.*, **688**, 143–153.
45. Kennedy, S.R., Schmitt, M.W., Fox, E.J., et al. 2014, Detecting ultralow-frequency mutations by *Duplex Sequencing, *Nat. Protoc.*, **9**, 2586–606.
46. Li, H. 2011, A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data, *Bioinformatics*, **27**, 2987–93.
47. Li, H., Handsaker, B., Wysoker, A., et al. 2009, The Sequence Alignment/Map format and SAMtools, *Bioinformatics*, **25**, 2078–9.
48. Barnett, D.W., Garrison, E.K., Quinlan, A.R., Stromberg, M.P. and Marth, G.T. 2011, BamTools: a C++ API and toolkit for analyzing and managing BAM files, *Bioinformatics*, **27**, 1691–2.
49. Zhang, J., Kobert, K., Flouri, T. and Stamatakis, A. 2014, PEAR: a fast and accurate Illumina Paired-End reAd mergeR, *Bioinformatics*, **30**, 614–20.
50. Li, H. and Durbin, R. 2010, Fast and accurate long-read alignment with Burrows-Wheeler transform, *Bioinformatics*, **26**, 589–95.
51. Robinson, J.T., Thorvaldsdottir, H., Winckler, W., et al. 2011, Integrative genomics viewer, *Nat. Biotechnol.*, **29**, 24–6.
52. Garwood, F. 1936, Fiducial limits for the poisson distribution, *Biometrika*, **28**, 437–42.
53. Patil, V.V. and Kulkarni, H.V. 2012, Comparison of confidence intervals for the poisson mean: some new aspects, *REVSTAT – Stat. J.*, **10**, 211–27.
54. Collins, A.R., Cadet, J., Moller, L., Poulsen, H.E. and Vina, J. 2004, Are we sure we know how to measure 8-oxo-7,8-dihydroguanine in DNA from human cells? *Arch. Biochem. Biophys.*, **423**, 57–65.
55. Diegoli, T.M., Farr, M., Cromartie, C., Coble, M.D. and Bille, T.W. 2012, An optimized protocol for forensic application of the PreCR Repair Mix to multiplex STR amplification of UV-damaged DNA, *Forensic Sci. Int. Genet.*, **6**, 498–503.
56. Briggs, A.W., Stenzel, U., Meyer, M., Krause, J., Kircher, M. and Paabo, S. 2010, Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA, *Nucleic Acids Res.*, **38**, e87.
57. Kunkel, T.A. 1984, Mutational specificity of depurination, *Proc. Natl. Acad. Sci. U S A.*, **81**, 1494–8.
58. Ehrlich, M., Norris, K.F., Wang, R.Y., Kuo, K.C. and Gehrke, C.W. 1986, DNA cytosine methylation and heat-induced deamination, *Biosci. Rep.*, **6**, 387–93.
59. Shen, J.C., Rideout, W.M., 3rd and Jones, P.A. 1994, The rate of hydrolytic deamination of 5-methylcytosine in double-stranded DNA, *Nucleic Acids Res.*, **22**, 972–6.