Reducing Sequence Artifacts in Amplicon-Based Massively Parallel Sequencing of Formalin-Fixed Paraffin-Embedded DNA by Enzymatic Depletion of Uracil-Containing Templates

Hongdo Do,1 Stephen Q. Wong,1 Jason Li,2 and Alexander Dobrovic1,3*

BACKGROUND: Formalin-fixed, paraffin-embedded (FFPE) tissues are routinely used for detecting mutational biomarkers in patients with cancer. A previous intractable challenge with FFPE DNA in genetic testing has been the high number of artifactual single-nucleotide changes (SNCs), particularly for the detection of low-level mutations. Pretreatment of FFPE DNA with uracil-DNA glycosylase (UDG) can markedly reduce these C:G→T:A SNCs with a small panel of amplicons. This procedure has implications for massively-parallel-sequencing approaches to mutation detection from DNA. We investigated whether sequence artifacts were problematic in amplicon-based massively parallel sequencing and what effect UDG pretreatment had on reducing these artifacts.

METHODS: We amplified selected amplicons from lung cancer FFPE DNAs using the TruSeq Cancer Panel. SNCs occurring at a frequency <10% were considered most likely to represent sequence artifacts and were enumerated for both UDG-treated and -untreated DNAs.

RESULTS: Massively parallel sequencing of FFPE DNA samples showed multiple SNCs, predominantly C:G→T:A changes, with a significant proportion occurring above the background sequencing error (defined as 1%). UDG pretreatment markedly reduced C:G→T:A SNCs without affecting the detection of true somatic mutations. However, C:G→T:A changes within CpG dinucleotides were often resistant to the UDG treatment as a consequence of 5-methyl cytosine being deaminated to thymine rather than uracil.

CONCLUSIONS: UDG pretreatment greatly facilitates the accurate discrimination of mutations in FFPE samples by use of amplicon-based approaches. This is particularly important when working with samples with low tumor purity or for the assessment of mutational heterogeneity in tumors.

The identification of predictive and prognostic mutational biomarkers for molecularly targeted therapies in cancer has greatly increased the demand for extensive profiling of individual tumors. Massively parallel sequencing (MPS) technologies have revolutionized the mutational profiling of tumors, allowing sequencing of multiple cancer-related genes in an efficient, high-throughput, and cost-effective manner. Whereas hybridization capture has been applied to detect somatic mutations in the exome (1–3) or in large panels of genes (4), amplicon-based approaches have commonly been adopted to profile sets of clinically important exons (5).

Although this amplicon-based approach has been successful with good-quality DNA sources such as blood or fresh frozen tissues, the use of DNA from formalin-fixed and paraffin-embedded (FFPE) tissues remains challenging owing to the limited amounts of DNA available, fragmentation of the DNA, and the presence of base damage leading to artifactual single-nucleotide changes (SNCs).

In particular, the increased sequence artifacts characteristic of FFPE DNA make the discrimination of true low-frequency genetic variants from artifactual changes extremely difficult (6, 7). An important factor...
that often leads to the more frequent detection of artifactual SNCs in FFPE DNA is the frequently limiting numbers of available templates. Limiting numbers of templates can cause damaged templates to become artifactualy overrepresented as a substantial fraction of the sequencing reads. Several technical or bioinformatic adaptations have been devised to address this problem (8–11), but eliminating damaged templates before PCR amplification is a preferable approach that involves only a simple enzymatic treatment of the FFPE DNA. Recently, using high-resolution melting (HRM) screening and Sanger sequencing of selected regions from the AKT1 (v-akt murine thymoma viral oncogene homolog 1), EGFR (epidermal growth factor receptor), KRAS (v-Ki-ras2 Kirsten rat sarcoma viral oncogene homolog), and Braf v-(raf murine sarcoma viral oncogene homolog B1) genes, we showed that the majority of artifactual SNCs in FFPE DNA are transitional C:G→T:A changes resulting from cytosine deamination to uracil DNA. Although we found that the number of detected sequence artifacts varied across FFPE samples, we also observed that C:G→T:A SNCs often accounted for >50% of all artifactual SNCs in FFPE DNA samples (12, 13). Importantly, we showed that in vitro treatment of FFPE DNA with uracil-DNA glycosylase (UDG) markedly reduced the C:G→T:A SNCs (13). UDG hydrolyses the N-glycosidic bond between the uracil base and the sugar phosphate backbone, thus generating abasic sites on the DNA strand at which certain DNA polymerases are likely to halt (14).

Here we extended our previous findings to an MPS mutation detection protocol using a large panel of amplicons covering recurrently mutated regions in human cancer.

Materials and Methods

PROCESSING OF FFPE DNA

We chose 3 formalin-fixed paraffin-embedded squamous cell lung carcinomas (SCCs) from a previously described panel (13, 15). For each sample, a tumor-enriched region was identified by a pathologist and a previously amplified in a single-tube reaction. We used 5 µL of each DNA sample (concentration 25 ng/µL) for the experiment according to the manufacturer’s instructions. We used the MiSeq system (Illumina) for paired-end sequencing with a v1 150-bp kit. The 3 lung pre- and posttreatment SCC samples were multiplexed into a single MiSeq run, achieving >100 000 reads per sample with a mean coverage of >2000×.

AMPLICON-BASED SEQUENCING

The TruSeq Amplicon Cancer Panel (Illumina) comprises 212 amplicons from 48 genes that are simultaneously amplified in a single-tube reaction. We used 5 µL of each DNA sample (concentration 25 ng/µL) for the experiment according to the manufacturer’s instructions. We used the MiSeq system (Illumina) for paired-end sequencing with a v1 150-bp kit. The 3 lung pre- and posttreatment SCC samples were multiplexed into a single MiSeq run, achieving >100 000 reads per sample with a mean coverage of >2000×.

BIIOINFORMATICS

Using Casava v1.8.2, we converted files generated from the MiSeq instrument to FastQ files containing short-read data. Global alignment, on the basis of the Needleman–Wunsch algorithm, was then performed between the reads and the amplicon reference sequences to identify sequence variations. The likely true variants (in the original biological sample) were operationally identified by a) VarScan2 and b) a variant frequency of >10% using Python scripts. Variants with a frequency <1% were not called but could be visualized with the Integrate Genomics Viewer (www.broadinstitute.org/igv).

Results and Discussion

BASELINE SEQUENCE ARTIFACTS FROM FFPE DNA

The current consensus is that there is an approximately 1% error rate in bases detected in commercial benchtop sequencers (17). Errors arise mainly as a result of sequencing chemistry, cluster amplification, cycle sequencing, and image analysis. Although these errors are sporadic and present in different degrees in all MPS systems, they are usually not considered a major issue in variant calling, given an acceptable level of quality and coverage in the amplicon sequenced. However, accurate detection of mutations at relatively low levels,
especially in the <10% range, when a 1% cutoff is used, is important in many clinical samples where tumor purities are low, even after macrodissection of tumor-enriched regions.

Using HRM and Sanger sequencing, we have previously demonstrated that uracil-derived FFPE sequence artifacts may occur at appreciable levels and, in some cases such as low DNA template amounts, can masquerade as potentially real variants (12). To gauge the level of these sequence artifacts in an MPS setting, we set out to sequence the DNA extracted from the FFPE blocks of 3 lung SCCs using a highly-multiplexed 212 amplicon-based panel from 48 cancer-associated genes (Illumina TruSeq Amplicon Cancer Panel).

The FFPE samples varied in time since fixation (SCC4, 16 years; SCC30, 10 years; and SCC39, 7 years). DNAs from these blocks had previously been tested in our pilot study of UDG treatment with HRM and Sanger sequencing (13) and were chosen for this study due to sufficient DNA being available for MPS. As with

---

**Fig. 1.** SNCs detected in 3 lung SCC FFPE DNAs. SNCs were filtered using lower and upper thresholds of >1% and <10% to minimize contamination by sequencing errors and true changes.

To calculate the relative prevalence of each type of SNC, the sum of correct base calls (A>A, C>C, G>G, and T>T) was used as the denominator for data normalization, and the normalized value was then multiplied by 10^6. In every case, C>G>T:A changes comprise the majority of the potentially artifactual SNCs.

**Fig. 2.** SNCs in 3 lung SCC FFPE DNAs before and after UDG treatment. The prevalence of each SNC detected with (gray) and without (black) UDG treatment is shown.

SNCs were filtered using lower and upper thresholds of >1% and <10% to minimize contamination by sequencing errors and true changes. To calculate the relative prevalence of each type of SNC, the sum of correct base calls (A>A, C>C, G>G, and T>T) was used as the denominator for data normalization, and the normalized value was then multiplied by 10^6.
the lowest sample (SCC30) and the highest sample (SCC39). When the prevalence of each SNC type was examined, C:G>T:A changes were by far the most frequent in all 3 samples, comprising as high as 70%–90% of the total SNCs (91.5% for SCC4, 69.9% for SCC30, and 77.9% for SCC39). The high frequency of C:G>T:A changes is strongly indicative that these changes are indeed artifactual and most readily explained as a consequence of cytosine being deaminated to uracil. These results also indicate that the extent of uracil artifact-inducing DNA damage can substantially differ in each FFPE DNA sample.

There has been 1 other report of predominant C:G>T:A substitutions in FFPE DNA. In a SOLiD-based whole genome sequencing of FFPE breast cancer samples (11), it was found that nucleotide alterations were substantially increased in FFPE tissues compared to matched germline DNA, and that C:G>T:A substitutions were relatively more numerous than the other types of substitutions (11). However, a mechanism for these changes was not proposed, and the proposed remedy was to apply bioinformatic correction.

### REDUCED SEQUENCE ARTIFACTS FROM FFPE DNA AFTER UDG TREATMENT

We found that artifactual C:G>T:A SNCs were markedly reduced after UDG treatment in all 3 samples: a 81% reduction for SCC4, a 60% reduction for SCC30, and a 75% reduction for SCC39 (Fig. 2). This marked reduction of C:G>T:A SNCs resulted in a 65% overall SNC reduction for SCC4, a 40% overall SNC reduction for SCC30, and a 50% overall SNC reduction for SCC39. In contrast, other types of single-base changes remained essentially unchanged. This substantial global reduction of SNCs by UDG treatment confirms that uracil lesions are present in FFPE DNA and are a cause of artifactual C:G>T:A SNCs in PCR-based se-
sequence analyses. Moreover, treatment of FFPE DNA with UDG is a powerful strategy for overall reduction of artifactual SNCs due to the very high contribution of uracil-driven C:G\rightarrow H11022 T:A SNCs.

Whereas there would be selective depletion of high C\rightarrow H11001 G-rich amplicons that are more susceptible to deamination of cytosine to uracil on account of the higher cytosine content, this depletion is preferable to dealing with difficult-to-interpret sequence artifacts. Examination of selected amplicons with the Integrative Genomics Viewer visually confirmed that the C:G\rightarrow H11022 T:A reduction was extensively and repeatedly detectable across different amplicons (see online Supplemental Fig. 1). This finding confirms and extends our pilot study results that C:G\rightarrow H11022 T:A SNCs comprise the major sequencing artifacts in FFPE DNA and are effectively reduced by UDG treatment.

DETECTION OF TRUE MUTATIONS IS UNAFFECTED

It was important to verify that UDG treatment did not obscure the identification of either the high- or low-frequency true variants that were present. We first examined the TP53 gene (tumor protein p53), since inactivating mutations in TP53 are the most common genetic alteration in SCC of lung and are present in nearly all samples of this tumor (1). Nonsynonymous TP53 mutations were detected in each of our samples (Table 1). All the mutations were detected before and after UDG treatment. Single TP53 mutations were detected in SCC30 (c.659A\rightarrow G, p.Y220C) and SCC39 (c.311_312del, p.Q104RfsX44). Interestingly, two different nonsynonymous TP53 mutations (c.445T\rightarrow C, p.S149P and c.725G\rightarrow T, p.C242F) were found in SCC4. The frequency of each TP53 mutation was quite different (60% for C242F and 11% for S149P mutation), suggesting that there is intratumoral heterogeneity for TP53 mutation status.

None of the detected TP53 sequence alterations had C:G\rightarrow T:A variants. Thus, to assess the effect of UDG treatment on detection of true C:G\rightarrow T:A variants, we examined the allelic frequency of the EGFR exon 20 rs1050171 single-nucleotide polymorphism (SNP) (c.2361G\rightarrow A) and of the PDGFRA (platelet-derived growth factor receptor, \alpha polypeptide) exon 18 rs2228230 SNP (c.2472C\rightarrow T) (Table 2). The genotype

![Table 2. Base calling at EGFR c.2361G\rightarrow A and PDGFRA c.2472C\rightarrow T SNPs before and after UDG treatment in 3 lung SCCs.]

<table>
<thead>
<tr>
<th>Sample</th>
<th>EGFR c.2361G\rightarrow A (rs1050171)</th>
<th>PDGFRA c.2472C\rightarrow T (rs2228230)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Reference reads, n (%)</td>
<td>Variant reads, n (%)</td>
</tr>
<tr>
<td>SCC4</td>
<td>G: 1458 (99) A: 12 (1)</td>
<td>C: 3806 (68) T: 1824 (32)</td>
</tr>
<tr>
<td>SCC4 + UDG</td>
<td>G: 1486 (100) A: 0 (0)</td>
<td>C: 2924 (70) T: 1278 (30)</td>
</tr>
<tr>
<td>SCC30</td>
<td>G: 687 (37) A: 1174 (63)</td>
<td>C: 2878 (99) T: 29 (1)</td>
</tr>
<tr>
<td>SCC30 + UDG</td>
<td>G: 849 (41) A: 1227 (59)</td>
<td>C: 3324 (100) T: 9 (0)</td>
</tr>
<tr>
<td>SCC39</td>
<td>G: 3 (0)</td>
<td>A: 1522 (100)</td>
</tr>
<tr>
<td>SCC39 + UDG</td>
<td>G: 3 (0) A: 1030 (100)</td>
<td>C: 1147 (100) T: 0 (0)</td>
</tr>
</tbody>
</table>

* Each of the 3 FFPE DNAs was examined for EGFR and PDGFRA SNPs before and after UDG treatment. The number of reads for each base call at the SNPs was determined from the Interactive Genomics Viewer program.
of both SNPs was identical in the sequencing results from before and after UDG treatment, indicating that true C:G>T:A sequence differences are not affected by UDG treatment. Notably, low-frequency SNCs (present at approximately 1%) found in homozygotes for these SNPs (EGFR c.2361G>A changes in SCC4 and PDGFRA c.2472C>T changes in SCC30 and SCC39) were abolished or substantially reduced after UDG treatment, indicating that they were artifacts.

C-TO-T CHANGES OCCURRING AT CPG DINUCLEOTIDES
Cytosines can be either methylated or unmethylated in DNA. 5-Methyl cytosine is almost exclusively found in the CpG dinucleotides. Whereas cytosine is deaminated to uracil, 5-methyl cytosine is deaminated to thymine. UDG is active only at uracil lesions, and thus UDG treatment will reduce only the uracils, causing C>T artifacts at deaminated cytosines, but will not affect the thymines, causing the C>T artifacts at 5-methyl cytosines.

Some C:G>T:A artifacts at CpG dinucleotides are resistant to UDG treatment as identified by limited-copy-number HRM and Sanger sequencing of 3 amplicons (13). Here we examined the prevalence of C>T changes at the 4 CpN (n = A, C, G, and T) dinucleotides over the entire set of 212 amplicons in the TruSeq panel. Compared to baseline, there was a substantial reduction in the overall C>T changes at CpN dinucleotides after UDG treatment, giving a 5.6-fold reduction in SCC4, a 4.4-fold reduction in SCC30, and a 4-fold reduction in SCC39 (Fig. 3). The marked reduction of C>T changes was confined to the CpA, CpC, and CpT dinucleotides, whereas the prevalence of C>T changes was almost unchanged at CpG dinucleotides. After UDG treatment, 36%, 44%, and 70% of the remaining C:G>T:A SNCs came from CpG sites compared with 9%, 10%, and 14% before UDG treatment.

This result not only confirms our previously reported study but also indicates that a large proportion of CpG sites in the examined coding sequences were methylated. It should be noted that most of the regions examined by the TruSeq cancer panel are not at gene promoters, where one would have the greatest likelihood of unmethylated CpG dinucleotides.

ACCURATE DISCRIMINATION OF LOW-LEVEL MUTATIONS
Accurate detection of low-level mutations is increasingly important for the management of cancer patients. In addition, a major challenge arising from the use of current molecularly targeted therapies for cancer treatment is the development of tumor resistance to treatment drugs (18, 19). For example, KRAS mutations are associated with negative response to anti-EGFR treatment in colorectal cancer (20). Diaz et al. (21) reported that KRAS-mutant tumor cells were already present in subclones of metastatic colorectal cancer lesions before the initiation of anti-EGFR treatment. Thus, accurate detection of low-level resistance mutations might be important in determining which patients are more likely to have an early relapse.

Table 3. C:G>T:A changes found at 6 mutation positions before and after UDG treatment.a

<table>
<thead>
<tr>
<th>Sample</th>
<th>KRAS</th>
<th>AKT1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>c.34G&gt;A</td>
<td>c.35G&gt;A</td>
</tr>
<tr>
<td>SCC4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A: 5 (0.5)</td>
<td>A: 22 (2.1)</td>
<td>A: 28 (2.7)</td>
</tr>
<tr>
<td>A: 0 (0)</td>
<td>A: 5 (0.7)</td>
<td>A: 1 (0.1)</td>
</tr>
<tr>
<td>SCC30</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A: 2 (0.1)</td>
<td>A: 8 (0.6)</td>
<td>A: 11 (0.8)</td>
</tr>
<tr>
<td>A: 1 (0)</td>
<td>A: 0 (0)</td>
<td>A: 1 (0)</td>
</tr>
<tr>
<td>SCC39</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A: 1 (0.1)</td>
<td>A: 3 (0.3)</td>
<td>A: 0 (0)</td>
</tr>
<tr>
<td>A: 0 (0)</td>
<td>A: 0 (0)</td>
<td>A: 0 (0)</td>
</tr>
</tbody>
</table>

a Data are n (%). Each of the 3 FFPE DNAs was examined for KRAS and AKT1 SNCs before and after UDG treatment. The number of reads for each base call at the SNCs was determined from the IGV program.
We counted the allelic calls for C:G>T:A changes at clinically relevant positions in the KRAS and AKT1 genes [KRAS G>A changes at c.34, c.35, c.38, and c.39 (codons 12 and 13)] and AKT1 c.49G>A (E17K mutation) using the Integrative Genomics Viewer. Most of the G>A changes at these positions were found at <1% (Table 3). However, in SCC4, apparent KRAS G>A changes were detected at 2.1% (c.35) and 2.7% (c.37). UDG treatment resulted in consistent reductions of those G>A SNCs, bringing the frequency of those SNCs to 0.7% and 0.1%, respectively.

Thus, artificial SNCs at clinically important positions can be present at above the current consensus MPS error rate of 1%. Our results clearly show the benefit of UDG treatment in eliminating artificial SNCs that mimic such potentially clinically important mutations. These results are also pertinent to the important issue of determining the role of tumor heterogeneity in the development of resistance mutations. The results indicate that one would need to be very careful in drawing any conclusions from FFPE samples and that fresh frozen tissue is desirable for tumor heterogeneity studies.

GENERAL APPLICABILITY OF OUR FINDINGS
It is likely that samples will vary extensively in their intrinsic deamination. We have seen very strong batch effects, in which samples from a given source tend to resemble each other in the degree of damage. For this study, we chose samples with a moderate degree of damage.

It is important to realize that the results will depend on the sequencing approach used. The critical differences are due to the properties of the DNA polymerase used and whether it can read through uracil and/or through abasic sites. In our case, the enzyme used in the TruSeq platform will read through uracil but does not appear to read through the abasic sites that are generated by UDG treatment. We have also shown that this is the case for Qiagen HotStarTaq (13).

In conclusion, the results presented here show that UDG pretreatment of FFPE DNA before PCR amplification can markedly reduce artificial SNCs in amplicon-based MPS, which has major implications for the detection of somatic mutations for personalized cancer treatment. Detection of true mutations was not affected by UDG treatment. Thus, UDG pretreatment is a simple and effective strategy for reduction of artificial C:G>T:A SNCs in amplicon-based, targeted deep sequencing of FFPE DNA. We consider UDG pretreatment to be a simple solution that should be universally applied in amplicon-based protocols.

Author Contributions: All authors confirmed they have contributed to the intellectual content of this paper and have met the following 3 requirements: (a) significant contributions to the conception and design, acquisition of data, or analysis and interpretation of data; (b) drafting or revising the article for intellectual content; and (c) final approval of the published article.

Authors’ Disclosures or Potential Conflicts of Interest: Upon manuscript submission, all authors completed the author disclosure form. Disclosures and/or potential conflicts of interest:

Employment or Leadership: None declared.
Consultant or Advisory Role: None declared.

Stock Ownership: None declared.

Research Funding: H. Do, Postdoctoral Fellowship from the Cancer Council of Victoria; A. Dobrovic, Cancer Council of Victoria, Cancer Australia, National Health and Medical Research Council (Australia).

Expert Testimony: None declared.

Role of Sponsor: The funding organizations played no role in the design of study, choice of enrolled patients, review and interpretation of data, or preparation or approval of manuscript.

Acknowledgments: We thank Paul Mitchell and Carmel Murone for the lung SCC samples. We also thank Andrew Fellowes, Anthony Bell, and Stephen Fox for the TruSeq reagent and access to the MiSeq instrument.

References

12. Do H, Dobrovic A. Limited copy number-high reso-
olution melting (LCN-HRM) enables the detection and identification by sequencing of low level mutations in cancer biopsies. Mol Cancer 2009;8:82.


