Reliable Next-Generation Sequencing of Formalin-Fixed, Paraffin-Embedded Tissue Using Single Molecule Tags

Astrid Eijkelenboom,* Eveline J. Kamping,1 Annemiek W. Kastner-van Raaij,* Sandra J. Hendriks-Cornelissen,* Kornelia Neveling,1 Roland P. Kuiper,1 Alexander Hoischen,12 Marcel R. Nelen,1 Marjolijn J.L. Ligtenberg,* and Bastiaan B.J. Tops*

From the Departments of Pathology* and Human Genetics1 and the Donders Centre for Neuroscience,1 Radboud University Medical Center, Nijmegen, the Netherlands

CME Accreditation Statement: This activity (“JMD 2016 CME Program in Molecular Diagnostics”) has been planned and implemented in accordance with the Essential Areas and policies of the Accreditation Council for Continuing Medical Education (ACCME) through the joint providership of the American Society for Clinical Pathology (ASCP) and the American Society for Investigative Pathology (ASIP). ASCP is accredited by the ACCME to provide continuing medical education for physicians. The ASCP designates this journal-based CME activity (“JMD 2016 CME Program in Molecular Diagnostics”) for a maximum of 36 AMA PRA Category 1 Credit(s)™. Physicians should claim only credit commensurate with the extent of their participation in the activity.

CME Disclosures: The authors of this article and the planning committee members and staff have no relevant financial relationships with commercial interests to disclose.

Accepted for publication
June 8, 2016.

Address correspondence to Bastiaan Tops, Ph.D., Radboud University Medical Center, PO Box 9101, 6500 HB, Nijmegen, the Netherlands. E-mail: bastiaan.tops@radboudumc.nl.

Sequencing of tumor DNA to detect genetic aberrations is becoming increasingly important, not only to refine cancer diagnoses but also to predict response to targeted treatments. Next-generation sequencing is widely adopted in diagnostics for the analyses of DNA extracted from routinely processed formalin-fixed, paraffin-embedded tissue, fine-needle aspirates, or cytologic smears. PCR-based enrichment strategies are usually required to obtain sufficient read depth for reliable detection of genetic aberrations. However, although the read depth relates to sensitivity and specificity, PCR duplicates generated during target enrichment may result in overestimation of library complexity, which may result in false-negative results. Here, we report the validation of a 23-gene panel covering 41 hotspot regions using single-molecule tagging of DNA molecules by single-molecule molecular inversion probes (smMIPs), allowing assessment of library complexity. The smMIP approach outperforms Sanger and Ampliseq-Personal Genome Machine—based sequencing in our clinical diagnostic setting. Furthermore, single-molecule tags allow consensus sequence read formation, allowing detection to 1% allele frequency and reliable exclusion of variants to 3%. The number of false-positive calls is also markedly reduced (>10-fold), and our panel design allows for distinction between true mutations and deamination artifacts. Not only is this technique superior, smMIP-based library preparation is also scalable, easy to automate, and flexible. We have thus implemented this approach for sequence analysis of clinical samples in our routine diagnostic workflow. (J Mol Diagn 2016, 18: 851–863; http://dx.doi.org/10.1016/j.jmoldx.2016.06.010)

The availability and requirement of molecular therapeutics in routine cancer treatment has greatly increased over the past decade. Combined with the stratification of patients amenable for targeted therapeutics based on the absence or presence of specific genomic aberrations, this has increased the requirement for genomic profiling of tumor specimens.1 Next-generation sequencing (NGS) allows for genomic characterization in a sensitive manner.2 Although whole genome or exome sequencing both provide extensive genomic information, targeted gene panels are currently best fit for tumor profiling in a routine clinical context. It
matches best with the billable costs, short turn-around time (TAT), and requirement for reliable detection of variants compatible with routinely obtained material [formalin-fixed, paraffin-embedded (FFPE) tissue specimens, fine-needle aspirates, and cytologic smears].\textsuperscript{3–5} This requires sufficient read depth to detect or exclude low-frequent variants that might be present due to tumor heterogeneity or a low tumor load in the tissue specimen and a robust and reliable bio-informatics pipeline. PCR-based amplification is commonly used to generate such targeted sequencing libraries for NGS.\textsuperscript{6,7} Multiple genomic regions can be amplified simultaneously (multiplex PCR) for analysis of multiple genes from limited tissue material with low-quality genomic DNA (gDNA), such as the FFPE samples routinely used in molecular diagnostics. However, as a consequence of this amplification, true library complexity cannot be determined, because PCR duplicate reads cannot be distinguished from independent reads originating from separate original template molecules. This could result in overestimation of the actual number of DNA molecules analyzed, risking false-negative calls, which is crucial in the context of poor-quality samples with a small amount of amplifiable DNA. Single-molecule tagging (SMT) has been developed to overcome this issue by marking PCR duplicates originating from the same template molecule, which allows both a genuine analysis of library complexity and the possibility to combine multiple sequencing reads from PCR duplicates to generate a single consensus read.\textsuperscript{8–10} The latter also allows filtering for errors originating during library amplification and sequencing that might result in false-positive calls due to jackpotting events.

We sought to develop an NGS-based targeted approach in a routine diagnostics setting for reliable detection of clinically relevant variants in tissue specimens from FFPE specimens, in which multiplex analysis could be combined with SMT technology. Recently, multiplex analysis and SMT technology have been combined in single-molecule molecular inversion probes (smMIPs) to detect low-frequent variants in FFPE-derived DNA isolates in a simple, scalable, and relatively cost-effective manner.\textsuperscript{11} In addition, the strand-specific nature of amplification by smMIPs can aid to distinguish genuine C:G\textrightarrow:T:A mutations from those induced by cytosine deamination, a common artifact in DNA recovered from FFPE-fixed material.\textsuperscript{12} Here, we describe the development, validation, and implementation of a single comprehensive smMIP-based Cancer Hotspot Panel (CHP) for mutation detection in clinically relevant genes.

**Materials and Methods**

**Sample Preparation**

Clinical specimens (generally from sections $3 \times 20 \mu m$) were digested at $56^\circ C$ for at least 1 hour in the presence of TET-lysis buffer (10 mmol/L Tris/HCl pH8.5, 1 mmol/L EDTA pH8.0, 0.01% Tween-20) with 5% Chelex-100 (143 to 2832; Bio-Rad, Hercules, CA), 15 µg/mL GlycoBlue (AM9516; Thermo Fisher, Waltham, MA), and 400 µg proteinase K (19133; Qiagen, Valencia, CA), followed by inactivation at 95°C for 10 minutes. The supernatant was transferred after centrifugation to reduce the total volume for the robotized protocol, cooled on ice and precipitated in the presence of 70% EtOH and 1/10 volume 3M NaAc (pH 5.2). Pellets were washed with cold 70% EtOH and dissolved in 80 µL Tris-EDTA, and DNA concentration was determined using the Qubit Broad Range Kit (Q32853; Thermo Fisher). Control DNA isolated from peripheral blood leukocytes was sonicated using a Covaris with a standard protocol to obtain 200-bp fragments and analyzed on a 1% agarose gel with 100-bp size ladder (15628-050; Invitrogen, Carlsbad, CA). The control NGS sample was obtained from Horizon Discovery (Waterbeach, UK; HD701).

**Preparation of the smMIP-Pool for Targeted Enrichment**

MIPs were designed using the procedure described\textsuperscript{13} for all hotspots (Table 1), in a tiling manner preferentially covering all hotspots with two independent smMIPs targeting both DNA strands. The sum of the targeting arms is 40 bp (extension plus ligation probe arms) and the gap-fill length was set to 112 bp. The targeting arms were joined by a common backbone sequence and a stretch of $8 \times N$ nucleotides was inserted between the backbone and ligation probe sequence (Supplemental Table S1). In case it was unavoidable to design smMIPs without common single nucleotide polymorphism variants in the respective capture arms, smMIPs were designed recognizing both alleles. Allelots of each oligonucleotide smMIP probe (produced by Integrated DNA Technologies, Leuven, Belgium) were mixed in an equimolar or corrected fashion (Supplemental Table S1) to form the CHP smMIP-pool. The smMIP pool was phosphorylated with 1 µL of T4 polynucleotide kinase (M0201; New England Biolabs, Ipswich, MA) per 25 µL of 100 µmol/L smMIPs and ATP-containing T4 DNA ligase buffer (B0202; New England Biolabs). The molecular ratio between gDNA and smMIPs was set to 1:3200 for every individual smMIP (and is thus independent of pool size), and the required quantity of the smMIP pool was determined for a standard input of 100 ng gDNA.

**Library Preparation**

In manual experiments, a total of 100 ng of genomic DNA was used as input in a 20-µL volume, unless otherwise specified, with a total capture volume of 25 µL, including the (diluted) phosphorylated smMIP pool, 1 unit of Ampligase DNA ligase (A0110K; EpiBio, Madison, WI) with Ampligase Buffer (A1905B, DNA ligase buffer), 3.2 units of Hemo Klentaq (M0332; New England Biolabs), and 8 µmol of dNTPs (28-4065-20/-12/-22/-32; GE Healthcare, Little Chalfont, UK). After denaturation (95°C for 10
minutes) the mix was incubated for probe hybridization, extension, and ligation at 60°C for 18 hours and cooled before exonuclease treatment. Exonuclease I (10 units; M0293; New England Biolabs) and III (50 units; M0206; New England Biolabs) and Ampligase Buffer (see above) were added to the capture volume, adding up to a total of 27 µL, and incubated for 45 minutes at 37°C, followed by inactivation at 95°C for 2 minutes. A total of 10 µL of the exonuclease-treated capture was used for PCR in a total volume of 25 µL with 25 nmol common forward primer and barcoded reverse primers and iProof high-fidelity master mix (1725310; Bio-Rad). The resulting PCR products were pooled before purification with 0.8 x volume of Agencourt Ampure XP Beads (A63881; Beckman Coulter, Brea, CA).

Semiautomated library preparation was performed as described (K. Neveling, A.R. Mensenkamp, R.C. Derks, et al, unpublished data). In short, hybridization reactions were pipetted using a Microlab STARplus robot (Hamilton, Reno, NV). The exonuclease treatment was performed manually to prevent exonuclease contamination in the

### Table 1: An Overview of All Regions Targeted by the smMIP Panel, Including Gene Name, Transcript IDs from Refseq and Ensembl, and Targeted Regions

<table>
<thead>
<tr>
<th>Gene</th>
<th>Transcript ID (RefSeq)</th>
<th>Transcript ID (Ensembl)</th>
<th>Exon number</th>
<th>Targeted codons</th>
<th>Positions analyzed for variants</th>
</tr>
</thead>
<tbody>
<tr>
<td>AKT1</td>
<td>NM_005163</td>
<td>ENST000000555528</td>
<td>03</td>
<td>E17</td>
<td>c.47-5 to c.86</td>
</tr>
<tr>
<td>BRAF</td>
<td>NM_004333</td>
<td>ENST00000288602</td>
<td>15</td>
<td>D594-K601</td>
<td>c.1742-5 to c.1845</td>
</tr>
<tr>
<td>CTNNB1</td>
<td>NM_001904</td>
<td>ENST00000349496</td>
<td>03</td>
<td>D32-S45</td>
<td>c.53 to c.146</td>
</tr>
<tr>
<td>EGFR</td>
<td>NM_005228</td>
<td>ENST00000275493</td>
<td>12</td>
<td>S492</td>
<td>c.1392 to c.1408+5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>18</td>
<td>E709-G719</td>
<td>c.2062-5 to c.2184+5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>19</td>
<td>All codons (G729-D761)</td>
<td>c.2185-5 to c.2283+5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>20</td>
<td>All codons (E762-K823)</td>
<td>c.2284-5 to c.2469+5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>21</td>
<td>L858-L861</td>
<td>c.2245 to c.2630</td>
</tr>
<tr>
<td>ERBB2</td>
<td>NM_004448</td>
<td>ENST00000269571</td>
<td>20</td>
<td>Y772-Y781</td>
<td>c.2308-1 to c.2357</td>
</tr>
<tr>
<td>GNA11</td>
<td>NM_002067</td>
<td>ENST00000078429</td>
<td>04</td>
<td>R183</td>
<td>c.495 to c.582</td>
</tr>
<tr>
<td>GNAQ</td>
<td>NM_002072</td>
<td>ENST00000286548</td>
<td>04</td>
<td>R183</td>
<td>c.540 to c.605+5</td>
</tr>
<tr>
<td>GNAS</td>
<td>NM_000516</td>
<td>ENST00000371085</td>
<td>08</td>
<td>R201</td>
<td>c.587 to c.635</td>
</tr>
<tr>
<td>H3F3A</td>
<td>NM_002107</td>
<td>ENST00000366815</td>
<td>02</td>
<td>K28, G35</td>
<td>c.49 to c.128+5</td>
</tr>
<tr>
<td>H3F3B</td>
<td>NM_005324</td>
<td>ENST00000254810</td>
<td>02</td>
<td>K37</td>
<td>c.38 to c.128+5</td>
</tr>
<tr>
<td>HRAS</td>
<td>NM_005343</td>
<td>ENST00000451590</td>
<td>02</td>
<td>G12, G13</td>
<td>c.17 to c.83</td>
</tr>
<tr>
<td>IDH1</td>
<td>NM_005896</td>
<td>ENST00000345146</td>
<td>04</td>
<td>R132</td>
<td>c.351 to c.410</td>
</tr>
<tr>
<td>IDH2</td>
<td>NM_002168</td>
<td>ENST0000030062</td>
<td>04</td>
<td>R140, R172</td>
<td>c.409 to c.518</td>
</tr>
<tr>
<td>JAK2</td>
<td>NM_004972</td>
<td>ENST00000381652</td>
<td>14</td>
<td>V617</td>
<td>c.1780 to c.1864+5</td>
</tr>
<tr>
<td>KIT</td>
<td>NM_000222</td>
<td>ENST00000288135</td>
<td>08</td>
<td>All codons (K412-Q448)</td>
<td>c.1232-5 to c.1346+5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>09</td>
<td>All codons (C450-K513)</td>
<td>c.1347-5 to c.1540+5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>11</td>
<td>All codons (K550-F591)</td>
<td>c.1648-5 to c.1774+5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>13</td>
<td>K642-N655</td>
<td>c.1880-1 to c.1990+5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>14</td>
<td>All codons (P665-S713)</td>
<td>c.1991-5 to c.2141+5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>17</td>
<td>D816-Y823</td>
<td>c.2395 to c.2484+5</td>
</tr>
<tr>
<td>KRAS</td>
<td>NM_004985</td>
<td>ENST00000311936</td>
<td>02</td>
<td>G12, G13</td>
<td>c.9 to c.71</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>03</td>
<td>A59, G61</td>
<td>c.122 to c.215</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>04</td>
<td>K117, A146</td>
<td>c.291-5 to c.357 and c.402 to c.450+5</td>
</tr>
<tr>
<td>MPL</td>
<td>NM_005373</td>
<td>ENST00000372470</td>
<td>10</td>
<td>W515</td>
<td>c.1526 to c.1565+5</td>
</tr>
<tr>
<td>MYD88</td>
<td>NM_002468</td>
<td>ENST000003696334</td>
<td>05</td>
<td>L265</td>
<td>c.776-5 to c.840</td>
</tr>
<tr>
<td>NRAS</td>
<td>NM_002524</td>
<td>ENST00000369535</td>
<td>02</td>
<td>G12, G13</td>
<td>c.-17-5 to c.64</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>03</td>
<td>A59, G61</td>
<td>c.161 to c.245</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>04</td>
<td>K117, A146</td>
<td>c.312 to c.450+5</td>
</tr>
<tr>
<td>PDGFRα</td>
<td>NM_006206</td>
<td>ENST00000257290</td>
<td>12</td>
<td>All codons (K552-L595)</td>
<td>c.1654-5 to c.1786+5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>14</td>
<td>All codons (T632-S667)</td>
<td>c.1892-5 to c.2002+5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>18</td>
<td>V824-D842</td>
<td>c.2440-5 to c.2544</td>
</tr>
<tr>
<td>PIK3CA</td>
<td>NM_006218</td>
<td>ENST00000263967</td>
<td>10</td>
<td>E542-Q546</td>
<td>c.1558 to c.1664+5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>21</td>
<td>M1043-G1049</td>
<td>c.3058 to c.3207+10</td>
</tr>
</tbody>
</table>

ID, identification; smMIP, single-molecule molecular inversion probe.
pre- and post-PCR environments. The post-hybridization PCR was performed using a MicroLab STARlet Replicator Robot (Hamilton). All samples were subsequently pooled on a MicroLab Starlet Replicator Robot (Hamilton). The automated workflow contained the following modifications compared with the manual protocol above: the maximum input volume of gDNA was reduced to 7 μL to allow reliable pipetting of other reagents, requiring precipitation of gDNA to accommodate for this smaller input volume. Samples with concentrations >30 ng/μL were diluted to obtain a final concentration within the 15 to 30 ng/μL range, resulting in approximately 100 to 200 ng of gDNA input per smMIP capture. A total of 20 μL from the exonuclease-treated capture mixture was used for PCR in a total volume of 50 μL. Note that the gDNA:smMIP ratio as previously described (K. Neveling, A.R. Mensenkamp, R.C. Derks, et al, unpublished data) was set to 800:1 for high-quality blood-derived gDNA, whereas 3200:1 was used for the FFPE-derived DNA from clinical specimens in the present study.

Sequencing

The purified libraries were denatured and diluted to a concentration of 1.2 pmol/L. Sequencing was performed on a NextSeq500 instrument (Illumina, San Diego, CA) according to the manufacturer’s protocol (300 cycles Mid Output sequencing kit, v2), resulting in 2 × 150 bp paired-end reads.

Data Analysis

Using the sample sheet, Bcl to fastq conversion and demultiplexing of barcoded reads was performed automatically (K. Neveling, A.R. Mensenkamp, R.C. Derks, et al, unpublished data) and uploaded to a server running the commercial analysis software (Sequence Pilot version 4.2.0; JSI medical systems, Ettenheim, Germany). Single-molecule-directed assembly of duplicate reads to generate consensus reads after alignment was performed in a package developed in collaboration with JSI medical systems (from version 4.2.0, build 503 and onward). The following settings were used for single-molecule-directed consensus: Tags active, yes; R1 tag length, 8; R2 tag length, 0; Min abs. cov. cons., 1; Min per. cov. cons., 50%; Ignore cons. read thresh., 30; Ignore N tags, yes, and Ignore low Qs tags, yes. Proper identification of sequence variants greatly relies on the variables for variant calling, such as the minimal number of mutant unique smMIPs (≥2 in our settings) and the minimal mutant allele frequency (≥1%). The following settings were used for variant calling using Sequence Pilot: Required Coverage/Min abs. cov., 40 combined; Mutations/Min abs. cov., 5 combined; Min % cov., 1% per dir. Note that one unique smMIP generates both a forward and a reverse read, thus providing a read depth of 2 reads using this software package. PIK3CA pseudogene reads were removed from the alignment and subsequent analysis. After variant calling using the commercial software, all variants were manually inspected and curated. GraphPad Prism 5 (GraphPad Inc., San Diego, CA) was used for graphical presentation. All smMIP analyses were performed in duplicate, resulting in similar results, of which one representative is shown.

Routine Sequencing Method for Parallel Analysis

The smMIP approach was compared with our previous routine sequencing approach, which consists of Sanger sequencing for CTNNB1, GNAS, IDH1, IDH2, and MYD88 (Supplemental Table S1) and Ampliseq-based NGS using the Personal Genome Machine (PGM; Thermo Fisher) for relevant codons in AKT1, BRAF, EGFR, ERBB2 (HER2), KRAS, NRAS, and PIK3CA (colon carcinoma/non-small cell lung carcinoma custom panel) or AKT1, BRAF, GNA11, GNAQ, HRAS, KIT, NRAS, and PDGFRα (melanoma/gastrointestinal stromal tumor custom panel) (Supplemental Table S1) according to the manufacturer’s protocol. Data were analyzed with the same commercial software (Sequence Pilot; JSI medical systems) and similar settings for variant calling, with the exception of the single-molecule-directed building of consensus reads.

Results

Validation Plan

We aimed to design a single CHP of smMIPs targeting all clinically relevant coding sequences in 23 genes, covering a total of 41 hotspot regions and surrounding sequences (adding up to a total of approximately 4.0 kb), to replace all our routine Ampliseq/PGM NGS approaches and Sanger-based sequencing analyses of oncogenes required for cancer diagnoses and/or predictive diagnostics (Table 1). We formulated the procedure and requirements before we started the implementation and validation, according to our quality guidelines. The following requirements were defined: i) >95% of all hotspot regions should be covered by a minimum of two independent smMIPs, preferably targeting both DNA strands; ii) analysis of all 41 hotspot regions should be performed in one single reaction, with uniform read depth distribution (the average read depth for >90% of targets may vary up to 1 order of magnitude); iii) the detection of variants should be comparable or better compared with our routine diagnostic Ampliseq/PGM and Sanger approach; iv) while the number of false-positive variant calls, and v) the number of sample drop-outs should be comparable or smaller; vi) in a routine diagnostic setting TAT of >90% of clinical requests should be ≤7 working days.

smMIP Panel Design

Targeted enrichment on isolated DNA using smMIPs relies on a three-step protocol, which can be completed within 24
hours (Figure 1, A and B), followed by sequencing using the NextSeq500 sequencer. The sequence reads are subsequently aligned, and PCR duplicates are joined to form a consensus when marked by an identical single-molecule tag (Figure 1C), which results in a unique read depth excluding PCR duplicates and provides a measure for library complexity. The single-molecule tag of eight random nucleotides can uniquely mark 65,536 molecules ($4^8$) per amplicon, which roughly equals a total of 200 ng of gDNA. To prevent allelic drop out due to genomic variation in the smMIP binding sites, smMIPs were designed to probe the target region in a tiling manner, preferably in both orientations, thus capturing both strands. Common single nucleotide polymorphisms in genomic sequences recognized by the probes were taken into account by designing two smMIPs recognizing either the major or the minor allele. Finally, smMIPs targeting X chromosomal AMELX and Y-chromosomal AMELY genes to provide a sex control were also included, which resulted in a panel of 120
smMIPs (Supplemental Table S1), targeting all relevant codons, with double tiling of 39 of 41 hotspot regions (95.1%), thus fulfilling requirement i.

smMIP Panel Performance

To evaluate the performance of our design, all smMIPs were pooled in an equimolar fashion and tested using archived gDNA isolated from four peripheral blood samples and 10 clinical FFPE specimens with DNA concentrations to 3 ng/mL (Supplemental Table S2). All targeted regions were indeed captured and sequenced, although with varying efficiency within and between samples (Figure 2A). Overall, the average read depth of all hotspot regions is within 1 order of magnitude for all FFPE samples combined (Figure 2B), and this also holds true for the 5% to 95% range of all hotspot regions for 8 of 10 FFPE-derived samples, with the two exceptions being the oldest samples, of which DNA has been stored at −20°C for 2 and 4 years, respectively (Figure 2A).

When testing the same smMIP pool on high-quality gDNA isolated from peripheral blood, a subset of smMIPs captured their targets less efficiently relative to the total smMIP pool. This was most prevalent in a subset of smMIPs with the highest guanine-cytosine content and can possibly be attributed to the fragmented nature of FFPE-isolated gDNA because it is corrected by fragmentation (Supplemental Figure S1). Although all hotspot regions are sufficiently covered in FFPE- and peripheral blood-derived gDNA, we attempted to boost the performance of 10 underperforming smMIPs by increasing their concentration in the smMIP pool. This indeed resulted in a significant increase in unique read depth for 5 of 10 smMIPs. Taken together, these data show that uniform sequence analysis of all 41 hotspot regions using smMIPs can be performed in a single capture reaction on gDNA of both poor and high quality (requirement ii).

We subsequently assessed if all types of clinically relevant variants are detected using the smMIP-based enrichment. Because the smMIPs target regions of exactly 112 bp, we anticipated that insertions or deletions (>10 bp) might be challenging to capture using smMIPs. We therefore

Figure 2  Single-reaction sequence analysis of all 41 hotspot regions using smMIPs. A: Sequencing libraries were prepared using an equimolar pool of all smMIPs on 10 clinical FFPE-derived gDNA samples (Supplemental Table S2), followed by next-generation sequencing. Consensus reads were generated and the minimal unique read depth at the relevant codons of all 41 hotspot regions was determined. B: The median unique read depth was set to 1 to correct for differences in sequencing depth per sample. Subsequently, the average median-corrected unique read depth of all FFPE samples was calculated for all hotspot regions. The graph depicts the normalized unique read depth per hotspot sorted by performance. Data are expressed as box plots with 5th to 95th percentile whiskers with the ratio between the 95th and 5th percentile values above each box plot (A). FFPE, formalin-fixed, paraffin-embedded; gDNA, genomic DNA; smMIP, single-molecule molecular inversion probe.
specifically tested 20 archived DNA specimens that were previously shown to harbor duplications (up to 42 bp) and deletions (up to 63 bp) in our routine diagnostic analysis. All 20 variations previously detected in these samples by other methods were confirmed using the smMIP analysis (Supplemental Table S3). Furthermore, during the process of implementation and validation we analyzed >200 samples using smMIPs parallel to Sanger sequencing or Ampliseq-based sequencing. During this period we verified all mutations, that is, a variety of substitutions, deletions, insertions, and complex rearrangements, using smMIPs (Figure 3A, Supplemental Table S3). Combined, our data show that clinically relevant genetic variants identified by different sequencing methods (Sanger-based and AmpliSeq-PGM–based) are also identified by the smMIP strategy (requirement iii). These include point mutations, duplications (up to 42 bp), deletions (up to 63 bp), and more complex deletion-insertion mutations.

Although these initial test variables were established using a manual protocol, further testing was performed in parallel with the implementation of a fully automated library preparation as described for blood-derived gDNA (K. Neveling, A.R. Mensenkamp, R.C. Derks, et al, unpublished data; with some adaptations described in Materials and Methods). Advantages of an automated strategy are that it facilitates library preparation of a larger number of specimens and minimizes manual errors during sample handling. In addition, a bioinformatics pipeline was implemented to automatically initiate data analyses on availability of raw sequencing data. The robotized smMIP strategy was first assessed for the sensitivity of mutation detection using a commercially available sample (Horizon Discovery standard) harboring 11 different mutations with varying allele frequencies. All of these variants were detected with nearly identical mutant allele frequencies as reported (Figure 3B and Supplemental Table S4). In addition, DNA isolated from two FFPE samples containing two different mutations were mixed in different ratios, thereby diluting the mutant alleles. The detected allele frequencies at different mix ratios nicely agree with the calculated allele frequencies (Figure 3C), demonstrating the accuracy and linear range in the detection of mutant alleles by the smMIP approach.

Dilation experiments to establish the optimal and minimal amounts of required input DNA for smMIP-based
sequencing showed an optimum in obtained unique read depth between 100 and 200 ng of input DNA (data not shown). As expected, serial dilutions of two FFPE-derived DNA samples in water showed a decrease in the number of captured unique (mutant) reads with decreasing amounts of total input DNA, indicating that the absolute amount of gDNA is limiting. However, although the amount of input DNA is decreased (to 19 ng of total input DNA in our protocol, which equals 2.7 ng/mL), the variants are consistently detected in these dilutions, with identical allele frequencies (Figure 3D). This demonstrates that smMIP-based sequencing is robust and also compatible with samples with low DNA concentrations (<10 ng/mL). It also shows that library complexity is proportionate to the amount of input DNA.

Validation

Validation of the performance of the smMIP-based enrichment, coupled with NextSeq 500-based sequencing (referred to as the smMIP-NextSeq500 approach), including an automated protocol in the routine diagnostic workflow, was performed in parallel to our routine NGS approach (the AmpliSeq-PGM strategy) and Sanger-based sequencing analysis. These methods were tested in parallel for 6 weeks, mainly to evaluate the smMIP strategy in the routine diagnostic workflow. All mutations identified in the Ampliseq-PGM/Sanger diagnostic setting (total of 42) were also identified using the smMIP-NextSeq500 approach. Two additional mutations were detected using the smMIPs for which the respective genes were not sequenced by our

Figure 4  False-positive variants in sequencing analysis. A: The number of false-positive variants per sample with an allele frequency ≥5% as identified by parallel sequencing analysis is depicted for both the Ampliseq-PGM and the smMIP-NextSeq500 approach. False-positive variants are defined as all variants excluding true positive variants, that is, known pathogenic variants (generally identified by both methods) and known nonpathogenic variants with established minor allele frequencies. Only samples with sufficient read depth for variant allele detection of all targets in single analysis using both next-generation sequencing methods were included. B: The number of false-positive variants per read as identified by the smMIP approach is determined for the previously described set of 10 FFPE gDNA samples (Figure 2 and Supplemental Table S2), either with and without the SMT feature used during data analysis. The bars thus indicate the number of false-positive variants per unique read (= SMT derived consensus) and all sequence reads, respectively. Above the bars the consequence of SMT usage depicted by a fold decrease in false-positive variant frequency. C: The strand-specific nature of target recognition by smMIPs is represented in the left illustration, with smMIPs recognizing either the forward (+ smMIPs) or the reverse DNA strand (− smMIPs). Five examples of transitions identified in clinical samples are depicted, including the affected gene, mutated position, and associated consequence for amino acid sequence. The percentage above the line represents the mutant allele frequency as detected by the + smMIPs, the percentage below represents the allele frequency from the − smMIP analysis. Because cytosine deamination induces sequencing artifacts only in the cytosine-containing strand, the imbalance can discriminate genuine C:G > T:A mutations (on the right) from artifacts (on the left). ****P < 0.0001. FFPE, formalin-fixed, paraffin-embedded; gDNA, genomic DNA; PGM, Personal Genome Machine; smMIP, single-molecule molecular inversion probe; SMT, single molecule tagging.
Speciation modes using SMT-derived unique read depth. Deletion in KIT (Figure 3E), which could be due to preferential amplification in the Ampliseq-PGM strategy. In addition, the results from the sex controls were fully concordant (data not shown). These data, together with the other sequencing data obtained during implementation (see above), show that the smMIP approach is at least as sensitive as the Ampliseq-PGM and Sanger-based sequencing approaches (requirement iii).

Although the detection of clinically relevant mutations is comparable, the number of false-positive variants with an allele frequency >5% is significantly reduced in the smMIP-NextSeq500 approach (on average >10-fold reduction, \( P < 0.0001 \)) (Figure 4A), although the total sequenced target region is >67% and 154% larger compared with the two Ampliseq panels. The false-positive rate and precision (Table 2) clearly show that the smMIPs-NextSeq500 approach outperforms the Ampliseq-PGM strategy, thereby meeting requirement iv. The near lack of false positives in the smMIP protocol is probably due to a combination of factors and may partly be attributed to the difference in sequencing technologies (Ion Torrent versus Illumina), but may also be due to the merging of PCR duplicates (sequence reads with identical SMTs) into consensus reads, thereby largely eliminating PCR and sequencing artifacts. As expected, consensus building using the SMT feature reduces the number of false-positive variants per read by approximately eightfold (Figure 4B). This clearly demonstrates the enormous advantage of SMTs in library preparation and data analysis.

Because of the relatively high level of false-positive calls with an allele frequency of 1% to 5%, the Ampliseq-PGM approach does not allow for the reliable detection of variants <5%. Because the SMTs in the smMIP-NextSeq500 strategy greatly reduce the number of false-positive variants, we analyzed whether this strategy is suitable to detect pathogenic variants in the lower frequency range. With the use of the smMIP procedure, the total number of identified variants

| Table 2 Specifications of the Ampliseq-PGM Strategy and the smMIP-NextSeq 500 Approach |
|-----------------|------------------|------------------|------------------|
| Criterion       | Ampliseq-PGM     | smMIP-NextSeq 500| Definitions and formulas |
| TP              | 98               | 194              | Variant calls: mutations and SNPs |
| FP              | 171              | 6                | FP calls |
| TN              | 98186            | 204185           | bp identical to reference genome |
| FN              | 0                | 0                | Missed variants |
| Sensitivity     |                   |                  |                             |
| TP rate         | 100.000%         | 100.000%         | \( TP/(TP+FN) \) |
| FP rate         | 0.175%           | 0.003%           | \( FP/(FP+TN) \) |
| Accuracy        | 99.826%          | 99.997%          | \( (TP+TN)/(TP+TN+FP+FN) \) |
| Precision       | 36.431%          | 97.000%          | \( TP/(TP+FP) \) |
| Specificity     |                   |                  |                             |
| TN rate         | 99.826%          | 99.997%          | \( TN/(FP+TN) \) |

The number of TP and FP variants was determined for both methods, as well as the number of base pairs identical to the reference genome (TNs) and number of variants missed when comparing both methods (FNs). Sensitivity and specificity variables were calculated with the formulas indicated in the right column.

FN, false negative; FP, false positive; PGM, Personal Genome Machine; smMIP, single-molecule molecular inversion probe; SNP, single nucleotide polymorphism; TN, true negative; TP, true positive.

Figure 5 Minimal read depth requirements using SMT-derived unique read depth. A: The graph depicts the cumulative binomial distribution to fulfill the minimal criteria for mutation calling (≥1% mutant allele frequency and ≥3 mutant unique reads), depending on a certain unique read depth (x axis). Each line indicates the probability to fulfill the variant calling criteria (the probability of success) with a fixed mutant allele frequency in the specimen using increasing read depth (the number of trials). The dashed line depicts the 95% probability level. B: Thresholds for the minimal number of unique smMIPs (represented by unique reads) covering a locus to reliably (approximate probability of a false-negative call <5%) exclude the presence of a mutation. The read depth levels are binned, and matched lower limits of mutation detection are shown per bin (eg, the read depth is too low to exclude the presence of a mutation). ND, not determined; smMIP, single-molecule molecular inversion probe; SMT, single molecule tagging.
in the lower frequency range (1% to 5%) is ≤1 in half of the samples tested (Supplemental Figure S2). In a subset of samples we identified a large number of variants in the 1% to 5% range. Nearly all variants are C>T transitions (10 samples in our series of 67 samples contain ≥10 C>G>T:A variants), that reflect formalin-induced deamination of cytosine residues and are therefore only detected with either the sense or antisense targeting smMIP (Figure 4C). These data show that it is also feasible to confidently detect low-level mutations (1% to 5%) using the smMIP strategy, as previously shown by Hiatt et al.11

There is an additional important advantage of the SMTs. When sequence analyses include PCR duplicates, this may result in false-negative results due to overestimation of the actual number of analyzed molecules. This is especially relevant in the context of poor-quality FFPE-derived gDNA. The use of SMTs to generate read depth in unique reads thus provides an exact measure of the number of analyzed input DNA molecules (library complexity). With the use of the binomial distribution, we calculated the probability to identify a mutation, depending on the unique read depth and an assumed mutant allele frequency in the specimen (Figure 5A). These probabilities provide an approximation of the minimum read depth thresholds to exclude mutations above a certain allele frequency, with a certainty of >95% (Figure 5B).

We used these minimum read depth levels to determine whether sufficient DNA molecules were sequenced to reliably analyze the hotspots relevant for the specific clinical request in our series of 67 samples. These data were also interpreted in the context of estimated tumor load (percentage of neoplastic cells) assessed by a pathologist. We performed this analysis for all hotspot regions relevant for the specific clinical request and generated a full report, which was

---

**Figure 6** Validation of the smMIP approach by parallel sequence analysis of 67 clinical samples in a clinical diagnostic setting over a period of 6 weeks. A: An overview of our routine diagnostic sequence analysis flow of cancer hotspot regions using sequence analysis based on Ampliseq-PGM (a total of two panels) or Sanger (five different amplicons), of which one is executed depending on the specific clinical request, and the smMIP-NextSeq500–based single-reaction set-up for all clinical requests. In Ampliseq-PGM and Sanger strategy, all steps are performed manually, whereas the smMIP-based diagnostics is semi-automated for library preparation and automated for variant detection on NGS analysis (indicated in orange). Clinical samples are batched and analyses are performed two or three times a week (blue arrows and text directly adjacent to flow). B: Parallel sequencing as described was performed in a clinical diagnostic setting for a total of 67 samples, also Figure 3E. The read depth at relevant hotspot positions was determined and judged sufficient or insufficient to reliably exclude the presence of mutations (cutoffs in Figure 5B), also taking the estimated tumor load into account. Partially insufficient represents samples for which the read depth was sufficient in only a subset of relevant hotspots, whereas the pathogenic mutation was identified. Sanger-based sequence analysis was judged sufficient in case a sequencing profile could be generated. C: The lower limit for reliable mutation detection was determined for all relevant hotspot positions in 67 clinical samples. For each sample, sorted by the year of tissue sampling, the DNA concentration is indicated on the y axis. The lower detection limit is shown by color. The lower limit for reliable mutation detection is indicated as ND when insufficient read depth was obtained. D: The turnaround time for all 67 clinical request is depicted for both the Ampliseq-PGM and Sanger-based approach and the smMIP-NextSeq500 strategy. ND, not determined; NGS, next-generation sequencing; PGM, Personal Genome Machine; smMIP, single-molecule molecular inversion probe.
compared with our clinical report generated by the routine Ampliseq/PGM- or Sanger-based approach (Supplemental Table S5). Figure 6A schematically depicts both our current diagnostic flow and the smMIP-based approach. Concordant reports were generated in 64 of 67 cases (96%), of which two generated insufficient reads in both methods to provide a conclusion. In our automated smMIP procedure library preparations are standard and performed in duplicate for all samples, which are normally independently analyzed. However, because these library preparations are independent, the sequence data may be merged (simultaneously analyzed). Merging of two to four replicates was required to obtain sufficient unique read depth in 6 of 64 concordant cases. Sufficient read depth was defined in the context of estimated tumor load and based on calculations in Figure 5B. For two clinical samples the AmpliSeq-PMG diagnostic analysis generated insufficient total read depth, whereas the smMIP-NextSeq500 analysis provided sufficient unique read depth to report results. In one case, insufficient unique reads were generated by the smMIP-NextSeq500 approach to analyze all relevant codons and reliably exclude the presence of a mutation, whereas total read depth in the AmpliSeq-PMG diagnostic approach was adequate (Figure 6F). With the use of these constraints the number of dropouts using the smMIP-NextSeq500 approach during the validation period is lower than the AmpliSeq-PMG approach (requirement v). Note that the minimal read depth requirements for the Ampiseq-PMG approach are estimations based on empirical (dilution) experiments, whereas those for the smMIP method are statistical calculations based on the actual number of unique molecules and therefore are more reliable.

To be able to predict upfront which samples require merged data analysis, thus requiring triplicate or even quadruplicate library preparations, we related the obtained unique read depth for all samples to the DNA concentration and age of the sample. A low concentration of extracted gDNA is predictive for suboptimal results, as well as an extended period between tissue sampling and DNA isolation (Figure 6C). From these result we decided to perform triplicate analyses for samples with a DNA concentration <5 ng/μL and the period between tissue sampling and DNA isolation of >2 years.

We compared the TAT of the smMIP-NextSeq500 and the AmpliSeq-PMG and Sanger approach during our validation period (from request until diagnostic report). The TAT on average increases with one working day using the smMIP-based approach in our diagnostic workflow (Figure 6D). This delay is due to the transfer of the sample for the robotized protocol (K. Neveling, A.R. Mensenkamp, R.C. Derks, et al, unpublished data), which is therefore less time-efficient than the manual protocol, but also due to the reduction to two NextSeq 500 sequencing runs per week, instead of three PGM runs. Using the smMIP strategy, 66 of 67 clinical requests (99%) were handled ≤7 working days. After official implementation of the smMIP approach for all our routine diagnostic mutational analyses (since December 2015), the turnaround time of 94% of 79 clinical requests in January 2016 is ≤7 working days, thus fulfilling our last requirement (vi).

Discussion

Over the past decade, there has been a rapid rise in the need for sequence analysis of hotspot regions in onco genes. Molecular diagnostic analysis of tumor tissue is used to refine diagnosis or to stratify patients for personalized treatment with targeted therapies. This requires sensitive and reliable sequence analysis of limited amounts of low-quality gDNA, isolated from FFPE samples. To analyze multiple relevant loci simultaneously, NGS approaches using hybridization- or PCR-based target enrichment methods have been developed. Although PCR-based methods generally require less input DNA, sequencing of PCR duplicates may result in an overestimation of library complexity and, thus, analytical sensitivity. To reliably detect and, equally important, to exclude the presence of mutations, accurate assessment of sequence library complexity is required. Here, we describe the results of the implementation and validation of SMT using smMIPs in routine molecular diagnostics of routinely processed FFPE tumor tissue. We developed a smMIP pool that targets >95% of all hotspot regions in 23 genes by a minimum of two separate smMIPs, analyzing all 41 hotspot regions in one single reaction. This approach fulfilled all our prospectively defined validation requirements and has proven to be superior to Sanger sequencing and Ampliseq-PMG sequencing. Therefore, we have now implemented this smMIP-based target enrichment in our routine sequence analyses pipeline in our laboratory. This also includes sequence analysis to identify somatic mutations in tumor suppressor genes such as BRCA1, BRCA2 (R.D.A. Weren, A.R. Mensenkamp, M. Simons, A. Eijkelenboom, et al, unpublished data shown), TP53, and CDKN2A (A.E. and B.B.J.T., unpublished data).

The smMIPs in our panel were designed to target genomic regions of exactly 112 bp. Including extension and ligation probes, the total targeted region is 152 bp. Larger regions are more efficient in probe design and sequencing, but, because most of our clinical sequence requests are performed on fragmented DNA isolated from FFPE tissue, a smaller target design is more beneficial. Although smaller regions may be more efficient using FFPE-derived DNA, designing these is more complex and the capture and detection of insertions and deletions may be complicated. Important in this respect is that reads spanning deletions and duplications should not be removed from the analysis due to too stringent alignment criteria, such as the absolute and relative number of matching bases, or a strict requirement of reads to include sequences of both the extension and ligation probe (i.e., with an NGS read length of 150 bp this requirement is not fulfilled in case of insertions or duplications >10 bp). We have extensively tested our approach and
demonstrated that our design is capable of reliably detecting insertions (42 bp) and deletions (63 bp).

During the entire implementation period we tested >200 samples using the smMIP approach, and all sequence results were concordant with other sequencing methods. In addition, we validated the smMIP approach side by side to our previous diagnostic sequencing strategies for all our diagnostic sequence request over a period of several weeks. We achieved a TAT of ≤7 working days for >94% of the clinical requests. In total, 42 mutations were identified in 67 clinical samples using our previous diagnostic strategies and confirmed using the smMIP approach. One of the true advantages of SMT is highlighted by the seven samples for which single capture events (library preparations) resulted in insufficient unique read depth to confidently call the absence of variants (true negative), indicating low library complexity for these samples. For six of these samples this could be remedied by combining the data of replicate assays. For the remaining sample combined analysis did not provide sufficient unique read depth, indicating that insufficient genomic molecules were present in this sample to generate a sufficiently complex sequence library. This sample was sufficiently sequenced using the Ampliseq-PGM approach (average total read depth of >500×), but this raises the question how many genomic molecules were amplified during PCR. The suboptimal sequence results of these seven cases would have gone unnoticed in the absence of SMTs and highlight the importance of unique read depth to assess library complexity versus total read depth.

smMIP approach not only provides a better sense of reliability but also results in a notable reduction of identified false-positive variants compared with the Ampliseq-PGM approach. In our diagnostic flow, all called variants are always manually reviewed (and curated) before a report is authorized. By generating consensus reads using the SMTs hardly any false-positive variants are called (Figure 4A), saving considerable amounts of hands-on analysis time per sample. This also allows for the detection of mutations in clinical FFPE samples with allele frequencies as low as 1% and exclude false negatives with a lower limit of 3%. These sensitivities are necessary for samples with a low tumor load (<20%, notably lung cancer biopsies), the detection of therapy-induced resistance mutations such as EGFR p.T790M, and some differential diagnostic requests (eg, mosaic somatic mutations in overgrowth syndromes). An additional benefit of the smMIPs is the DNA strand-specific capture, because it allows for discrimination of genuine C:G>T:A transition mutations and those occurring due to deamination after sample collection. The latter only occur on the cytosine-containing strand and are therefore detected differentially by individual smMIPs targeting both DNA strands. The distinction between these artifacts and genuine C:G>T:A mutations is relevant for FFPE samples that may contain a substantial number of these artifacts, especially in DNA isolated from older tissue specimens.

Besides a superior performance in variant detection, the smMIP approach provides additional advantages. First, existing gene panels are easily adaptable by adding smMIPs targeting new regions to an already optimized smMIP pool. Second, because the smMIP library preparation protocol is simple, it allows for straightforward library preparation automation. This and our automated mapping and variant calling solution (K. Neveling, A.R. Mensenkamp, R.C. Derks, et al, unpublished data) reduces hands-on time, the risk of sample handling errors, and makes this method scalable to accommodate a further increase in clinical requests. Third, although not our main incentive, the smMIP approach is highly cost-effective. Whereas prices for commercial library preparation kits are fixed per sample, the smMIP method only has relatively high initial costs (151 smMIP probes × approximately €10 per probe = approximately €1500) with low additional costs. A single order of smMIP probes is sufficient for >106 samples, and additional reagents required for library preparation are relatively inexpensive, adding up to a total of <€10 per sample for library preparation (depending on ordered volumes). Therefore, smMIP-based library preparations seem especially suited for routine sequence analysis (on a large number of samples), that justify a high initial investment to order and optimize the smMIP probes. In our workflow, most costs required to analyze a sample are the sequencing costs, and those are minimized in our setting by combining the requests of the here described smMIP-based CHP with other smMIP-based sequence analyses, such as hereditary and somatic testing of BRCA1/2 [(K. Neveling, A.R. Mensenkamp, R.C. Derks, et al, unpublished data) and (R.D.A. Weren, A.R. Mensenkamp, M. Simons, A. Eijkelenboom, et al, unpublished data)]. This also demonstrates that not only hotspot positions but also entire (tumor suppressor) genes can be reliably sequenced using smMIPs, as has been published for hereditary testing using MIPs. Finally, preliminary results and published data also suggest that the smMIP approach can be extended beyond sequence analysis to detect copy number variants in our set up (K. Neveling, A.R. Mensenkamp, R.C. Derks, et al, unpublished data), which has been published for MIPs in general and known clinically relevant fusion transcripts using a similar automated workflow (results not shown).

Conclusions

We show that target enrichment using MIPs in combination with single-molecule tags is highly sensitive and provides a true sense of reliability of sequence results. This is especially important when analyzing low amounts of DNA or DNA of poor quality that is daily practice in the routine diagnostics of FFPE tissue. Other (clinical) applications, like the sequencing of cell-free tumor DNA isolated from blood plasma (liquid biopsies), will undoubtedly benefit from target enrichment strategies using single-molecule tags.
Acknowledgments

We thank all technicians of the Molecular Diagnostic Group and other members of the Laboratory of Tumor Genetics, especially Arjen Mensenkamp and Hicham Ouchene; the core sequencing facility, especially Ronny Derks, Marloes Tychon, Bart van Lier, Ermanno Bosgoed, Dimitra Zafeiriopoulou, and Steven Castelein; Joachim Strub, Marc Gissler, and Thorsten Kurz (JSI medical systems); the members of the Genomic Disorders Group and Developmental Genomics Group of our Genetics department for critical discussions and sharing of MIP protocols, especially Marloes Steehouwer; and all pathologists of our Pathology department.

Supplemental Data

Supplemental material for this article can be found at http://dx.doi.org/10.1016/j.jmoldx.2016.06.010.

References

4. Gray PN, Dunlop CL, Elliott AM: Not all next generation sequencing diagnostics are created equal: understanding the nuances of solid tumor assay design for somatic mutation detection. Cancers 2015, 7:1313–1332