TITLE

A Robust Targeted Sequencing Approach for Low Input and Variable Quality DNA from Clinical Samples

Austin So[1,*], Anna Vilborg[1,*,§], Yosr Bouhlal[1], Ryan T. Koheler[1], Susan M. Grimes[2], Yannick Pouliot[1], Daniel Mendoza[1], Federico Goodsaid[1], Mike Lucero[1], Francisco M. De La Vega[1,3], Hanlee P. Ji.[2,4,§]

[1]TOMA Biosciences, Foster City, CA, USA

[2]Stanford Genome Technology Center, Palo Alto CA, USA

[3]Department of Biomedical Data Science, Stanford University School of Medicine, Stanford CA, 94305

[4]Division of Oncology, and Stanford University School of Medicine, Stanford CA, USA

*These authors contributed equally to this work.

[§]Corresponding authors

Correspondences:

Anna Vilborg

Email: anna@tomabio.com

Hanlee P. Ji

Email: genomics_ji@stanford.edu

ABSTRACT

Next-generation sequencing is being adopted as a diagnostic test to identify actionable mutations in cancer patient samples. However, clinical samples such as formalin-fixed, paraffin-embedded specimens frequently provide low quantities of degraded, poor quality DNA. To overcome these issues, many sequencing assays rely on extensive PCR amplification leading to an accumulation of bias and artifacts. Thus, there is a need for a targeted sequencing assay that performs well with DNA of low quality and quantity without relying on extensive PCR amplification. We evaluated the performance of a targeted sequencing assay based on Oligonucleotide Selective Sequencing. This assay enables on to sequence and call variants from low amounts of damaged DNA. This assay utilizes a repair process developed to sequence clinical FFPE samples, followed by adaptor ligation to single stranded DNA and a primer-based capture technique. This approach generates sequence libraries of high fidelity without relying heavily on PCR amplification, and facilitates the assessment of copy number alterations across the target regions. Using an assay designed to capture the exons of a panel of 130 actionable cancer genes, we obtain an on-target rate of >50% and high uniformity across targeted regions at starting input DNA amounts of 10ng per sample. We demonstrate the performance of this targeted sequencing assay using a series of reference DNA samples, and in variant identification from low quality DNA samples originating from different tissue types.

INTRODUCTION

Next-generation sequencing **(NGS)** with targeted gene panels has seen general adoption as a diagnostic and screening tool for a wide variety of disorders.[1] Clinical applications include identifying germline variants, such as single nucleotide polymorphisms **(SNPs)** and structural variants **(SV)** related to hereditary disorders, as well as somatic mutations and other genetic aberrations in cancer that may have implications for treatment and prognosis.[2] The use of targeted gene panels has multiple advantages in all of these cases. Firstly, deep sequencing of genes and other clinically-actionable genomic targets results in higher read coverage, oftentimes in the thousands, and as a result, improves the confidence of variant calling and the limit of variant allele detection.[3] This aspect of targeted sequencing is particularly valuable for analyzing clinical samples that are composed of genetic mixtures, such as cancers that have mixed normal stromal components in addition to multiple clones of tumor cells. In these instances, cancer mutations may occur at low variant allelic fractions **(VAF)** and are more difficult to detect from biopsy samples that are intermingled with normal tissue.[4]

A major challenge for diagnostic sequencing is the variable quality of the genomic DNA obtained from clinical samples. This variability arises from the processing that these samples undergo adversely affecting the integrity of DNA.[5] The vast majority of clinical tumor biopsies undergo formalin fixation and paraffin embedding **(FFPE)** to facilitate histopathologic examination**.** This archival process modifies nucleotides, generates chemical crosslinks and can lead to degradation of the DNA over time into shorter fragments and a higher proportion of damaged and single stranded molecules.[6] As a result, molecular diagnostics based on FFPE DNA often require significant optimization and assay failures are significantly higher compared to when higher quality DNA is used.[5] Indeed, many methods now employ DNA quality control criteria to reject samples and thus mitigate test failures due to sample quality.[7] While increasing

the success of the diagnostic assay, these exclusion criteria will eliminate some samples that may be of clinical significance and value.

To address these challenges, we developed a targeted sequencing approach termed Oligonucleotide-Selective Sequencing **(OS-Seq)** that has multiple features facilitating its application to diagnostic targeted sequencing of DNA from a variety of clinical samples.[8,9] As an extension of OS-Seq, we developed an in-solution version that maintains its streamlined and efficient process for targeted sequencing but without the need for flow cell modification as required in the original version. This assay was optimized to enable the sequencing of clinical samples of variable quality such as what is seen in ancient DNA samples.[10,11] In particular, in-solution OS-Seq involves a repair step that excises damaged bases without corrective repair followed by a highly efficient single stranded adapter ligation process. The efficient ligation allows for the conversion of all nucleic acid species irrespective of quality and quantity into partial sequencing libraries with a single adapter. Following this ligation process, target-specific multiplexed primer annealing and extension of the genomic targets on different strands completes the library for sequencing.[8,9,12]

Herein, we demonstrate the performance of in-solution OS-Seq using a variety of reference DNA samples and show its broader applicability to clinical samples that include FFPE biopsies. Using a 130-cancer gene panel, we confirm the technical reproducibility and high performance of the in-solution OS-Seq assay in terms of on-target coverage, uniformity and ability to detect SNVs, indels and copy number variation from as little as 10 ng of input material. Finally, we demonstrate that the assay works on DNA samples from a variety of different sources including matched samples from individuals with Stage III lung and colorectal cancer. These tissue samples include peripheral blood mononuclear cells **(PBMCs)**, FFPE solid tumor samples, and plasma.

RESULTS

Overview of in-solution OS-Seq

The in-solution version of OS-Seq involves three general steps: a mild repair consisting of excision of damaged bases, ligation of adapters to single-stranded DNA, which high conversion rate even for damaged DNA eliminates the need for whole genome amplification, and "capture" or selection of genomic targets using massively multiplexed pools of target-specific primer oligonucleotides **(Figure 1)**.

We developed a 130-gene panel as a diagnostic test for clinical samples **(Supplementary Table 1).** The panel (i.e. TOMA COMPASS assay) is composed of cancer genes that are established tumors suppressors, oncogenes and other cancer drivers and are known to contain clinically actionable cancer mutations across different malignancies **(Methods)**. All exons of these genes are targeted by primers sets as described in the Methods, making up a region-of-interest **(ROI)** of 419.5 kb.

Analysis of the reference genome NA12878

As a preliminary assessment of performance, we conducted a mass titration experiment using the Coriell sample NA12878. This DNA sample that has been sequenced extensively under a variety of NGS platforms. It is used a reference genome for the Genome-in-a-Bottle **(GIAB)** and the National Institute of Standards (NIST) testing material. Because of this availability of a high confidence list of ground truth variants, NA12878 is widely used for assessment of germline variant detection accuracy.[13] We performed four independent technical replicates of the assay across DNA inputs of 300 ng, 100 ng, 30 ng and 10 ng, each with uniquely barcoded sample adapters. Following library quantification with ddPCR, the same number of molecules across these libraries were pooled and sequenced.

Analysis of the sequencing results revealed that the average on-target coverage was high across all samples regardless of DNA input quantity. At 300 ng of input DNA, we observed an on-target average coverage of 3,097X ± 125 across all four technical replicates. The depth of coverage was maintained even at 10 ng of DNA input, where the average on-target coverage was 2,700X ± 125 **(Table 1)**. The fraction of on-target reads was also high regardless of the starting amount of DNA. At its highest, an on-target fraction of 85% with no discernible variance was achieved across all replicates at an input amount of 300 ng. More significantly, at an input quantity of 10 ng, the on-target fraction of reads was still high at 67 ± 3% across all replicates **(Supplementary Figure 1a)**.

Coverage uniformity across ROIs was examined using the fold 80 base penalty metric[14] and the cumulative fraction of ROI bases achieving a series of coverage thresholds. The fold 80 base penalty is an indicator of the overall sequencing coverage needed to increase 80% of the sequenced bases to the mean coverage level for a given set of genomic targets. The lower the value, the less variability that occurs among the coverage of the individual targets, the perfect score being 1. We noted that a high level of uniformity was achieved across the range of input quantities, where fold 80 base penalty values ranged from 1.77 (SD=0.01) for 300 ng to 3.57 (SD=0.33) for 10 ng of input material **(Table 1 and S2, Supplementary Figure 1b)**. This compares favorably with published high quality exome sequence data sets, where the fold 80 base penalty typically ranges between 2 to 4, with the lower numbers typically achieved using up to μg DNA input material[15-17]. Importantly, the uniformity in coverage results in a high fraction of targeted ROI bases being covered at thresholds of 100× or more, with 98% covered at 300 ng, and 92% at 10 ng **(Supplementary Table 2 and Supplementary Figure 1c).** Thus, the targeting uniformity remained consistently high even at lower DNA starting amounts.

The performance of variant detection was assessed by examining 137 high confidence variants from the GIAB results that overlap with the target regions-of-interest **(ROI)** of our assay - this included 128 single nucleotide variants (SNVs), and 9 insertions and deletions (indels). We determined that SNVs could be detected with high confidence (95 ± 1% of GIAB) at the highest input quantity of 300 ng. With 10-fold less material (30 ng), 91 ± 2% of the SNVs could still be detected (**Supplementary Figure 2, Table 2**).

Detection of variants at different variant allelic fractions

We assessed the ability of the 130-gene assay in detecting a range of variants present at different variant allele fractions **(VAF**), mimicking the distribution of VAFs expected for somatic mutations in tumor tissue samples. We used a set of reference materials derived from either mixtures of engineered cell lines or synthetic DNAs spiked into a reference background genome. All of these DNA samples have known somatic variants at pre-validated allelic fractions. First, we analyzed the STMM-Mix-II reference standard, which includes 37 known cancer somatic variants within the genes covered by the 130-gene assay. These variants are spiked-in at known VAFs within the background of the NA24385 genome. We obtained a dilution series of STMM-Mix-II at 5, 10, 15, and 25% VAFs for all the 37 mutations where the VAF for each somatic variant was validated with droplet digital PCR **(ddPCR)** by the manufacturer **(Supplementary Table 4)**. Because the GIAB has produced a list of high confidence ground truth germline variants for the genome of NA24385, we could account for whether any variant detected was either a somatic, germline, or false positive. This allowed us to calculate sensitivity and specificity for SNV detection.

From each DNA mixture, we used 100 ng per each assay. Overall, the sequencing metrics from each DNA mixture at 100 ng input were comparable to that observed with NA12878, with on-

target average coverage being greater than 1,690X across all samples and replicates **(Supplementary Table 2)**. The assay demonstrated a specificity of ~100% regardless of the VAF from the STMM-Mix-II samples (**Table 4**). Sensitivity was consistently high with VAFs at 10% or greater having a sensitivity of more than 90.0%. At a VAF of 5%, a sensitivity of 83.78% was observed, indicating the general high performance of the assay for low allelic fractions.

To test the performance of the assay on DNA of compromised quality, we used the FFPE reference material HD200, which includes validated and frequently occurring cancer mutations at VAFs lower than 50%. These mutations have been validated with ddPCR by the manufacturer. This sample consists of a mixture of the colorectal cancer cell lines HCT116, RKO and SW48 at defined ratios. This cell line mixture has been subject to FFPE processing as a surrogate for archival tissue. Twenty-four of the nonsynomyous mutations within this sample are covered within the 130-gene assay. Overall, the sequencing metrics were similar to what was observed with NA12878 **(Table 1** and **Supplementary Table 2)**. Average on-target coverage ranged from 4,509 ± 1,312X at 100 ng to 1,721 ± 214X for 10 ng. The fraction of on-target bases was greater than 50% regardless of the amount of input DNA across replicates. Moreover, the average fold 80 base penalty ranged between 1.85 at 100 ng to 2.22 at 10 ng. This was slightly better than observed for NA12878, underlining the consistency and uniformity of coverage at the lowest input amount of FFPE DNA. In aggregate, 82% of all variants were detected in four out of four replicates, and 94% were found in at least three replicates **(Figure 2 a**, **Supplementary Table 3)**.

Next, we used the reference material HD753 to test the performance of the assay in identifying copy number alterations in addition to somatic SNVs/indels. Similar to HD200, the HD753 DNA contains validated copy number variants, translocations, and large insertions/deletions engineered into the genomes of a set of background cell lines. This reference sample also has

18 validated cancer somatic mutations. Of these 18 somatic variants, 13 overlap with the target ROI within the 130-gene assay **(Supplementary Table 4, Figure 2 b)**. General sequencing metrics obtained from input quantities ranging from 300 to 10 ng were found to be equivalent to that found with the other samples analyzed **(Table 1** and **Supplementary Table 2)**. Across the entire range of starting DNA input, 78% of all variants were found in all three replicates, and 95% were found in at least two replicates. Even at 10 ng of starting DNA, we detected all of the variants with the one exception being an insertion mutation of *EGFR* (V769D770insASV).

Finally, we assessed the performance of the 130-gene panel in identifying copy number variations **(CNVs)** as the HD753 cancer cell line has two previously characterized CNVs in cancer drivers *MET* and *MYC.* Both of these genes are represented in the 130-gene assay. A range of DNA input amounts including 100 ng, 30 ng, and 10 ng were tested across three technical replicates and using NA12878 as a normal diploid DNA control (**Figure 3., Table 3**). We used two methods to determine CNV values: Varscan2[18] and a custom method that identified outliers in the log2 ratios of the median coverage depth across all ROIs between the test and negative control samples **(Methods)**. Both *MYC* and *MET* amplifications were identified with either method at the expected ratios and across all the input amounts tested. Additionally, an *ALK* gene amplification was identified that was not previously reported in this material **(Figure 3)**. To verify this amplification, ddPCR was used with commercial CNV assays to the *ALK* gene with the *RPP30* gene as a reference, and confirmed both the presence and the magnitude of the *ALK* amplification as determined by OS-Seq **(Table 3)**.

Analysis of clinical FFPE tumor and matched normal DNA

Given the observed performance on the above reference materials, we evaluated the ability to detect variants from DNA extracted from a variety of clinical samples with the 130-gene assay. Commercially sourced matched blood and tissue samples from Stage II and III lung and

colorectal patients from the time of operation were used **(Supplementary Table 6)**. Cell-free DNA (cfDNA) was isolated from plasma, and genomic DNA was isolated from both peripheral monocytes and archival FFPE tumor tissue. After repair, 100 ng of FFPE- and PBMC derived DNA, and 40 ng of cfDNA, was inputted to adapter ligation. For assessing sequencing quality, the results using DNA extracted from PBMCs, a high quality DNA source, was compared to results from the FFPE-extracted DNA.

Using the metrics as described above, the OS-Seq panel exhibited similarly robust performance on both PBMC and FFPE samples compared to the performance observed with high quality genomic DNA extracted from cell lines **(Supplementary Table 2)**. On-target coverage was 2,300X at its lowest to over 5,600X at its highest. The fraction of on-target reads was consistently high at greater than 50%, regardless of whether the starting DNA originated form PBMCs or FFPE samples.

First, we identified germline variants from the matched pair of PBMC and FFPE samples **(Figure 4)**. Germline variant annotations involved comparison with highly quality population genotypes available from the Exome Aggregation Consortium **(ExAC)** and the 1,000 Genome Project.[19,20] High sequencing data quality would result in the majority of called SNPs being annotated. Overall, over 90% of the SNV calls from all of the PBMC and FFPE DNA sources were also reported SNVs in the ExAC and 1000 Genome Projects **(Supplementary Table 7)**. This result suggests that the sequencing data was of sufficiently high quality for accurately calling germline SNPs regardless of the DNA source. In addition, we compared the SNV overlap between DNA sample from matched PBMC and FFPE tissues. As we noted, FFPE DNA has chemical modifications that can compromise sequencing data quality and leads to issues with variant calling. Importantly, we observed a large overlap of germline variants called in FFPE and PBMCs, ranging from 79 to 91% of variants being found in both FFPE and PBMC

samples.

Somatic mutations within the FFPE sample were identified using the matched normal DNA derived from PBMCs.[21] The somatic mutations identified are reported in **Supplementary Table 8** and we have included a variety of annotations from the COSMIC database that include how frequently these mutations occur in colorectal and lung cancers.[22] Of the detected somatic variants, 17% had previously been reported in other tumors. In Patient 1's colorectal cancer, we detected a well-documented cancer driver mutation introducing a stop codon in *APC*, an essential tumor suppressor gene involved in colorectal cancer pathogenesis.[23] We identified mutations in *ERBB2*, which have recently been found to be mutated in CRC and may represent a gene for targeted therapy in this cancer.[24] Furthermore, we find mutations in *NF1*, also reported mutated previously in CRC.[25]

Patients 2, 3 and 4 were diagnosed with non-small cell lung cancer **(NSCLC)**. For Patient 2, we discovered mutations in genes previously reported to be mutated in NSCLC. These included a mutation in *JAK3* as well as an *ABL2* mutation. *ABL2* is reported mutated in 4% of NSCLC[26], and mutations in the JAK family of kinases are reported in 1.5-2% of NSCLCs.[27] For Patients 3 and 4, genes with mutations included *FGFR3* and *ABL1* (in distinct sites for each patient). *FGFR3* has been reported mutated with a frequency of 1.2% in NSCLC and *ABL1*, with a frequency of 1.5% of NSCLC.[26] For Patient 4, we also find a splice site mutation in the *EGFR* gene. Mutations in this gene are well-known druggable drivers in NSCLC.[28]

Finally, as a proof of concept, matched cell-free DNA **(cfDNA)** samples from the same patient were sequenced using the same protocol. The sequencing data demonstrated equivalent sequencing metrics to those found with both reference DNA and genomic DNA from clinical samples (**Supplementary Table 2**). The ability to detect overlapping germline variants in

cfDNA was determined in both the DNA samples derived from PBMCs and FFPE **(Supplementary Figure 3)**. When comparing with FFPE and PBMC variants, we find a high proportion of cfDNA variants found also in the PBMC and FFPE samples: 75-91% of cfDNA variants were called in all three sample types and 79 to 94% were found in at least one additional sample type.

DISCUSSION

There are several challenges when sequencing tumor tissue samples from clinical biopsies. First, tumor tissues are a complex mixture of adjacent normal cells and potentially multiple clones of cancer cells. As tumor purity in clinical specimens can be lower than 20%, deep sequencing is required to detect somatic mutations present in VAFs down to 5%, motivating the use of targeted sequencing of actionable cancer genes for both sensitivity and cost-effectiveness. Second, the most readily available clinical tumor samples are archived in FFPE blocks, in a process aimed towards enabling histological evaluation but unfortunately leads to DNA damage. Third, sample abundance in clinical samples can be low, limiting both the amount of sample that can be analyzed and resulting in small DNA inputs to the sequencing assay. Finally, genetic biomarkers that can inform therapy decision or prognosis not only include SNV and indel mutations, but also involve copy number alterations that are more difficult to detect with commonly used targeted assays.[29]

Clinical implementation of targeted sequencing assays commonly involve positive selection of ROIs through PCR amplification (amplicon-based approaches) or through affinity purification of these regions through hybridization with long oligonucleotides (bait hybridization).[12] However, both approaches are ill-suited to addressing the challenges associated with processing FFPE clinical samples. Amplicon-based targeting approaches require that both forward and reverse

primers are able to hybridize to capture the ROI. This is particularly challenging with fragmented material that typify FFPE clinical samples, requiring either a reduction in the amplicon footprint to increase the likelihood that both primers are able to hybridize, and/or requires extensive amplification to generate sufficient amounts of sequencing library material from the fraction of molecules upon which both primer can hybridize. Moreover, in the presence of damaged bases, PCR amplification efficiency is reduced particular when using DNA polymerases with proof-reading capability that have poor tolerance to base modifications such as those generated through deoxy-cytosine deamination to deoxy-uracil or through depurination to abasic sites.[6]

In contrast, bait hybridization based approaches mitigate the need for two primers to capture a ROI by instead capturing any fragment within the ROI that can hybridize to the bait oligonucleotide. However, more extensive enzymatic and technical processing of the material is required resulting in a complex workflow and intricacies of preparation that are more prone to experimental error. In particular, following traditional library preparation methods, double stranded adaptors are first ligated to the fragmented DNA, requiring blunt DNA ends in preparation for ligation. This is more problematic for FFPE tissue samples. The extracted DNA can not only have a high proportion of damaged bases, but also have a high proportion of unligated, single-stranded molecules that are effectively eliminated from traditional sequencing library preparations[6]. Furthermore, the use of long oligonucleotides to capture ROIs through hybridization increases the likelihood of capturing off-target regions[30], and requires extensive washing to ensure specificity. This increases the differential efficiency in retention of regions of varying GC-content, affecting uniformity of coverage across genomic targets, potentially resulting in false negative results.[31]

Both of these approaches require extensive PCR amplification to generate sufficient quantities of sequencing library molecules, and as a means to mitigate the poorer efficiencies that are introduced because of the properties of the sample and/or extensive processing steps. Extensive amplification exacerbates biases associated with GC-content and length that can skew the representation of the original sample within the sequenced library. This results in reduced sensitivity to the identification of structural variants such as copy number alterations.[32,33] Overall, these issues affect the diagnostic accuracy of NGS assays.

To address these issues and create a clinically oriented NGS-based assay, we developed an in-solution targeted sequencing assay that is based on a distinct enzymology that is significantly different than current methods previously described. This method relies on primer annealing and extension of single stranded DNA.[8,9] As we have demonstrated, this assay has been optimized for high performance on low input quantities and compromised nucleic acid quality from clinical specimens. Specifically, by sampling degraded and fragmented single stranded DNA with high efficiency, and with high on-target rates and limited PCR amplification, our results show that this assay demonstrates highly uniform coverage at inputs down to 10 ng of DNA, enabling a cost-effective means of performing deep sequencing of target regions with low false negatives for somatic and germline variants.

As another added feature, the dense tiling of specific capture primers across both DNA strands and optimized hybridization conditions provides a combination of fold 80 base penalty and percentage of ROI covered at >= 100X with high performance compared to other assays.[15,17,21] This performance is seen for amounts of DNA less than 100 ng. As a result, one see fewer genomic regions of low coverage where false negatives can occur, provides resilience when analyzing poor samples with limited DNA inputs, and enables a cost-efficient lab operation without the need to overshoot depth of sequencing to compensate for low coverage regions.

Adding new regions of interest or improving coverage in a few problematic regions can be readily achieved by selection of additional probes and / or replacement of poorly performing probes.

The success of precision oncology hinges on its ability to obtain comprehensive molecular profiles of tumor specimens already available from cancer patients. When validating clinical tests for tumor sequencing, there has been significant emphasis on controlling the false positive rate.[34-36] In addition, validation has been done using normal cell lines, which are not representative of the allelic distribution of somatic variants in tissues.[36] These shortcomings are partly a result of the lack of reference material that accurately replicate the complexity of tumor DNA. As a result, it is difficult to make an accurate estimation of the false negative rate. This is problematic, as the false negative rate is equally – if not more – important than the false positive rate. Gaps in variant detection can lead to less information underlying therapy selection for patients, as well as poor diagnostic rate.[37,38]

In addition to improving the false negative rate, another approach to identify targeted therapies for a greater proportion of patients is the use of larger pan-cancer panels. Such panels generate more information, permitting clinicians to consider off label and investigational drugs.[39] With this 130-gene panel, we provide a library preparation method for targeted NGS that is validated on reference materials mimicking the lower VAFs of clinical samples. Additionally, the 130-gene panel covers most relevant target regions of a comprehensive pan-cancer gene panel at a high and uniform coverage that permits identifying clonal and potentially sub-clonal (lower VAF) somatic mutations of even low cellularity tumors. This test is specifically well suited for the poor quality of real clinical specimens, and could increase the yield of actionable variants in clinical testing to inform cancer therapy decisions.

MATERIALS AND METHODS

DNA samples and preparation

Purified genomic DNA **(gDNA)** from the NA12878 Coriell cell line were obtained from the Coriell Institute for Medical Research (Camden, NJ). Purified DNA from the structural multiplex reference standard HD753 was obtained from Horizon Diagnostics (Cambridge, UK). The SeraCare STMM-Mix-II standard was acquired from SeraCare (Milford, MA). Blocks of FFPE cell line mixtures (HD200) with defined allelic frequencies were obtained from Horizon Diagnostics (Cambridge, UK), and sectioned into 15-20 µm curls. Anonymous matched plasma, buffy coat and FFPE solid tumor samples from stage II+/III lung and colorectal cancer patients were purchased from Indivumed GMBH (Hamburg, Germany). Blood components were shipped on dry ice and stored at -80 °C until ready for processing.

gDNA was purified from two 10-20 µm FFPE curls using the ReliaPrep FFPE gDNA Miniprep System (Promega, Sunnyvale, CA), with the following modifications: FFPE curls were incubated for 16 hrs overnight with proteinase K at 65 °C in lysis buffer. Following a 1h incubation at 90 °C, tubes were flash cooled, and the entire mixture transferred to a microfiltration device equipped with a 0.45 µm cellulose acetate filter (Corning COSTAR, Corning, NY). Upon centrifugation for 15 min at 4 °C at 16,000 ×*g* to remove particulates, the filtrate was processed according the manufacturer's guidelines.

Buffy coat samples were gently resuspended in 500 µL phosphate-buffered saline and transferred to a 15 mL conical tube. Residual red blood cells were then lysed by the addition of 4.5 mL of ACK lysis buffer (ThermoFisher Scientific, Carlsbad, CA) and incubation for 10 minutes with inversion at room temperature. Peripheral blood mononuclear cells (PBMCs) were then pelleted via centrifugation for 10 min at 1,600 ×*g*. Pelleted cells were then resuspended in

400 µL of cell lysis buffer (50 mM Tris·HCl, 50 mM Na·EDTA, 0.1% Triton-X100 1.0% sodium dodecyl sulfate, pH 8.0) with 20 µL of >600 mAU/mL proteinase K (Qiagen) and 20 µL of 100 mg/ml RNase A (Qiagen). Following incubation for 1 hr at 65°C, ~0.7 volumes (350 µL) of neat isopropanol was added and the solution mixed by gentle inversion. After incubation for 30 min at -20°C, samples were centrifuged at 16,000×*g* for 15 min, and the supernatant removed. Pellets containing genomic DNA were then washed once with 1 mL of freshly prepared 70% Ethanol, and air-dried for 5 min at room temperature, foillowed by resuspension in 300 µL IDTE buffer (Integrated DNA Technologies, Coralville, IA).

Cell-free DNA (cfDNA) was purified from 3 mLs of plasma by weight using the QIAamp Circulating Nucleic Acid Kit (Qiagen, Redwood City, CA) according to the manufacturer's recommended guidelines. All samples, with the exception of cfDNA samples, were mechanically sheared prior to input into to the TOMA OS-Seq protocol. Briefly, up to 1 µg of DNA was sheared either with a Covaris E210R (Covaris, Woburn, MA) or a ST30 (Microsonic Systems, San Jose, CA) sonicator to a target base pair peak of 600 bp according to the manufacturers' recommendations.

DNA quantification

DNA samples were quantified with ddPCR using the *RPP30* gene as a surrogate for the number of genomic equivalents. For each sample to be analyzed, ddPCR reactions were prepared using 11 µl of Droplet PCR Supermix for probes, 1.1 µl of HEX-labeled PrimePCR™ ddPCR™ Copy Number Assay: RPP30, Human (Assay ID: dHsaCP2500313; BioRad, Hercules, CA), 2.2 µl gDNA, and nuclease free water to a final volume of 22 µl. 20 uL of this reaction mixture was then processed and analyzed on the QX200™ Droplet Digital™ PCR System according to the manufacturer's recommended guidelines using QuantaSoft v1.7.4.0917 (BioRad, Hercules, CA). Values were converted from copies/µL to ng/µL using 30 ng per 10,000 copies of genome

equivalents.

OS-Seq assay

The OS-Seq assay uses a repair process wherein damaged bases are removed from genomic DNA isolated from FFPE samples by excision only, without implementing a corrective repair step. Next, the DNA sample is fully denatured to single-stranded DNA followed by single-stranded ligation of the adapter. This approach ensures that all DNA species, whether they are present in single-stranded or double-stranded form, can be interrogated regardless of starting material quality and quantity. Finally, the "capture" or selection of genomic targets occurs with massively multiplexed pools of target-specific primer oligonucleotides that are designed to tile both stands of the regions of interest at an average spacing of 70bp. Following targeting primer hybridization, the primer provides a start site for polymerase extension of the captured DNA molecule, which incorporates the second sequencing adapter and completing the library for sequencing. As a consequence of the high efficiency of both the ligation and capture steps, the use of PCR is limited to a post-capture step only, where a 15 cycles PCR reaction is performed to expand the library, providing sufficient quantities to load onto the sequencer. In the case of paired end sequencing, the first read (Read 1) covers the synthetic target-specific primer-probes and then adjacent genomic target sequence. The second read (Read 2) comes from the universal adaptor end of the fragment. In-solution OS-Seq capture primers **(Figure 1)** possess a 5'-end moiety corresponding to the Illumina flow cell oligonucleotide and a 3'-end moiety designed to target unique genome sequences within a region-of-interest**.**

The TOMA COMPASS 130-gene kit (TOMA Biosciences, Foster City, CA) includes a set of 14,050 OS-Seq primers designed to cover 2,111 ROIs encompassing the exons of 130 cancer genes. Briefly**,** to select the set of targeting sequences, a melting temperature compatible with the annealing temperature was selected to delineate candidate primers considering the

annealing buffer composition, and sequences were scored with an empirical scheme that accounted for both intrinsic features of the primer sequence, such as G+C content, homopolymers, and secondary structure, as well as genomic features such as the presence of SNPs identified within the dbSNP database, relative target position, the anticipated contribution to ROI coverage, and the predicted specificity of the primer across the genome. Finally, potential interactions between primers in the same pool were evaluated. After evaluation, candidate sequences with scores below a threshold were discarded, and the highest scoring sequences were selected to target each ROI.

Samples were processed using the TOMA COMPASS 130 library preparation kit according to manufacturer's recommendation (TOMA Biosciences, Foster City, CA). Briefly, up to 1 µg of DNA was used for the TOMA repair. After DNA repair, DNA concentrations were measured via ddPCR as described and an appropriate amount of DNA was used as input to ligation. Adapter ligation, target capture, and library expansion were then carried out according to the TOMA COMPASS 130 library preparation kit. Serial dilutions (10^{-4}, 10^{-6}, 10^{-8}) of the resulting libraries were performed in TE buffer and the 10^6 dilutions were then quantified via ddPCR using the TOMA ILQ assay, using the following PCR cycling parameters: 95°C 10 min; 30 s at 94°C, 30 s at 55°C, 60 s at 70°C, 40 cycles; followed by 5 min at 70°C. The TOMA ILQ assay measures P7 (labeled by FAM) and P5 (labeled by HEX) and uses the linkage value - indicating the number of molecules with both P7 and P5 labeling – to calculate number of library fragments per ul.

Based on the library quantification results, 1.0 to 1.4 billion total library fragments were loaded onto the NextSeq 500 (Illumina, San Diego, CA) according to the manufacturer's recommendations with the following adjustments. Briefly, libraries to be run were pooled, and volume adjusted to 20 µl with TE buffer. The pooled library was denatured by adding 1 µl of freshly prepared 0.5 M NaOH and incubating for 5 minutes at room temperature. Chilled HT1

buffer (1280 µl) was then added to the library and the entire mixture loaded into the Illumina NextSeq 500/550 High Output v2 kit (300 cycle) sequencing cartridge. The sequencing primers were diluted and used as indicated in the TOMA COMPASS 130 Library preparation kit protocol. Libraries were then sequenced as paired-ends (2x150 bp).

Analysis of sequencing data

Read Mapping and performance metrics

Before aligning reads, we pre-processed FastQ files to remove bases where the quality value was less than 28. We used two algorithms for mapping and aligning reads to the human genome reference assembly (hg19 with decoys). We used BWA (v7.1.5) with default settings, or alternatively, we mapped the reads with RTG map v3.7 (Real Time Genomics Ltd., New Zealand). We relied on Samtools[40] or Picard for additional sequence processing and coverage analysis. We determined the OS-Seq primers that generated the read based on a probe metadata file, and tagged the alignment file with the primer. We evaluated paired end reads, and for those sequences with the correct OS-Seq primer sequence we identified those sequences that were located within the ROI targeted by the primer and in correct orientation (plus/minus strand). Sequence reads were called as being off-target when they aligned with an insert size larger then 1.5 Kb between sequence read and primer probe.

Variant calling in NA12878 and matched samples

For the targeting assay, we created a series of bed files for target regions; this involved using the location of the primer probes and then enlarging the interval by 50 bases on each flank. To eliminate synthetic sequences from the primer probe, we only used sequence reads that did not overlap with the primer sequence. For germline variant calling in NA12878 and PBMC samples of patients, we utilized either GATK (v3.4.6) using published best practices[41] or RTG snp (v3.7) using default parameters. For calling somatic mutation in tumor/normal pair samples, Samtools

was used to create mpileup files (settings -B -d100000000 –q 15), and subsequently, Mutect (1.1.4) with settings: --min-coverage 10 --min-var-freq 0.15 --min-avg-qual 25 --p-value 0.05 for mpileup2snp with addition of --somatic-p-value 0.05).[42]

Somatic mutation calling for HD200, HD753, and SeraCare STMM-Mix-II.

To call somatic mutations in absence of matched normal sample such as in the case of the reference materials, we used a modification of the Bayesian network variant caller previously described for family pedigrees[43], describing a tumor/normal network where the tumor node inherits variants from the germline and incurs de novo somatic mutations (Irvine et al, in preparation). In the absence of normal data the germline variants were to be imputed. Germline and somatic priors from the ExAC[19] and COSMIC[22] databases were used to score the variants into putative somatic calls. The final VCF files generated were examined for the expected variants. Afterwards, we compiled the sequencing depth and %VAF. In addition, the corresponding BAM files were visually inspected and the depth and %VAF was recorded. The average and standard deviation of depth and %VAF was calculated for each cell line and DNA input amount, and is presented in **Supplementary Table 3** for HD200 and **Supplementary Table 4** for HD753.

Benchmarking of variant calls

To evaluate sensitivity and specificity of variant calling with reference materials we compared the test VCF with a ground truth reference VCF using the vcfeval utility of the RTG Tools package (Real Time Genomics, LTD, New Zealand;[43]). In the case of the germline calls for NA12878, the ground truth file was the reference dataset released by the GIAB for the high confidence regions that overlap the regions of the 130-gene panel ROIs (v 3.2.2 [13]). In the case of somatic reference materials (HD200, HD753, SeraCare STMM-Mix-II), we created a synthetic VCF with the corresponding calls as provided by the COSMIC database (v77)[22] VCF, and used

vcfeval with the –suash-polidy option to only consider allele matches. Received operator curves (ROC) were created with the rocplot utility of RTG Tools.

Measuring copy number variation

To identify copy number variations we normalized coverage depth of the aligned data across each ROI in the assay by the median across all of the ROIs for the test sample and a negative control diploid cell line (NA12878). We then calculated $\log_2$ ratios of the test sample and the negative control at the ROI and then at the gene level, to eliminate region specific biases. To establish if a$\log_2$ ratio value for a given ROI was significantly different from the rest of the population, we applied the Thompson Tau test for outliers (t = 2.629; 2-tailed inverse t-distribution at a = 0.01 and df = 129) across all the gene's ratios. Genes that were deemed significant are reported as changed, either deletions or amplifications. As an additional method, we used Varscan 2 with default settings to determining copy number[18]. We used the Integrated Genome Viewer **(IGV)** to visually inspect sequence reads and variant positions.[44]

Digital PCR confirmation of ALK amplification

The ddPCR assay was performed as described above using probes for *ALK* and *RPP30* (BioRad, Hercules, CA) for HD753 and NA12878 as control. The *ALK* result was first normalized to that of RPP30 for each sample, and then the normalized *ALK* ratio was compared for HD753 versus NA12878 to calculate a final ratio.

ACKNOWLEDGEMENT

We would like to thank Dr. Lincoln Nadauld for his valuable contributions to selecting genes included in the TOMA OS-Seq 130 gene panel, as well as Greg Jensen and Wolfgang Daum for critical discussion and leadership. This work was partly supported by a National Institutes of

Health / National Cancer Institute award from the Innovative Molecular Analysis Technologies program (R33CA174575).

CONTRIBUTIONS

AS, AV, ML, FG, FDLV and HPJ designed the experiments for the study. AS, AV, DM and YB did the experiments. FDLV, RK, SMG, YP, AV and HPJ analyzed the data. AS, AV, FDLV and HPJ wrote the manuscript.

COMPETING INTERESTS

Stanford University holds a patent related to this work where HPJ is listed as a co-inventor. AS, AV, YB, RK, DM, YP, FG, ML, and FDLV are or were employees of TOMA Biosciences at the time this study was carried out.

REFERENCES

1. Gargis, A.S. *et al.* Good laboratory practice for clinical next-generation sequencing informatics pipelines. *Nat Biotechnol* **33**, 689-93 (2015).
2. Frampton, G.M. *et al.* Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. *Nat Biotechnol* **31**, 1023-31 (2013).
3. Sims, D., Sudbery, I., Ilott, N.E., Heger, A. & Ponting, C.P. Sequencing depth and coverage: key considerations in genomic analyses. *Nature reviews Genetics* **15**, 121-132 (2014).
4. Andor, N. *et al.* Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. *Nat Med* **22**, 105-13 (2016).
5. Ivanov, M. *et al.* Towards standardization of next-generation sequencing of FFPE samples for clinical oncology: intrinsic obstacles and possible solutions. *J Transl Med* **15**, 22 (2017).
6. Do, H. & Dobrovic, A. Sequence artifacts in DNA from formalin-fixed tissues: causes and strategies for minimization. *Clin Chem* **61**, 64-71 (2015).
7. Araujo, L.H. *et al.* Impact of Pre-Analytical Variables on Cancer Targeted Gene Sequencing Efficiency. *PLoS One* **10**, e0143092 (2015).
8. Hopmans, E.S. *et al.* A programmable method for massively parallel targeted sequencing. *Nucleic Acids Res* **42**, e88 (2014).
9. Myllykangas, S., Buenrostro, J.D., Natsoulis, G., Bell, J.M. & Ji, H.P. Efficient targeted resequencing of human germline and cancer genomes by oligonucleotide-selective sequencing. *Nat Biotechnol* **29**, 1024-7 (2011).
10. Meyer, M. *et al.* A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338**, 222-6 (2012).

11. Gansauge, M.T. & Meyer, M. Single-stranded DNA library preparation for the sequencing of ancient or damaged DNA. *Nat Protoc* **8**, 737-48 (2013).

12. Myllykangas, S. & Ji, H.P. Targeted deep resequencing of the human cancer genome using next-generation technologies. *Biotechnol Genet Eng Rev* **27**, 135-58 (2010).

13. Zook, J.M. *et al.* Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nat Biotechnol* **32**, 246-51 (2014).

14. Lander, E.S. & Waterman, M.S. Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* **2**, 231-9 (1988).

15. Van Allen, E.M. *et al.* Whole-exome sequencing and clinical interpretation of formalin-fixed, paraffin-embedded tumor samples to guide precision cancer medicine. *Nat Med* **20**, 682-8 (2014).

16. Cibulskis, K. & Kernytsky, A. Quality Assessment of Hybrid Selection Experiments (2010).

17. Bonfiglio, S. *et al.* Performance comparison of two commercial human whole-exome capture systems on formalin-fixed paraffin-embedded lung adenocarcinoma samples. *BMC Cancer* **16**, 692 (2016).

18. Koboldt, D.C. *et al.* VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res* **22**, 568-76 (2012).

19. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285-91 (2016).

20. Genomes Project, C. *et al.* A global reference for human genetic variation. *Nature* **526**, 68-74 (2015).

21. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol* **31**, 213-9 (2013).

22. Forbes, S.A. *et al.* COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res* **43**, D805-11 (2015).

23. Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339**, 1546-58 (2013).

24. Pectasides, E. & Bass, A.J. ERBB2 emerges as a new target for colorectal cancer. *Cancer Discov* **5**, 799-801 (2015).

25. Ahlquist, T. *et al.* RAS signaling in colorectal carcinomas through alteration of RAS, RAF, NF1, and/or RASSF1A. *Neoplasia* **10**, 680-6, 2 p following 686 (2008).

26. Testoni, E. *et al.* Somatically mutated ABL1 is an actionable and essential NSCLC survival gene. *EMBO Mol Med* **8**, 105-16 (2016).

27. Reungwetwattana, T. & Dy, G.K. Targeted therapies in development for non-small cell lung cancer. *J Carcinog* **12**, 22 (2013).

28. Russo, A. *et al.* A decade of EGFR inhibition in EGFR-mutated non small cell lung cancer (NSCLC): Old successes and future perspectives. *Oncotarget* **6**, 26814-25 (2015).

29. Ulahannan, D., Kovac, M.B., Mulholland, P.J., Cazier, J.B. & Tomlinson, I. Technical and implementation issues in using next-generation sequencing of cancers in clinical practice. *Br J Cancer* **109**, 827-35 (2013).

30. Samorodnitsky, E. *et al.* Evaluation of Hybridization Capture Versus Amplicon-Based Methods for Whole-Exome Sequencing. *Hum Mutat* **36**, 903-14 (2015).

31. Tewhey, R. *et al.* Enrichment of sequencing targets from the human genome by solution hybridization. *Genome Biol* **10**, R116 (2009).

32. Benjamini, Y. & Speed, T.P. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res* **40**, e72 (2012).

33. Zhao, M., Wang, Q., Wang, Q., Jia, P. & Zhao, Z. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC Bioinformatics* **14 Suppl 11**, S1 (2013).

34. Garcia-Garcia, G. *et al.* Assessment of the latest NGS enrichment capture methods in clinical context. *Sci Rep* **6**, 20948 (2016).

35. Lih, C.J. *et al.* Analytical Validation and Application of a Targeted Next-Generation Sequencing Mutation-Detection Assay for Use in Treatment Assignment in the NCI-MPACT Trial. *J Mol Diagn* **18**, 51-67 (2016).

36. Rennert, H., Eng, K., Zhang, T., Tan, A. & Xiang, J. Development and validation of a whole-exome sequencing test for simultaneous detection of point mutations, indels and copy-number alterations for precision. *npj Genomic Medicine* **1**, 16010 (2016).

37. Le Tourneau, C. *et al.* Molecularly targeted therapy based on tumour molecular profiling versus conventional therapy for advanced cancer (SHIVA): a multicentre, open-label, proof-of-concept, randomised, controlled phase 2 trial. *Lancet Oncol* **16**, 1324-34 (2015).

38. Le Tourneau, C. & Kurzrock, R. Targeted therapies: What have we learned from SHIVA? *Nat Rev Clin Oncol* **13**, 719-720 (2016).

39. Stockley, T.L. *et al.* Molecular profiling of advanced solid tumors and patient outcomes with genotype-matched clinical trials: the Princess Margaret IMPACT/COMPACT trial. *Genome Med* **8**, 109 (2016).

40. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-60 (2009).

41. Van der Auwera, G.A. *et al.* From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics* **43**, 11 10 1-33 (2013).

42. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-303 (2010).

43. Cleary, J.G. *et al.* Joint variant and de novo mutation identification on pedigrees from high-throughput sequencing data. *J Comput Biol* **21**, 405-19 (2014).

44. Thorvaldsdottir, H., Robinson, J.T. & Mesirov, J.P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* **14**, 178-92 (2013).

FIGURE LEGEND

Figure 1. Overview of in-solution OS-Seq process.

Damaged bases are removed by excision only, without implementing a corrective repair step The DNA is then denatured followed by adapter ligation to single stranded DNA. The single stranded approach allows for adapter ligation to DNA regardless of to starting material quality, making the adapter ligation step highly efficient for damaged material. Because of the efficient adapter ligation, no whole genome amplification is required. Capture is performed using primer-probes in solution in ~2h time and is followed by second strand extension. Finally, the ready sequence library is generated by PCR expansion. 5' and 3' ends indicated, P7 and P5 indicate either the P7 and P5 parts of adapters and probes, respectively, that are required for clustering on the Illumina flow cell, or in the "expansion" section, they indicate PCR primers complementary to the P7 and P5 parts of the adapters and probes, respectively. Ix stands for index sequence, and SP for sequencing primer binding site.

Figure 2. Analysis of variant allelic fraction.

Detection rate of spiked-in somatic variants in HD200 (a, n=4) and HD753 (b, n=3).

Figure 3. Detecting copy number variation.

Normalized coverage for all genes in the 130 gene panel for each replicate of HD753 (target) plotted vs normalized coverage for all genes in NA12878 (control, the same control is used for comparison with each target replicate) for 100 ng (a), 30 ng (b) and 10 ng (c) DNA input. Each replicate is shown in a different color. The three amplified genes are shown as diamonds (*MYC*), squares (*MET*) and triangles (*ALK*).

Figure 4. Variant overlap between DNA from PBMCs versus FFPE tissue.

Overlap of variants called by GATK in the FFPE and PBMC samples from the four patients included in the matched sample study. A-D shows Patient 1-4, respectively.

Table 1. Sequencing metrics for control DNA samples

Cell line	Input DNA (ng)	Number of replicates	Mean target coverage	SD target coverage	Mean % on target bases	SD % on target bases	Mean Fold 80 penalty	SD Fold 80 penalty
NA12878	300	4	3097	125	85.0%	0.0%	1.77	0.01
	100	4	3028	149	79.0%	0.0%	1.96	0.01
	30	4	2342	161	78.0%	1.0%	2.20	0.04
	10	4	2735	289	67.0%	3.0%	3.57	0.33
HD753	100	3	6941	739	56.0%	1.0%	1.85	0.01
	30	3	3920	301	56.0%	1.0%	1.87	0.08
	10	3	4045	727	51.0%	2.0%	2.22	0.16
HD200	300	4	4441	1312	73.0%	1.0%	1.96	0.05
	100	4	4509	1073	66.0%	1.0%	2.05	0.10
	30	4	2766	507	62.0%	1.0%	2.29	0.06
	10	4	1721	214	61.0%	1.0%	2.64	0.11

SD: standard deviation

Table 2. Detection of SNV and indel variants from NA12878

DNA input (ng)	Replicate number	TP	FP	FN	% of expected	Average per input amount	Standard deviation per input amount
300	1	131	54	6	0.96	0.95	0.01
	2	129	65	8	0.94		
	3	131	35	6	0.96		
	4	129	52	8	0.94		
100	1	125	144	12	0.91	0.93	0.01
	2	129	133	8	0.94		
	3	128	133	9	0.93		
	4	128	146	9	0.93		
30	1	125	231	12	0.91	0.91	0.02
	2	120	212	17	0.88		
	3	125	208	12	0.91		
	4	128	267	9	0.93		
10	1	100	443	37	0.73	0.75	0.06
	2	113	431	24	0.82		
	3	95	400	42	0.69		
	4	103	465	34	0.75		

All variants (n=137)

Table 3. CNV calling from a control DNA sample

Amount of DNA (ng)	Gene	Read depth CNV calling			Varscan2 CNV calling		
		Expected ratio*	Observed ratio (mean)	Observed ratio (SD)	Expected ratio (log scale)*	Observed ratio (log scale) (mean)	Observed ratio (log scale) (SD)
100	*MYC*	4.90	4.09	0.04	2.29	2.51	0.15
	MET	2.25	1.87	0.04	1.17	1.24	0.14
	ALK	1.32	1.41	0.04	0.40	0.66	0.14
30	*MYC*	4.90	3.67	0.05	2.29	2.93	0.08
	MET	2.25	1.51	0.03	1.17	1.61	0.08
	ALK	1.32	1.40	0.03	0.40	1.11	0.09
10	*MYC*	4.90	4.52	0.24	2.29	3.26	0.11
	MET	2.25	1.64	0.07	1.17	1.49	0.20
	ALK	1.32	1.49	0.04	0.40	1.16	0.20

SD: standard deviation

Table 4. Detection of variants from control DNA mixtures

DNA Input (ng)	Expected variant allelic fraction (VAF)	Calls			Accuracy	
		FN	TP	FP	Sensitivity	Specificity
100	25%	2	35	6	94.59%	99.9995%
	15%	1	36	6	97.30%	99.9998%
	10%	1	36	5	97.30%	99.9998%
	5%	6	31	3	83.78%	99.9986%

Figure 1

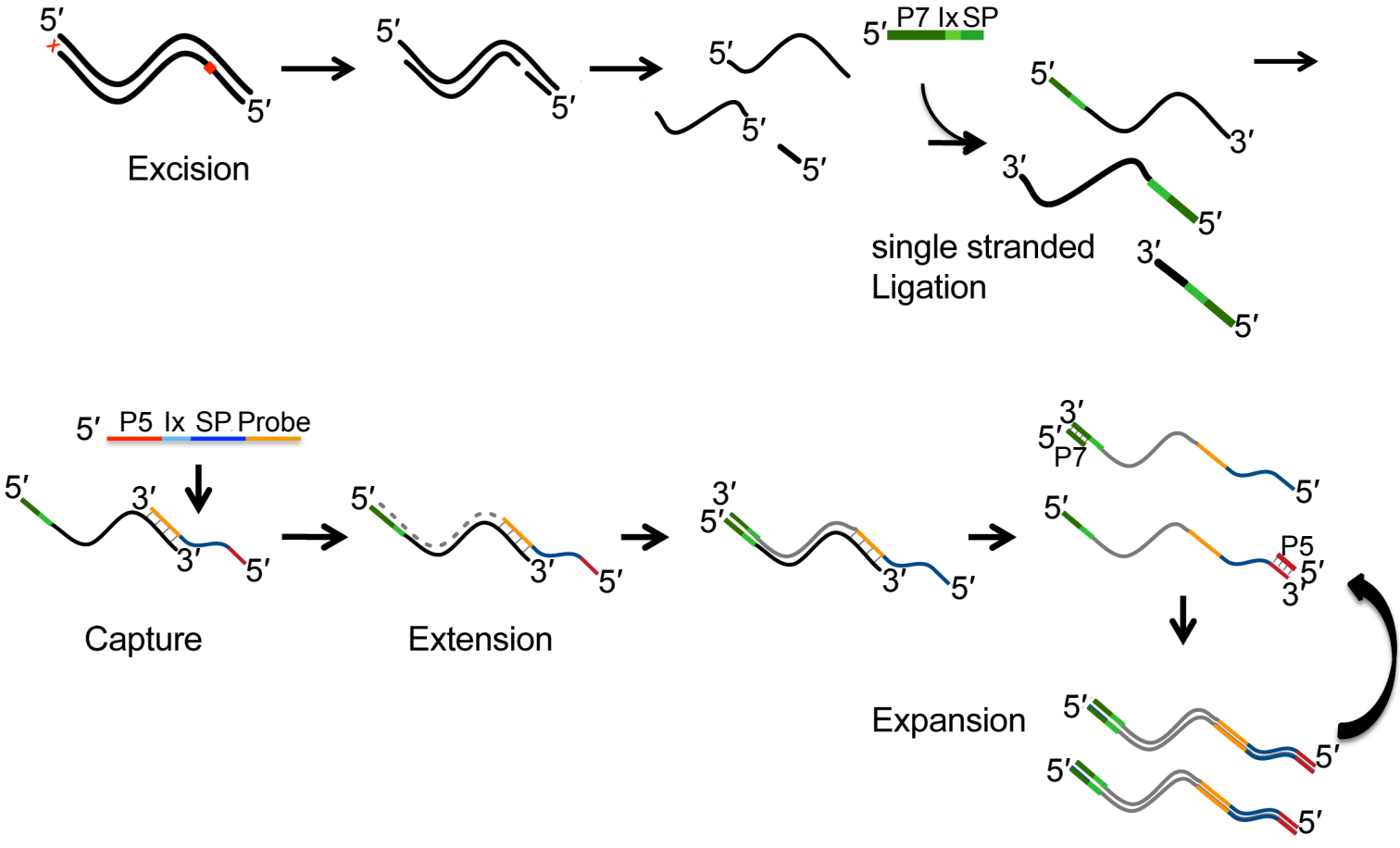
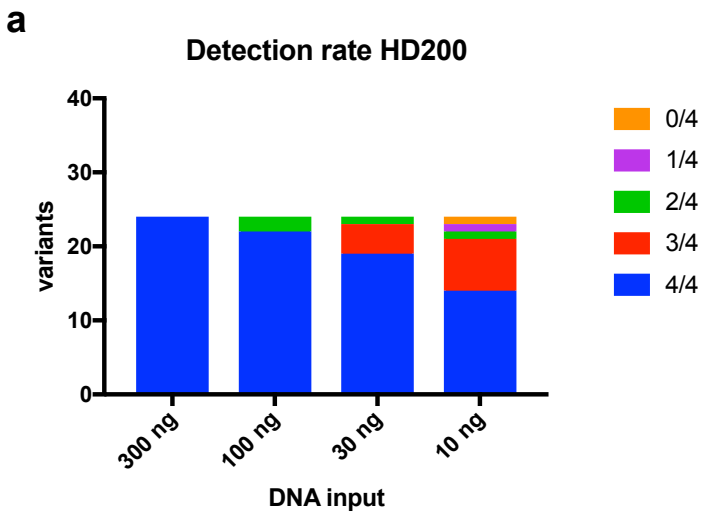

5′
5′
Excision
5′
5′
5′
5′
5′
5′
P7 Ix SP
5′
3′
3′
5′
single stranded
Ligation
3′
5′
5′ P5 Ix SP Probe
5′
3′
3′
5′
Capture
5′
3′
5′
Extension
3′
5′
3′
5′
3′
5′
P7
5′
5′
P5
5′
3′
Expansion
5′
5′
5′
5′

Figure 2


a
Detection rate HD200
variants
40
30
20
10
0
300 ng
100 ng
30 ng
10 ng
DNA input
0/4
1/4
2/4
3/4
4/4

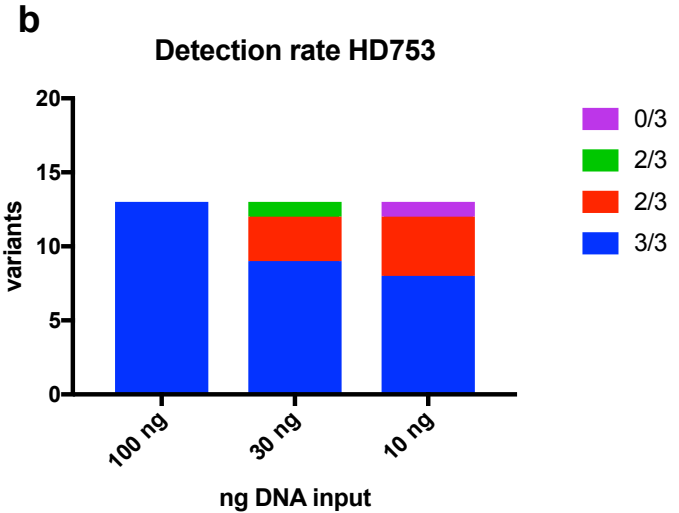

b
Detection rate HD753
variants
20
15
10
5
0
100 ng
30 ng
10 ng
ng DNA input
0/3
2/3
2/3
3/3

Figure 3

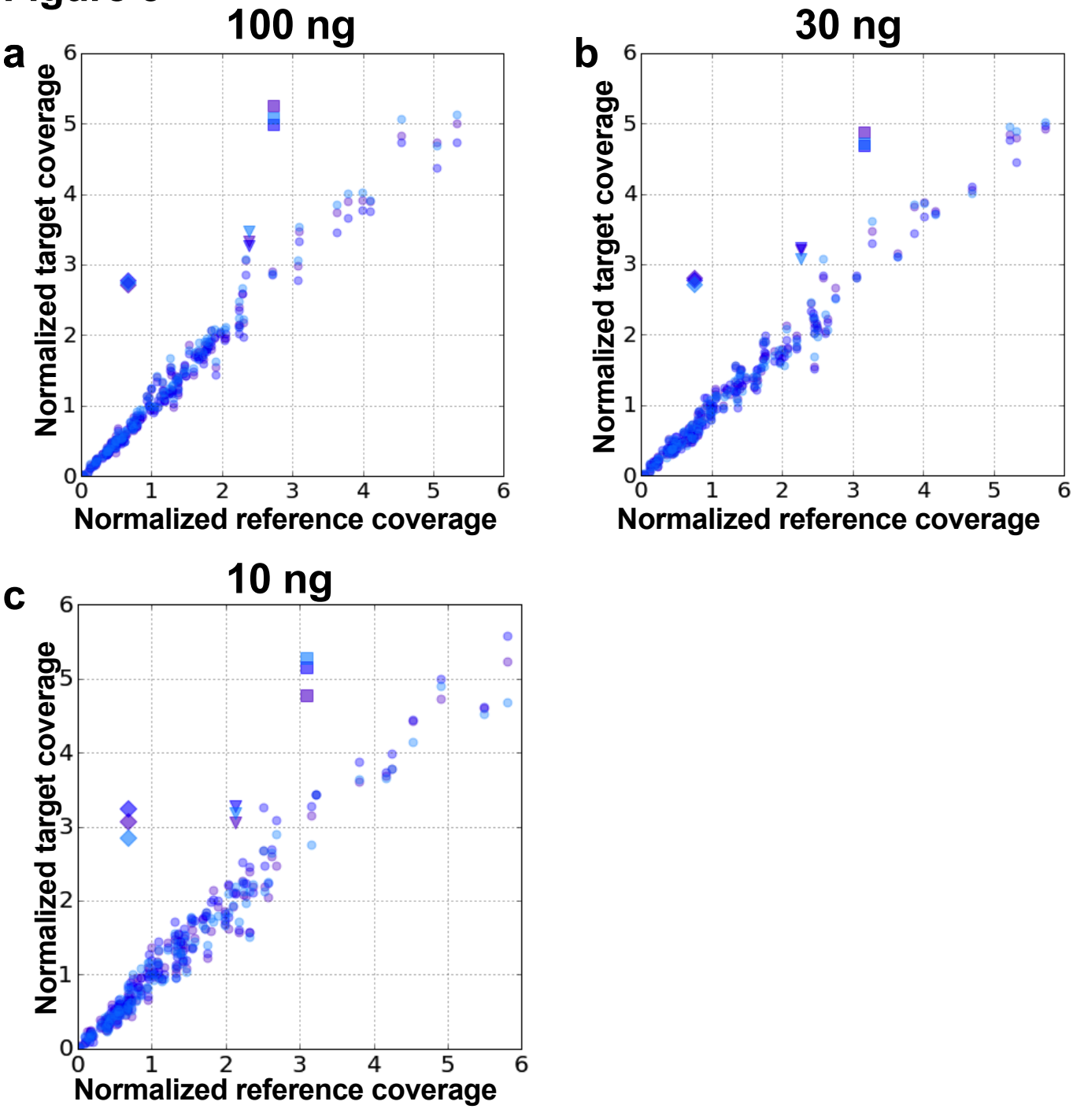
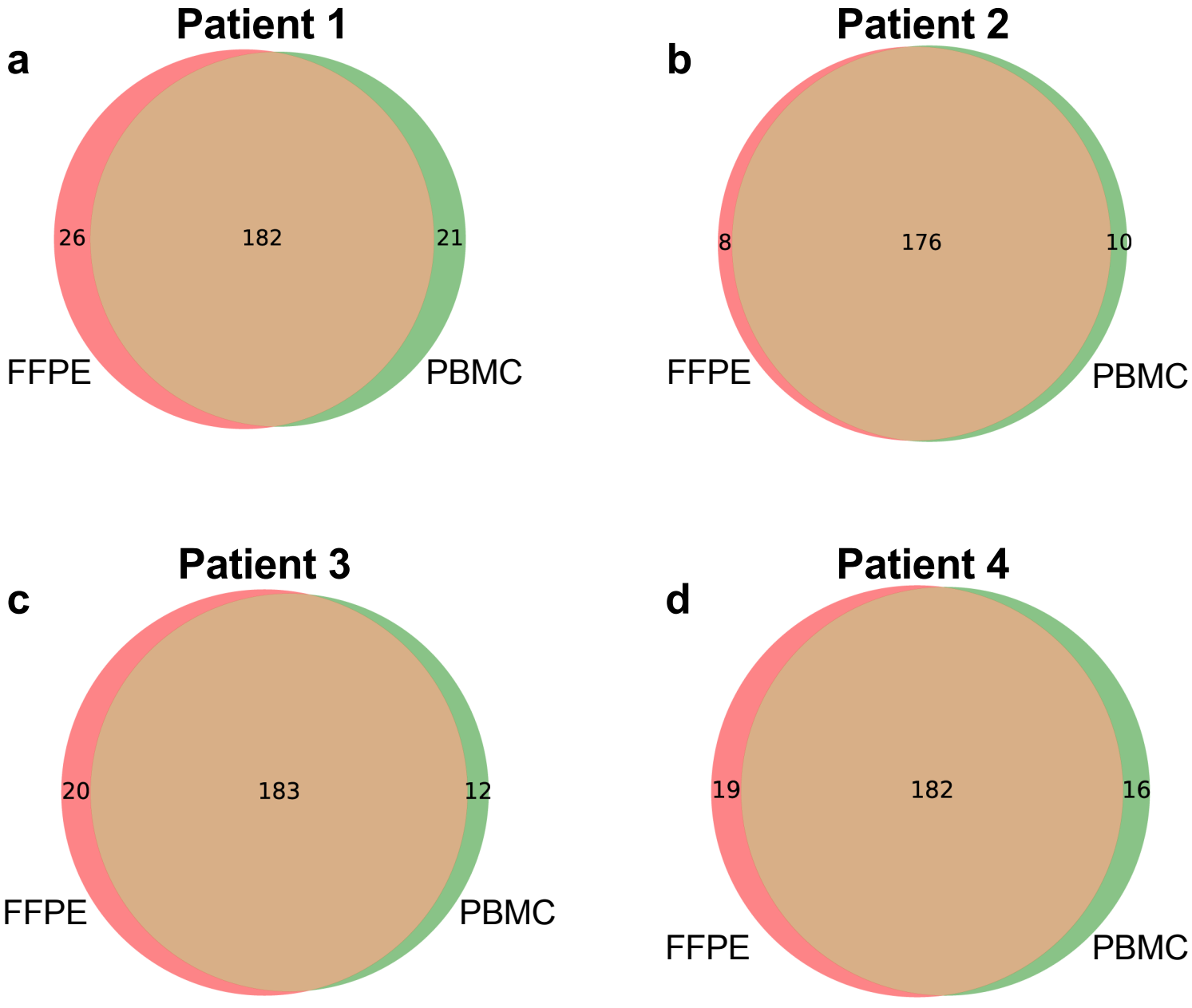

a
100 ng
Normalized target coverage
Normalized reference coverage
b
30 ng
Normalized target coverage
Normalized reference coverage
c
10 ng
Normalized target coverage
Normalized reference coverage

Figure 4


a
Patient 1
26
182
21
FFPE
PBMC
b
Patient 2
8
176
10
FFPE
PBMC
c
Patient 3
20
183
12
FFPE
PBMC
d
Patient 4
19
182
16
FFPE
PBMC