View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Apollo

Somatic mutation screening using archival formalin-fixed paraffin-embedded tissues by Fluidigm multiplex PCR and Illumina sequencing

Ming Wang,[1,*] Leire Escudero-Ibarz,[1,*] Sarah Moody,[1] Naiyan Zeng,[1] Alexandra Clipson,[1] Yuanxue Huang,[1] Xuemin Xue,[1] Nicholas F Grigoropoulos,[1] Sharon Barrans,[2] Lisa Worrillow,[2] Tim Forshew,[3] Jing Su,[3] Andrew Firth,[4] Howard Martin,[5] Andrew Jack,[2] Kim Brugger,[5] Ming-Qing Du[1]

[1]Division of Molecular Histopathology, Department of Pathology, University of Cambridge, Cambridge, UK;
[2]Haematological Malignancy Diagnostic Service, St. James's Institute of Oncology, Leeds, UK;
[3]Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge, UK;
[4]Division of Virology, Department of Pathology, University of Cambridge, Cambridge, UK;
[5]Department of Molecular Genetics, Addenbrooke's Hospital, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK.

*The authors should be regarded as joint first authors.

Number of text page: 27
Number of figures: 5 figures, 6 supplementary figures
Number of tables: 2 supplementary tables

Running title: Somatic mutation screening using FFPE tissues

Funding: The research was supported by grants [LLR10006, LLR13006] from Leukaemia & Lymphoma Research, U.K. and Kay Kendal Leukaemia Fund. SM is a PhD student supported by Medical Research Council, Department of Pathology, University of Cambridge, and Addenbrooke's Charitable Trust. LEI is a PhD student supported by the Pathological Society of UK & Ireland. XX was supported by a visiting fellowship from the China Scholarship Council, Ministry of Education, P.R. China. NG was supported by a Kay Kendal Leukaemia Fund [KKL649] and an Addenbrooke's Charitable Trust fellowship.

Corresponding author:
Professor Ming-Qing Du,
Division of Molecular Histopathology,
Department of Pathology,
University of Cambridge,
Level 3 Lab Block, Box 231,
Addenbrooke's Hospital,
Hills Road, Cambridge, CB2 2QQ, UK.
Email: mqd20@cam.ac.uk
Tel.: 00 44 1223 767092
Fax: 00 44 1223 586670

Conflict of interest: The authors declare no conflict of interest.

ABSTRACT

High throughput somatic mutation screening using formalin-fixed paraffin-embedded (FFPE) tissues is a major challenge due to a lack of established methodology and validated variant calling algorithms. We aimed to develop a targeted sequencing protocol by Fluidigm multiplex PCR and Illumina sequencing, and to establish a companion variant calling algorithm. The experimental protocol and variant calling algorithm were first developed and optimised against a series of somatic mutations (147 substitutions, 12 indels ranging 1-33bp) in 7 genes, previously detected by Sanger sequencing of DNA from 163 FFPE lymphoma biopsies. The optimised experimental protocol and variant calling algorithm were further ascertained in two separate experiments by including the 7 genes as a part of larger gene panels (22 or 15 genes) using FFPE and high molecular weight lymphoma DNAs respectively. We showed that most false positives were due to DNA degradation, deamination and Taq polymerase errors, but they were non-reproducible and could be efficiently eliminated by duplicate experiments. A small fraction of false positives appeared in duplicate, but they were at low alternative allele frequencies and could be separated from mutations when appropriate threshold value was used. In conclusion, we established a robust practical approach for high throughput mutation screening using archival FFPE tissues.

INTRODUCTION

The advent of next generation sequencing (NGS) technology has transformed the landscape of life science research and has led to unprecedented discoveries. In the field of cancer research, NGS has already uncovered a catalogue of extensive somatic mutations and continues to extend this ever growing list of genetic changes in human cancer. In a majority of human malignancies, somatic mutations are found in a wide spectrum of diverse oncogenes and tumour suppressor genes at variable frequencies. For example, in diffuse large B-cell lymphoma (DLBCL), somatic mutations are found in more than 300 cancer genes, and on average each lymphoma harbours ~30 pathogenic mutations.[1-4] Most of these pathogenic mutations occur in <20% of cases, but different somatic mutations may affect a common molecular pathway.[1-4] One of the major challenges is to investigate somatic mutations in these newly identified cancer genes, investigate their potential value in diagnosis, prognosis and treatment stratification, and translate the relevant research findings into clinical practice using routine formalin fixed paraffin embedded (FFPE) diagnostic tissue biopsies.

There are several target enrichment approaches for high throughput mutation screening by NGS, for example, hybrid capture with Agilent SureSelect or NimbleGen SeqCap products, and PCR using HaloPlex or RainDance technology. These targeted re-sequencing approaches were originally developed based on high molecular weight (HMW) DNA samples and have now been successfully applied to those from FFPE tissues and circulating cell-free tumour DNA.[5-7] Several commercial NGS-based assays have been developed for detection of well-characterised somatic alterations, particularly the hotspot mutations, in cancer genes, and these assays, particular those by Ion Torrent, can be applied to a minute amount of DNA extracted from FFPE tissue biopsies.[8-10] Nonetheless, there is no established protocol for discovery research, i.e. detecting unknown mutations in the newly identified cancer genes using FFPE tissue specimens, particularly small biopsies. Many of the caveats for NGS-based mutation screening using FFPE tissue DNA such as artefacts due to poor DNA quality and sequencing errors, false negatives due to inadequate target enrichment and suboptimal performance of variant calling algorithm, the cut-off value of variant allele frequency for diagnosis of somatic mutation, minimal DNA quantity and quality required for successful NGS etc, have not been properly investigated.

Among the various target enrichment methods currently available, PCR using the microfluidic technology (Fluidigm Access Array™ System) represents a practical alternative for high throughput mutation screening using routine FFPE tissue biopsies. The Fluidigm Access Array™ System offers several distinct advantages including 1) being amenable to small amounts (50ng) of DNA samples; 2) allowing parallel amplification of 48 samples with

48 pairs of PCR primers; and 3) offering great flexibility in choice of primers and genetic targets. The system has been successfully used for targeted sequencing using HMW DNA, plasma DNA and very recently FFPE tissue DNA.[11-13] However, all the caveats as mentioned above, including the strategies to eliminate false positives and the cut-off value of variant allele frequency for diagnosis of somatic mutation remain to be established. In the present study, we developed a protocol for high throughput mutation analysis by multiplex PCR with Fluidigm Access Array™ System using DNA samples from FFPE tissues, followed by Illumina MiSeq sequencing. We also developed and validated an in-house variant calling algorithm against a wide range of known mutations. More importantly, we have addressed the above issues through a series of designed experiments and data analysis.

MATERIALS AND METHODS

Tumour materials and DNA extraction: FFPE lymphoma specimens from 163 cases of DLBCL were retrieved from the Haematological Malignancy Diagnostic Service (HMDS) at St James's University Hospital, Leeds, and Addenbrooke's Hospital, Cambridge. Local ethical guidelines were followed for the use of archival tissues for research with the approval of the ethics committees of the involved institutions.

Haematoxylin and eosin slides were reviewed and crude microdissection was performed in each case to enrich tumour cells, ensuring that a tissue area containing >60% of tumour cells was used for DNA extraction. DNA was extracted using the QIAamp DNA Micro Kit (QIAGEN, Crawley, UK) and quantified using Qubit® Fluorometer (Life Technologies, UK). In addition, DNA was extracted from FFPE reactive tonsils, and used for validation of various PCR conditions.

Assessment of DNA quality by conventional PCR. This was performed by PCR of variably sized genomic fragments,[14] using 2ng template DNA in a 10µl reaction mixture for 40 cycles under a standardised protocol (Supplementary Table S1).

Quantification of DNA copy number by real time PCR (qPCR): This was performed on a Quantstudio 6 instrument (Life Technologies, UK) using a custom TaqMan® assay of a 195bp fragment of the *PPIA* gene, which was chosen as there is no evidence of *PPIA* gene copy number change in lymphoma. Primers and probe were designed using the Primer Express software, Version 3.0.1 (Life Technologies, UK). The sequences of the primers (Thermo Scientific, USA) and the probe (Life Technologies, UK) are as follows: forward primer 5'-TATGGCTGTCAGGAGCAGTTCTT-3', reverse primer 5'-AAATGGACCAACCTGCTGTCTT-3' and probe 5'-ACTAAGCAACAAAATAAGCA-VIC-3'. The qPCR conditions and performance were systematically tested and validated prior to data collection. The real time PCR was performed in a volume of 10µl reaction containing 5µl TaqMan® gene expression master mix (Life Technologies, UK), 0.9µl of each primer (900nM final concentration), 0.25µl of probe (250nM final concentration), 1.95µl PCR certified water (Teknova, UK) and 1µl of template genomic DNA. PCR cycling conditions were: 50°C for 2 minutes, 95°C for 10 minutes and 40 cycles of 95°C for 15 seconds, 60°C for 1 minute. For each sample to be quantified, DNA concentration was measured by Qubit dsDNA HS assay (Life Technologies, UK) and serial dilutions were performed to give 10ng/µl, 5ng/µl and 2.5ng/µl solutions. A 10-point standard curve with DNA quantity ranging from 10-0.020ng was prepared using high quality human genomic DNA (Promega, USA) (Supplementary Figure S1). TaqMan® qPCR was carried out in a batch of 38 DNA samples together with negative control and standard curve in triplicate. The estimated copy number was then calculated and expressed as the % of functional DNA copies relative to the standard curve, with an average of the three dilutions taken as the final result (Supplementary Figure S1).

Sanger sequencing: Mutations in 7 genes including *CARD11, CD79A, CD79B, MYD88, TNFAIP3, PRDM1* and *TP53* were first investigated by PCR and Sanger sequencing in 163 DLBCLs using the primers and conditions detailed in

Table S1. PCR products were routinely sequenced using the BigDye Terminator 3.1 System (Applied Biosystems, UK) on an ABI 3730 instrument (Applied Biosystems, UK). In each case, sequence change was confirmed by at least two independent PCR and sequencing experiments. The somatic nature of mutations was ascertained by excluding germline changes through SNP database search and sequencing DNA samples from the microdissected normal cells.

PCR product cloning and sequencing: To confirm mutations that were detected by Illumina MiSeq but not by conventional Sanger sequencing, the relevant PCR products were cloned into the pCR^{TM}2.1-TOPO vector (Invitrogen) and then transformed into TOP10 competent cells. Colonies were screened by PCR using vector primers and up to 30 positive clones were routinely sequenced by the Sanger method as above.

Primer design and validation for PCR with Fluidigm Access Array: PCR primer pairs were re-designed for the above 7 genes and a further 15 genes using Primer3 (http://frodo.wi.mit.edu) based on hg19 of the human genome. A set of criteria were followed for the primer design and these included: a) targeting a small segment of the coding sequence with all amplicons in the range of 144-213bp, thus amenable to DNA samples from FFPE tissues; b) covering the entire coding sequence or all the regions known to be mutated in human malignancies; c) giving a Tm value at 60±3°C; d) where possible avoiding any known SNPs and GC rich sequence region. The specificity of the primers designed and their potential formation of primer dimers were checked with Primer Blast (www.ncbi.nlm.nih.gov/tools/primer-blast/), then further assessed by In-Silico PCR (http://genome.ucsc.edu/cgi-bin/hgPcr?command=start) and AutoDimer program (http://www.cstl.nist.gov/strbase/AutoDimerHomepage/AutoDimerProgramHomepage.htm) (Figure S2).

For each primer pair designed, the forward and reverse primer were tagged with a common sequence 1 (CS1: 5'-ACACTGACGACATGGTTCTACA-3') and common sequence 2 (CS2:5'-TACGGTAGCAGAGACTTGGTCT-3') respectively. All primer pairs were purchased from Thermo Fisher Scientific GmbH and then experimentally validated by PCR using DNA samples from FFPE tonsils. Any primer pairs which failed to yield satisfactory amplification of the expected PCR product or gave rise to a non-specific product were redesigned. In total, 343 primer pairs for 22 genes were successfully designed and validated, and used for PCR with Fluidigm Access Array (Supplementary Table S2).

Preamplification to enrich template target: For PCR with Fluidigm Access Array using DNA samples from FFPE tissues, it was necessary to perform a preamplification with gene specific primers to enrich the template targets before PCR with Fluidigm Access Array (Figure 1). Our initial experiments showed that it was not feasible to include all primer pairs in a single preamplification reaction and achieve a uniform amplification of all the targets due to overlapping primers and primer dimer interactions. We then separated the primer pairs that might potentially give rise to the above issues based on In-Silico PCR and AutoDimer analyses, and performed two separate preamplifications for each sample accordingly.

For each DNA sample, the preamplification and Fluidigm Access Array PCR were performed in duplicate. The preamplification was carried out in a 10µL FastStart High Fidelity Reaction mixture containing 50ng genomic DNA from FFPE tissues (or 20ng HMW DNA from fresh frozen tissues), 50nM of each primer, 4.5 mM $MgCl_2$, 5% DMSO, 200 µM dNTPs, 1x FastStart High Fidelity Reaction Buffer with $MgCl_2$ and 1U of FastStart High Fidelity Enzyme, under the following conditions: 95°C for 10 minutes, 2 cycles of 95°C for 15 seconds and 60°C for 4 minutes, and 13 cycles of 95°C for 15 seconds and 72°C for 4 minutes. The preamplified products were routinely treated with 4 µL ExoSAP-IT enzyme (Affymetrix, UK) to eliminate the unincorporated primers and dNTPs. The efficacy of preamplification was then validated by conventional PCR (Figure S3A).

Massive parallel PCR with Fluidigm Access Array: This was carried out essentially according to the manufacturer's instructions. Briefly, a sample mixture was prepared by mixing 1µl of the 5-fold diluted pre-amplified product with 4µl FastStart High Fidelity Reaction Buffer containing 4.5 mM $MgCl_2$, 5% DMSO, 200µM each dNTPs, and 0.25U FastStart High Fidelity Enzyme. Separately, a primer mixture was prepared for each primer pair or multiple primer pairs where indicated, with 6µM of each primer and 1x Access Array Loading Reagent in a final volume of 50µl. The Fluidigm 48.48 Access Array was loaded with the sample and primer mixtures via the appropriate inlets using an IFC controller. The array chip was then placed in the Fluidigm Thermal Cycler and PCR was performed under the default conditions of the manufacturer (Table S2). The amplified products for each sample were harvested together using an IFC controller.

At the initial stage of the methodology development, the 7 genes were screened for mutations in each of the 163 lymphoma samples by singleplex PCR with Fluidigm Access Array. While at the late stage of the methodology validation, the 7 genes were included as a part of the 22 genes panel and screened for mutation in 142 cases of the above lymphoma samples where sufficient DNA was available by multiplex PCR with Fluidigm Access Array. In both experiments, the preamplification and Fluidigm PCR for each DNA sample were carried out in duplicate.

In a separate parallel study, the 7 genes were included as a part of the 13 genes panel and screened for mutation in 38 cases of splenic marginal zone lymphoma by multiplex PCR with Fluidigm Access Array using high molecular weight (HMW) DNA.[15] This experiment was similarly carried out in duplicate. The novel variants identified in these samples were further verified by a totally independent experiment. The sequence data from these HMW DNA were analysed in parallel as a comparison.

Barcoding and Illumina MiSeq sequencing: Barcoding was carried out in a 20µL reaction mixture containing 1µL of the 100-fold diluted harvested Fluidigm PCR products and 400nM barcode primers (Fluidigm) in FastStart High Fidelity reaction buffer. The reaction was performed on a conventional PCR thermal cycler under following conditions: 95°C for 10 minutes, 15 cycles of 95°C for 15 seconds, 60°C for 30 seconds and 72°C for 1 minute, with a completion step at 72°C for 3 minutes.

The barcoded PCR products from various samples were pooled and purified using AMPure XP beads (Beckman Coulter, UK) following the manufacturer's instructions. A ratio of bead to library at 0.8:1 efficiently removed non-specific products, commonly <200 bp (Figure S3D). Purified PCR product library was quantified using a Qubit® Fluorometer. Purified libraries were routinely sequenced on an Illumina MiSeq sequencer using 250-base paired-end sequencing protocol.

MiSeq sequence Data analysis: The fastq conversion from BCL and demultiplexing were carried out using the MiSeq Reporter software. The adaptor sequence (TGTAGAACCATGTCGTCAGTGT) was removed using cutadapt.[16] The reads were aligned to the target sequences using BWA aln and sampe with the "-e 50" parameter for the latter.[17] The coordinates of the aligned reads were transposed into GRCh37/HG19 coordinates using an in-house perl program, and transformed to a bam file using samtools.[18] Variants were identified using an in-house developed variant caller python program, which was specially designed to identify variants by Fluidigm PCR and MiSeq sequencing, and systematically validated against a large number of known mutations from 7 genes (Figure 2). The identified variants were annotated using the ensembl human database, using the ensembl Variant Effect Predictor,[19] and the result was transformed into an excel sheet using a bespoke perl script. After filtering baseline sequence errors and germline changes through SNP database search, novel variants seen in both replicates of the same sample were recorded.

Sequence search for features potentially associated with false positive variants: For each type of nucleotide substitutions, the 21bp sequence flanking the nucleotide change was extracted. These sequences were aligned together and the position weight matrices were calculated and displayed by WebLogo.[20] The *de novo* enriched motifs were discovered from these sequences using the MEME suite.[21]

RESULTS

In the initial study, the experimental protocol and variant calling algorithm were developed and validated against the somatic mutations in 7 genes detected by Sanger sequencing of DNA samples from a total of 163 FFPE diffuse large B-cell lymphoma biopsies. In the subsequent study, the above optimised experimental protocol and variant calling algorithm were tested in two sets of independent experiments with larger panels of genes.

1) Development of multiplex PCR with Fluidigm Access Array

As DNA samples from FFPE tissues are highly fragmented and inefficient for direct PCR with Fluidigm Access Array, the template targets were first enriched by preamplificaton of each DNA sample with gene specific primers. The major challenges for preamplification are to design primers that can work efficiently in the presence of a large number of other primer sets and yield a uniform target enrichment with minimal non-specific products. We started with the 7 genes and designed 111 primer pairs, covering 21kb sequence. Despite a meticulous effort in primer design, a uniform target enrichment could not be achieved in a single preamplification reaction due to undesired amplification by overlapping primers and poor amplification with a small proportion of primer pairs due to primer dimer interaction. To resolve this, we separated the primer pairs that potentially gave rise to these problems into two independent preamplifications based on In-Silico PCR and AutoDimer analyses (Figure 1, Figure S2). The preamplified products were first validated by conventional PCR, and then by Fluidigm Access Array PCR and Illumina MiSeq sequencing (Figure S2). Illumina MiSeq sequencing confirmed adequate depth of read for each of the 111 amplicons.

The standard protocol for Fluidigm Access Array allows PCR with 48 pairs of primers. To increase capacity, we tested a range of multiplex PCR (2-10 primer pairs) with Fluidigm Access Array. The combination of various primer pairs for multiplex PCR was guided by In-Silico PCR and AutoDimer analyses. Illumina MiSeq sequencing of the Fluidigm amplified products showed adequate depth of coverage for each of the amplicons by multiplex PCR with up to 4 primer pairs (Figure S4), but unsatisfactory coverage for some of the amplicons by multiplex PCR with 5 or more primer pairs.

In addition to the strategies outlined above, a series of quality control measures were established at various steps of Fluidigm PCR/MiSeq sequencing including quality control assessment of template DNA, pre-amplification, Fluidigm PCR, barcode labelling and library purification (Figure S3).

2) Development of strategy and variant calling algorithm for mutation detection

PCR using FFPE DNA is prone to generate sequence errors due to a variety of reasons, such as DNA base modification/damage, few copies of intact templates for PCR, Taq polymerase error, etc. Most of these errors are likely to be random, thus not reproducible and could be efficiently eliminated by performing Fluidigm PCR / MiSeq sequencing analyses in duplicate. As expected, the vast majority of these non-reproducible changes were found at lower AAF, particularly <10% (Figure 3A). Nonetheless, a very small fraction of non-reproducible changes were seen at much higher AAF, even up to 100% of all reads, indicating errors introduced at the very

early steps of the amplification procedure. The level of these non-reproducible variants was also dependent on DNA quality (Figure 3A).

After elimination of non-reproducible changes and SNPs, the remaining variants represented those seen in both replicates and were designated as “reproducible variants”. The absolute number of reproducible changes was also much higher at lower AAF (Figure 3A). However, the percentage of these reproducible variants was minimal at lower AAF, but increased steadily at >10% AAF, particularly for HMW DNA, then followed by FFPE tissue DNA samples amenable to PCR of 400bp or 300bp. In contrast, the percentage of reproducible variants was consistently low in FFPE tissue DNA samples amenable to PCR of up to 200bp (Figure 3B). To quantify the number of functional copies adequate for PCR, we performed TaqMan real time PCR in a series of representative DNA samples (Figure 3C, Figure S1). This was successful in all HMW and FFPE tissue DNA samples amenable to PCR of 300bp or above, but only in 4 of the 7 DNA samples amenable to PCR of up to 200bp. Of the 4 samples amenable to PCR of up to 200bp, which were successfully assayed by TaqMan PCR, the average percentage of functional copies was only 1.1% (Figure 3C). For these reasons, and with further evidence of high baseline sequence errors from later analysis, we excluded the DNA samples amenable for PCR of only up to 200bp from subsequent mutation analysis.

The reproducible variants at high AAF were likely true genetic changes, while those at lower AAF were probably a mixture of false positives and subclonal genetic changes. To permit comparison of mutations detected between Fluidigm PCR / MiSeq sequencing and conventional PCR/Sanger sequencing, we thus initially chose 10% AAF as a cut-off value as the variants above this value can be validated by conventional PCR and Sanger sequencing, or by cloning and sequencing where necessary.

The reproducible variants with AAF>10% in both replicates were then cross-examined with known somatic mutations detected by Sanger sequencing of the 7 genes in a total of 163 DNA samples from FFPE lymphoma tissues (Figure 2, Figure S5). At first, the performance of an in-house variant calling algorithm was assessed and "tuned" against 114 known mutations including 106 substitutions and 8 indels (ranging 1-33bp). While assessing the variant calling algorithm, additional novel variants were identified by Fluidigm PCR/MiSeq sequencing, and these novel variants were further validated by PCR and Sanger sequencing, or where indicated by cloning and sequencing of the PCR products (n=15). The resulting sequence data were used to further fine-tune the algorithm, until the algorithm was able to detect all mutations detected or confirmed by Sanger sequencing (Figure 2), without both false negatives and false positives. Taken together, a total of 159 Sanger sequencing confirmed somatic mutations including 147 substitutions and 12 indels (ranging 1-33bp) were used to optimise the algorithm.

3) Testing the optimised experimental protocols and variant calling algorithm and determining the cut-off value of AAF for somatic mutation detection

To further ascertain the performance of the above optimised experimental protocol and variant calling algorithm, we performed the following two sets of independent experiments (Figure 4).

In one set of experiments, the above 111 PCR primer pairs for the 7 genes were further investigated as a part of a total of 343 PCR primer pairs for 22 genes covering 65kb sequence using the same cohort of FFPE lymphoma DNA samples as above. These independent experiments confirmed the characteristics of non-reproducible and reproducible changes for the 7 genes as presented above, and also showed little difference in these profiles between the 7 genes and 15 additional genes. Cross examination of novel reproducible changes in the 7 genes between the two sets of independent experiments showed that the concordance in mutation detection critically depended on the cut-off value of AAF (Figure 4B). With 10% AAF as a cut-off value, a 98.8% concordance was

observed, while with cut-off value below 10% AAF, concordances progressively deteriorated, and therefore unreliable for mutation detection.

In the other set of experiments, the same 111 PCR primer pairs for the 7 genes were analysed as a part of 157 PCR primer pairs for 13 genes in an additional cohort of 38 HMW DNA samples from splenic marginal zone lymphoma,[15] and this experiment was performed twice independently. Cross examination of novel reproducible changes from the 7 genes between the two sets of independent experiments showed 100% concordance at 7% AAF or above (Figure 4B). Thus, the best cut-off value for HMW DNA was defined as 7%.

4) Distinct difference in the nature of false positives between FFPE tissue and HMW DNA

To understand the potential factors underpinning false positives, we examined the nature of non-reproducible and reproducible variants in both FFPE tissue and HMW DNA. Separate analyses of data from the 7 genes and others showed no apparent difference and the data were thus combined and presented together.

For non-reproducible changes, there was a broad similarity in the pattern of nucleotide changes between FFPE tissue and HMW DNA samples, and both showed frequent C:G>T:A and A:T>G:C alterations, with other base changes being at relatively low frequencies (Figure 5). However, there were marked differences in the frequencies of these changes between FFPE tissue and HMW DNA samples, with the frequencies of C:G>T:A change being remarkably higher in the FFPE tissue (Figure 5). There was neither an apparent difference in the frequency of indels between FFPE tissue and HMW DNA, nor any association between the nature of non-reproducible changes and their AAF (Figure S6).

For reproducible changes, we further subdivided them according to the cut-off AAF as those above this value were true genetic changes, while those below this value were a mixture of sub-clonal changes and false positives. In contrast to non-reproducible changes, the spectrum of the reproducible changes above the cut-off value in both FFPE tissue and HMW DNA was broad, without apparent bias toward any particular nucleotide changes (Figure 5). The slightly more variations of the spectrum of the reproducible changes in the HMW group are most likely due to a small number (**) of mutations in this group.

Finally, we also searched for sequence features that might be potentially associated with non-reproducible or reproducible variants in both FFPE tissue and HMW DNA using the MEME suite,[21] but the analyses did not identify any sequence features associated with false positives or true mutations.

DISCUSSION

In this study, we have developed and validated a robust high throughput mutation screen using DNA samples from archival FFPE tissues. Experimentally, we have established a practical protocol with various quality control steps for multiplex PCR with Fluidigm Access Array, providing a uniform amplification of target genes for Illumina MiSeq sequencing. Bioinformatically, we have generated an in-house variant calling algorithm, and fine-tuned its performance against somatic mutations detected by Sanger sequencing. In addition, we have established a strategy to maximally eliminate false positives, enabling detection of known as well as novel mutations. Our study also highlights several critical issues for application of PCR-based target enrichment and next generation sequencing to DNA samples from FFPE tissues.

Potential sources of false positives

There are many potential causes leading to a false positive sequence change. Apart from those associated with Illumina sequencing, the major causes for false positivity in the context of the current study are poor quality of DNA and errors of *Taq* polymerase.

It is known that poor quality of DNA is prone to PCR and sequencing errors. We showed in the present study that the extent of false positives as measured by non-reproducible changes between the two replicates of the same DNA sample depended on DNA quality, with the poorer quality DNA samples displaying higher rates of false positives. The propensity of the DNA samples from FFPE tissues to generate false positives is most likely due to DNA damage and few copies of intact templates for PCR. In comparison with HMW DNA, those from FFPE tissues showed a remarkably high incidence of C:G>T:A, accounting for ~40% of non-reproducible changes. This extraordinarily high false positive rate is most likely due to deamination of cytosine during tissue formalin fixation and storage.[13,22,23]

Due to degradation, only a small fraction of a DNA sample from FFPE tissue is adequate to serve as templates for PCR despite the fact that primers were designed to amplify short fragments (200bp) of genomic sequences. By TaqMan real time PCR, we showed that only 1.1% of genomic DNA from FFPE tissues was adequate for PCR of 200bp of genomic sequences. For a DNA sample amenable to conventional PCR of up to 200bp, 50ng DNA contains only ~170 functional copies adequate for PCR of 200bp genomic sequences. As few functional templates are available for PCR, any errors introduced at the early steps of the amplification process would appear in a substantial proportion of the amplified products. In line with this, the other major non-reproducible changes are A:T>G:C alterations, which are likely the result of *Taq* polymerase errors.[24]

Despite that HMW DNA samples are far better in quality than those from FFPE tissues, these samples also gave rise to considerable false positives at low AAF. In contrast to FFPE tissue DNA, the majority of false positives in HMW DNA samples as measured by non-reproducible changes were A:T>G:C changes, being far more frequent than C:G>T:A alterations.

Strategies to eliminate false positives

We have established several practical means to eliminate false positives, allowing highly efficient and specific detection of somatic mutations.

1) Assessing DNA quality to select those with adequate quality and quantity. By quality control PCR and further supported by TaqMan real time PCR, we found that DNA samples amenable to PCR of 300bp or above were adequate for mutation screening with Fluidigm PCR and Illumina MiSeq sequencing. Under the protocols described in this study, 50ng FFPE tissue DNA or 20ng HMW DNA yielded excellent results for sequencing of 343 amplicons covering 65kb. However, the amount of template DNA may be subjected to change depending on the number of primer pairs used and the size of the amplicons.

2) Investigating each DNA sample in duplicate. The vast majority of false positives are not reproducible, and thus can be efficiently eliminated by analysis of each DNA sample in duplicate. Under the experimental conditions described, duplicate analyses are sufficient and there is no need to further increase the number of replicates. Theoretically, this approach is potentially capable of eliminating all types of random false positives resulting from poor quality DNA or Taq polymerase errors. An alternative approach to reduce false positives is treatment of FFPE tissue DNA with uracil glycosylase.[13,25] This can significantly reduce false positives resulting from deamination of cytosine, however uracil glycosylase is active only at uracil lesions, but not thymine lesions resulting from deamination of 5-methyl cytosine. In addition, the C:G>T:A artefact at CpG dinucleotides is

resistant to uracil glycosylase treatment.[25] Thus, duplicate experiments offer broader efficacy in elimination of false positives.

3) Choosing appropriate cut-off value of AAF for reliable detection of somatic mutations. Based on the concordance of reproducible variants between two sets of independent experiments, together with known somatic mutations by Sanger sequencing, we suggest to use 10% and 7% as the optimal cut-off value of AAF for mutation detection in FFPE tissue and HMW DNA respectively. For detection of well-characterised hotspot mutations, it is possible to go below these cut-off values with caution. However, for detection of unknown mutations, it is impossible to distinguish somatic mutations from baseline sequence errors when AAF is below the cut-off value. These findings further emphasise the importance of DNA preparation from specimens with high tumour cell content, or microdissected tumour cells.

4) A fully validated in-house variant calling algorithm. This was developed and fine "tuned" by assessing its performance on detection of a large number of known somatic mutations including a variety of indels. We also tested this in-house variant calling algorithm in two independent ongoing studies including one on solid tumours with different gene panels, and confirmed its excellent performance as judged by correlation with known mutations by Sanger sequencing. In comparison with commercial software, the validated in-house variant calling algorithm gave much better performance particularly in indel calling.

Detection of sub-clonal mutations

Based on the above established protocols, Fluidigm PCR and Illumina MiSeq sequencing are much more sensitive in somatic mutation screening than conventional Sanger sequencing. Nearly a third of the mutations (45/159=28%) detected by Fluidigm PCR and Illumina MiSeq sequencing were missed by original PCR and Sanger sequencing, albeit confirmed by further Sanger sequencing or cloning and sequencing of the PCR products. In line with our findings, Bodor et al also demonstrated an improvement of 39% in mutation detection by amplicon based next generation sequencing in comparison with conventional PCR and Sanger sequencing.[26] A proportion of the somatic mutations additionally detected by Fluidigm /MiSeq sequencing may represent subclonal genetic changes. Nonetheless, subclonal somatic mutations particularly uncharacterised changes at a frequency below the cut-off value cannot be reliably identified as these changes are not distinguishable from baseline sequence errors despite being technically detectable by the method. Importantly, it is the cut-off value of AAF, rather than the technical sensitivity of the next generation sequencing, which determines how low a subclonal mutation can be reliably detected.

In conclusion, we have established a practical protocol for high throughput mutation analysis using DNA samples from archival FFPE tissues by Fluidigm multiplex PCR / Illumina MiSeq sequencing, and an in-house variant calling algorithm. The strategies used to eliminate false positives and identify somatic mutations provide a practical solution for high throughput mutation screening using routine FFPE tissue biopsies.

Acknowledgements: The authors would like to thank David Withers for his help in DNA sequencing. We would also like to thank Ruth Littleboy, Fay Rodger, Antje Schulze Selting for technical assistance, Mark Ross, Stefano Berri and Anthony Rogers for helpful discussion on data analysis and Shubha Anand for critical reading of the manuscript.

Author contributions: Experimental design, data collection and analysis: MW, LEI, SM, NZ, AC, YH, NFG, TF, JS, HM; Case contribution: SB, LW, AJ; Illumina sequencing analysis and variant calling: KB, XX, AF; Manuscript

preparation and writing: MQD, MW, LEI, SM, AC. Study design and coordination: MQD. All authors commented on the manuscript and approve its submission for publication.

Supplementary data:

Supplementary Figures:
Figure S1. Quantification of functional copies of DNA by TaqMan real time PCR.
Figure S2. Stratification of PCR primer pairs for different pre-amplification reactions and multiplex PCR by In-Silico PCR and AutoDimer analysis.
Figure S3. Quality control at various steps of Fluidigm Access Array PCR and Illumina MiSeq sequencing.
Figure S4. Examples of depth of coverage by Fluidigm Access Array PCR and Illumina MiSeq sequencing. Adequate depth of coverage can be achieved for each of the amplicons by multiplex Fluidigm Access Array PCR with up to 4 primer pairs.
Figure S5. Examples of somatic mutations detected by Fluidigm Access Array PCR and Illumina MiSeq sequencing.
Figure S6. The nature of non-reproducible variants is independent of AAF.

Supplementary Tables:
Supplementary Table S1: Primers and PCR conditions used for conventional PCR and Sanger sequencing.
Supplementary Table S2: Primers and conditions used for multiplex PCR with Fluidigm Access Array.

References:

1. Morin RD, Mendez-Lago M, Mungall AJ, Goya R, Mungall KL, Corbett RD, Johnson NA, Severson TM, Chiu R, Field M, Jackman S, Krzywinski M, Scott DW, Trinh DL, Tamura-Wells J, Li S, Firme MR, Rogic S, Griffith M, Chan S, YAkovenko O, Meyer IM, Zhao EY, Smailus D, Moksa M, Chittaranjan S, Rimsza L, Brooks-Wilson A, Spinelli JJ, Ben Neriah S, Meissner B, Woolcock B, Boyle M, McDonald H, Tam A, Zhao Y, Delaney A, Zeng T, Tse K, Butterfield Y, Birol I, Holt R, Schein J, Horsman DE, Moore R, Jones SJ, Connors JM, Hirst M, Gascoyne RD, Marra MA: Frequent mutation of histone-modifying genes in non-Hodgkin lymphoma. Nature 2011, *476*:298-303.
2. Pasqualucci L, Trifonov V, Fabbri G, Ma J, Rossi D, Chiarenza A, Wells VA, Grunn A, Messina M, Elliot O, Chan J, Bhagat G, Chadburn A, Gaidano G, Mullighan CG, Rabadan R, Dalla-Favera R: Analysis of the coding genome of diffuse large B-cell lymphoma. Nat Genet 2011, *43*:830-837.
3. Lohr JG, Stojanov P, Lawrence MS, Auclair D, Chapuy B, Sougnez C, Cruz-Gordillo P, Knoechel B, Asmann YW, Slager SL, Novak AJ, Dogan A, Ansell SM, Link BK, Zou L, Gould J, Saksena G, Stransky N, Rangel-Escareno C, Fernandez-Lopez JC, Hidalgo-Miranda A, Melendez-Zajgla J, Hernandez-Lemus E, Schwarz C, Imaz-Rosshandler I, Ojesina AI, Jung J, Pedamallu CS, Lander ES, Habermann TM, Cerhan JR, Shipp MA, Getz G, Golub TR: Discovery and prioritization of somatic mutations in diffuse large B-cell lymphoma (DLBCL) by whole-exome sequencing. Proc Natl Acad Sci U S A 2012, *109*:3879-3884.
4. Zhang J, Grubor V, Love CL, Banerjee A, Richards KL, Mieczkowski PA, Dunphy C, Choi W, Au WY, Srivastava G, Lugar PL, Rizzieri DA, Lagoo AS, Bernal-Mizrachi L, Mann KP, Flowers C, Naresh K, Evens A, Gordon LI, Czader M, Gill JI, Hsi ED, Liu Q, Fan A, Walsh K, Jima D, Smith LL, Johnson AJ, Byrd JC, Luftig MA, Ni T, Zhu J, Chadburn A, Levy S, Dunson D, Dave SS: Genetic heterogeneity of diffuse large B-cell lymphoma. Proc Natl Acad Sci U S A 2013, *110*:1398-1403.

5. Teer JK, Bonnycastle LL, Chines PS, Hansen NF, Aoyama N, Swift AJ, Abaan HO, Albert TJ, Margulies EH, Green ED, Collins FS, Mullikin JC, Biesecker LG: Systematic comparison of three genomic enrichment methods for massively parallel DNA sequencing. Genome Res 2010, *20*:1420-1431.
6. Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, Howard E, Shendure J, Turner DJ: Target-enrichment strategies for next-generation sequencing. Nat Methods 2010, *7*:111-118.
7. Murtaza M, Dawson SJ, Tsui DW, Gale D, Forshew T, Piskorz AM, Parkinson C, Chin SF, Kingsbury Z, Wong AS, Marass F, Humphray S, Hadfield J, Bentley D, Chin TM, Brenton JD, Caldas C, Rosenfeld N: Non-invasive analysis of acquired resistance to cancer therapy by sequencing of plasma DNA. Nature 2013, *497*:108-112.
8. Kanagal-Shamanna R, Portier BP, Singh RR, Routbort MJ, Aldape KD, Handal BA, Rahimi H, Reddy NG, Barkoh BA, Mishra BM, Paladugu AV, Manekia JH, Kalhor N, Chowdhuri SR, Staerkel GA, Medeiros LJ, Luthra R, Patel KP: Next-generation sequencing-based multi-gene mutation profiling of solid tumors using fine needle aspiration samples: promises and challenges for routine clinical diagnostics. Mod Pathol 2014, *27*:314-327.
9. Singh RR, Patel KP, Routbort MJ, Reddy NG, Barkoh BA, Handal B, Kanagal-Shamanna R, Greaves WO, Medeiros LJ, Aldape KD, Luthra R: Clinical validation of a next-generation sequencing screen for mutational hotspots in 46 cancer-related genes. J Mol Diagn 2013, *15*:607-622.
10. Tsongalis GJ, Peterson JD, de Abreu FB, Tunkey CD, Gallagher TL, Strausbaugh LD, Wells WA, Amos CI: Routine use of the Ion Torrent AmpliSeq Cancer Hotspot Panel for identification of clinically actionable somatic mutations. Clin Chem Lab Med 2014, *52*:707-714.
11. Forshew T, Murtaza M, Parkinson C, Gale D, Tsui DW, Kaper F, Dawson SJ, Piskorz AM, Jimenez-Linan M, Bentley D, Hadfield J, May AP, Caldas C, Brenton JD, Rosenfeld N: Noninvasive identification and monitoring of cancer mutations by targeted deep sequencing of plasma DNA. Sci Transl Med 2012, *4*:136ra68
12. Halbritter J, Diaz K, Chaki M, Porath JD, Tarrier B, Fu C, Innis JL, Allen SJ, Lyons RH, Stefanidis CJ, Omran H, Soliman NA, Otto EA: High-throughput mutation analysis in patients with a nephronophthisis-associated ciliopathy applying multiplexed barcoded array-based PCR amplification and next-generation sequencing. J Med Genet 2012, *49*:756-767.
13. Bourgon R, Lu S, Yan Y, Lackner MR, Wang W, Weigman V, Wang D, Guan Y, Ryner L, Koeppen H, Patel R, Hampton GM, Amler LC, Wang Y: High-throughput detection of clinically relevant mutations in archived tumor samples by multiplexed PCR and next-generation sequencing. Clin Cancer Res 2014, *20*:2080-2091.
14. Liu H, Bench AJ, Bacon CM, Payne K, Huang Y, Scott MA, Erber WN, Grant JW, Du MQ: A practical strategy for the routine use of BIOMED-2 PCR assays for detection of B- and T-cell clonality in diagnostic haematopathology. Br J Haematol 2007, *138*:31-43.
15. Clipson A, Wang M, de Leval L, Ashton-Key M, Wotherspoon A, Vassiliou G, Bolli N, Grove C, Moody S, Escudero-Ibarz L, Gundem G, Brugger K, Xue X, Mi E, Bench A, Scott M, Liu H, Follows G, Robles EF, Martinez-Climent JA, Oscier D, Watkins AJ, Du M: KLF2 mutation is the most frequent somatic change in splenic marginal zone lymphoma and identifies a subset with distinct genotype. Leukemia 2014
16. Martin M : Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet journal 2011, *17*:10-12.
17. Li H, Durbin R: Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009, *25*:1754-1760.
18. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009, *25*:2078-2079.
19. McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F: Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics 2010, *26*:2069-2070.
20. Crooks GE, Hon G, Chandonia JM, Brenner SE: WebLogo: a sequence logo generator. Genome Res 2004, *14*:1188-1190.

21. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS: MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res 2009, *37*:W202-W208.
22. Do H, Dobrovic A: Dramatic reduction of sequence artefacts from DNA isolated from formalin-fixed cancer biopsies by treatment with uracil- DNA glycosylase. Oncotarget 2012, *3*:546-558.
23. Hofreiter M, Jaenicke V, Serre D, von Haeseler A, Paabo S: DNA sequences from multiple amplifications reveal artifacts induced by cytosine deamination in ancient DNA. Nucleic Acids Res 2001, *29*:4793-4799.
24. Bracho MA, Moya A, Barrio E: Contribution of Taq polymerase-induced errors to the estimation of RNA virus diversity. J Gen Virol 1998, *79 (Pt 12)*:2921-2928.
25. Do H, Wong SQ, Li J, Dobrovic A: Reducing sequence artifacts in amplicon-based massively parallel sequencing of formalin-fixed paraffin-embedded DNA by enzymatic depletion of uracil-containing templates. Clin Chem 2013, *59*:1376-1383.
26. Bodor C, Grossmann V, Popov N, Okosun J, O'Riain C, Tan K, Marzec J, Araf S, Wang J, Lee AM, Clear A, Montoto S, Matthews J, Iqbal S, Rajnai H, Rosenwald A, Ott G, Campo E, Rimsza LM, Smeland EB, Chan WC, Braziel RM, Staudt LM, Wright G, Lister TA, Elemento O, Hills R, Gribben JG, Chelala C, Matolcsy A, Kohlmann A, Haferlach T, Gascoyne RD, Fitzgibbon J: EZH2 mutations are frequent and represent an early event in follicular lymphoma. Blood 2013, *122*:3165-3168.

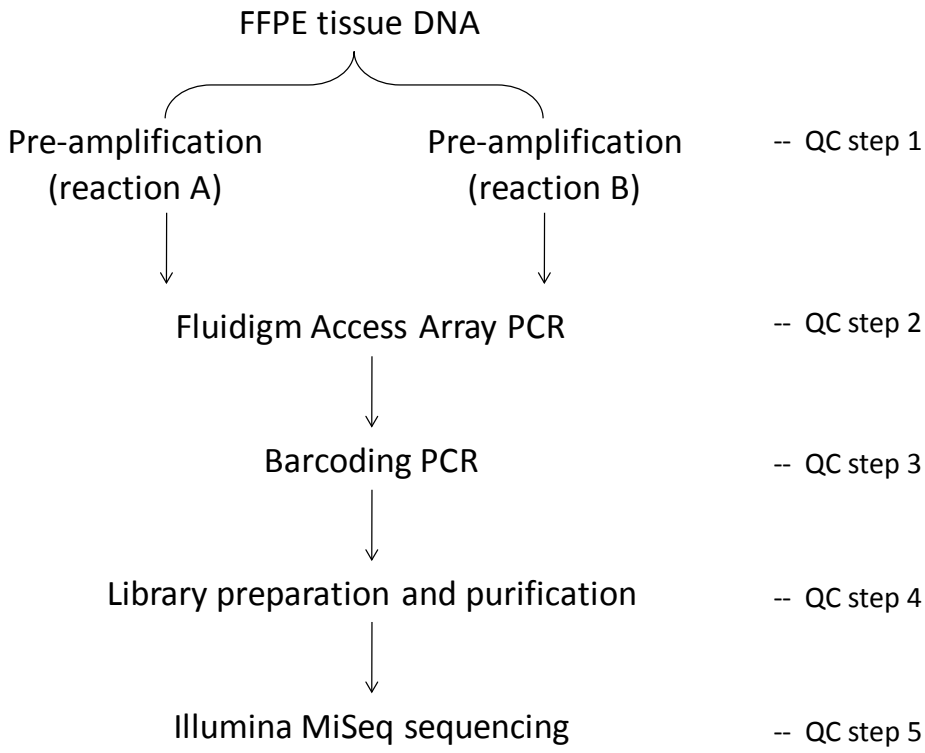


Figure 1. Outline of experimental design for mutation screening using Fluidigm Access Array PCR and Illumina MiSeq sequencing.

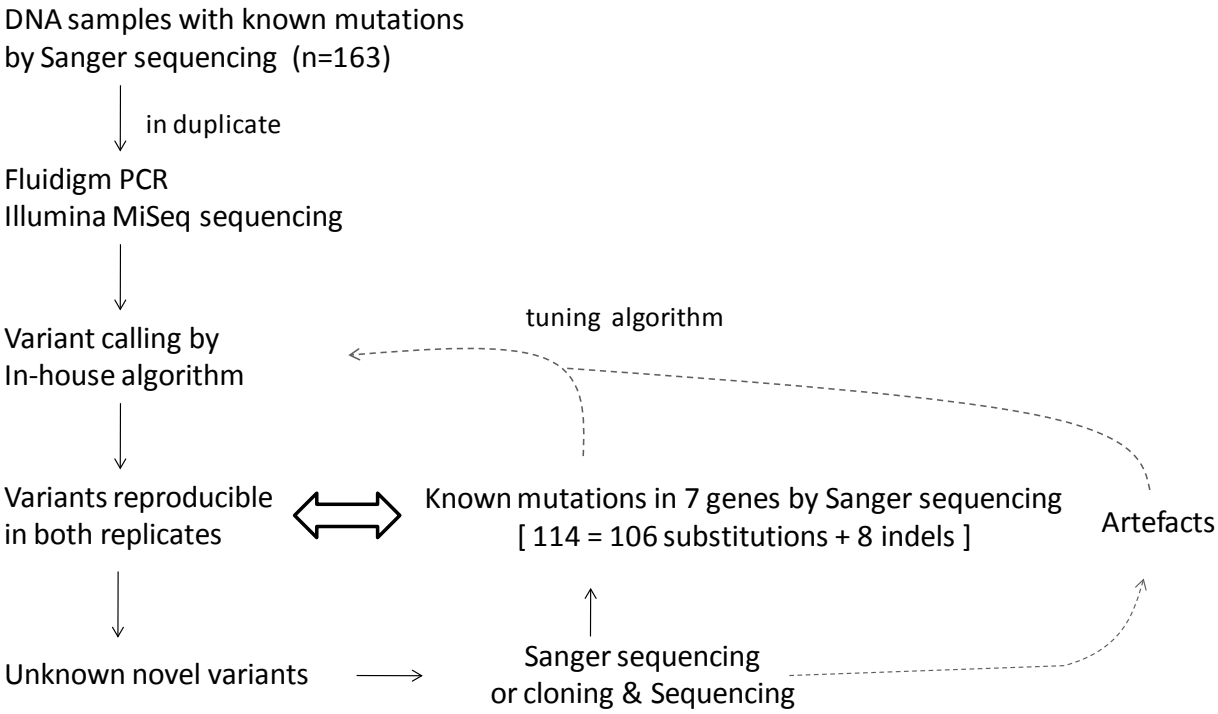


Figure 2. Strategies for development and improvement of in-house variant calling algorithm. At first, the performance of the in-house variant calling algorithm was assessed and "tuned" against 114 known mutations by Sanger sequencing. The additional novel variants identified by Fluidigm PCR/MiSeq sequencing were further validated by PCR and Sanger sequencing, where necessary by cloning and sequencing of the PCR products. The resulting sequence data were used to further fine-tune the algorithm. Taken together, a total of 159 Sanger sequencing confirmed somatic mutations including 147 substitutions and 12 indels (ranging 1-33bp) were used to optimise the algorithm.

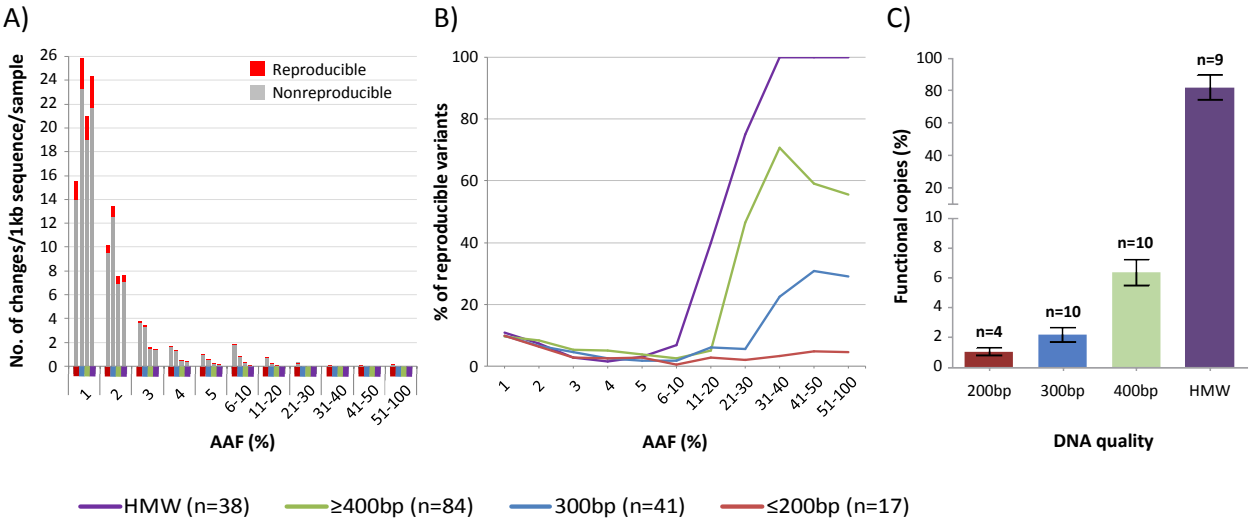


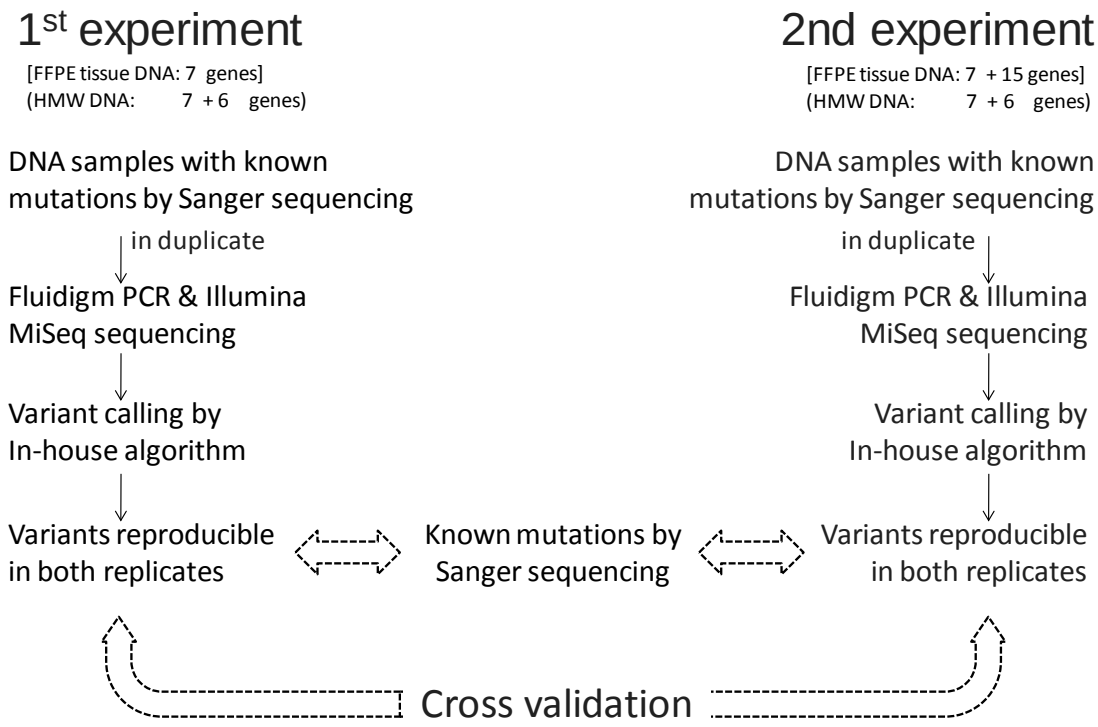
Figure 3. Impact of DNA quality on background noise by Fluidigm Access Array PCR and Illumina MiSeq sequencing.

A) The level of both reproducible (seen in both replicates) and non-reproducible (seen only in one replicate) variants according to AAF and DNA quality. FFPE DNA samples are further divided in to subgroups according to the size of genomic sequence amenable for amplification by conventional multiplex quality control PCR. The data from FFPE tissue DNA are based on 142 cases investigated by two sets of independent experiments (Figure 4) and are shown from one set of the experiments. The level of both reproducible and non-reproducible variants, particularly the latter, are remarkably high toward lower AAF. Due to small numbers of data points above 5% AAF, these data are combined and presented in groups as indicated.

B) The percentage of reproducible variants of the total (reproducible plus non-reproducible variants) according to AAF and DNA quality as above. Non-reproducible variants are background noise. Reproducible variants at high AAF are likely true mutations, but those at lower AAF are probably a mixture of false positives and subclonal genetic changes. Thus, the proportion of reproducible variants could indicate the level of background noise, and a putative threshold level of AAF to be used for detection of somatic mutation. The percentage of reproducible variants critically depends on AAF and DNA quality, being minimal at lower AAF, but increasing steadily at >10% AAF in HMW and FFPE tissue DNA samples amenable for PCR of ≥300bp genomic fragments, but not in those only supporting PCR of up to 200bp.

C) Quantification of functional copy number of DNA by TaqMan real time PCR. The level of functional copies amenable to PCR critically depends on DNA quality, being 80% in HMW DNA, but only ~1% in DNA samples amplifiable for PCR of up to 200bp.

A



B

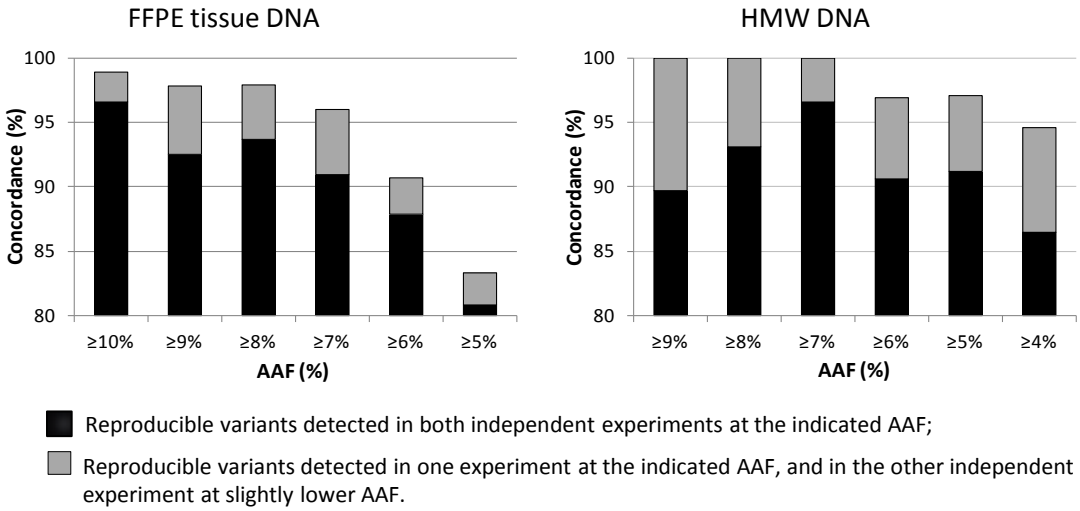


Figure 4: Determining the cut-off value of AAF for somatic mutation detection. A) Experimental strategy: two sets of independent experiments were performed and the reproducible variants detected in the 7 genes were cross validated, together with the known mutations by Sanger sequencing. B) Comparison of the reproducible variants from the 7 genes between the two independent experiments shows that the concordances critically depend on the level of AAF. For DNA samples from FFPE tissues, a cut-off value of ≥10% AAF yields 98.8% concordance, while for HMW DNA, the cut-off value can be as low as ≥7% AAF, generating 100% concordance.

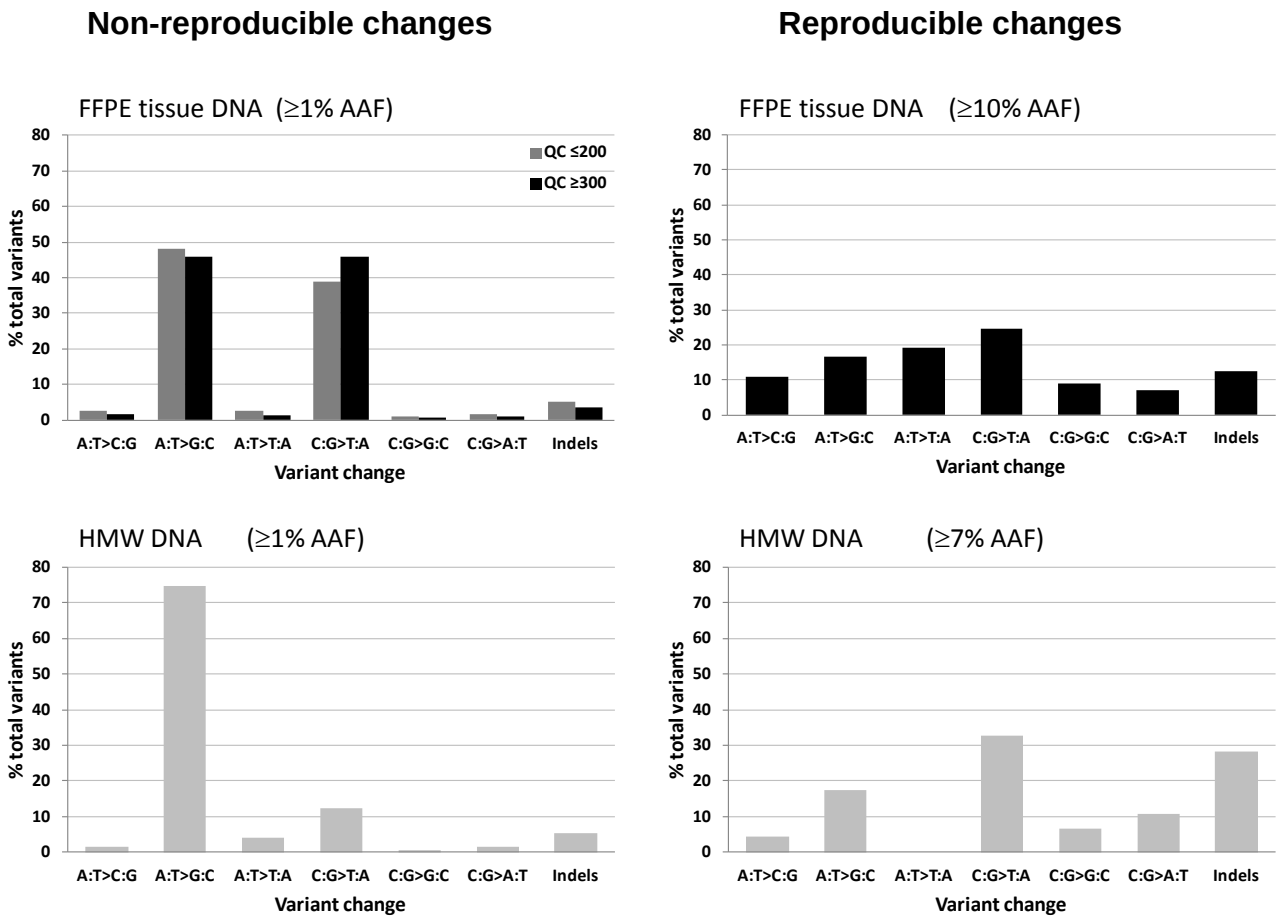


Figure 5: The nature of non-reproducible and novel reproducible changes in FFPE tissue and HMW DNA samples. For non-reproducible changes, there are marked differences in the frequencies of base changes between FFPE tissue and HMW DNA samples, with the frequencies of C:G>T:A change being remarkably higher in the FFPE tissue DNA. For novel reproducible changes, the spectrum of base changes between FFPE tissue and HMW DNA is similar, being broad without major bias toward any particular changes.

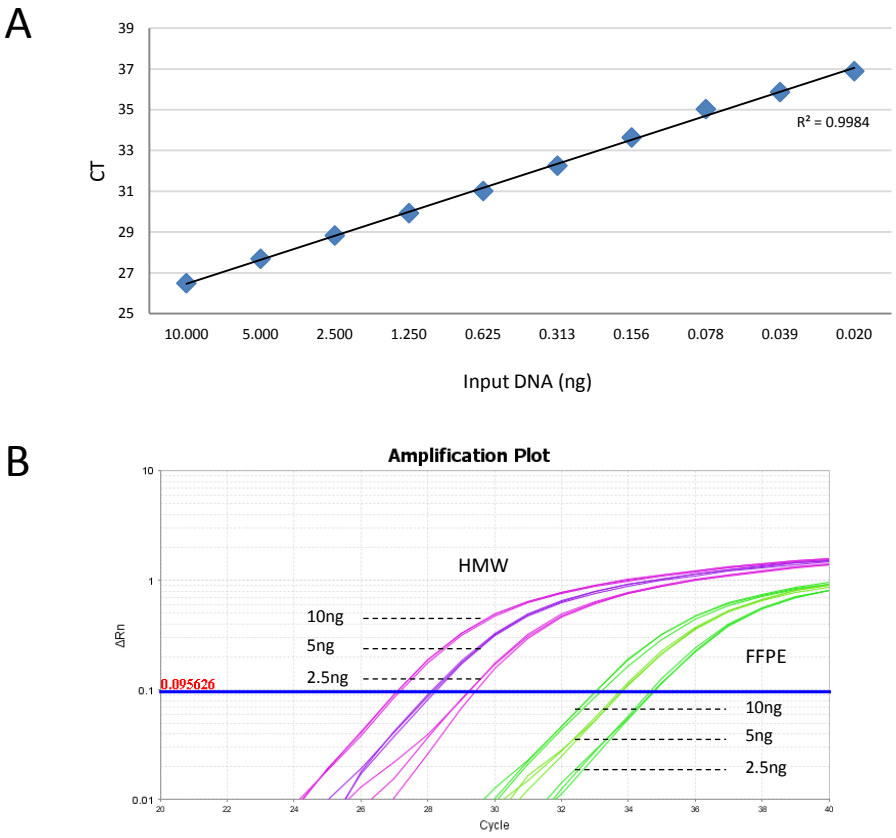
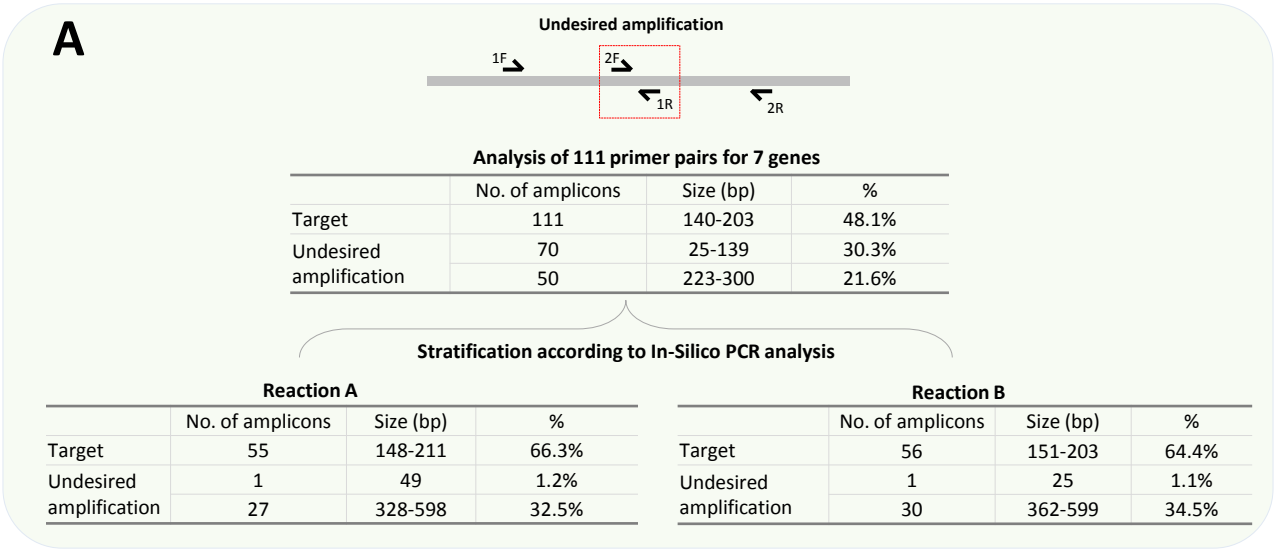


Figure S1. Quantification of functional copies of DNA by TaqMan real time PCR.

A) Standard curve. High quality human genomic DNA (Promega) was serially diluted, and input DNA ranging from 10ng to 0.020ng was quantified in triplicate by TaqMan real time PCR of a 195bp fragment of the PPIA gene to generate a standard curve.

B) Example of TaqMan real time PCR. For each sample, input DNA of 10ng, 5ng and 2.5ng was quantified in triplicate as above. The average estimated % of functional copy number relative to the standard curve was then calculated.



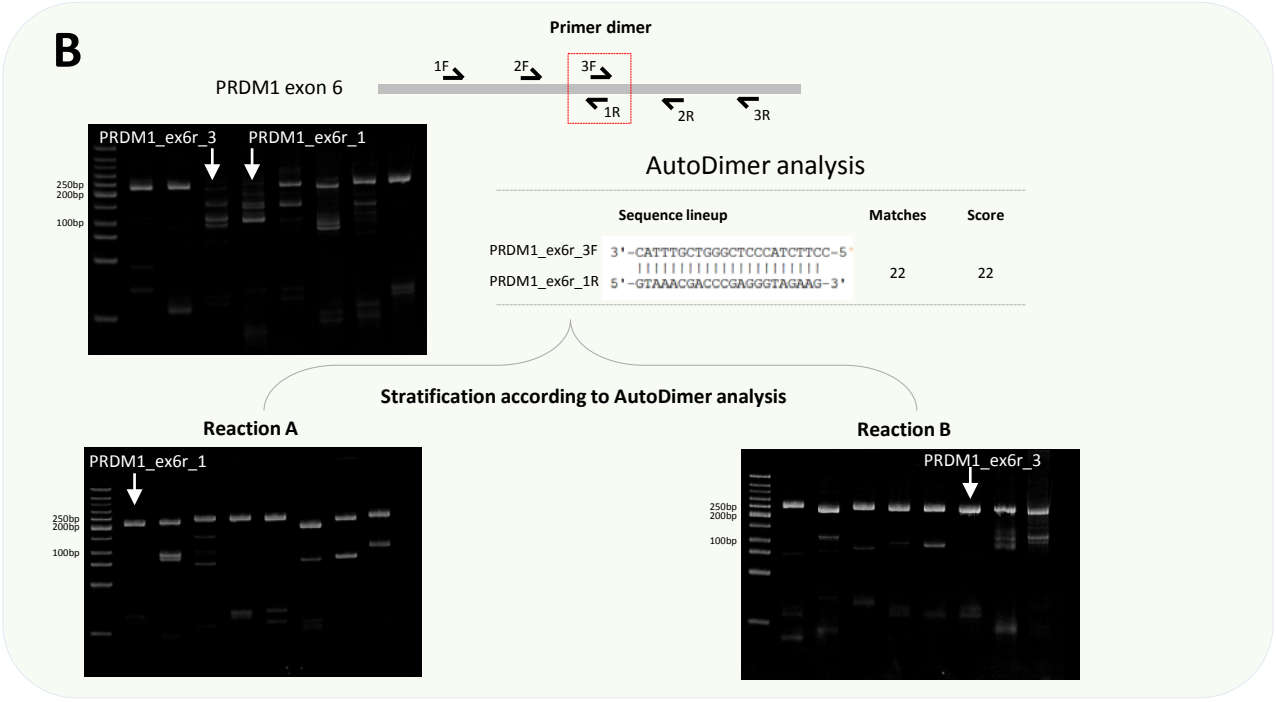
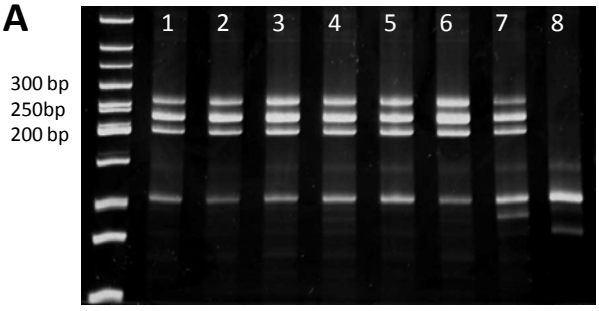


Figure S2. Stratification of PCR primer pairs for different pre-amplification reactions and multiplex PCR by In-Silico PCR and AutoDimer analysis.

A. In-Silico PCR analysis allows identification of primer pairs that give rise to undesired and non-specific amplification, and stratification of them in separate preamplification and multiplex PCR reactions. For example, analysis of the 111 primer pairs for the 7 gene panel identifies a number of putative undesired amplifications largely due to overlapping primers. The large sized putative undesired amplification (>300bp) are normally not amplified under the experimental conditions used and thus do not impose any problem. However, the small sized undesired amplifications are seen experimentally and these undesired amplifications cause failure of amplification of a proportion of primer pairs in Fluidigm Access Array PCR. These undesired amplifications can be efficiently prevented by placing the relevant primer pairs in different preamplification reactions as guided by In-Silico PCR analysis. After purification, the two separate preamplified products can then be pooled and used as template for multiplex PCR with Fluidigm Access Array.

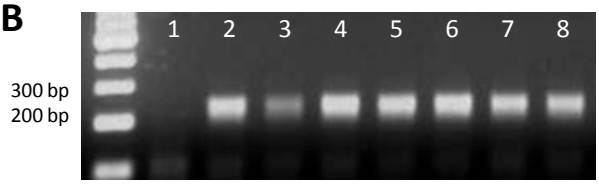
B. AutoDimer analysis identifies primers potentially interacting with each other. Any primers with an AutoDimer score above 12 should be placed in separate reactions. As shown here, PRDM1_ex6r_3 and PRDM1_ex6r_1 have a high degree of complementary sequences and prevent amplification of the targeted sequences when present together. By separating them into two separate reactions, both targeted genomic fragments can be efficiently amplified. PCR products were run on 10% PAGE gel.

A



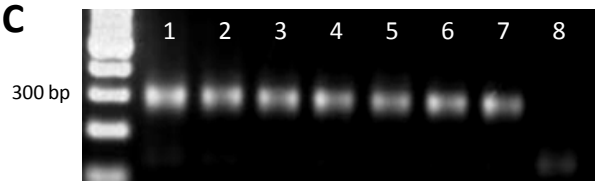
QC step 1: Assessment of preamplified products by conventional multiplex PCR and 10% polyacrylamide gel electrophoresis. 1µl of 5-fold diluted pre-amplified product was amplified with 3-4 primers pairs grouped based on In-Silico PCR and AutoDimer analyses. PCR was carried out in a 5µl reaction mixture containing 4.5mM $MgCl_2$, 5% DMSO, 200µM each dNTPs, 1.5µM of each primer, and 0.25U FastStart High Fidelity Enzyme Blend. The PCR conditions are identical to those used for Fluidigm Access Array (Table S2). Lanes 1-7: representative samples; Lane 8: negative control from preamplification reaction without template DNA.

B



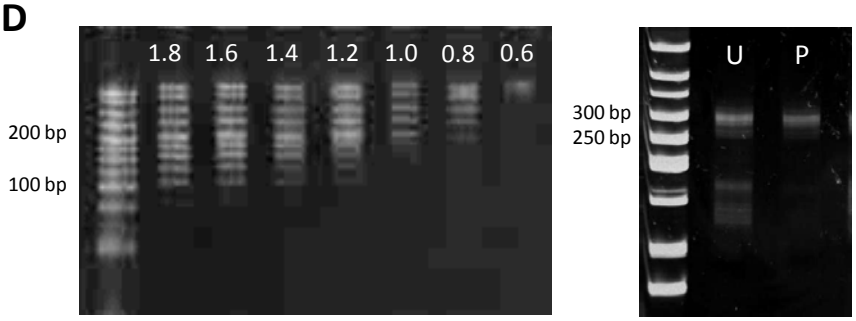
QC step 2: Assessment of Fluidigm Access Array PCR products by 2% agarose gel electrophoresis. Lane 1: a negative control went through preamplification and Fluidigm PCR; Lanes 2-8: individual samples.

C



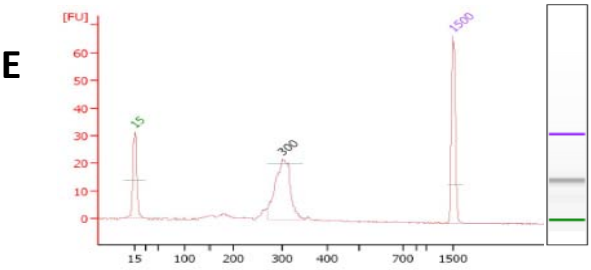
QC step 3: Assessment of barcoded PCR products by 2% agarose gel electrophoresis. Lanes 1-7: individual samples; Lane 8: a negative control went through preamplification, Fluidigm PCR and barcoding.

D



QC step 4: Assessment of library purification. Left panel: optimisation of AMPure XP beads:sample volume ratio using DNA ladder and the ratio at 0.8 gives the best purification, effectively removing DNA fragments <200bp. Right panel: The barcoded PCR products from various samples were pooled and purified as above. U: un-purified product; P: purified product.

E



QC step 5: Assessment of library quality by Bioanalyser**.** The above purified library was routinely assessed using Agilent DNA 1000 Bioanalyser. The 300 bp peak represents the purified library products, while the 15 bp and 1500 bp peaks are ladder marker. FU (Fluorescence Units).

Figure S3. Quality control at various steps of Fluidigm Access Array PCR and Illumina MiSeq sequencing.

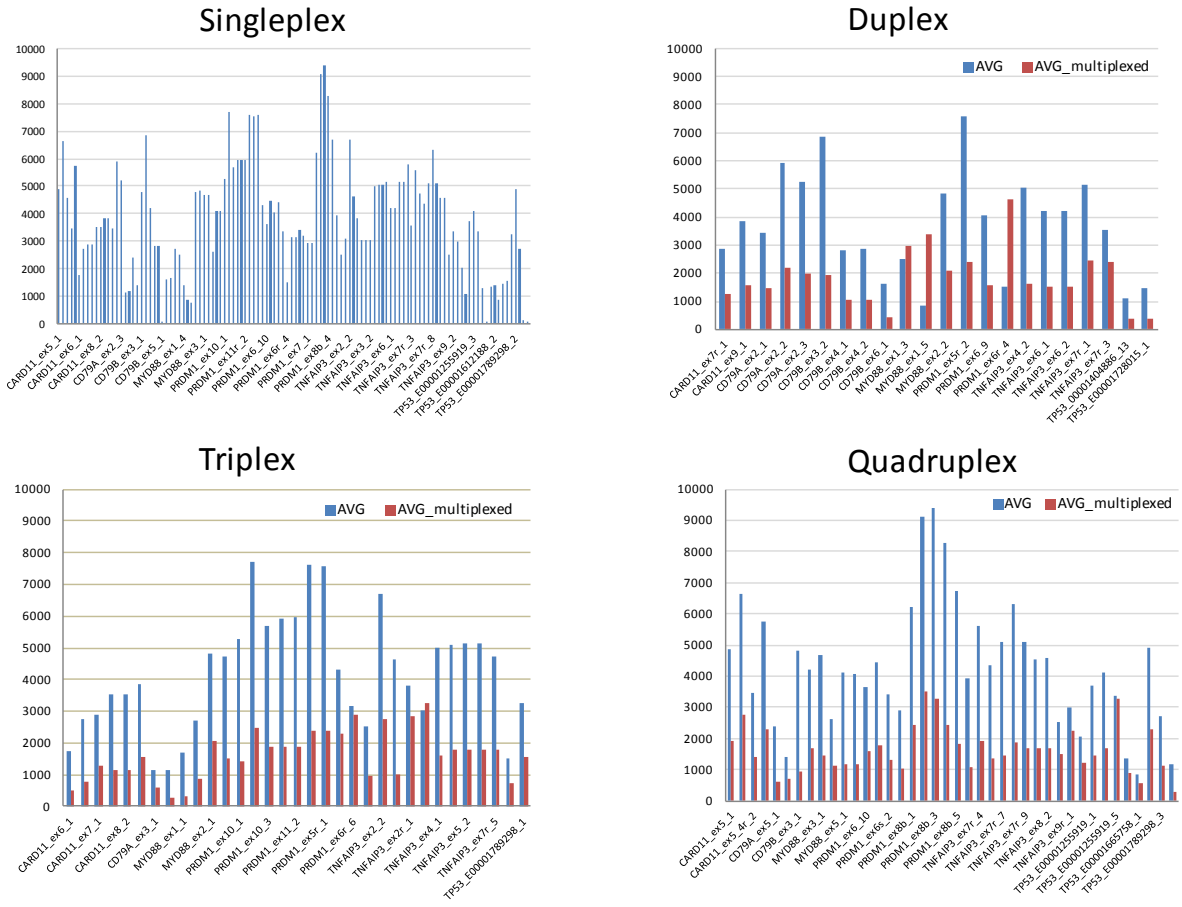
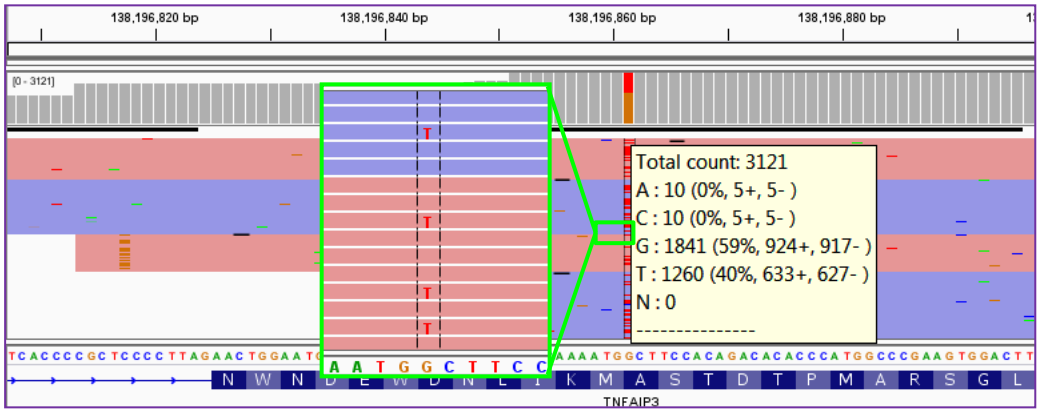
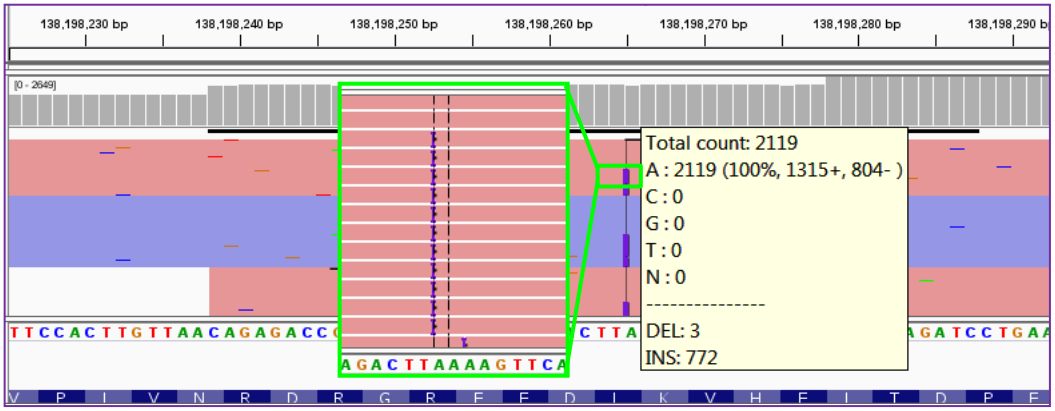


Figure S4. Examples of depth of coverage by Fluidigm Access Array PCR and Illumina MiSeq sequencing. Adequate depth of coverage can be achieved for each of the amplicons by multiplex Fluidigm Access Array PCR with up to 4 primer pairs. AVG: average.

Substitution:
L0092
TNFAIP3
c.589G>T



Insertion:
L0092
TNFAIP3
c.923 T>TAA



Deletion:
L0112
TNFAIP3
c.973 TTAAA>T

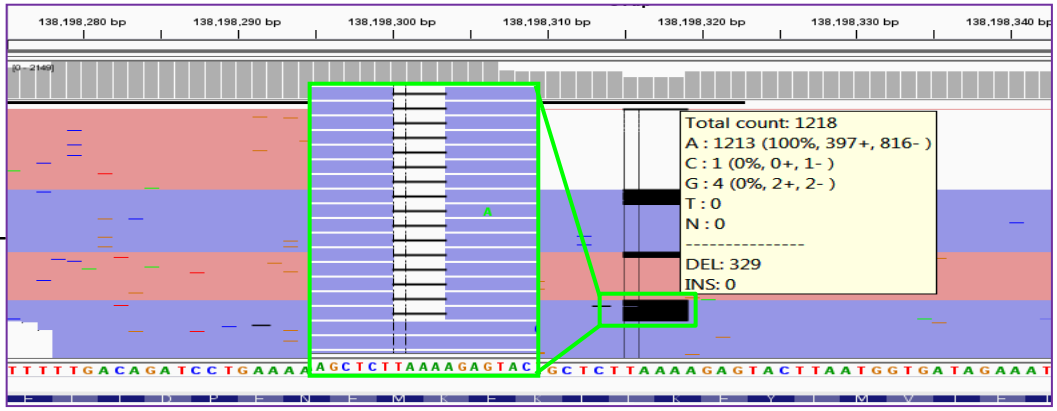


Figure S5. Examples of somatic mutations detected by Fluidigm Access Array PCR and Illumina MiSeq sequencing.

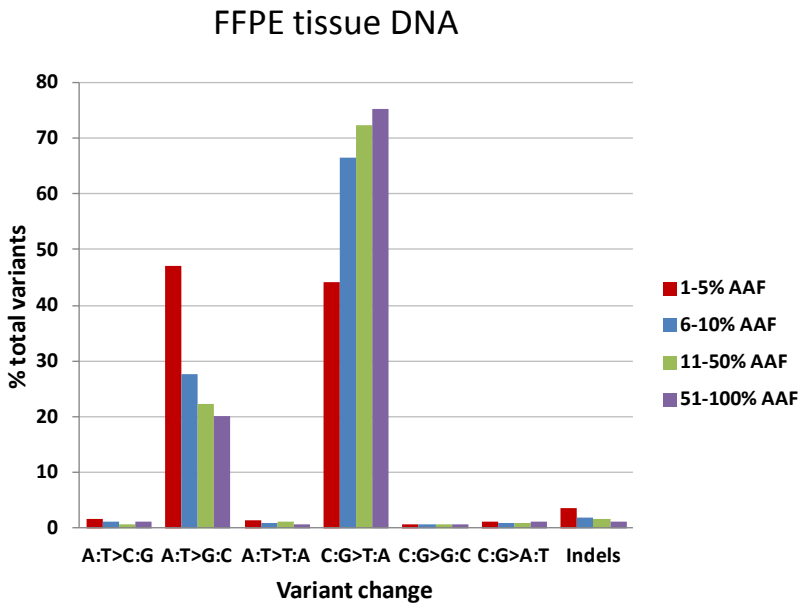


Figure S6. The nature of non-reproducible variants is independent of AAF.