Mutation Screening Using Formalin-Fixed Paraffin-Embedded Tissues: A Stratified Approach According to DNA Quality

Francesco Cucco,[1*] Alexandra Clipson,[1*] Hannah Kennedy,[1] Joe Sneath Thompson,[1] Ming Wang,[1] Sharon Barrans,[2] Moniek Van Hoppe,[2] Eguzkine Ochoa Ruiz,[1] Josh Caddy,[3] Debbie Hamid,[3] Thomas Cummin,[4] Cathy Burton,[2] Andrew J Davies,[4] Peter Johnson,[4] Ming-Qing Du[1,5]

*These authors contributed equally to this study.

[1]Department of Pathology, University of Cambridge, Cambridge, United Kingdom;
[2]HMDS, Leeds Cancer Centre, Leeds Teaching Hospitals NHS Trust, Leeds, United Kingdom;
[3]Southampton Clinical Trials Unit, University of Southampton, Southampton, United Kingdom;
[4]Cancer Research UK Clinical Centre, University of Southampton, Southampton, United Kingdom
[5]Department of Histopathology, Addenbrooke's Hospital, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK;

Running title: Targeted sequencing using FFPE tissues
Key words: targeted sequencing, mutation analysis, FFPE tissue, DNA quality

Correspondence to
Professor Ming-Qing Du,
Division of Cellular and Molecular Pathology,
Department of Pathology
University of Cambridge
Box 231, Level 3, Lab Block
Addenbrooke's Hospital,
Hills Road
Cambridge, CB2 2QQ
United Kingdom
Tel: +44 (0)1223 767092
Fax: +44 (0)1223 586670
Email: mqd20@cam.ac.uk

ABSTRACT

DNA samples from formalin-fixed paraffin-embedded tissues are highly degraded with variable quality, and this imposes a big challenge for targeted sequencing due to false positives, largely caused by PCR errors and cytosine deamination. To eliminate false positive, a common practice is to validate the detected variants by Sanger sequencing or perform targeted sequencing in duplicate. Technically, PCR errors could be removed by molecular barcoding of template DNA prior to amplification as in the HaloPlexHS design. Nonetheless, it is uncertain to what extent variants detected using this approach should be further validated. Here, we addressed this question by correlating variant reproducibility with DNA quality using HaloPlexHS target enrichment and Illumina HiSeq4000, together with an in-house validated variant calling algorithm. The overall sequencing coverage, as shown by analyses of 70 genes in 266 cases of large B-cell lymphoma, was excellent (98%) in DNA samples amenable for PCR of ≥400bp, but suboptimal (92%) and poor (80%) in those amenable for PCR of 300bp and 200bp respectively. By mutation analysis in duplicate in 93 cases, we demonstrated that 20 alternative allele depth (AAD) was an optimal cut-off value for separating reproducible from non-reproducible variants in DNA samples amenable for PCR of ≥300bp , with 97% sensitivity and 100% specificity. By cross validation with a previously established targeted sequencing protocol by Fluidigm-PCR and Illumina MiSeq, the HaloPlexHS protocol was shown to be highly sensitive and specific in mutation screening. To conclude, we proposed a stratified approach for mutation screening by HaloplexHS and Illumina HiSeq according to DNA quality. DNA samples with good quality (≥400bp) are amenable for mutation analysis with a single replicate, with only variants at 15-20 AAD requiring for further validation, while those with suboptimal quality (300bp) are better analysed in duplicate with reproducible variants at >15 AAD regarded as true genetic changes.

INTRODUCTION

Next generation sequencing (NGS) has made an unprecedented contribution to discoveries in the biomedical research field. One of its many applications is targeted sequencing to screen mutations in a panel of interesting genes. In view of the rapid discovery in cancer research and huge impact from mega genome sequencing initiatives such as the 100k genome project, it is imperative to establish highly sensitive, specific and also robust targeted sequencing protocol that is amenable to degraded DNA from formalin fixed paraffin embedded (FFPE) tissues/cells, allowing translational research as well as routine clinical application. There are several target enrichment methods available, based on hybridisation capture (Agilent SureSelect or NimbleGen SeqCap products), PCR (Fluidigm Access Array PCR, RainDance technology) or a combination of both hybridisation capture and PCR (HaloPlexHS). The enriched target sequence library can then be sequenced by an Illumina or Ion Torrent platform. Among these different approaches, PCR-based targeted enrichment is commonly used due to its easy application to minute amounts of template DNA. One of the major issues in such an application is the uncertainty whether the variants detected are true genetic changes or false positives. A common practice is to further verify the variants detected by an independent method such as conventional Sanger sequencing or droplet PCR followed by next generation sequencing.[1]

In a recent study, we have established a targeted sequencing protocol by Fluidigm Access Array PCR and Illumina MiSeq sequencing together with an in house validated variant calling pipeline that was optimised against a large number of various known mutations.[2] False positives were eliminated by performing the targeted sequencing and data analyses in duplicate with only the variants, which appeared in both replicates and above the experimentally defined cut-off value of alternative allele frequency (AAF), being regarded as true genetic changes.[2] Detailed analyses of false positives revealed their distinct nature and origins between high molecular weight DNA from fresh frozen tissues and degraded DNA from FFPE tissues. For high molecular weight DNA, the majority of false positives are derived from PCR/sequencing errors, while for FFPE tissue DNA, the false positives are from both PCR/sequencing errors and cytosine deamination, which is caused by tissue formalin fixation and storage.[2-6]

The more recent HaloPlexHS target enrichment design incorporates molecular barcode in hybridisation probes, and this allows removal of PCR errors during sequence data analysis. Apart from this, the HaloPlexHS target enrichment also offers several other advantages including easy probe design, highly flexible in the number of genes to be investigated, and a streamlined experimental protocol. Nonetheless, it remains to be investigated whether mutation analysis can be reliably performed in a single replicate with DNA samples from FFPE tissues by the HaloPlexHS target enrichment protocol. In the present study, we have optimised a protocol for high throughput mutation screening by HaloPlexHS target enrichment and Illumina sequencing, and established a practical strategy for reliable mutation detection using DNA samples from FFPE tissues. The strategy allowed stratification of DNA samples into different protocols of mutation analyses according to their quality, with good quality DNA investigated in a single replicate, while sub-optimal quality DNA was analysed in duplicate.

MATERIALS AND METHODS

Tumour materials and DNA extraction:

FFPE lymphoma specimens were retrieved from 266 cases of diffuse large B-cell lymphoma (DLBCL) enrolled to the REMoDL-B and MaPLe trials. Local ethical guidelines were followed for the use of these tissue materials for research with the approval of the ethics committees of the involved institutions.

Haematoxylin and eosin slides were reviewed and tumour rich areas (>40%) in each specimen were isolated by crude microdissection for DNA extraction. DNA was extracted using the QIAamp DNA Micro Kit (QIAGEN, Crawley, UK) and quantified using Qubit® Fluorometer (Life Technologies, UK).

Assessment of DNA quality by conventional PCR.

This was performed by PCR of variably sized genomic fragments using 2ng template DNA in a 10μl reaction mixture for 40 cycles under a standardised protocol as described previously.[2]

Targeted sequencing by Fluidigm multiplex PCR and Illumina MiSeq sequencing

This was used to investigate mutation in 22 genes in 60 cases of DLBCL as described previously.[2] Each DNA sample was simultaneously investigated in duplicate. Briefly, 50ng genomic DNA was used for preamplification and Fluidigm Access Array PCR, followed by barcoding and Illumina MiSeq sequencing. Variants were identified using an in-house developed and validated variant caller python program.[2] After filtering baseline sequence errors and germline changes through SNP database search, novel variants seen in both replicates of the same sample were recorded and those above 10% AAF (alternative allele frequency) were regarded as true changes as defined previously.[2]

Gene panel and target enrichment by HaloPlexHS

A total of 70 genes (~205kb sequence) that are recurrently mutated in aggressive B-cell lymphomas were included in the HaloPlexHS target enrichment design, and they included the 22 genes that were investigated by Fluidigm multiplex PCR and Illumina MiSeq sequencing as outlined above.[2] The HaloPlexHS target enrichment design incorporates molecular barcodes in the hybridisation probes, thus allowing the removal of PCR errors during sequence data analysis (Agilent Technologies). HaloPlexHS target enrichment was performed essentially according to the manufacturer's instructions for FFPE tissue samples. Briefly, 100ng of genomic DNA was digested with restriction enzymes, and hybridized to the above customised HaloPlexHS probe library. The probe-target DNA hybrids were ligated and circularized with HS DNA ligase, then purified with streptavidin beads and finally amplified by PCR.

Library purification and Illumina Sequencing

The above amplified target library was purified twice using AMPure XP beads (Beckman Coulter, Pasadena, CA) to remove fragments below the expected target size. The purified target library from each sample was then validated and quantified using the 4200 TapeStation (Agilent Technologies), and pooled with appropriate adjustment according to their concentration. The pooled libraries were then sequenced on one of the following Illumina platforms: MiSeq (2x250bp end sequencing protocol), HiSeq2500 (Rapid Run Mode 2x150bp end sequencing protocol), or HiSeq4000 (2x150bp end sequencing protocol).

Variants calling and data analysis

Demultiplexing and conversion from bcl to fastq were performed using bcl2fastq v2.19, followed by read trimming and adaptor sequence removal with SurecallTrimmer from the Agilent Genomics NextGen Toolkit v4.0.1 (AGeNT). The reads were aligned to hg38 using bwa mem v0.7.17 and deduplication was carried out using the AGeNT LocatIt tool. The resultant sam files were converted to bam with samtools v1.3.1, which was also used for sorting and indexing of bam files.[7]

For SNV detection, bam files were processed using a pipeline based on GATK v3.6 best practices including indelrealigned and recalibration steps. The calling variant was run using UnifiedGenotyper with 10000 to prevent downsampling.[8] As GATK was unable to call SNVs at <8% AAF reliably, MuTect2 was additionally employed for detection of hotspot mutations at low AAF values. Indel detection was separately carried out on the recalibrated bam files using Pindel v0.2.5,[9] which allowed detection of indels as low as 2% AAF.

Variant call files were concatenated to produce one library vcf each for the SNV and Indel pipelines. These library files were then filtered using a combination of vcftools v0.1.15 and bedtools v2.25 for read depth, quality score, and known PCR/sequence artefacts.[10,11] Further filtering was accomplished using an in-house script to remove variants in intronic regions outside essential splicing sites, SNPs with a minor allele frequency <1% and synonymous changes. The resulting novel variants were further scrutinised by reviewing bam file to eliminate any potential PCR/sequence artefacts.

The above in house variant calling pipeline was optimised and validated using two virtual sequence libraries containing a large number of various known somatic mutations: one contained lymphoma associated mutations (60 SNVs, 19 indels and 6 splicing variants), while the other included brain tumour associated mutations (5 SNVs, 11 indels), largely unrelated to the current study.

RESULTS

1) Optimisation of experimental protocols

In the initial experiments, a series of testing experiments were performed in duplicate to optimise the experimental protocols. First, various amounts of template DNA (50ng, 100ng) from representative FFPE DLBCL tissue specimens (n=9) with known mutations in 22 of the 70 genes investigated were used for HaloPlexHS target enrichment, followed by Illumina sequencing. Based on the coverage and depth of sequencing, and the efficacy of detecting known mutations, 100ng DNA was considered as the minimal optimal amount of template DNA for the targeted sequencing (Figure S1A). Second, we compared the sequence coverage among different Illumina sequencing platforms, namely MiSeq, HiSeq2500 and HiSeq4000 with compatible amounts of target enrichment library according to their sequencing capacity. The results showed that HiSeq4000 yielded the highest overall sequence coverage and depth reads (Figure S1B). Hence, Illumina HiSeq4000 was chosen for all the subsequent experiments, with all the data presented below being derived from optimised HaloPlexHS target enrichment and Illumina HiSeq4000 sequence unless otherwise specified.

2) Impact of DNA quality on sequence coverage and variant reproducibility

A total of 266 FFPE DLBCL tissue specimens were investigated by targeted sequencing of 70 genes using HaloPlexHS target enrichment and Illumina HiSeq4000. As expected, the quality of DNA samples had a major impact on the target library quantity, overall sequence coverage as well as the read depth, with better quality of DNA samples clearly showing higher performance of these parameters (Figure 1A-D). In addition, the DNA samples amplifiable for ≥400bp genomic fragments showed a much lesser extent of variation in their sequence coverage than those only amplifiable for up to 300bp (Figure 1D).

To examine whether DNA quality had any impact on the reproducibility of variant detection, we next focused on the 93 cases where duplicate experiments were carried out. The variants identified by the in house variant calling pipeline were filtered to remove variants that were deemed unacceptable: <50 total read depth (TD), <5 alternative allele read depth (AAD), <2% alternative allele frequency (AAF), as well as known PCR/sequencing artefacts. The resulting novel variants were interrogated between the two replicates, and recorded as reproducible or non-reproducible changes, with the latter group being most likely false positive (Figure S2). As shown in Figure 1E, the proportion of non-reproducible variants clearly depended on the quality of DNA, being much higher in DNA samples amplifiable for ≤300bp than those amplifiable for ≥400bp genomic fragment. Importantly, a high proportion of these non-reproducible variants had a high AAF, limiting its value to separate reproducible from non-reproducible changes (Figure 1F).

3) Determining cut-off parameters for reliable variant detection

The HaloPlexHS target enrichment design incorporates a molecular barcode, allowing removal of PCR duplicates, thus PCR errors. In view of this, the number of reads that bear novel variants, i.e. alternative allele depth (AAD), would represent the copy number of "mutant" template that are successfully captured and sequenced. Theoretically, this would be a good parameter to distinguish true genetic changes from false positives as the higher the AAD, the higher the probability of a variant originating from template DNA rather than experimental artefact. As expected, all non-reproducible variants in DNA samples amplifiable for ≥400bp or up to 300bp genomic fragment were at low AAD values with mean plus 2SD being <15 in both groups (Figure 2A&B). To ensure highly specific mutation detection, we used 20 AAD as the cut-off value for dichotomy between reproducible and non-reproducible variants. This allowed detection of 97% reproducible variants with 100% specificity. As all non-reproducible variants are below this cut-off value, theoretically a single replicate could be sufficient for reliable mutation detection if this threshold proved to be highly efficient and specific. To verify this, we next examined the concordance in mutation detection between HaloPlexHS /Illumina HiSeq and Fluidigm/Illumina MiSeq approaches.

Non-reproducible variants in DNA samples amplifiable for only up to 200bp genomic fragment were rather dispersed, with a high proportion showing relatively high AAD values (Figure 2C).

4) Validation of mutation detection by HaloPlexHS /Illumina HiSeq sequencing

Among the cases investigated by HaloPlexHS /Illumina HiSeq4000, 60 (22 in duplicate and 38 in a single replicate) were also investigated for mutation in 22 genes by Fluidigm Multiplex PCR/Illumina MiSeq in duplicate. The novel variants that were considered as true genetic changes were identified independently by their respective protocols and thresholds,[2] and then compared between the two different targeted sequencing approaches.

Within the 22 cases that were investigated in duplicate by both Fluidigm and HaloPlexHS approaches, 60 and 61 novel variants were identified in the common region of 22 genes covered by these methods respectively, with 58 variants being mutually detected by both methods (Figure 3A). The 2 variants detected by Fluidigm but not HaloPlexHS approach were found by the HaloPlexHS approach but not called as both variants were at the end of their amplicon, while the 3 variants detected by HaloPlexHS , but not Fluidigm approach were due to low AAF value.

Of the 38 cases that were investigated in duplicate by the Fluidigm, but a single replicate by the HaloPlexHS approach, 96 and 98 novel variants were identified in the common region of 22 genes covered by these methods respectively, with 91 variants being mutually detected by both methods (Figure 3B). The 5 variants detected by Fluidigm but not HaloPlexHS approach were due suboptimal AAD (6 and 12) in 2, low AAF in 2, and at the end of its amplicon in 1, while the 7 variants detected by HaloPlexHS , but not Fluidigm approach were due to low AAF value.

Taken together, the above findings indicate that the target sequencing protocol by HaloPlexHS /Illumina HiSeq was highly sensitive and specific in mutation screening, compatible to the Fluidigm/Illumina MiSeq approach established previously.[2] This was further supported by the similar mutation frequencies in DLBCL between the present and published studies.

5) Nature of false positives from FFPE tissue DNA by HaloPlexHS target enrichment

For non-reproducible changes, C:G>T:A and C:G>A:T alterations accounted for the majority of single substitution changes, with other base changes including A:T>G:C (a feature of PCR errors) being at relatively low frequencies (Figure S3). In contrast, a wide spectrum of substitution changes was seen for the reproducible changes above the 20 AAD cut-off value. There was no apparent correlation between the type of substitution changes and the quality of DNA samples.

DISCUSSION

In this study, we have developed a targeted sequencing protocol for mutation screening using HaloPlexHS target enrichment, Illumina HiSeq platform and an in house variant calling algorithm that was validated against virtual sequence libraries containing a large number of known mutations. By performing the targeted sequencing in duplicate and validating the detected variants against an independent targeted sequencing protocol, namely Fluidigm Access Array PCR and Illumina MiSeq sequencing,[2] we have established a stratified approach for mutation screening using DNA samples from FFPE tissues (Figure 4).

Stratified targeted sequencing approach according to DNA quality

Under the defined experimental and data analysis protocol and the proposed 20 AAD cut-off value, reliable mutation screening could be achieved by a single replicate for DNA samples amenable for good amplification of ≥400bp genomic fragments. This is evident by the excellent overall sequence coverage, read depth and highly sensitive and specific detection of known mutations in these samples.

For DNA samples showing good amplification of up to 300bp genomic fragments, reliable mutation detection could still be obtained by a single replicate in the sequence regions that were adequately captured and sequenced. This is evident by absence of any non-reproducible variants above the 20

AAD cut-off value. However, the overall sequence coverage in these samples is relatively low, with an average of ~8% (ranging 4-16%) targeted sequences not adequately captured and sequenced when only a single replicate was carried out (Figure 1D). In addition, the extent of non-reproducible variants is much higher than those amenable for PCR 0f >400bp (Figure 1E). In view of these concerns, these samples are best sequenced in duplicate, which not only improves the sequence coverage moderately, but also enables mutation detection based on reproducible variants, further ensuring the detection specificity (Figure S4).

For DNA samples that are amplifiable for only up to 200bp genomic fragments, reliable mutation screening could not be achieved due to poor sequence coverage and substantial non-reproducible variant above the proposed AAD cut-off value (Figure 1B-D, Figure 2C). Such samples are best excluded from targeted sequencing analysis.

It is important to note that the above AAD cut-off value for dichotomy between reproducible and non-reproducible variants and the proposed stratified targeted sequencing approach according to DNA quality were based on the experimental and data analysis protocols defined in this study. The AAD cut-off value most likely varies depending on the experimental protocol and data analysis pipeline, and thus should be experimentally determined for the methodology to be employed.

Although the cost for next generation sequencing is decreasing, targeted sequencing is still costly due to expensive target enrichment kit and relatively high cost for running fast sequencing platforms such as Illumina MiSeq and HiSeq2500. The above stratified approach would improve the cost effectiveness. Based on our ongoing works, 92% of DNA samples from FFPE DLBCL tissue biopsies recruited to an ongoing MaPLe clinical trial would be adequate for targeted sequencing using a single replicate. Nonetheless, the proportion of such good DNA samples was relatively lower in an early REMoDL-B trial (tissue specimens stored for an average 4 years, range 2-6 years) (87%) and archival population based DLBCL (tissue specimens stored for an average 9 years, range 5 -13 years) (44%), most likely reflecting variation and deterioration of DNA quality during tissue storage.

Detection of low burden mutations

As shown in Figure 2A&B, there was a considerable overlap among the reproducible and non-reproducible variants below the 20 AAD cut-off value, with those below 15 AAD being largely non-reproducible changes. For the variants between 15-20 AAD, it is not possible to ascertain their nature, i.e. true genetic change or false positive if targeted sequencing was performed in a single replicate. Nonetheless, these variants were small in number, could be validated by an independent approach such as Sanger sequencing (Figure 4). A high proportion of these variants could be subclonal genetic changes, and may be beyond the sensitivity of Sanger sequencing detection. Practically, these variants could be amplified by conventional PCR, selectively pooled and barcoded, then sequenced with an Illumina MiSeq platform.

To maximise the sensitivity and specificity of mutation detection, it is critical to use specimens with high tumour cell content for DNA extraction, and where necessary to enrich tumour cell population by microdissection, as this will improve the separation of true genetic changes from background noise. We compared the extent of reproducible variants below 20 AAD according to tumour load, and found that this accounted for only 2% of the total reproducible variants in DNA samples with 90%

tumour cell content, but 5% of the total reproducible variants in those with 20-50% tumour cell content.

Nature of non-reproducible variants

In the previous study by Fluidigm Access Array PCR and Illumina MiSeq, the vast majority of non-reproducible variants in DNA samples from FFPE tissues were PCR errors (A:T>G:C changes) and changes resulting from cytosine deamination (C:G>T:A).[2] In contrast to the Fluidigm approach, the present HaloPlexHS protocol showed a low level of A:T>G:C changes among non-reproducible variants, indicating that the incorporation of a molecular barcode in the HaloPlexHS target enrichment design was highly efficient in removing PCR errors. Importantly, this has made it possible for reliable mutation screening by targeted sequencing with a single replicate when the quality of DNA sample is adequate.

In summary, we have established a protocol with defined cut-off values for highly sensitive and specific mutation screening by HaloPlexHS target enrichment and Illumina sequencing, and provided a practical and stratified approach for mutation analysis using DNA samples from FFPE tissues according to their quality. We are currently using this established protocol to perform mutation profiling in DLBCL recruited to the REMoDL-B and MaPLe trials.

Acknowledgements: The authors would like to thank Shubha Anand and Yuanxue Huang for their assistance with using TapeStation, and Graeme Clark and Ezequiel Martin for their assistance with Illumina sequencing.

Sources of Research support: The research was supported by grants from Bloodwise (13006, 15002, 15019) UK, and Kay Kendal Leukaemia Fund (KKL582), UK.

There is no conflict of interest to declare.

Author contributions: Experimental design, data collection and analysis: FC, AC, HK, MW, SB, MVH; Illumina sequencing analysis and variant calling: JST, EOR; Case contribution: SB, MB, JC, DH, TC, CB, AJD, PJ; Manuscript writing and preparation: MQD, FC, HK, JST; Study design and coordination: MQD, PJ, AJD, SB, TC. All authors commented on the manuscript and approve its submission for publication.

REFERENCES:

1. Robasky K, Lewis NE, and Church GM. The role of replicates for error mitigation in next-generation sequencing. Nat Rev Genet 2014;15:56-62.

2. Wang M, Escudero-Ibarz L, Moody S, et al. Somatic Mutation Screening Using Archival Formalin-Fixed, Paraffin-Embedded Tissues by Fluidigm Multiplex PCR and Illumina Sequencing. J Mol Diagn 2015;17:521-532.

3. Bracho MA, Moya A, and Barrio E. Contribution of Taq polymerase-induced errors to the estimation of RNA virus diversity. J Gen Virol 1998;79 (Pt 12):2921-2928.

4. Do H and 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.

5. Hofreiter M, Jaenicke V, Serre D, et al. DNA sequences from multiple amplifications reveal artifacts induced by cytosine deamination in ancient DNA. Nucleic Acids Res 2001;29:4793-4799.

6. Bourgon R, Lu S, Yan Y, et al. 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.

7. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009;25:2078-2079.

8. DePristo MA, Banks E, Poplin R, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 2011;43:491-498.

9. Ye K, Schulz MH, Long Q, et al. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics 2009;25:2865-2871.

10. Danecek P, Auton A, Abecasis G, et al. The variant call format and VCFtools. Bioinformatics 2011;27:2156-2158.

11. Quinlan AR and Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 2010;26:841-842.

FIGURE LEGENDS

Figure 1. Impact of DNA quality on sequence coverage and variant detection reproducibility.

A) Assessment of DNA quality by PCR of variably sized genomic fragments under a standardised protocol.[2]

B) Average reads before and after removal PCR duplicate according to DNA quality. For DNA samples amenable for PCR of >400bp, an average of at least 1300 reads after deduplication is achieved.

C) Overall sequence coverage according to DNA quality. Sequence coverage is calculated based on per nucleotide and the percentage of the targeted sequences that are covered by more than 50 reads is given. DNA samples that support PCR of variable genomic fragment are indicated accordingly.

D) Correlation among DNA quality, target library quantity and overall sequence coverage (>50 reads). In general, DNA samples that support PCR of >400bp genomic fragment yield much higher quantity of target sequence libraries and excellent sequencing coverage, while those that support PCR of only up to 300bp genomic fragment generate much lower quantity of target sequence libraries and suboptimal sequencing coverage. Mean and SD of sequencing coverage are provided according to DNA quality.

E) Extent of non-reproducible variants according to DNA quality. The analysis is based on variants after initial filtering to remove those deemed unacceptable: <50 total read depth, <5 alternative allele read depth, <2% alternative allele frequency, as well as known PCR/sequencing artefacts.

F) Comparison of alternative allele frequency (AAF) between reproducible and non-reproducible variants according to DNA quality. There is considerable overlap in AAF between reproducible and non-reproducible variants regardless of their DNA quality, hence limiting its value in identification of true genetic changes.

Figure 2. Alternative allele depth (AAD) as an effective parameter in dichotomy between reproducible and non-reproducible variants. As the HaloPlexHS target enrichment design incorporates molecular barcode to remove PCR duplicate, hence PCR errors, the number of reads that bear novel variants, i.e. AAD, would represent the number of "mutant" templates successfully captured and sequenced. In general, the AAD value for non-reproducible variants is low, particularly in DNA samples that support PCR of 300bp or more genomic fragments, with mean plus 2SD being less than 15 AAD. To ensure highly specific mutation detection, a cut-off value of 20 AAD is used for dichotomy between reproducible and non-reproducible variants.

Figure 3. : Concordance in mutation detection between the HaloPlexHS and Fluidigm target enrichment approach under their respective protocols.

Figure 4. Proposed stratified approach for mutation screening by HaloplexHS target enrichment and Illumina HiSeq sequencing according to DNA quality. The DNA samples that support PCR of ≥400bp genomic fragment are amenable for mutation analysis with a single replicate, with only variants at 15-20 AAD required for further validation, while those with suboptimal quality (PCR of up to 300bp) are better analysed in duplicate with reproducible variants >15 AAD regarded as true

genetic changes. DNA samples that support PCR of only 200bp genomic fragment are not suitable for targeted sequencing under the condition specified.

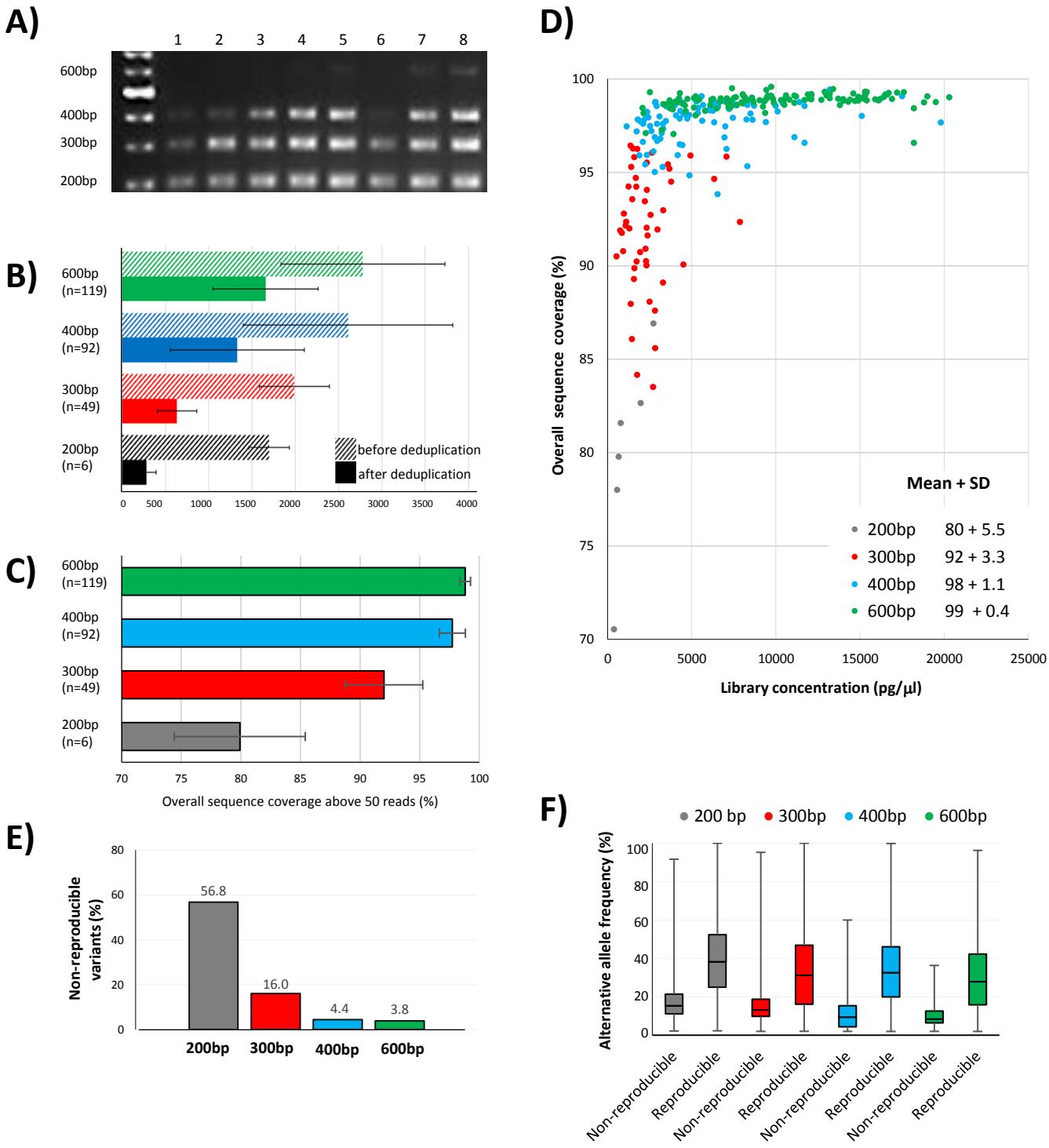


Figure 1: Impact of DNA quality on sequence coverage and variant reproducibility.

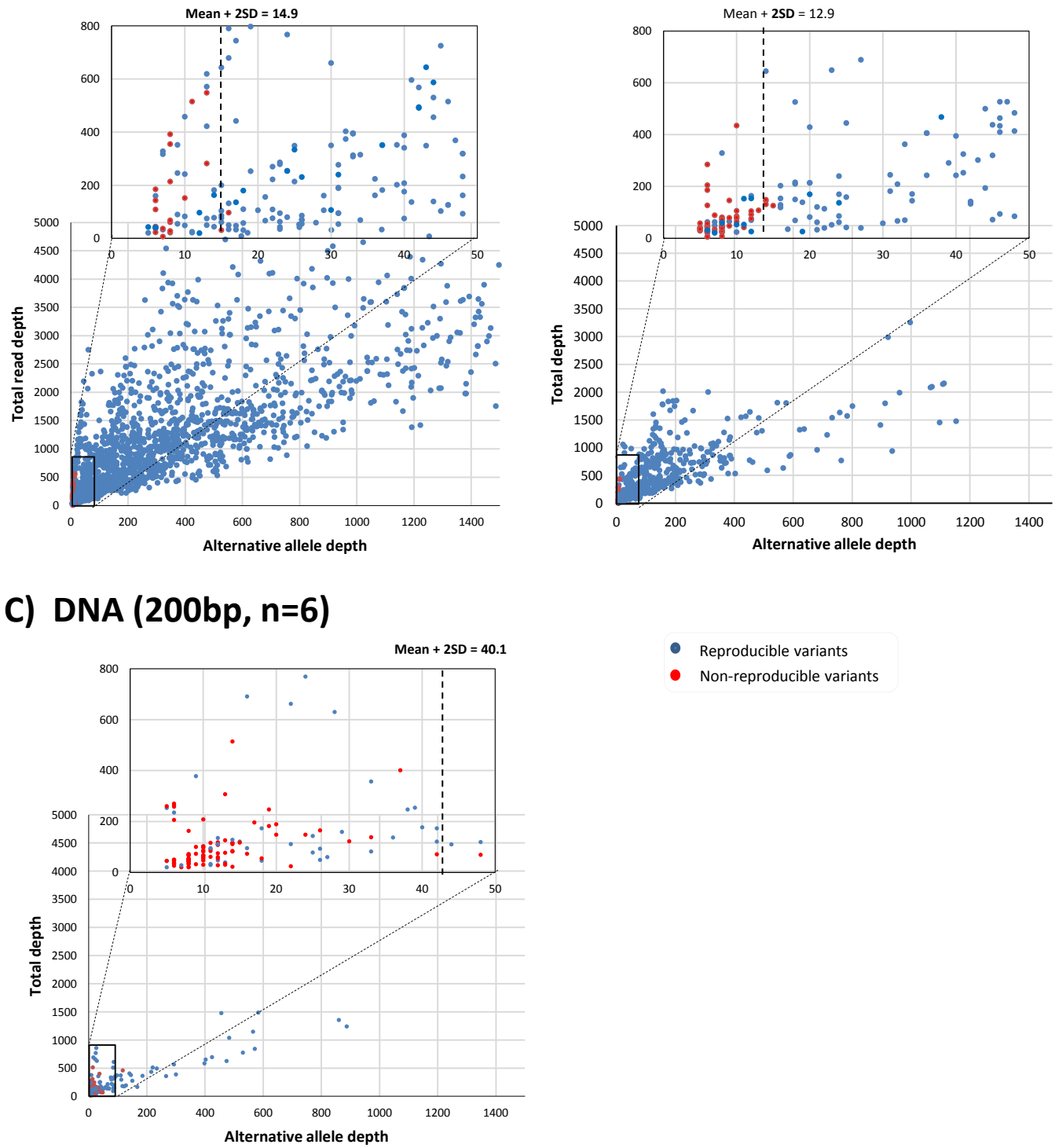
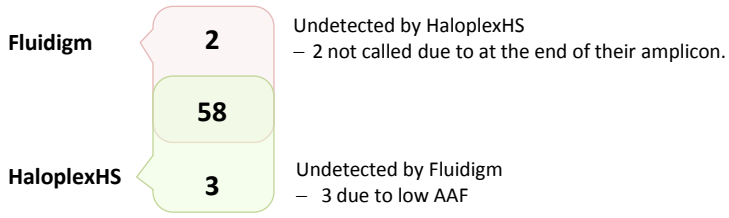


Figure 2: Alternative allele depth as an effective parameter in dichotomy between reproducible and non-reproducible variants.

A) Investigated in duplicate by both HaloplexHS and Fluidigm methods (n=22).



B) Investigated in duplicate by Fluidigm, but single replicate by HaloplexHS method (n=38)

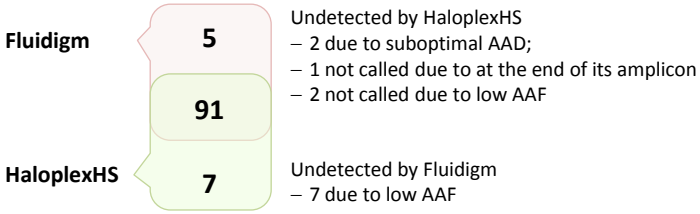


Figure 3: Concordance in mutation detection between the HaloPlexHS and Fluidigm target enrichment approach.

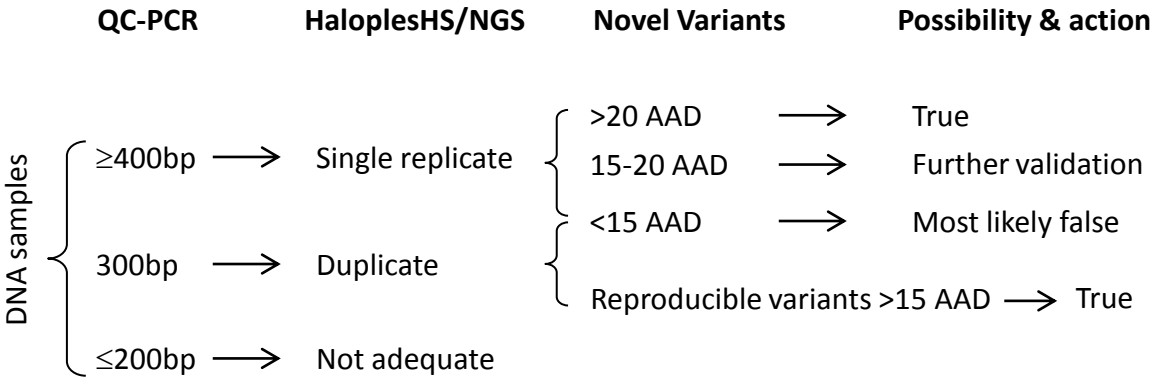
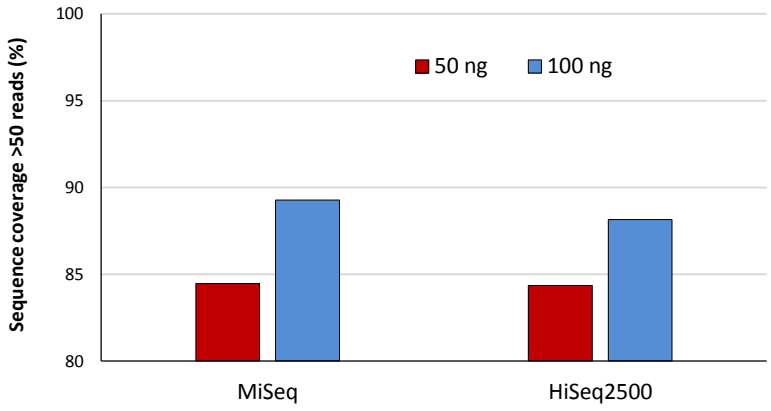


Figure 4: Proposed stratified approach for mutation screening by HaloplexHS target enrichment and Illumina HiSeq sequencing according to DNA quality.

A) Comparison of different amounts of input DNA



B) Comparison of different Illumina sequence platforms

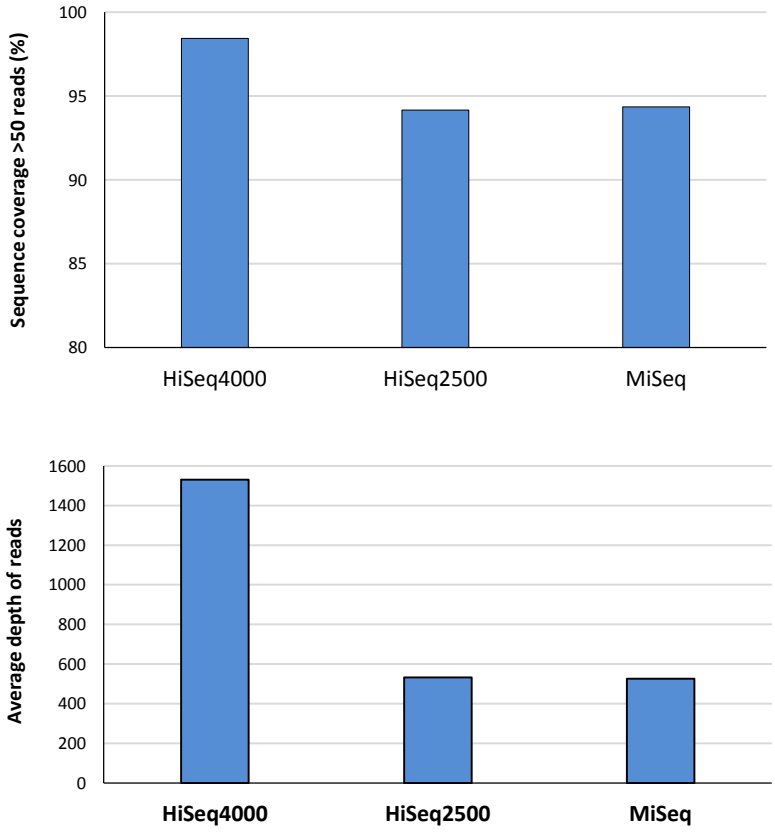


Figure S1: Optimisation of experimental conditions for targeted sequencing with HaloplexHS target enrichment and Illumina sequencing.

A) Reproducible changes

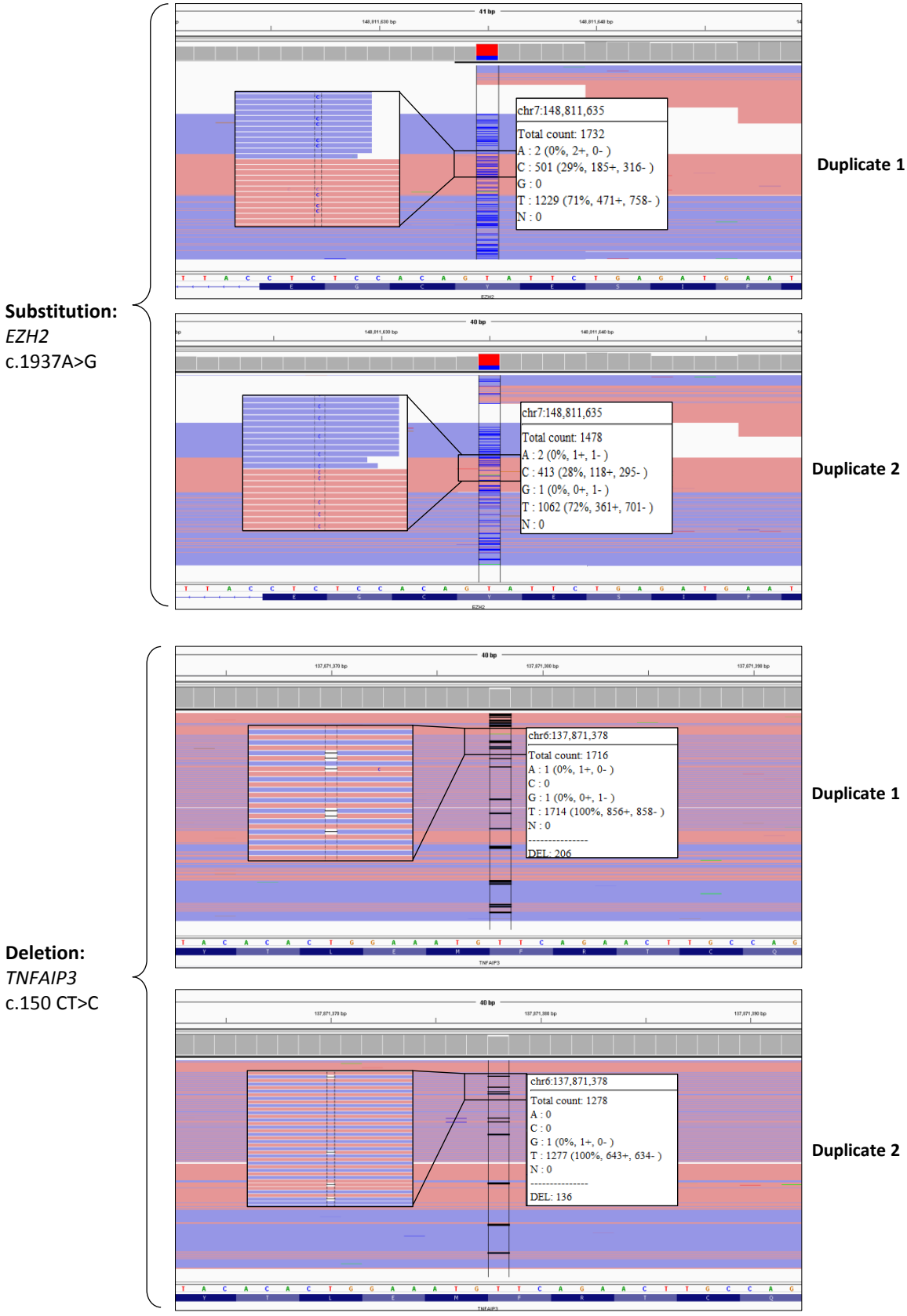


Figure S2: continue to the next page

B) Non-reproducible changes

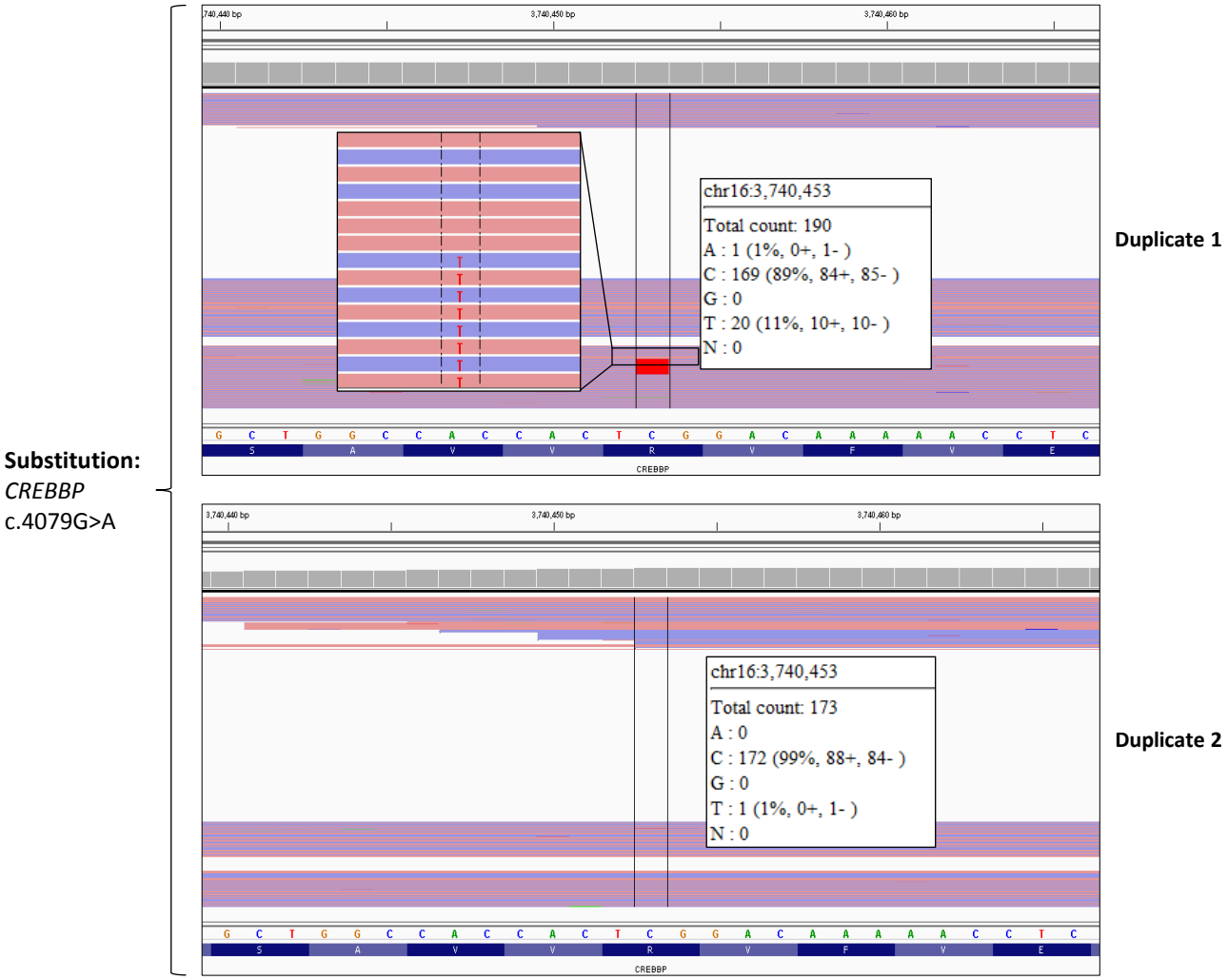


Figure S2: Examples of reproducible and non-reproducible variants detected by HaloplexHS target enrichment and Illumina sequencing.

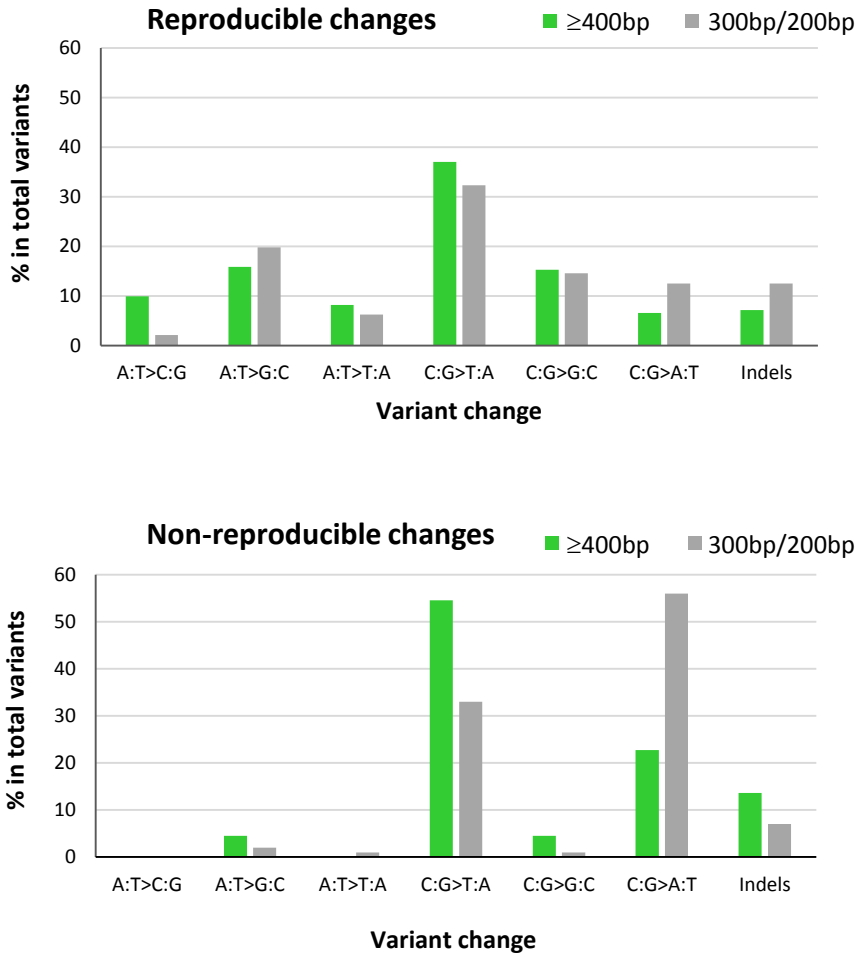


Figure S3: Comparison of the nature of reproducible and non-reproducible variants by HaloPlexHS target enrichment and Illumina HiSeq4000 sequencing. In contrast to targeted sequencing by Fluidigm Access Array PCR and Illumina MiSeq,[2] the present HaloPlexHS protocol show a low level of A:T>G:C changes among non-reproducible variants, indicating that the incorporation of molecular barcode in the HaloPlexHS design is highly efficient in removing PCR errors.

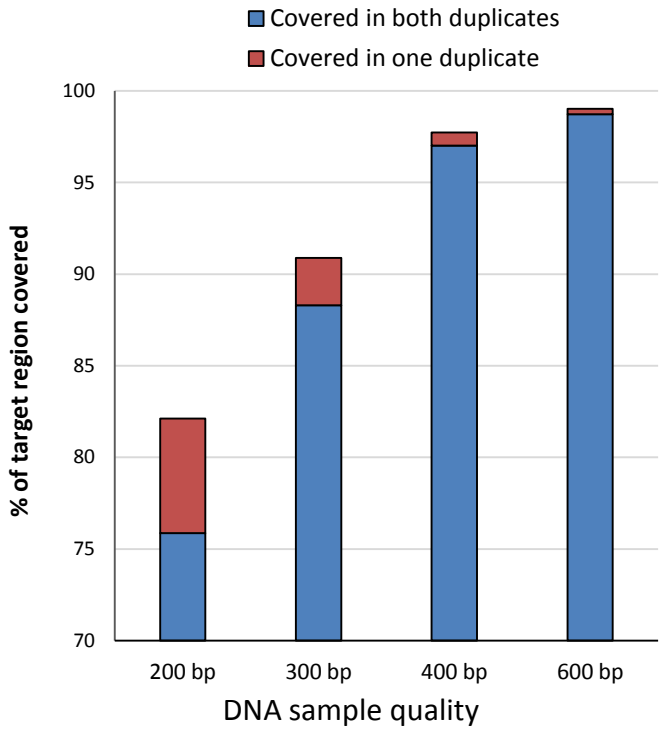


Figure S4: Improving sequence coverage by duplicate experiments.