SureSeq

A Sysmex Group Company

Application Note

Selecting the best NGS enrichment assay for your needs

Introduction

Next generation sequencing (NGS) is now in routine use for a broad range of research and clinical applications. The rapid rate of adoption has been facilitated by falling reagent costs, benchtop instruments, improved chemistries and improved data analysis solutions. However, the cost and complexity of data analysis still remain significant hurdles — particularly for whole genome sequencing. In the majority of cases, targeted approaches, such as custom NGS panels, are more cost-effective and generate significantly less, but equally meaningful data in a much shorter timescale.

Targeted sequencing requires an initial sequence enrichment step, which, if poorly designed, can be a source of bias and error in the downstream sequencing assay[1]. This article discusses the main strategies employed to optimise the enrichment step, depending on the type of assay chosen.

Which enrichment assay?

Two broad categories of enrichment assays exist: amplicon (PCR) and hybridisation (Figure 1). As a very general rule, hybridisation-based assays, when designed well, offer superior performance[2].

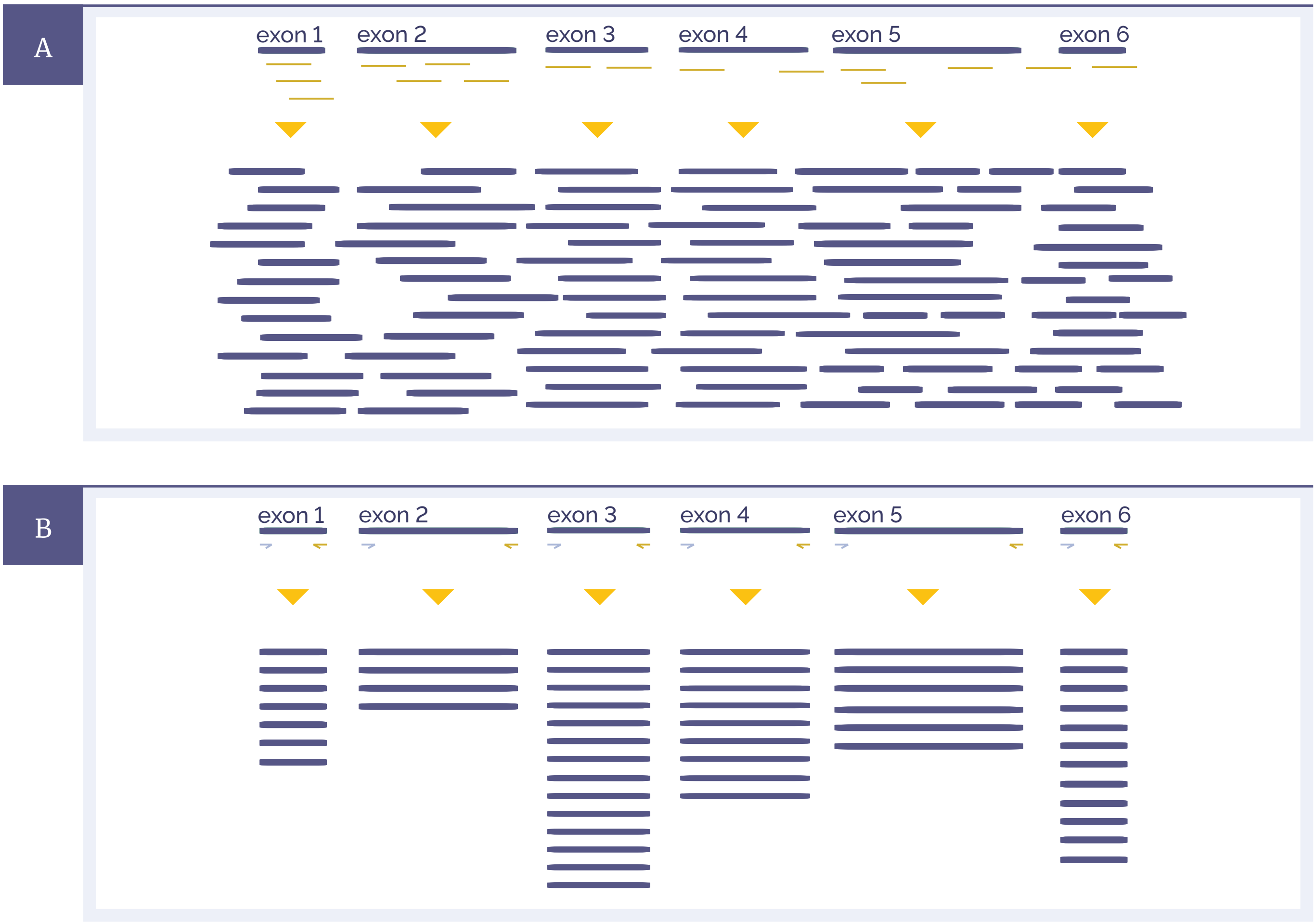

Figure 1: Schematic representations of amplicon and hybridisation enrichment approaches. A Hybridisation assays begin with random shearing of the genomic DNA, followed by capture using long oligonucleotide baits. Because of this random shearing, fragments captured are overlapping and unique. Baits can be tiled, overlapped and positioned to overcome challenges of repetitive sequences etc. With advanced design, capture can be made very uniform. B Amplicon assays provide less flexibility in the positioning and design of primers – primer pairs need to flank the region to be targeted. All fragments generated from a single primer set are identical, with the disadvantage that assay artefacts cannot be distinguished from genuine variation. Primer competition and preferential amplification of some regions over others will lead to non-uniform enrichment.

Hybridisation protocols start with random shearing of the DNA, followed by “capture” of the randomly sheared overlapping fragments with long oligonucleotide (oligo) baits. This allows independent sequencing of a large number of unique fragments. Any duplicates (assay artefacts) can be easily identified and removed, leaving high-quality data for analysis. Because the fragments are randomly sheared they should not align perfectly with one another and if they do, they are most certainly duplicates. In addition, enrichment of challenging regions such as GC-rich regions or internal tandem repeats can be optimised by careful positioning and design of baits. Long oligo baits can tolerate sequence variation, so that all alleles of a heterogeneous mix can be captured equally. Amplicon assays require design of primers flanking the region to be amplified. The resulting amplification products are identical, such that duplicates cannot be distinguished from unique products.

In making the choice between hybridisation and amplicon approaches, there are several factors which are worth considering (Figure 2).

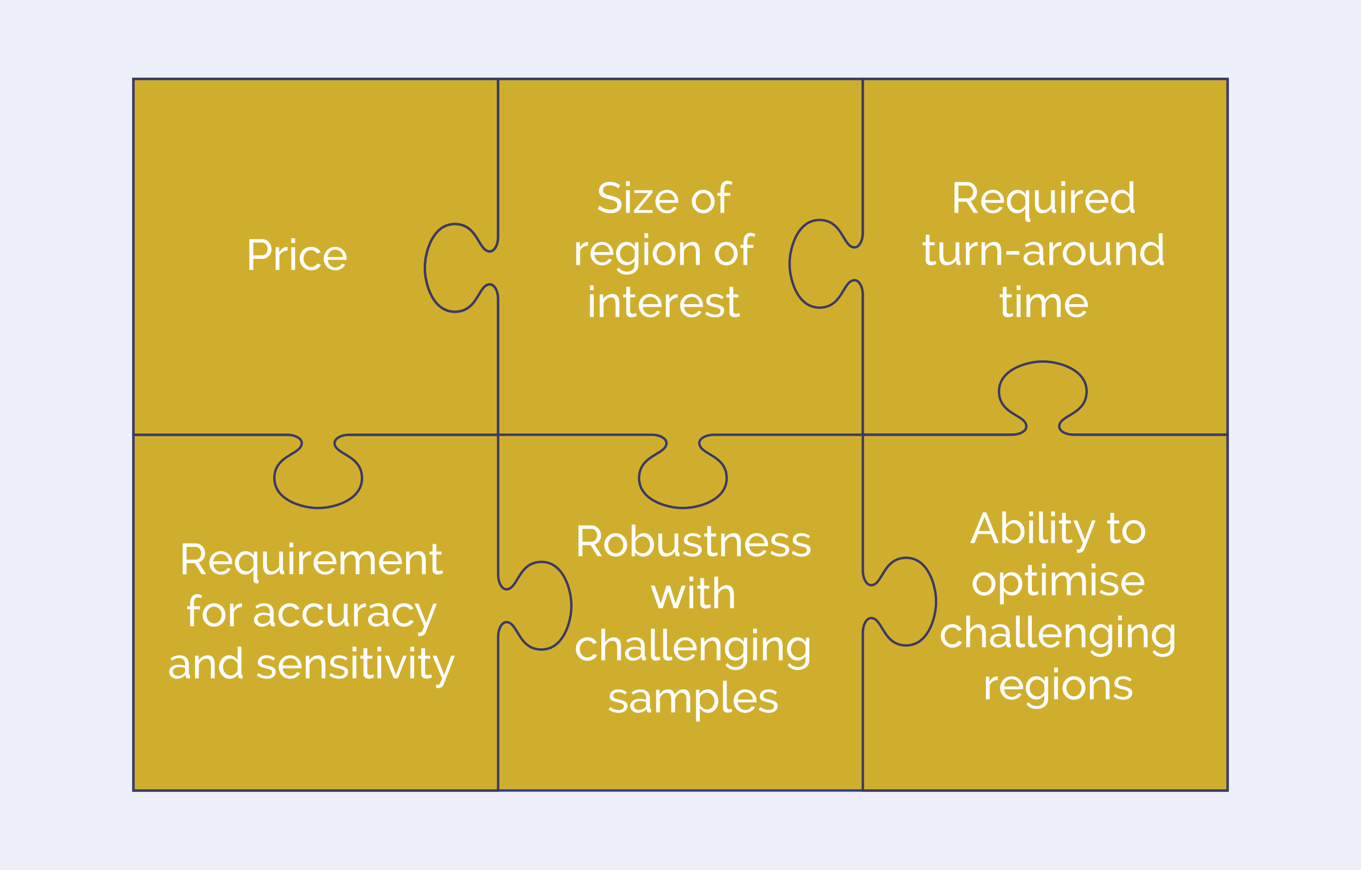
Price
Size of region of interest
Required turn-around time
Requirement for accuracy and sensitivity
Robustness with challenging samples
Ability to optimise challenging regions

Figure 2: A number of key factors are important in selection of the most appropriate enrichment assay for a given application.

1. Size of the region to be targeted

- Hybridisation-based assays are amenable to any size of target region, from very small to very large (i.e. whole exome).
- Amplicon assays, although ideal for small numbers of well-defined regions, are challenging to multiplex to any great extent. As the degree of multiplexing and/or the number of PCR cycles increases, so does the tendency towards bias and error. Primer competition and non-uniform amplification of target regions caused by varied GC content (Figure 3) or amplicon length for example in the presence of an insertion (under-represented) or deletion (over-represented) (Figure 4), contribute to variation in amplification efficiency. Some platforms attempt to overcome this by performing hundreds to thousands of single-plex reactions which are then combined prior to sequencing.

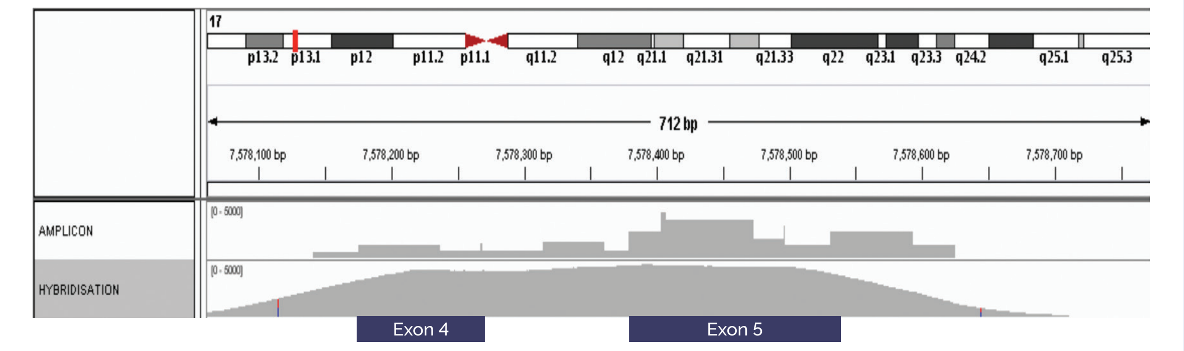

Figure 3. Hybridisation-based enrichment delivers more uniform coverage of GC-rich regions. Comparison of amplicon and hybridisation-based enrichment of the GC-rich exons 4 and 5 of the *TP53* gene.

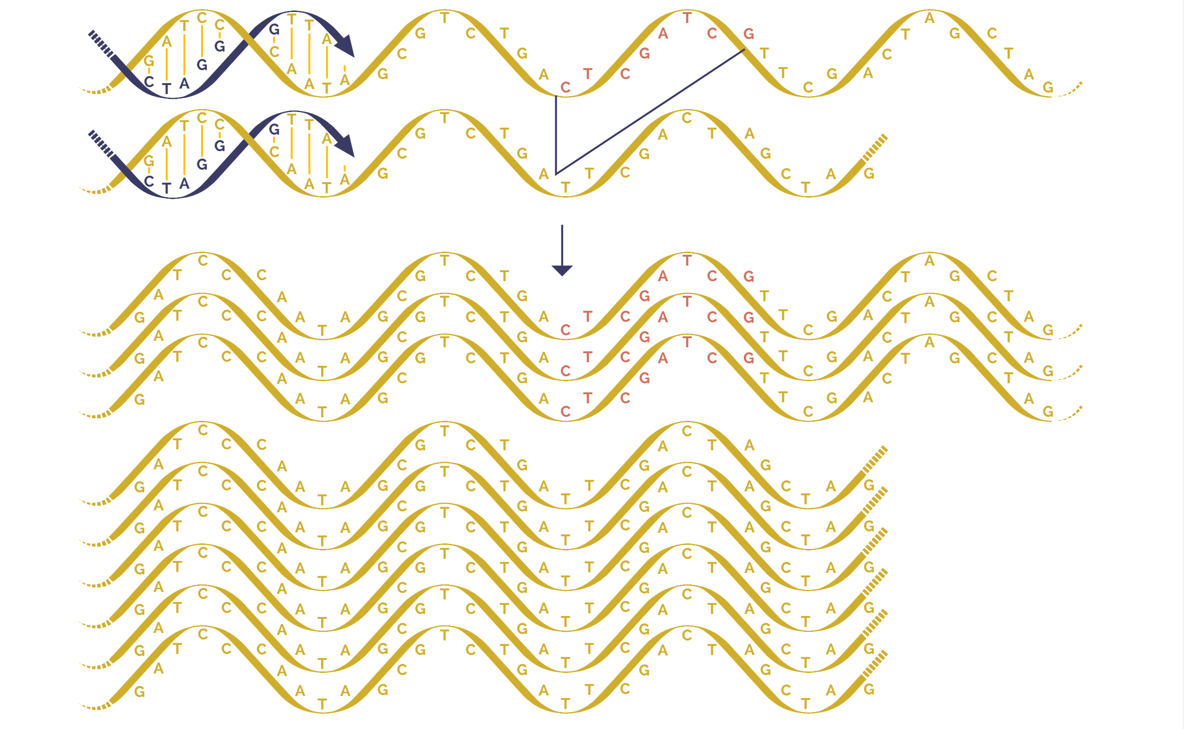

Figure 4. A deletion within the amplicon can cause preferential amplification of smaller fragments.

2. Required turnaround time

- Hybridisation assay protocols are more time consuming than amplicon approaches. However, hybridisation protocols that once took two to three days, and require large numbers of manual steps are now much more streamlined. OGT, as an example, in its library preparation protocol has introduced a short enzymatic fragmentation step, combined the end-repair and adaptor ligation steps, and optimised the hybridisation step itself to just 30 minutes. Meaning that hybridisation assay protocols can now take users from sample to sequencer in a single day.
- For speed and simplicity, PCR-based approaches are normally faster, only a few hours in duration, with fewer steps. It is important to note, however, that PCR itself is the most common source of bias and error in any enrichment assay, so this advantage of speed may need to be balanced with the requirement for high-quality data, and the additional time required to validate potential false positive results.

3. Ability to optimise for challenging regions

- Some genes such as *FLT3* contain internal tandem duplications that are challenging to target using amplicon approaches because they are by nature repetitive and can be very long and are generally masked in most designs. To provide optimal results, OGT employs sophisticated bait design strategies to create additional probes both up- and down-stream of the repetitive region. The length of these probes is customised to provide the same isothermal properties as other probes in the panel, leading to improved performance. This facilitates the capture of sequences around the repeat masked areas, and because of the read lengths, can read through short (up to 100 bp) repetitive regions.
- Other challenges arise for amplicon assays when novel variants are present within the primer site (Figure 5), as these can result in strand or allelic bias, or even drop-out of that region altogether. Hybridisation assays, on the other hand, are less restricted by variant position and can still enrich all strands and alleles equally even in the presence of multiple novel variants.

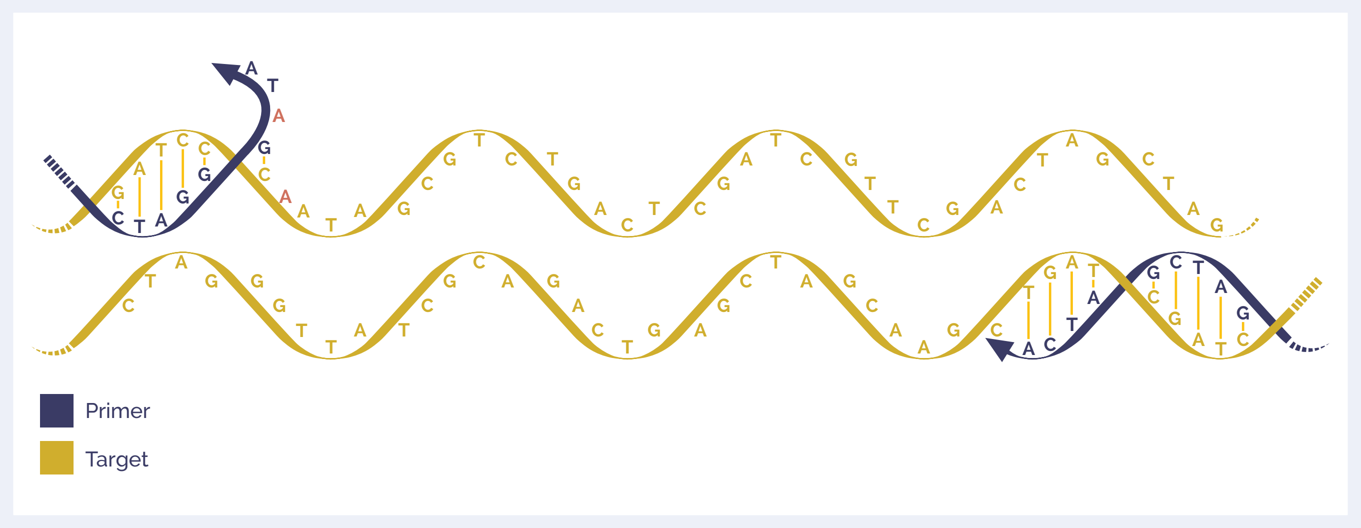

Figure 5. Single nucleotide variants (SNVs) in primer sites can lead to allelic bias and drop-out.

- Similarly, with regions that are GC-rich, hybridisation baits can still be designed to capture efficiently giving much more uniform coverage than amplicon assays. An example of this is given in Figure 6, where the high (up to 90%) GC content of the *CEBPA* gene has made it notoriously difficult to sequence and analyse reliably with NGS[3]. However, OGT's bait design expertise can overcome these difficulties, providing excellent depth of coverage and uniformity.

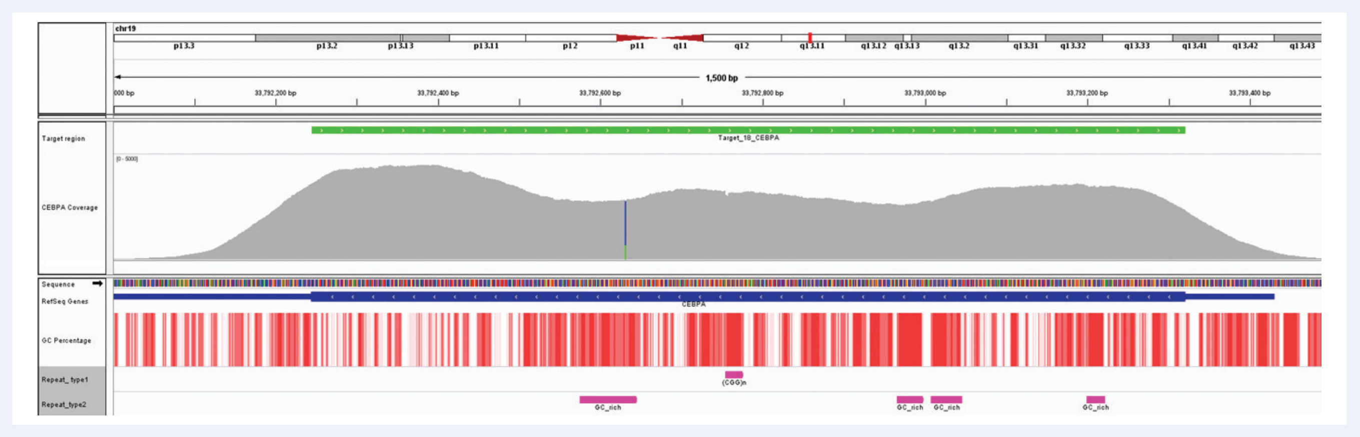

Figure 6. Despite the very high GC content of the *CEBPA* gene and the number of repeat regions, hybridisation enrichment coupled with expert bait design can still achieve excellent results. Depth of coverage per base (grey). Targeted region (green). Gene coding region as defined by RefSeq (blue). GC percentage (red). Repeat regions, and those rich in GC (pink). Data generated using SureSeq myPanel™ custom content.

4. Robustness with challenging samples

Samples vary in quality and quantity so any assay must be able to deal with a wide range of input DNA types and input quantities:

- Formalin-fixed, paraffin-embedded (FFPE) samples: Both hybridisation and amplicon assays can be optimised to perform well with FFPE, though PCR methods, in particular, are susceptible to contaminants found in FFPE material. Assay performance can be substantially improved by the usage of an upstream FFPE repair step (Figure 7), which can remove a broad range of damage, such as nicks and gaps, oxidised bases and cytosine to uracil deamination.

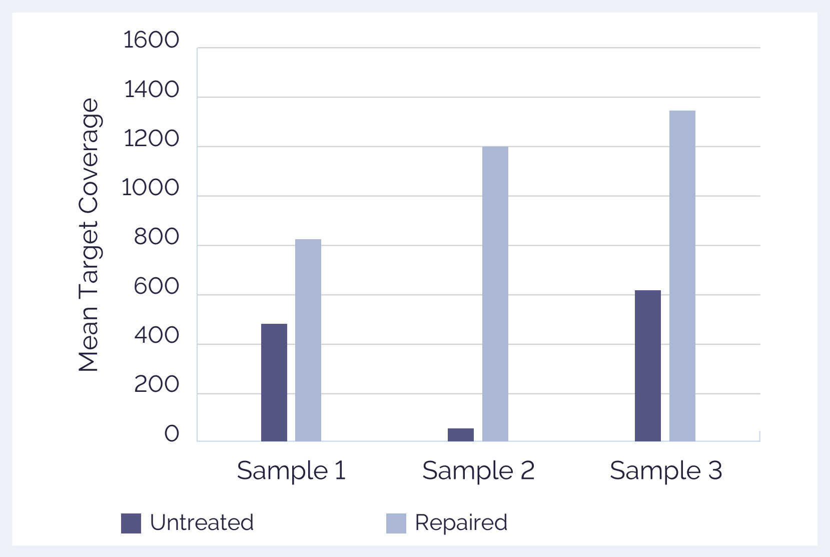

Figure 7. The SureSeq FFPE DNA Repair Mix significantly improves mean target coverage resulting in more confident calls. Data obtained using 500 ng of FFPE DNA from ovarian and colon cancer samples; 16 samples per MiSeq® lane.

- Starting quantity of DNA: Amplicon assays offer a slight advantage in being able to work with smaller quantities of input DNA, often down to 10ng. Hybridisation assays generally require more input DNA — typically ~500 ng — although well designed hybridisation assays can utilise significantly less input DNA (Figure 8), where using the OGT SureSeq™ FFPE Repair Mix a reduction in the amount of starting material down to 100ng or lower is possible, depending on the required depth of coverage required.

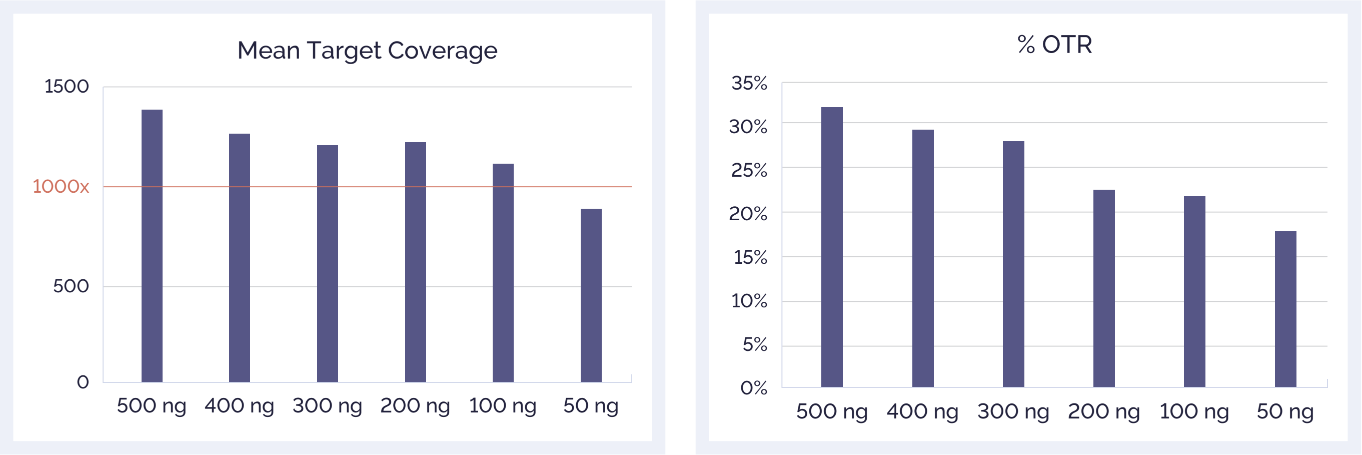

Figure 8. Effect of reduced amount of DNA input on mean target coverage A and %OTR B.

- **Duplicates:** It is important to note that with small starting quantities, the actual number of templates available is relatively low, so duplication rates can increase significantly. With hybridisation assays, these can be removed computationally to leave clean, high-quality data. With amplicon assays, this is not possible without the use of molecular barcodes, and the resulting data may be skewed by the over-amplification of a small number of fragments.

5. Requirement for accuracy and sensitivity

Performance should be a key requirement for all applications: the confidence that the assay will detect all variants present in any region of interest, while avoiding false negatives and false positives. Hybridisation assays offer a number of key benefits which enhance performance:

- **Reduced false negatives:** The most common reason for false negatives in a targeted sequencing assay is poor coverage at the locus. Hybridisation assays, when well designed, can deliver superior uniformity of enrichment and excellent coverage of all loci, thereby reducing the incidence of false negatives.
- **Reduced false positives:** The most common cause of false positives are artefacts introduced by PCR polymerases, even when using proofreading enzymes. Hybridisation assays use very few PCR cycles, in comparison to amplicon assays, and therefore the data is less "noisy".
- **Higher detection sensitivity:** The reduced "noise" in the hybridisation assay also delivers higher detection sensitivity of variants present at low frequency in the sample.
- **Broader scope for discovery of novel as well as known variants:** It can be challenging to validate an assay for the discovery of novel variants across large numbers of genes. However, because the ability to detect variants with sensitivity and accuracy largely relies on the depth of coverage at the locus, as well as quality and accuracy of the data, hybridisation-based assays can offer greater confidence in detection of all variants present.

6. Price

- Price is always an important factor. For larger regions, hybridisation-based panels are very cost effective. For smaller regions, amplicon assays can be more cost effective because the cost of a small number of primers is low. Price must always be considered in parallel with performance requirements for the application of interest and with the cost of any additional sequencing required downstream.

Uniformity of enrichment as a key metric of performance

The ultimate goal of any sequencing assay is to discover all variants present. Uniformity of enrichment means that all regions are represented more equally, and that variants present in any region will be called (Figure 9). It also allows much lower average sequencing depths to be used, enabling larger numbers of samples to be multiplexed in a run, and significant cost savings.

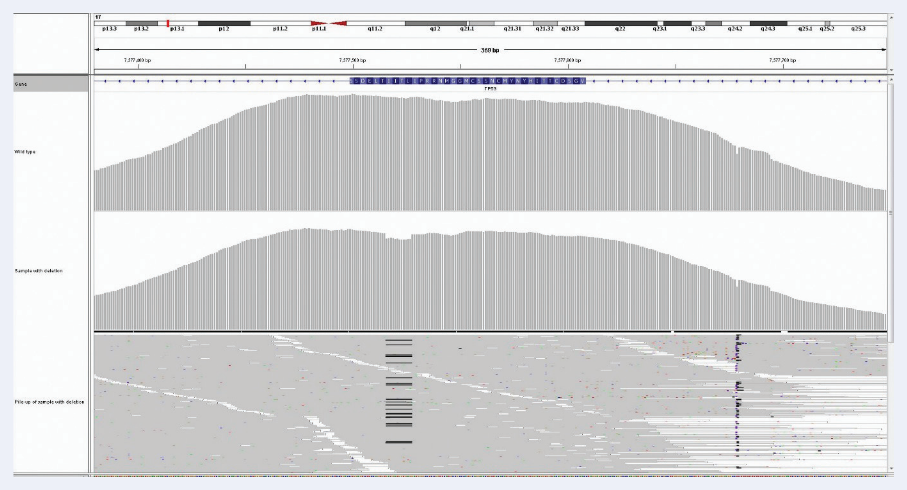

Figure 9. High uniformity of coverage allows the reliable detection of lowfrequency somatic indels even in FFPE derived DNA. Example is of 12 bp deletion (c.754_765delCTCACCATCATC) in *TP53*, 6% deletion, mean target coverage >1400, 12 samples per MiSeq lane, using the SureSeq Ovarian Cancer Panel.

Uniformity is particularly important when looking at heterogeneous samples. For example tumour mixed with normal tissue, or somatic variants present only within a single clone in a heterogeneous tumour sample where it is essential to have enough reads to confidently call a variant at any given position.

The importance of optimising the enrichment assay

As NGS moves forwards, the aspiration is clear: confident calling of all variants present with no false negatives and no false positives. The most common reason for missing variants (i.e. false negatives) is lack of coverage at the variant locus due to non-uniform enrichment, the importance of which is illustrated with the example of exon 9 of *CALR* gene, an important exon for interrogation of myeloproliferative disorders (Figures 10 and 11).

The highlighted box below provides a summary of all of the potential sources of error and bias. As a very broad rule, hybridisation-based assays offer greater opportunity for optimisation through probe design and placement, and they can offer better uniformity of coverage, fewer false positives, and superior variant detection due to fewer PCR cycles. Hybridisation-based assays also offer greater scope in terms of the number of genes and regions that can be targeted.

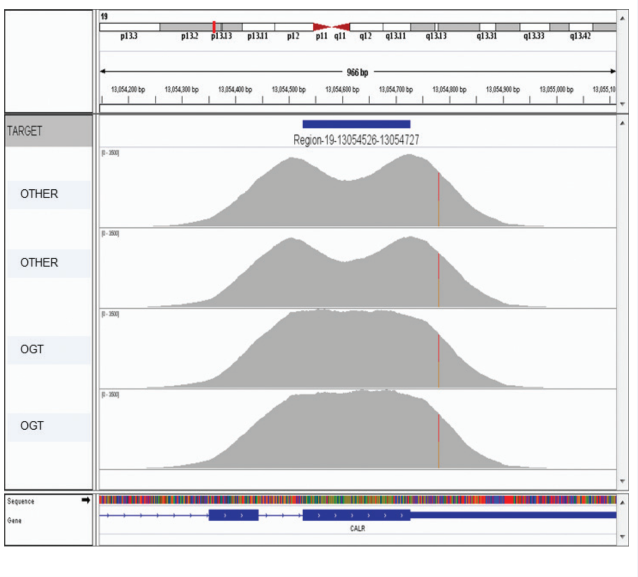

Figure 10. Advanced bait design strategies deliver uniform enrichment, reducing the likelihood of false negative results.
The top two captures have been completed using baits designed with standard commercially available software. They have a considerable dip in coverage in the middle of the exon due to the fact it presents a low complexity region with low nucleotide diversity. Most algorithms would avoid such regions in the design. However, OGT's superior bait design can increase the evenness of coverage of such regions.

Conclusions

The choice of enrichment assay for targeted sequencing assays is an important consideration. No enrichment technology is perfect for every application. The choice will depend largely on the size of region to be targeted, cost of sequencing and the required coverage uniformity and sensitivity of the assay. Optimisation is also important and hybridisation-based assays offer more scope for superior performance through optimisation of bait design.

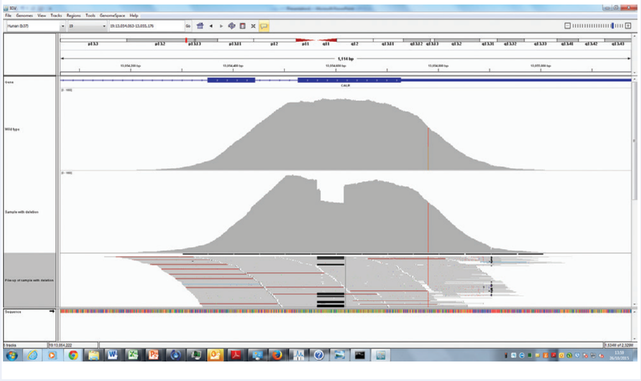

Figure 11. Superior uniformity of coverage allows reliable detection of indels even as big as 1/3 of a read length. Shown here 52bp deletion in exactly the same region as illustrated in Figure 10, exon 9 of the *CALR* gene. 23% deletion (c.1092_1143del_52bp). Mean target coverage >1000, 24 samples per MiSeq lane, using the SureSeq Myeloid panel.

Spotlight: Potential sources of bias and error in sequence enrichment assays

PCR artefacts

Even proof-reading PCR polymerases introduce errors. The likelihood of artefacts (as well as the rate of duplication) increases with increasing PCR cycles. Amplicon-based assays are reliant entirely on PCR, and are potentially more susceptible to artefacts than hybridisation-based assays, which aim to minimise the number of PCR cycles.

Duplicate reads from a single template

Duplicates are amplification artefacts, normally arising during the library preparation stage.

It is highly desirable to remove them prior to data analysis, otherwise some regions may be massively over-represented. In a hybridisation-based assay, duplicate reads can be identified easily and removed. However, in a PCR assay, not only is it impossible to remove duplicates during the analysis step, but it is also true that there are generally more duplicates due to the enrichment step. Duplicates are a particular problem when input quantities of DNA are limiting, because higher numbers of PCR cycles need to be used.

Bias in PCR amplification — PCR drift and PCR selection

PCR is difficult to multiplex where consistent and uniform amplification of each region is desired. There are thought to be two main processes that introduce amplification bias during PCR, PCR selection and PCR drift[4]. In PCR selection, some amplicons are favoured and therefore overrepresented due to intrinsic properties of the target sequence, flanking sequences or genome composition. Key contributors to this type of variation include preferential denaturation and amplification of low GC content templates, higher binding efficiency of GC-rich primers (particularly when using degenerate primers) and direct correlation between amplification efficiency and gene copy numbers. PCR drift is assumed to be caused by random interaction of the components of the mix early on in the amplification when the original genomic material is still the main source of template. This type of variation is variable between reactions and more difficult to control. There are ways to optimise PCR that will reduce bias[1] including increasing the amount of template, reducing the number of cycles, optimising the instrumentation and performing the multiplex in a number of discrete, lower-plex reactions which are then pooled after amplification.

Variant positional bias in amplicon assays

Variant positional bias is a particularly important consideration for amplicon assays where there is some constraint in choice of position for primer sites. If a SNP or variant happens to fall within the primer site itself, it will not be detected. This is less likely to be an issue with hybridisation-based methods that use long oligonucleotides that can tolerate target sequence variation and can be tiled across the region to be enriched.

Repeat regions and pseudogenes

Repeat regions and pseudogenes represent significant challenges for all enrichment technologies. Depending on the size of the region, hybridisation-based assays may allow better targeting of these regions, by allowing design of baits to flanking regions.

SureSeq

NGS Enrichment Assay

Ordering information

UK +44 (0) 1865 856800

US +1 914 467 5285

contact@ogt.com

ogt.com

References

1. Aird D, *et al.*, (2011) Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome Biology 12:R18 doi:10.1186/gb-2011-12-2-r18
2. Samorodnitsky E, *et al.*, (2015) Evaluation of hybridization capture versus amplicon-based methods for whole-exome sequencing. Hum Mutat 36(9), 903-915
3. Behdad A, *et al.*, (2015) A clinical grade sequencing-based assay for CEBPA mutation testing: report of a large series of myeloid neoplasms. J Mol Diag 17(2), 76-84
4. Wagner A, *et al.*, (1994) Surveys of gene families using polymerase chain reaction: PCR selection and PCR drift. Syst Biol 43, 250–261

ogt

A Sysmex Group Company

**What binds us,
makes us.**

Oxford Gene Technology Ltd., Begbroke Science Park, Woodstock Road, Begbroke, Oxfordshire, OX5 1PF, UK

SureSeq: For Research Use Only; Not for Diagnostic Procedures. This document and its contents are © Oxford Gene Technology IP Limited – 2021. All rights reserved. Trademarks: OGT™, SureSeq™ and myPanel™ (Oxford Gene Technology); Agilent® (Agilent Technologies Inc.); MiSeq® (Illumina Inc.).