TARGETED DNA SEQUENCING:

Probing for Answers

Custom Publishing From

TheScientist

EXPLORING LIFE, INSPIRING INNOVATION

Sponsored By

Roche

BETTER BY DESIGN

KAPA Target Enrichment Portfolio for Targeted NGS

Answer your most challenging research questions by leveraging more than a decade of probe-design experience and an improved manufacturing process.

Roche's **NEW KAPA Target Enrichment Portfolio** offers fully customizable target enrichment panels for hybridization-based capture before next-generation sequencing.

Design your custom target enrichment panels by either (1) using our online, user-friendly HyperDesign tool or (2) working directly with one of our expert designers.

Improved specificity and uniformity deliver better target coverage

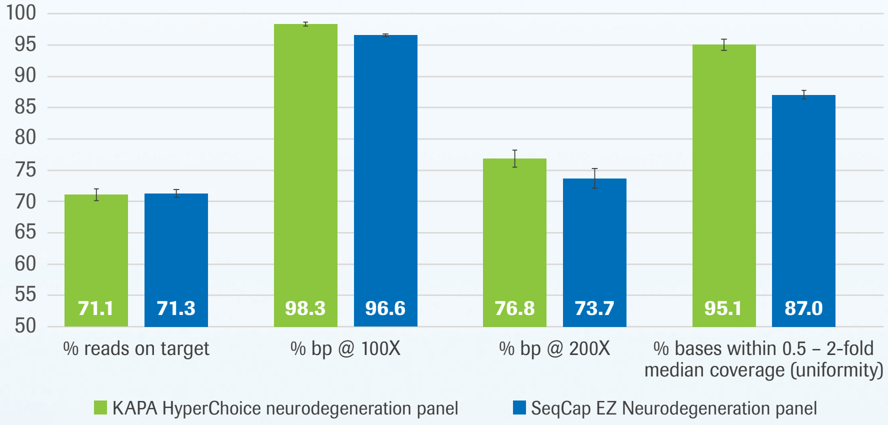


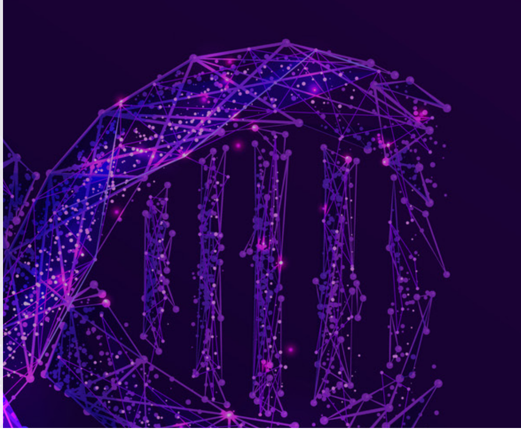
Figure 1. The new KAPA HyperChoice neurodegeneration probe panel outperforms the SeqCap EZ Neurodegeneration panel with higher uniformity and better target coverage, without the need for rebalancing. Performance was compared between the SeqCap EZ Neurodegeneration panel and a new KAPA HyperChoice panel covering similar regions. For each panel, 8 target-enriched DNA libraries were prepared from Coriell control DNA using the appropriate workflow: KAPA HyperCap v2 (SeqCap EZ) or KAPA HyperCap v3 (KAPA HyperChoice). Libraries were captured in 8-plex reactions and sequenced on an Illumina MiSeq instrument (2 x 100 bp). The KAPA HyperChoice panel was used out-of-the-box, while the SeqCap EZ panel was empirically rebalanced to improve performance.

Learn more and begin designing your custom targeted panel at: go.roche.com/HyperChoice

Project: NGHC, Roche Sequencing Solutions, Pleasanton, CA 2019
For Research Use Only. Not for use in diagnostic procedures.

A577 9/20

INTRODUCING TARGETED DNA SEQUENCING



The sequencing of the human genome in 2003 marked the beginning of what sequencing technology could accomplish for the scientific community. Since that monumental moment, next-generation sequencing (NGS) techniques have made sequencing exponentially faster, more accurate, and less costly. NGS accomplishes this by first fragmenting input DNA into short segments. The segments are ligated to adapters containing indexing information, creating a library. If every sample has a different set of indexed or barcoded adapters, multiplexing—the simultaneous processing of multiple samples—can be done. Libraries are then amplified via PCR and sequenced, with sequencing reads reassembled to establish genes, regions, and genomes. This data is evaluated using several quality metrics, including sequence quality, sequence uniformity, and depth of coverage. The read depth for an individual position is the number of times it has been read (once is 1X depth, twice is 2X depth, and so on), and averaging all of the individual read depths within a library provides the average read depth.

NGS has made whole genome sequencing (WGS) and whole exome sequencing (WES) commonplace, resulting in not only a more comprehensive understanding of the sequences that make up the human genome, but also the identification of numerous genetic drivers and biomarkers of disease.

The Need for a More Focused Approach

Although powerful, WGS and WES are not without their drawbacks. Sequencing full genomes or exomes produces large datasets—sometimes as big as 350GB.[1] The sheer size of these datasets can make analysis a time- and labor-intensive process. Additionally, since sequencing efforts must be spread out across an entire genome or exome, overall average depth is reduced. This has a negative impact when it comes to large-scale reconstruction, because read depth is integral to algorithmic mapping and arrangement of reads and is critical for establishing high confidence in the results—something that is invaluable when examining rare variants.[2] In light of these challenges, scientists seek to strike a balance between the ability to identify specific mutations/variants and the cost and data burdens involved with large-scale sequencing.

What is Targeted Sequencing?

Targeted sequencing is a strategy where researchers target specific genomic regions of interest rather than an entire genome or exome. There are two principle methods for targeted sequencing: amplicon sequencing and hybridization capture. Amplicon sequencing uses specifically designed primers to preferentially amplify selected regions directly from genomic DNA during PCR. In contrast, hybridization capture starts with an NGS library, then uses probes to bind to library molecules containing desired sequences. The resulting complexes are then isolated, creating enriched samples for amplification and sequencing.[1] Hybridization capture is a more involved process than amplicon sequencing, but it offers greater sensitivity, better sequencing uniformity, fewer PCR artifacts, and the ability to examine millions of targets simultaneously.[1]

The exact size of a region of interest covered by a probe panel varies depending on researcher objectives, but it typically falls between 10^5 and 10^7 base pairs—orders of magnitude fewer than WES (6 x 10^7 bp) and WGS (3 x 10^9 bp).[1] Targeted panels focusing on specific organelles or diseases can be even smaller: cancer gene panels can range from 3.5 x 10^3 to 5.4 x 10^4 bp in size depending on the number of cancer-associated regions covered, while mitochondrial gene panels can be much smaller. This allows targeted sequencing to offer much deeper coverage without becoming cost-prohibitive. In contrast to WGS (30-60X) and WES (150-200X), the typical targeted sequencing experiment provides depths of 200X to 1000X.[1] However, it is possible to reach depths of 10,000X or much deeper coverage sequencing, something very useful for profiling low-quality clinical samples, identifying sub-clonal mutations, and detecting minimal or early-onset pathogenesis.[1] This is tremendously useful in situations where sample scarcity is a problem, such as with clinical samples. Moreover, added sequencing depth allows targeted sequencing to characterize and profile heterogeneous cell populations, either to identify key mutations present only within small sub-clonal cell fractions or to detect low levels of variant alleles that may be a hallmark of early onset pathogenesis.[1]

For all of these reasons, targeted sequencing approaches have rapidly become popular for disease research and diagnostics, especially for oncology.

References

1) F. Bewicke-Copley et al., "Applications and analysis of targeted genomic sequencing in cancer studies," *Comput Struct Biotechnol J*, 17:1348-59, 2019.

2) D. Sims et al., "Sequencing depth and coverage: key considerations in genomic analyses," *Nat Rev Genet*, 15:121-32, 2014.

COVERING YOUR BASES

How Much Depth and Coverage is Enough?

Next-generation sequencing can achieve ultra-high speeds and throughputs because of its massively parallel nature. In NGS, researchers fragment DNA and add sample-specific indexed adapters to create sequencing libraries. They then amplify and read these fragments. The sequences obtained from these "short reads" are then either mapped to existing genome sequence information or combined to construct de novo genomes when working with non-human research models. Naturally, it is critical to make sure that the sequence information captured by these short reads encompasses the entire region of interest, whether that is a gene, an exon, or an entire genome.

What is Coverage?

The number of unique reads containing sequencing data for a given nucleotide position is referred to as the "coverage" or "depth" for that position, with the terms generally used interchangeably.[1] Depth is usually written as "X": for example, a nucleotide position read twice has 2X read depth. Sufficient coverage depth increases data robustness, which is important for determining whether a changed sequence variation is a true mutation or variation or is the product of sequencing error. Coverage depth is also critical for proper short-read assembly, especially for de novo genomic information where no prior scaffold exists. However, increasing coverage means increasing read counts, which in turn means increased reagents, sequencing time, and dataset complexity. As such, the desired depth changes from experiment to experiment as researchers balance coverage with logistics.

How Much Coverage is Necessary?

There are numerous factors that scientists must consider when deciding how much depth is sufficient. Some of these are technical, while others depend on the researchers' scientific goals. First and foremost, the sequencing error rate needs to be considered.[2] Because coverage depth directly affects the level of confidence for a base call at a particular position, higher error rates will necessitate increased coverage depth to compensate. Second, what is the goal of the study? If researchers are investigating rare mutations, rare cell types, or both (i.e. sub-clonal mutations), then deep, or possibly even ultra-deep coverage is necessary to avoid false negatives.[3] Alternatively, shallower depths are sufficient for evaluating copy number variations, and is more efficient when screening a larger number of samples and/or regions. That said, more depth gives greater confidence to the integrity of the sequence, something that is important in scenarios where data will be shared and standardized across laboratories and research groups.

How to Plan Coverage

With all this in mind, planning and designing a targeted sequencing experiment can be complicated. Fortunately, numerous assists are available to researchers. A variety of software options evaluate desired probe coverage, ranging from calculators for determining appropriate depth to genome browsers to find areas missing probe coverage. Beyond that, commercially available pre-designed targeted sequencing panels cover key genes for specific research areas such as oncology, inherited conditions, and cardiovascular diseases. Commercial panels can also focus deeper within a research area; for example, different panels exist for researchers looking at either tumor detection or longitudinal tumor monitoring.

Every project is unique, and therefore might require coverage of regions that are not present in pre-designed panels. To that end, commercial manufacturers offer custom panel design options where researchers simply input their gene targets and experimental organism, and algorithmic applications generate custom probe designs that are created with target coverage, capture efficiency, and binding uniformity in mind. These custom designs can also be fine-tuned by users for particularly challenging or variable regions. Custom panels are also uniquely suited for targeting genomic regions that are difficult to access.

References

1) D. Sims et al., "Sequencing depth and coverage: key considerations in genomic analyses," *Nat Rev Genet*, 15:121-32, 2014.

2) F. Pfeiffer et al., "Systematic evaluation of error rates and causes in short samples in next-generation sequencing," *Sci Rep*, 8(1):10950, 2018.

3) M.W. Schmitt et al., "Detection of ultra-rare mutations," *PNAS*, 109(36):14508-13, 2012.

WHAT SHOULD I LOOK FOR IN A PROBE?

High-quality probes are integral to hybridization capture targeted sequencing, as the best-designed panel will fail if the probes cannot efficiently and specifically capture their targets. Probes (also referred to as baits) are ~120 nucleotide-long biotin-labeled oligonucleotide sequences designed to bind specific genomic regions of interest. In contrast to primers, which are short (~20-30 nt) unlabeled oligonucleotides used during amplicon sequencing, baits are not designed with amplification in mind. Instead, baits are designed to bind to specific nucleic acid fragments within a sequencing library. These complexes are then pulled from the sample using streptavidin-coated beads, which bind to the biotin on the probes, and washed. The target-enriched library sample can then be amplified using primers that bind to the sequencing adapters on each library molecule.

Advantages to Using Probe-based Capture

Probe design for hybridization capture mitigates several problems present in amplicon sequencing. The fact that oligonucleotide baits serve exclusively to capture fragments and do not serve as amplification scaffolds affords them more design flexibility. They can be engineered to straddle known variation regions, for example. In this way, they can capture fragments containing potential sequence variants of interest while avoiding potential "drop out"—where primers have difficulty annealing properly to binding sites containing variations. Applying the same principle to long sequence repeats gives researchers the ability to delineate between naturally occurring genomic sequences and overamplification bias, something that is difficult to do with amplicon sequencing. Placing probe targets both up- and downstream of targeted regions helps ensure that fragments of interest are properly captured and subsequently read.

Hybridization capture probes can also take advantage of the randomness of genomic DNA fragmentation to reduce bias. In particular, randomness also makes it easy to isolate and remove amplification duplicates; as most individual reads do not overlap perfectly, those that do are likely be products of amplification bias or duplicate reads.

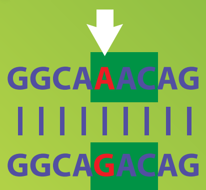
General Probe Considerations

Sequence composition affects probe specificity and selectivity, so probes should be designed so that each probe binds to a unique nucleotide sequence, ensuring that one probe does not bind in multiple locations. This is necessary to avoid non-specific binding, which would capture off-target fragments alongside genomic regions of interest and thereby reduce the effectiveness of targeted sequencing. Beyond that, all of the probes used in a given experiment should have fairly similar GC composition proportions, if possible. This affords them similar thermodynamic properties to ensure a narrow range of annealing temperatures across all baits, thereby facilitating more uniform hybridization capture. Finally, the probe panel, in conjunction with other factors such as sample multiplexing and sequencing technique selection, should help ensure that the entire region of interest is covered to a sufficient depth. This helps to reduce the number of amplification cycles necessary to generate sufficient read depth, decreasing PCR-induced amplification. Genome browsers can map out probe targeting sites along a region of interest and help flag low-coverage regions. As genome information is progressively updated through newer builds, scientists should be sure to coordinate build selection with previously established data or to align with collaborators.

The precise number and location of probes and their positioning for a given experiment will vary depending on researcher needs. Is it necessary to capture low-complexity or repetitive regions? Are the low-coverage regions being flagged important for achieving experimental goals? Is more depth necessary for difficult-to-capture regions or low-frequency variants? To help address the challenges stemming from the answers to these questions, researchers can turn to the literature, the academic community, as well as commercial entities with established track records in probe design.

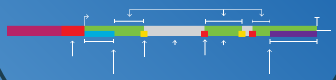
PROBE DESIGN 101
WHAT ARE THE KEY QUESTIONS?

What is the purpose of the study?



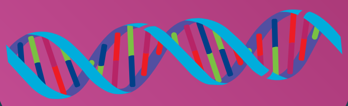
- Is the aim to detect known gene variations?
- Is there an established genome sequence to serve as a reference?
- How important is standardization?

What sequence information should be prioritized?



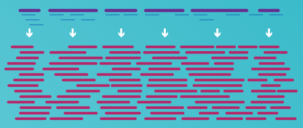
- Do entire genes need to be sequenced, or just specific regions?
- Do non-coding sequences (UTRs, introns, etc.) need to be captured?
- Is there a specific biotype of interest?
- What sequencing depth do you need?

Is the sample degraded?



- Probe binding affinity may be reduced in damaged or degraded samples
- Damaged nucleotides can negatively impact binding affinity
- Additional probe coverage may be necessary to compensate

Where can I turn to for help?



- Academic literature and collaborators can offer tried-and-true sequencing panels and protocols
- Off-the-shelf and custom panel options are available from commercial vendors
- Experienced vendors can provide in-depth design expertise
- Probe design software and genome browsers allow for full individual control

Next-Gen Sequencing

Roche

Primer Extension Target Enrichment for NGS

with the KAPA HyperPETE Portfolio

Combine the performance of hybrid capture with the speed and simplicity of amplicon-based workflows.

- Save valuable time with an efficient, single-day, automatable workflow
- Achieve superior performance and coverage uniformity
- Uncover critical genomic information from a wide variety of sample types, including FFPET and cfDNA
- Reliably enrich challenging, previously inaccessible genomic regions

Optimized to detect all major somatic variants in cfDNA, FFPE, and RNA samples, including SNVs, short indels, CNVs, MSI, and fusion transcripts (novel and known).

Learn more at:
go.roche.com/KAPAHyperPETE

For Research Use Only. Not for use in diagnostic procedures.