WHITE PAPER

Ion AmpliSeq technology

The importance of coverage: advantages of amplicon-based approaches in next-generation sequencing

In this paper, we highlight:

- Why genomic coverage is important
- The advantages of targeted amplification compared to hybridization enrichment
- The benefits of Ion AmpliSeq™ technology and its comprehensive solutions for targeted next-generation sequencing

Introduction

Next-generation sequencing (NGS) has proven to be a disruptive technology, furthering our scientific knowledge and opening research opportunities faster than anyone could have envisioned even just 10 years ago. During this time, massively parallel short-read sequencing has decreased sequencing costs faster than Moore's Law [1], enabling researchers to study entire genomes at an unprecedented scale and capacity. Fundamental advances in genomics enabled by NGS have made precision medicine a reality, with medical decisions and treatments now tailored to individuals and their genetic makeup.

While whole-genome sequencing has advanced discovery and human health, challenging regions of the genome are difficult to analyze using this approach, resulting in sequencing bias, and existing databases are noted to be neither complete nor accurate [2]. For many research applications the cost of whole-genome sequencing can still be a burden, particularly when you take into account the computational processing and informatics needs for analysis. This added cost and complexity would be of little benefit when studying a specific genomic region for disease and translational research applications. To help address this issue, many researchers have adopted a targeted sequencing approach to improve coverage, simplify analysis and interpretation, and lower their total sequencing workflow costs.

The importance of coverage

Coverage, as the word implies, describes the number of sequencing reads that are uniquely mapped to a reference and "cover" a known part of the genome. Ideally, the sequencing reads that are uniquely aligned are uniformly distributed across the reference genome and hence provide uniform coverage. In reality, coverage is not uniform and may be underrepresented in genetic regions of interest due to a variety of factors, including genomic complexity (Table 1). The genome contains an assortment of coding and noncoding DNA, repetitive sequences, and other elements that can make it difficult to align the sequencing reads to the proper genomic coordinates.

The number of sequencing reads that map to a known region is also an important part of coverage. A sufficient number of properly mapped reads is required to find and correctly identify genetic mutations. With high sequencing coverage, researchers can identify low-frequency mutations or discover mutations in a heterogeneous sample such as a tumor biopsy. Poor coverage, whether due to an insufficient number of reads or sequencing reads that are mapped incorrectly, will result in the inability to detect variants of interest.

Table 1. Potential reasons for poor sequencing coverage and uniformity.

Reason for poor coverage	Why this can affect coverage
Sample quality	Degraded samples are more difficult to prepare and result in shorter sequencing reads that are more difficult to map to the correct region since they may be less unique.
Sample input	Enough sample may not be available to sequence, and the DNA is not representative of the entire genome.
Homologous regions	Homologous regions have similar sequences, making it more difficult to map the reads to the correct positions of the reference genome.
Regions of low complexity	Sequence reads with low complexity may be mapped to the wrong part of the genome, resulting in coverage bias.
Hypervariable regions	Due to the high number of variants, the sequencing read will look very different compared to the reference genome and may not be mapped appropriately.
GC content	Potential sequencing bias due to the percentage of guanine and cytosine nucleotides.

ThermoFisher SCIENTIFIC

Coverage is clearly important to ensure that the genomic region of interest can be studied with high confidence. For regions with low coverage, researchers frequently increase the sequencing throughput for their studies and try to increase coverage by brute force. However, this method is inefficient, increases costs, and does not address the underlying reasons for the poor coverage itself. By increasing throughput, genomic regions with sufficient coverage will be overrepresented and the reads effectively wasted. Areas with zero coverage before may not have sufficient coverage just by sequencing more sample. A more efficient way to address coverage is a targeted sequencing approach. Through targeted sequencing, researchers can focus on just their regions of interest instead of sequencing the entire genome. This helps to reduce costs and ensure sufficient coverage, including in parts of the genome that may not have been accessible previously.

Targeted sequencing

Leveraging current genomic knowledge, a targeted NGS approach utilizes molecular biology methods to enrich for specific genetic sequences, allowing researchers to focus their studies on individual genes or genomic regions. Obtaining sequence coverage of challenging genomic regions is now possible, including regions from difficult sample types such as samples with degraded DNA or RNA, or circulating cell-free DNA from blood. By only sequencing what you need, there are cost benefits beyond more efficient computational processing and informatics. Focusing on specific regions of interest allows researchers to sequence at a much higher depth of coverage for rare variant discovery. Many more samples can also be processed simultaneously in a single sequencing run, for faster time-to-results from individual and cohort samples.

Hybridization capture and amplicon-based enrichment are two general techniques used in sequencing to enrich for specific genetic regions (Figure 1). Hybridization capture can be either solution based or performed on a solid substrate such as a microarray. Both require the use of synthesized oligonucleotide probes (also known as baits) that are complementary to the genetic sequences of interest. In solution-based methods, the probes are biotinylated and added to the genetic material in solution to hybridize with the desired regions of interest. Magnetic streptavidin beads are used to capture and isolate the hybridized sequences from the unwanted genetic material. With array-based capture, the probes are attached directly to a solid surface. The genetic material is applied to the microarray, and target regions hybridize to the surface. Any unbound material is washed away, leaving the desired target regions isolated on the substrate. For both methods, the isolated and purified target regions are subsequently amplified and prepared for sequencing. Amplicon-based enrichment uses carefully designed PCR primers to flank targets and specifically amplify regions of interest. The amplified products are then purified from the sample material and used for sequencing, bypassing the need for enrichment by hybridization.

Advantages of amplicon-based approaches

There are many advantages in using amplicon-based enrichment techniques compared to hybridization capture methods (Table 2). Amplicon-based approaches offer a simpler, faster workflow with unmatched PCR specificity, allowing enrichment for target gene regions from low sample input amounts. Genetic material from limited sample sources, such as fine-needle aspirates or circulating tumor DNA, can be sequenced for biomarker discovery.

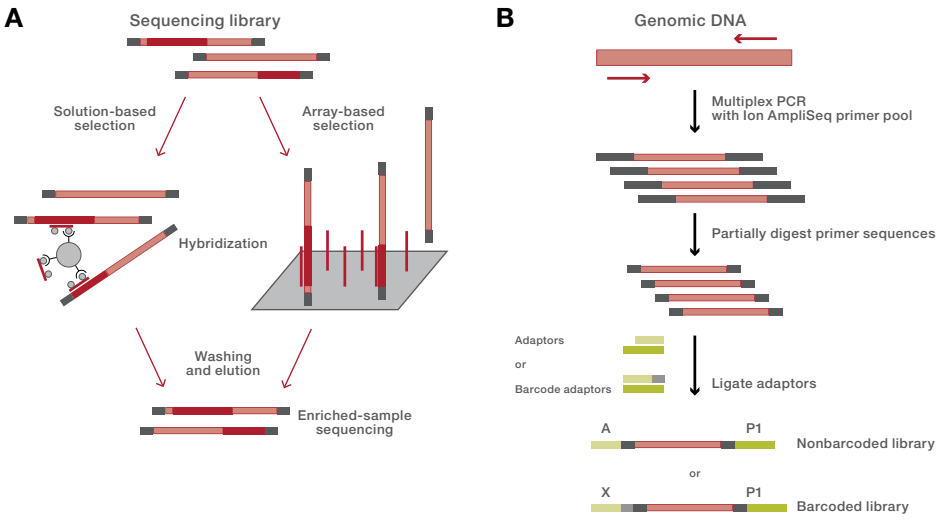

Figure 1. Workflows for target enrichment. (A) Hybridization capture can be performed in solution or on a solid substrate. **(B)** Amplicon-based target enrichment using Ion AmpliSeq technology.

Table 2. Advantages of amplicon-based enrichment compared to hybridization capture methods.

Advantage	Example	Why is it important?
Low sample input	Liquid biopsies and fine-needle aspirates	Enables sequencing of limited samples, ensuring appropriate sequencing coverage.
Better enrichment for homologous regions	Pseudogenes	Frequently studied genes such as *PTEN* have associated pseudogenes [3].
	Paralogs	Monogenic disease genes often have functionally redundant paralogs [4].
Targeting of hypervariable regions	T cell receptor	Predictive biomarker discovery for immunotherapy [5].
Target and enrich low-complexity regions	Di- and tri-nucleotide repeats	Microsatellite instability studies for diseases such as prostate cancer [6].
Fusion detection	*ETV6-NTRK3*	Known fusion oncogene for secretory cancers [7] and other cancer types. *NTRK3* inhibitors may be therapeutic for cancers exhibiting this biomarker.

A targeted gene of interest may share homology with another segment of the genome. Hybridization capture may have difficulty distinguishing between the two regions, resulting in nonspecific enrichment. This is not an issue with amplicon-based enrichment, where PCR primers can be uniquely designed to target the desired region. For example, the *PTEN* gene is a known tumor suppressor gene and is one of the most commonly disrupted suppressor genes in cancer. *PTENP1* is a processed pseudogene very similar in sequence, but a mutation prevents translation of the normal *PTEN* protein. Being able to distinguish and target the right gene and avoid the pseudogene clearly is important in cancer research. The same concept applies when targeting low-complexity regions that are prevalent in whole genomes, such as di- and tri-nucleotide repeats.

Since capture hybridization requires development of complementary capture probes against a known reference genome, genetic mutations of interest can potentially disrupt the hybridization itself and result in failure to enrich for the target region of interest. Again, since PCR primers are uniquely designed to flank and then amplify target regions, amplicon-based enrichment can better detect known and novel insertions, deletions, and fusion events. This is particularly true in genetic regions that have many variants in close proximity to one another. Hypervariable regions such as the T cell immune repertoire can be sequenced using amplicon-based enrichment, providing translational researchers a tool to discover predictive biomarkers.

Ion AmpliSeq technology

Thermo Fisher Scientific has further advanced amplicon-based enrichment with Ion AmpliSeq technology. Unique to this technology is the ability to multiplex up to 24,000 primer pairs in a single reaction, allowing researchers to sequence hundreds of genes from multiple samples in a single sequencing run with faster turnaround time and lower cost. Targeted enrichment can be performed on as little as 1 ng of low-quality DNA or RNA, with more full-length PCR products and thus better coverage compared to other amplicon-based methods. The analytical validity of Ion AmpliSeq technology has been demonstrated in thousands of peer-reviewed publications across a broad range of applications, including oncology, inherited disease, and microbial research.

Ion AmpliSeq technology is powered by a designer (Figure 2) that leverages our capabilities in both PCR amplification and comprehensive NGS. The Ion AmpliSeq Designer allows us to rapidly design and build custom assays that address the needs of each researcher and enable sequencing and analysis in as little as one day.

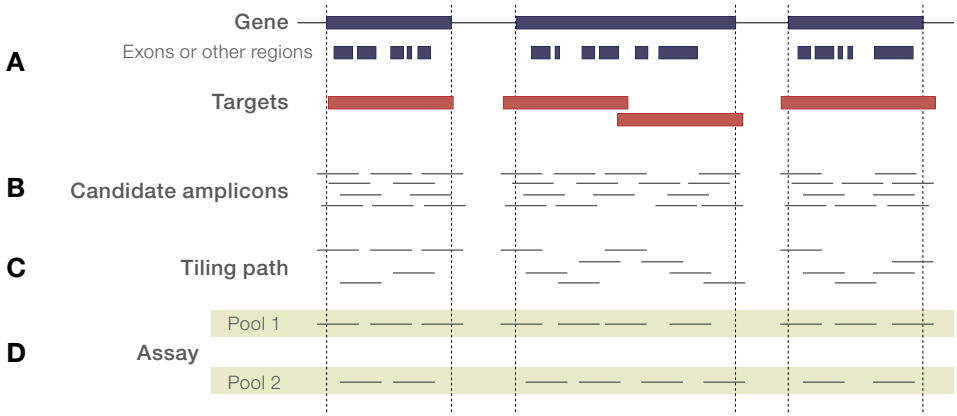

Figure 2. Schematic of the Ion AmpliSeq Designer. (A) The designer starts with identification of the genome of interest and targets of interest. For human DNA designs, the targets of interest may be genes, genomic regions, or hotspots (SNVs or small indels of interest). The designer then **(B)** generates a large number of candidate amplicons, and **(C)** tiles the amplicons so that they can be optimally combined in the available pools, maximizing the coverage of the requested targets. **(D)** The designer selects the best amplicons that fully cover the targets of interest, in the number of pools desired. For gene and region designs, two pool designs are often preferred. For hotspot designs, one pool design may be preferred.

ion torrent

Recent advances in the designer allow greater sequencing efficiency with less sample in difficult-to-enrich genetic regions (Figure 3). The faster designer has improved *in silico* coverage, resulting in assay designs with higher coverage and greater uniformity. This means that genetics targets can be analyzed more efficiently, lowering costs, reducing throughput requirements, and decreasing turnaround time from initial design to obtaining sequencing data.

Summary

Appropriate sequencing coverage across the genome is needed to help ensure that the scientific community has complete and accurate genomic data to further discovery and improve human health with precision medicine. Researchers have turned toward targeted enrichment approaches to help sequence specific regions of interest at much higher depths of coverage and at lower cost. Compared to hybridization capture, amplicon-based enrichment methods can target difficult genomic regions with lower input amounts of DNA and RNA to enable discovery. Ion AmpliSeq technology is the industry leader in target enrichment, providing higher multiplexing and coverage in a single assay than other methods. Our comprehensive solutions, including our unique design pipeline, offer a robust, customizable set of tools designed to work together to help answer your unique and important scientific questions faster and with less effort.

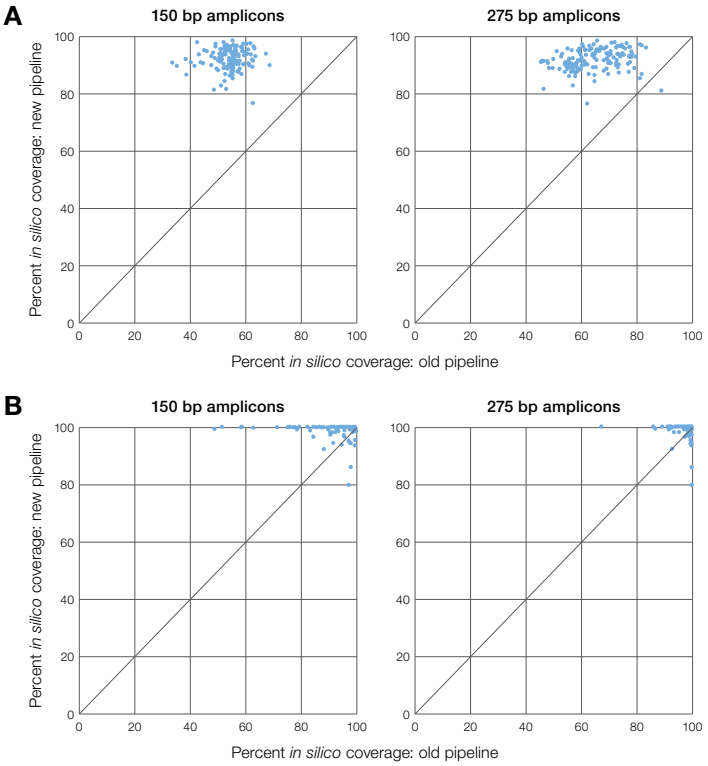

Figure 3. Percent *in silico* coverage comparison of new pipeline vs. old pipeline. The comparison was performed for **(A)** single-pool assay designs and **(B)** two-pool assay designs, for 150 target genes. For each pool, designs were created for 150 bp and 275 bp amplicon lengths. The 150 target genes selected for this analysis are the most common genes designed with our Ion AmpliSeq Designer, and each gene is represented by a single dot in each plot. The diagonal line in each plot represents equivalent coverage performance between the new and old designers. Overall design performance is improved with the new pipeline, with most genes having higher coverage as indicated by the dots being to the left of and above the diagonal. In particular, the improvement in coverage is significant for single-pool assay designs and for designs that require shorter amplicons. For researchers, this translates to greater sequencing efficiency and higher confidence in sequencing lower-quality samples.

References

1. National Human Genome Research Institute. The cost of sequencing a human genome. https://www.genome.gov/27565109/the-cost-of-sequencing-a-human-genome/
2. Watson CT et al. (2017) Comment on "A database of human immune receptor alleles recovered from population sequencing data". *J Immunol* 198:3371–3373.
3. Chu EC et al. (2004) *PTEN* regulatory functions in tumor suppression and cell biology. *Med Sci Monit* 10:RA235–241.
4. Chen W-H et al. (2013) Human monogenic disease genes have frequently functionally redundant paralogs. *PLoS Comput Biol* 9:e1003073.
5. Zhang L et al. (2018) Peripheral blood TCRB repertoire convergence and clonal expansion predict response to anti-CTLA-4 monotherapy for cancer. *J Immunother Cancer* 6(Suppl 1):P82.
6. Le DT et al. (2017) Mismatch repair deficiency predicts response of solid tumors to PD-1 blockade. *Science* 357:409–413.
7. Stenman G (2013) Fusion oncogenes in salivary gland tumors: molecular and clinical consequences. *Head Neck Pathol* 7:S12–19.

Find out more at **thermofisher.com/ampliseq**

ThermoFisher SCIENTIFIC