Targeted NGS coverage metrics that matter

Optimizing on-target rate, molecular complexity, and target coverage uniformity for your research

ABSTRACT

Targeted next generation sequencing helps obtain genetic information without the cost and time needed to sequence the entire genome. Hybridization capture is an effective targeted sequencing method. Sequencing efficiency of hybridization capture can be measured by three main metrics: on-target rate, complexity of captured molecules, and uniformity of target sequencing coverage. Here, we explain these key sequencing metrics and highlight that while sequencing metrics are useful indicators for NGS performance in research applications, ultimately what matters for targeted sequencing is reliable variant calling which comes from reliable target coverage.

INTRODUCTION

Next generation sequencing (NGS) has been widely adopted in research applications ranging from sensitive detection of low-abundance targets to profiling genomes in normal and disease states. Regardless of the scientific question being addressed, the NGS assay must reach sufficient sequencing read depth to inspire confidence about the observed sequence variations (variants). Ideally, the NGS assay has high sequencing efficiency which means achieving the required sequencing coverage across all intended targets of interest with minimal amount of total sequencing data.

While whole genome sequencing (WGS) can be used to obtain information from all 3 billion base pairs in the haploid human genome, the cost of sequencing can become inhibitory depending on the sequencing depth needed to detect the variants of interest. A more cost-effective method for researchers is to focus on the exome or a subset of the exome since coding regions make up <2% of the genome [1]. Compared to WGS, whole exome sequencing (WES) allows researchers to get similar coverage for more samples using the same total sequencing data size or increased coverage for the same number of samples using the same total sequencing data size (Figure 1).

Hybridization capture is an effective and proven method to carry out targeted next generation sequencing (Figure 1) [2]. Nucleic acid samples (DNA or RNA) are typically converted to WGS libraries with sequencing platform-specific adapter sequences in a process called library preparation. In the subsequent hybridization capture step, libraries are incubated with biotinylated oligonucleotides called probes which are complementary to the target sequence. If sample-specific barcode sequences are incorporated during library preparation, multiple libraries may be pooled together to carry out multiplexed hybridization capture. The hybridization reaction also includes blocking oligonucleotides (dark blue in Figure 1B) that prevent non-specific interactions among adapter sequences that can cause off-target capture. Streptavidin is used to capture the biotinylated probe-DNA hybrids from solution, and non-specific interactions are washed away. Target-enriched libraries are PCR-amplified to obtain sufficient quantity of each target for sequencing. Many hybridization-capture assays aim to achieve similar read depth across all targets with the use of a single pool of probes, typically called a panel. In some cases, multiple panels may be combined during hybridization to achieve the different levels of coverage for different targets in a single sequencing run.

Sequencing efficiency of targeted NGS assays can be characterized by three main variables: on-target rate, complexity of captured molecules, and uniformity of target sequencing coverage. This paper reviews these key sequencing metrics. While sequencing metrics are useful indicators for NGS assay performance, what matters for targeted sequencing is reliable variant calling which requires reliable target coverage.

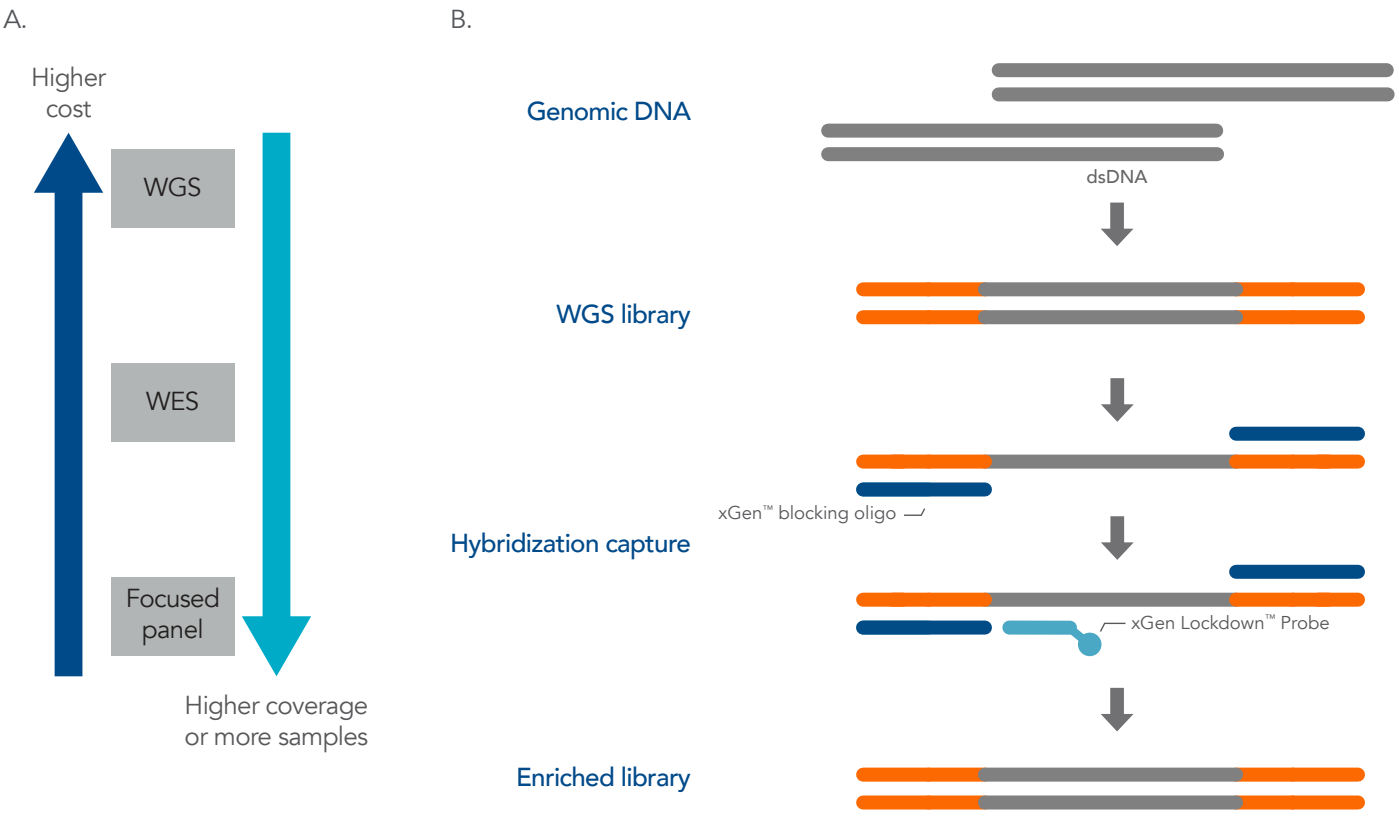


Figure 1. Schematic of hybridization capture workflow which reduces sequencing cost per sample compared to whole genome sequencing (WGS). (**A**) WGS covers the entire genome but comes with higher cost per sample than targeted sequencing. As sequencing becomes more targeted from WGS to whole exome sequencing (WES) smaller panels, more samples and/or higher coverage can be delivered with the same amount of sequencing data. (**B**) Hybridization capture enriches target sequences from WGS libraries. Using targeted sequencing requires adding more steps compared to WGS in the workflow but saves substantial sequencing cost and downstream analysis time.

High on-target rate increases the amount of usable sequencing data

In a hybridization capture workflow, probes are designed to be complementary to the targeted regions so that they will hybridize to the targets of interest. The resulting probe and target DNA duplex is captured on streptavidin-immobilized beads; the fragments that are not bound to a probe are washed away. High specificity between the probe and target DNA during capture reduces sequencing of unintended targets and makes a significant impact upon the amount of usable data coming from the sequencer. For example, if 60% of the sequencing reads are specific to the target regions of interest, and 40% of the reads are non-specific, then 40% of the sequencing data are not useful for the intended goal of the research experiment.

Sequencing reads that are specific to the target region are called "on-target," and reads that are from regions which were not targeted are called "off-target." Target capture specificity can be measured by the on-target rate which is defined by the ratio of sequenced reads that are specific to the targeted region of interest to the total number of reads that are mapped to the genome (Figure 2). Picard HsMetrics offers a standard on-target rate metric (PCT_SELECTED_BASES) that can be conveniently referenced by NGS users to get a standard estimate of the on-target rate [3].

On-target rate is important in optimizing sequencing efficiency because it maximizes the generation of usable sequencing data (illustrated in Figure 3B). Hybridization-based capture workflow solutions offered by various manufacturers have on-target rates as low as 50% and as high as 90% [4, 5]. Depending on the sequencing platform, even a seemingly moderate 10% increase in on-target rate can translate to a significant increase in usable sequencing data.

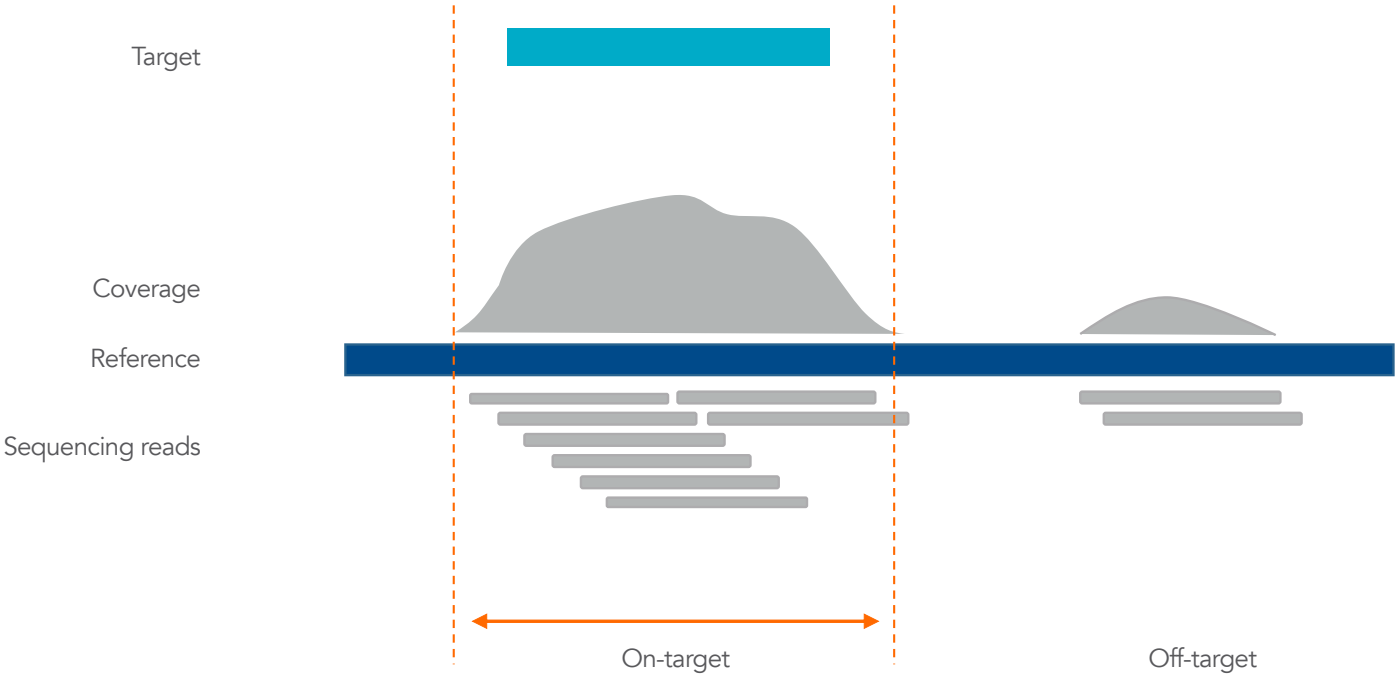


Figure 2. Usable sequencing data is generated from on-target reads. On-target sequencing reads map to or near the target region (light blue). Unintended target reads that map to regions in the genome that were not targeted are called off-target reads which lower sequencing efficiency.

Higher complexity maximizes the usability of the on-target reads

Another NGS metric, called complexity, increases the accuracy of identifying sequence variants. Complexity refers to the number of unique molecules from the sample that are represented in the final sequencing data (Figure 3A). There are several factors that contribute to diversity of unique molecules in the final sequencing data. Two key factors are the conversion of initial input DNA into libraries and on-target rate during hybridization capture.

The number of unique molecules in the targeted portion of a sample library can be estimated using the Picard metric: HS_LIBRARY_SIZE. Since this estimate is an extrapolation from the number of on-target reads for the sample, HS library size measurements can be used to compare panels of equal size with similar total sequencing depth.

Duplication rate (percent duplication) from Picard Duplication Metrics, among others, can be used to estimate how many reads in the sample do not represent unique molecules. Duplicates are reads with the same start and stop sites, and duplication rate is calculated as the fraction of duplicates in the total number of mapped reads. An elevated duplication rate could be the result of a low-complexity library, but it could also arise from deeper sequencing. High duplication from deep sequencing is useful in error correction analysis that helps differentiate between sequence artifacts and real mutations present at low abundance. For applications that do not require error correction, duplicate sequencing reads have little use, and therefore, a lower duplication rate correlates with higher complexity and a more efficient use of the sequencing data. On-target rate and complexity metrics can be considered together to assess how much of the sequencing data is usable (Figure 3B). For more information on duplication rate, see the article, **Minimizing duplicates and obtaining uniform coverage in multiplexed target enrichment sequencing**. For more information about error correction, see the analysis guideline, **xGen™ Dual Index UMI Adapters*—Tech Access: Processing sequence data with unique molecular identifiers (UMIs)**.

It is helpful to look at both the HS library size and duplication rate to monitor complexity because they are not always inversely correlated. Higher duplication rate can also come from higher on-target rate which increases target sequencing depth for the same amount of total sequencing data. For example, hybridization captures performed with or without blocking oligos resulted in higher on-target rate and lower on-target rate, respectively (Figure 3C). When blocking oligos are used to increase on-target rate, the HS library size does not change, but duplication rate increases. The mean target coverage metric confirms the increase in target sequencing depth, and it is this sequencing efficiency that leads to the higher duplication rate. This example highlights the importance of interpreting the two complexity metrics in context with other sequencing metrics.

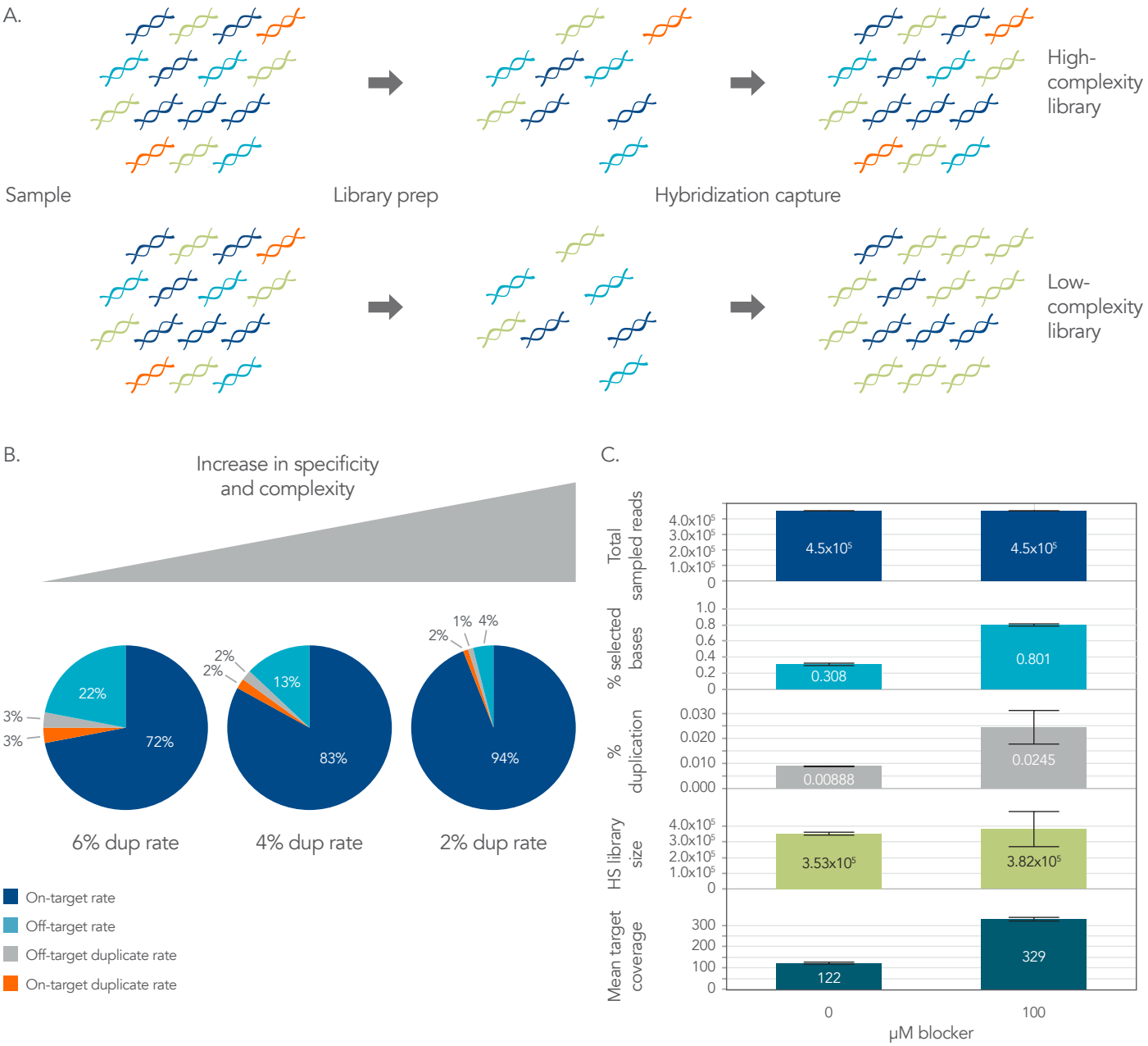


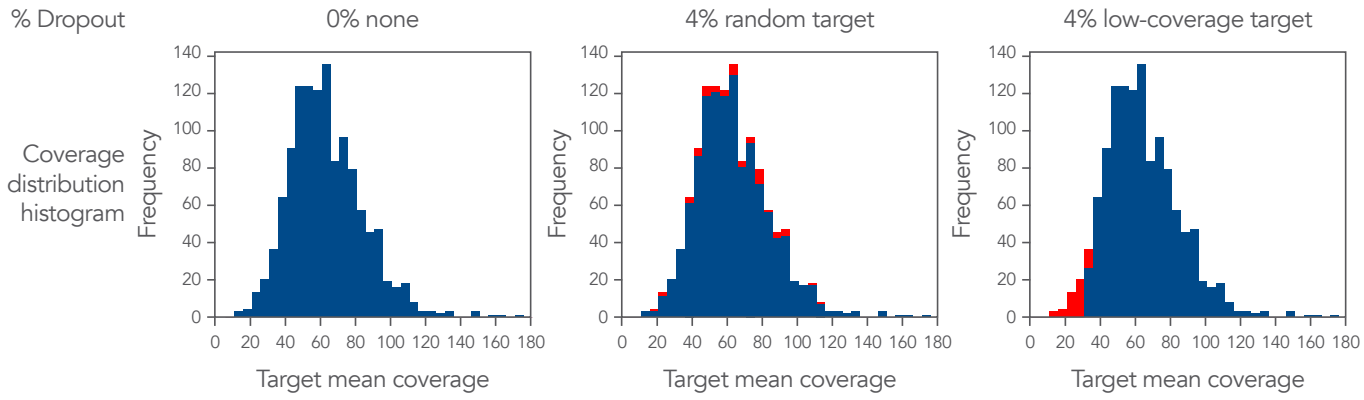
Figure 3. High on-target rate and high complexity yields more usable sequencing data. (**A**) Diagram of high- and low-complexity libraries generated by targeted sequencing workflows. Sequencing a low-complexity library would miss detection of low-abundance targets depicted in orange (lost during library prep) and light blue (lost during hybridization capture). (**B**) High on-target rate and low duplication rate together increase the usability of sequencing data. (**C**) Increased duplication rate does not always indicate lower complexity.

High uniformity decreases the amount of usable sequencing data needed per sample

Uniformity describes the distribution of observed coverage values across the entire population of targets. Different target sequences will have different enrichment efficiencies in the hybrid capture workflow which results in some targets with high coverage and some targets with low coverage. Although uniformity depends on the entire distribution of observed coverage values, it is impractical to describe every data point in the distribution to convey uniformity. Therefore, uniformity calculations attempt to simplify the distribution of observed coverage values across the entire population of targets into a single value. Uniformity is calculated after off-target and duplicated sequences are removed.

- Picard analysis tools offer a simple estimate using the Fold-80 Base Penalty, which describes the fold of additional sequencing required to ensure that 80% of the targeted bases achieve at least the mean coverage. Fold-80 is calculated by the mean target coverage divided by the low 20th percentile coverage, and this method emphasizes the low coverage regions but omits the zero-coverage regions (Figure 4).
- An alternative method assesses a broader range of the coverage distribution by calculating the uniformity as a percentage of targeted bases that are observed within a defined range in normalized coverage, such as 50–200% of the mean target coverage (percent bases at 0.5–2.0X mean coverage). This can be calculated using the output from the same Picard analysis (HS Metrics).

It is important to understand the limitations of uniformity measurements. For example, the fold-80 metric has the caveat that zero-coverage targets are completely excluded from its calculation. Fold-80 is also based on just two values out of the entire coverage distribution histogram. Uniformity estimated by the fold-80 metric may remain the same or even improve with additional drop-outs or zero-coverage regions in the hybridization capture assay (Figure 4). To illustrate this, data simulation for a panel targeting the ACMG59 genes (reference https://www.ncbi.nlm.nih.gov/clinvar/docs/acmg/) was carried out. This model panel had a starting uniformity of 1.41 fold-80 and 95% bases 0.5–2.0X mean coverage. 50 exons were selected randomly or by picking the ones with lowest coverage, and their coverage was converted *in silico* to zero. The fold-80 metric remained the same and did not reflect the complete loss in coverage for both random and lowest-coverage targets. The uniformity calculation using the percent bases 0.5–2.0X mean coverage was more representative of the actual sample; it was sensitive to random target coverage loss but was still insensitive to loss of lowest-coverage targets (Figure 4). Therefore, uniformity measurements can obscure unexpected loss in coverage and cannot be used as a sole indicator for NGS assay performance. The goal of targeted sequencing is to provide complete coverage of all targets, including the low-abundance targets which may be lost when on-target rate and complexity are low.



	0% none	4% random target	4% low-coverage target
Fold-80 base penalty	1.41	1.41	1.40
Percent bases 0.5–2.0X mean coverage	95%	92%	95%

Figure 4. Loss of targets is challenging for uniformity measurements to detect. Comparison of coverage distribution and uniformity calculations for a simulated panel with target coverage loss. Target loss is highlighted in red in the histograms.

Optimizing uniformity allows more targets to achieve the required coverage without increasing the total sequencing depth. With lower uniformity, targets with highest coverage will consume more sequencing data than necessary at the cost of targets with lowest coverage. With higher uniformity, a greater proportion of the targeted regions will achieve required coverage level without the need for additional sequencing. As a result, more target bases can meet the minimum required coverage level without increasing the total sequencing depth. Depending on the size of the targeted regions, the coverage requirement, and the sequencer capacity, the optimization of uniformity may or may not allow additional samples to be sequenced in the same run. If the sequencing capacity saved by lower sequencing data size per sample is enough to cover additional sample(s), then higher uniformity allows a more efficient use of sequencing run.

CONCLUSION

Achieving sufficient coverage across all intended targets of interest with the least amount of total sequencing data is the ultimate goal of targeted sequencing. Using a targeted NGS assay in your research with optimal sequencing efficiency maximizes specific capture of as many unique molecules of input DNA as possible but also uniformly across all targets. Efficiency improvements can be made in many steps of the hybridization capture workflow to improve on-target rate, complexity, and uniformity, but relying too heavily on a single metric can become counterproductive to sequencing efficiency optimization efforts. While all three sequencing metrics combined can provide guidance to researchers on sequencing efficiency or the sequencing cost, the best fit for the targeted sequencing solution depends on the researcher and the research project. Variable hybridization times, multiplexing, customizable panel designs and workflow protocols, and reliability of reagent quality are critical factors not captured even when analyzing multiple sequencing metrics. Flexibility in workflow setup and handling are key in accommodating the varied and changing needs of researchers using targeted NGS. All these metrics must be taken into consideration when choosing the best NGS products for your investigations.. For more information about NGS workflows, methods, and applications, download our **Targeted sequencing guide**. To learn about the on-target rates and coverage of the **xGen Exome Research Panel v2*** during internal feasibility studies, download our white paper, **Consistent, comprehensive efficient: An improved human exome sequencing solution**.

METHODS

To demonstrate NGS metrics, investigational hybridization capture reactions using a custom 57 kb xGen hybridization capture panel* were set up with or without the addition of xGen Blocking Oligos*. Sequencing data was subsampled to 450K reads before analysis using the Picard suite of tools described in the discussion sections.

ACMG59 gene list [6] was used to extract the coding sequences from the RefSeq 109 database as targets in the simulation. Sequencing reads were simulated for the 1217 extracted targets in ACMG59 based on internal existing data. The sequencing reads overlapping with 50 randomly selected targets within ACMG59 were removed from the original simulation data to model the random dropout case. The sequencing reads overlapping with 50 targets with the lowest coverage level were removed from the original data to a simulate low-coverage dropout situation. The Fold_80_BASE_PENALTY and the percent of target bases within 50–200% of the mean target coverage were calculated using Picard HsMetrics [3]. The target coverage histogram was plotted for each of the three simulated situations, with blue denoting the actual coverage and red denoting the coverage being dropped (Figure 4).

REFERENCES

1. Schwarze K, Buchanan J, Taylor JC, et al. Are whole-exome and whole-genome sequencing approaches cost-effective? A systematic review of the literature. Genet Med. 2018;20(10):1122-1130.
2. Samorodnitsky E, Datta J, et al. (2015) Comparison of custom capture for targeted next-generation DNA sequencing. J Mol Diagn 17(1): 64–75.
3. Institute B (2019) Picard Toolkit. Broad Institute, GitHub repository http://broadinstitute.github.io/picard/.
4. Clark MJ, Chen R, et al. (2011) Performance comparison of exome DNA sequencing technologies. Nat Biotechnol 29(10): 908–914.
5. Sulonen AM, Ellonen P, et al. (2011) Comparison of solution-based exome capture methods for next generation sequencing. Genome Biol 12(9): R94.
6. Coriell Institute. ACMG59 genes. Accessed June, 2021.

Technical support: applicationsupport@idtdna.com