The Importance of Coverage Uniformity Over On-Target Rate for Efficient Targeted NGS

Yehudit Hasin-Brumshtein, Ph.D., Maria Celeste M. Ramirez, Ph.D., Leonardo Arbiza, Ph.D., Ramsey Zeitoun, Ph.D.

INTRODUCTION

Next-generation sequencing (**NGS**) has become the technique of choice for variant detection in both research and clinical settings. Although the cost of sequencing is steadily decreasing, large-scale, whole genome sequencing is still prohibitively expensive, so investigations often focus on specific genes and loci using targeted sequencing (Dillon, et al. 2018).

Targeted sequencing relies on enrichment of genomic regions of interest prior to sequencing. In exome sequencing, for example, biotinylated synthetic DNA probes are designed to hybridize to exon regions. Following hybridization with a genomic DNA sample, probes are purified to produce a sample that is enriched for the exon regions. Although target enrichment can reduce sequencing costs and make experiments more feasible and focused, it also introduces biases that compromise the efficiency of the sequencing effort (Goldfeder et al. 2016, Meynert et al. 2013, 2014).

While some inefficiency is unavoidable due to the stochastic nature of targeted NGS, much of it is inherent to the design and production of target enrichment probe panels (Warr et al. 2015). Some probes cross-hybridize to non-target regions, leading to “off-target” (non-specific) capture. Probe panels may also have imbalances in capture efficiency (lack of uniformity) that lead to over-enrichment of some targets and under-enrichment of others. To ensure high-confidence data, researchers must increase the amount of sequencing to boost coverage of areas with low read depth. This strategy, however, leads to over-sequencing of otherwise adequately covered regions, which in turn results in higher sequencing costs and reduced efficiency.

The extent of this “wasted sequencing” is reflected in the uniformity and on-target rate, two metrics that describe the overall efficiency of targeted sequencing. In this paper, we use ranges of on-target rate and uniformity typical of commercial exome kits to mathematically model the relative impacts of both metrics on overall efficiency. We demonstrate that, though most commercial probe panels cite only on-target rate in their specifications, uniformity has a more significant contribution to the efficiency of targeted sequencing.

EVALUATING SEQUENCING REQUIREMENTS

When designing a sequencing experiment, a fundamental task is to determine how many reads are required per sample for actionable data (**read coverage**). The answer determines the costs, feasibility, number of samples to include, and ultimately the study power to reach meaningful conclusions. Different applications require different read coverage: for example, whereas information from ten reads that align over a given position (10x coverage) may suffice for a call of germline variation in a research setting, this number would be inadequate for a confident call of somatic mutation in a clinical setting. We refer to the desired coverage as $\mathbf{C_D}$ and the mean coverage actually observed in the experiment as $\mathbf{C_M}$.

An ideal sequencing experiment would generate reads that are distributed equally and exclusively across target regions (perfect uniformity and on-target capture, respectively). The rest of the genome would be devoid of reads (**Figure 1A**). In this ideal scenario, sequencing efficiency would be 100%, and C_M would equal C_D. Non-uniform and off-target capture are inevitable, however, and they lead to variable coverage (**Figure 1B**).

To ensure coverage of most targeted regions reaches C_D, the amount of sequencing is often increased such that $C_M >> C_D$ (**Figure 1B**). This strategy, however, wastes a considerable fraction of sequencing reads. The C_M/C_D ratio represents the amount over-sequencing needed to ensure a certain percentage of targets reach C_D: the larger the ratio, the more over-sequencing will be required to get enough usable data. Optimizing the efficiency of targeted NGS, therefore, involves minimizing the C_M/C_D ratio without compromising results.

FIGURE 1

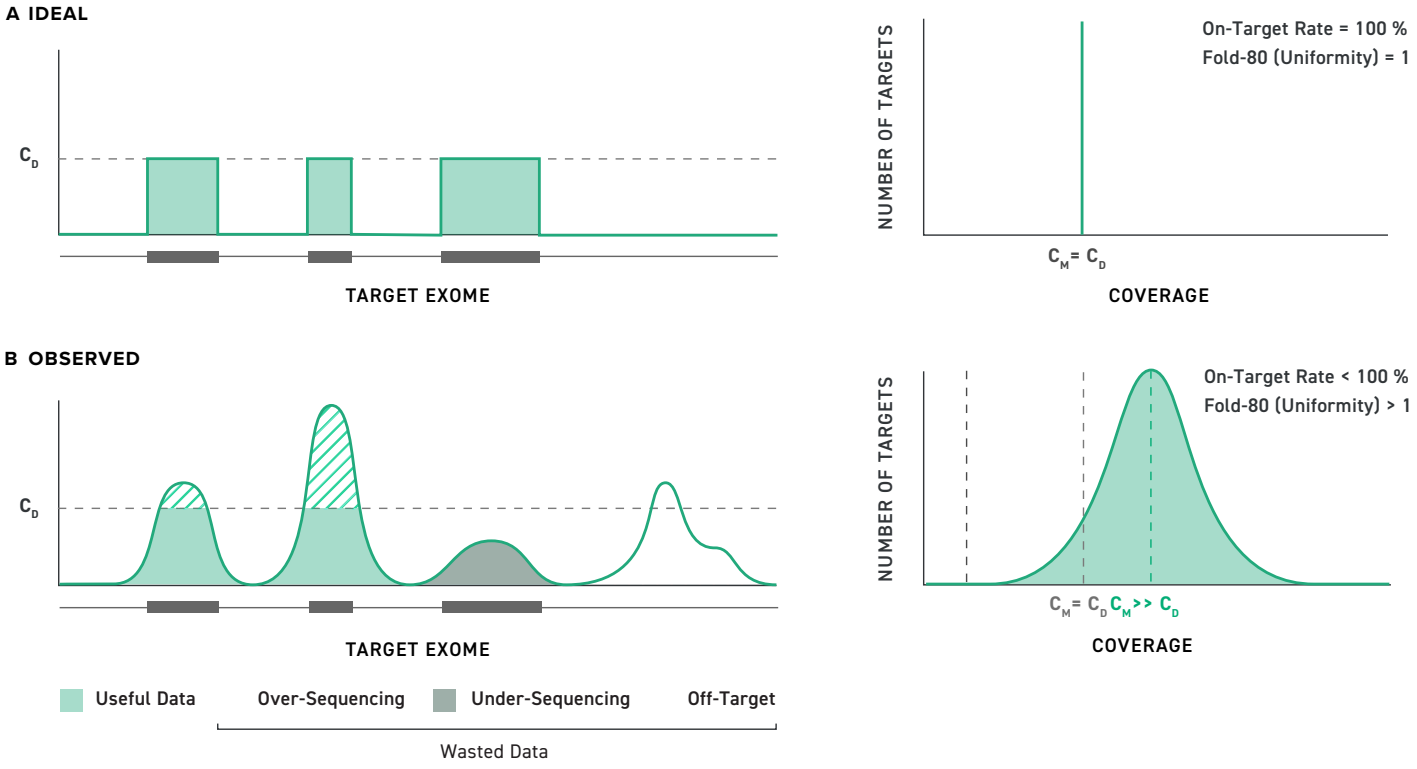

Figure 1. Read distribution. A. Read distribution in an ideal experiment, where all targets have specific and equal read depth, and non-target regions are free of reads. In this situation, C_M=C_D. **B.** Representation of a realistic distribution of coverage, where some targets are under-sequenced, others are over-sequenced, and off-target regions are also captured.

UNIFORMITY AND THE FOLD-80 METRIC

Uniformity describes the read distribution along target regions of the genome. Uniform coverage reduces the amount of sequencing required to reach a sufficient depth of coverage for all regions of interest. Uniformity is a measure of the spread around the C_M and is estimated from the mean and quantiles of the read distribution (**Figure 2**).

A convenient metric for uniformity is the **fold-80** base penalty (fold-80 for short). Calculated by the widely used Picard[1] pipeline, fold-80 is the fold of additional sequencing required to ensure that 80% of the target bases achieve C_M. For example, if one million reads produce a C_M of 30x, a fold-80 of 2.0 means two million reads would be required to ensure that 80% of the targeted bases reach 30x coverage. A fold-80 of 1.4 would mean that increasing sequencing to 1.4 million reads would achieve the same goal.

Assuming a normal distribution, fold-80 is proportional to the coefficient of variation (the ratio of the standard deviation to the C_M) and is greater than 1.0 (a fold-80 of 1.0 would indicate perfect uniformity and no variance, Figure 1A). Higher fold-80 scores correspond to wider coverage distribution and low uniformity, and lower fold-80 scores indicate high uniformity (all target bases are sequenced with similar coverage).

FIGURE 2

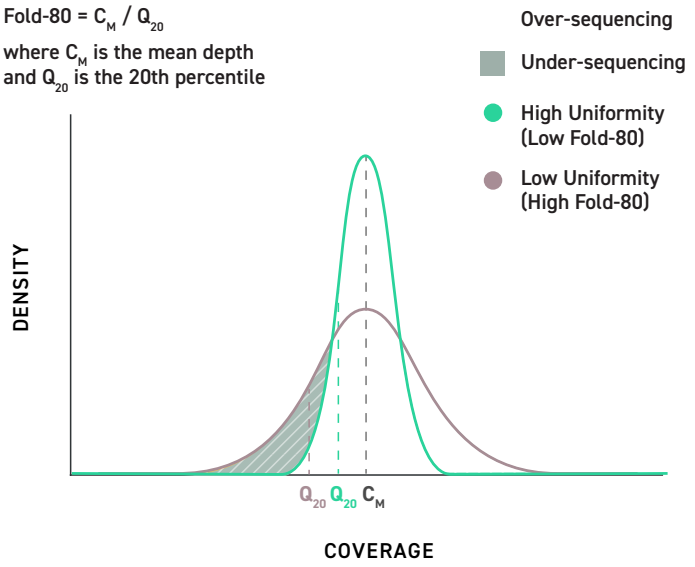

Figure 2. Uniformity reflects distribution shape. Two different hypothetical read distribution profiles showing low (green) and high (gray) fold-80 scores and the relative abundance of reads mapping back to over- and under-sequenced regions. Lowering the fold-80 score (gray curve to green curve) both rescues under-sequenced regions and reduces the fraction of over-sequenced regions for more efficient read utilization. In reality, poor uniformity often shows less symmetric distributions.

ON-TARGET RATE

On-target rate describes the percentage of sequencing data that maps to target regions; conversely, off-target rate refers to the sequencing data that maps to other regions (**Figure 1B**). It is typically expressed as the ratio of the number of sequenced bases covering the target regions to the total number of mapped bases output by the sequencer (**Figure 3**). Some off-target sequencing is inevitable; a considerable proportion of it is probe panel-specific and can be due to promiscuous hybridization.

FIGURE 3

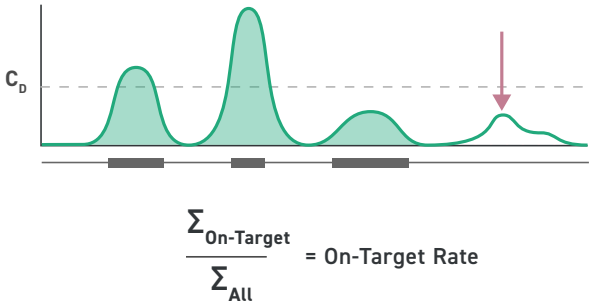

Figure 3. On-target rate is the proportion of the sequencing effort that maps to targeted regions. In calculating on-target rate, the entire sequencing effort (Σ_{ALL}) is represented by the area under the sequencing curve, and the on-target area ($\Sigma_{On\text{-}Target}$) is represented by the green area. Here, off-target sequencing is indicated by the arrow.

[1] *https://broadinstitute.github.io/picard/*

RELATIVE IMPACTS OF OPTIMIZING UNIFORMITY AND ON-TARGET RATE

Uniformity (fold-80) and on-target rate both define the efficiency of targeted sequencing. But how much impact does each metric have?

As long as library preparation conditions for the probe panel are consistent, on-target rates tend to vary only a little and can be considered a "tax" on the sequencing effort (Chilamakuri et al. 2014). When uniformity is perfect (fold-80 is 1.0), the on-target rate and C_M are inversely proportional. For example, assuming a desired coverage (C_D) of 10x and perfect uniformity, an on-target rate of 80% would mean one should aim for a C_M of 12.5x:

$$\mathbf{C_M = C_D \,/\, on\text{-}target\ rate} = 10 \,/\, 0.8$$
$$C_M = 12.5x$$

Conversely, even small improvements in fold-80 can significantly improve efficiency. Improving uniformity reduces coverage of over-sequenced targets and increases coverage of under-sequenced targets.

To examine the relative effects of on-target rate and uniformity, we simulated 3,003 normal distributions[2] with varying uniformity, mean coverage, and on-target rates. Improving the on-target rate while maintaining constant uniformity (**Figure 4A**) shifts the coverage distribution toward higher mean (C_M) values, increasing the proportion of bases covered above the desired coverage (C_D). Improving fold-80 scores, as stated earlier, improves read utilization by both rescuing under-sequenced regions and reducing the fraction of over-sequenced regions (**Figure 4B**). In this case, although mean coverage (C_M) values remain constant, the proportion of bases covered above the desired coverage (C_D) increases. In both figures, the differences in the number of actionable bases are represented by the areas between the curves, below the C_D.

Figure 4C illustrates the combined impacts of changing on-target rates, fold-80 scores, and mean coverage. Each colored curve represents a different fold-80, and the width of the curve represents the percentage of actionable bases recovered when on-target rates are between 80% (bottom limit of each curve) and 100% (top). In each curve, when C_M is 30x, improving on-target rate from 80% to 100% — essentially eliminating all off-target sequencing — increases the fraction of actionable bases by 1–2%. In contrast, improving fold-80 from 1.7 to 1.4 increases this number more dramatically, by 5–6%.

The data demonstrate that improvements to fold-80 scores (uniformity) have a much more significant impact on the efficiency of targeted NGS than do improvements to on-target rates, even if the off-target rate could be reduced to zero.

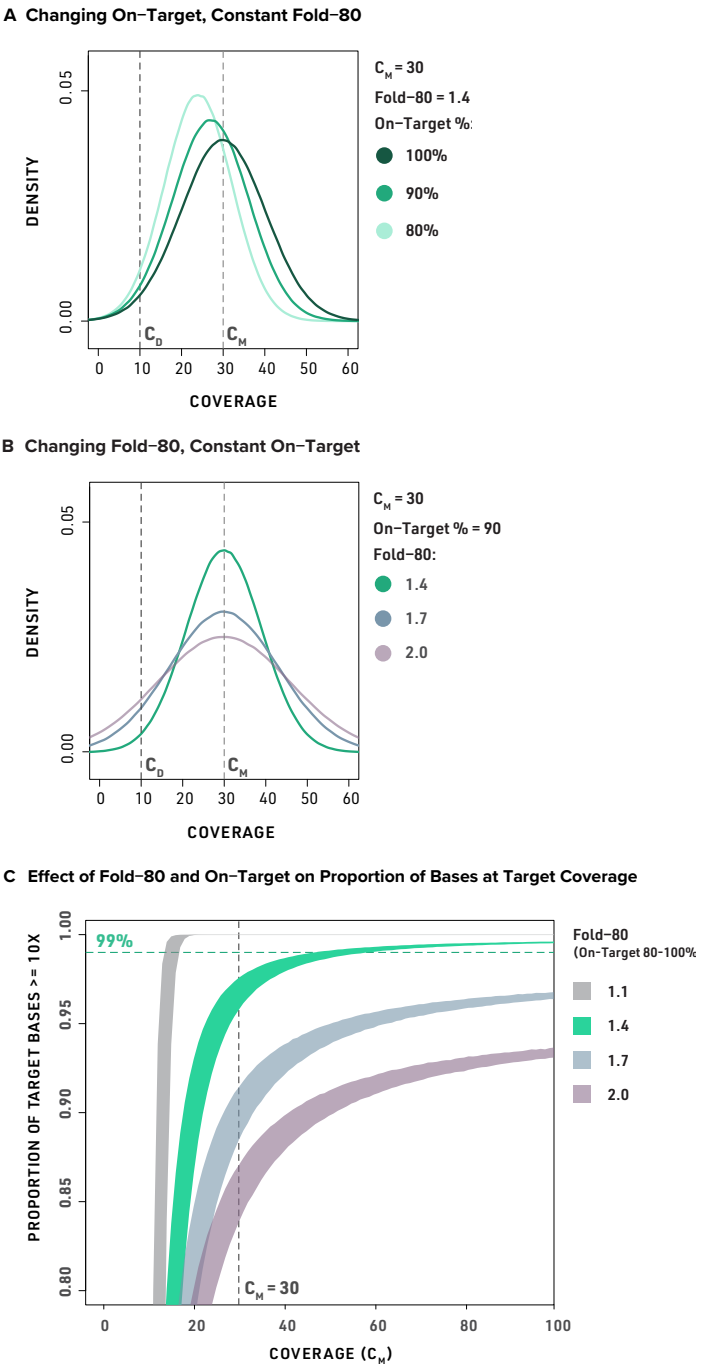

Figure 4. Effect of uniformity versus on-target rate on required depth of sequencing. Simulation results assuming desired coverage (C_D) = 10x, normal distribution of coverage depth, and varying mean coverage (C_M), on-target rate (0.8–1.0) and fold-80 (1.1–2.0). **A.** Simulated depth of coverage distributions with changing on-target rates but constant fold-80 and C_M (1.4 and 30, respectively). Improvements in on-target rate increase the mean coverage, shifting the distribution to the right. **B.** Simulated depth of coverage distributions with changing fold-80 and constant on-target rate and C_M (0.9 and 30 respectively). Improving (reducing) fold-80 scores reduces coverage of over-sequenced targets and increases coverage of under-sequenced targets. **C.** Proportion of target bases covered at 10x or higher for changing on-target rates, fold-80 scores, and mean coverage.

[2] *Normal distributions were used for an intuitive illustration of the concepts of on-target rate and uniformity. Though actual coverage distributions do not usually follow a normal distribution, the general conclusions of our analysis extend to the distributions typically observed in NGS (exact numerical values may be different).*

CONCLUSIONS AND PERSPECTIVES

In targeted NGS, uniformity (fold-80) and on-target rate are important metrics for evaluating efficiency of the sequencing effort. These two metrics are largely intrinsic properties of the probe panels themselves, and optimizing them can reduce the amount of sequencing needed to obtain high-confidence data.

Choosing the most efficient target enrichment system requires carefully weighing the actual range of uniformity and on-target rate offered. While on-target rate is important, we demonstrate here that improvements to fold-80 scores (uniformity) have a much more significant impact on the efficiency of targeted NGS.

REFERENCES

Chilamakuri CSR, Lorenz S, Madoui M-A, Vodák D, Sun J, Hovig E, Myklebost O, Meza-Zepeda LA (2014) Performance comparison of four exome capture systems for deep sequencing. BMC Genomics 15(1): 449.

Dillon OJ, Lunke S, Stark Z, Yeung A, Thorne N, Gaff C, White SM, Tan TY (2018) Exome sequencing has higher diagnostic yield compared to simulated disease-specific panels in children with suspected monogenic disorders. Eur J Hum Genet 26(5): 644–651.

Goldfeder RL, Priest JR, Zook JM, Grove ME, Waggott D, Wheeler MT, Salit M, Ashley EA (2016) Medical implications of technical accuracy in genome sequencing. Genome Med. 8(1): 24.

Meynert AM, Bicknell LS, Hurles ME, Jackson AP, Taylor MS (2013) Quantifying single nucleotide variant detection sensitivity in exome sequencing. BMC Bioinformatics. 14: 195.

Meynert AM, Ansari M, FitzPatrick DR, Taylor MS (2014) Variant detection sensitivity and biases in whole genome and exome sequencing. BMC Bioinformatics. 15: 247.

Warr A, Robert C, Hume D, Archibald A, Deeb N, Watson M. (2015) Exome Sequencing: current and future perspectives. G3: *Genes|Genomes|Genetics*. 5(8):1543–1550.