Optimizing Coverage for Targeted Resequencing

How to determine the total amount of sequence data needed to achieve the desired level of coverage in a typical enrichment study.

Introduction

The Illumina TruSeq™ Exome Enrichment technology provides the simplest, most cost-effective targeted resequencing solution available with integrated DNA sample preparation and pre-enrichment sample pooling. To maximize the efficiency of targeted resequencing studies and ensure that sufficient coverage is obtained for highly sensitive variant calling, three key factors should be taken into account:

1. **Sum length of targeted regions**
2. **Enrichment efficiency (percentage of reads passing filter and mapping to targeted regions)**
3. **Distribution of coverage depth for targeted regions**

This document discusses these key parameters in detail and provides a method for precalculating the amount of sequencing and mean coverage required to fully optimize any targeted sequencing study.

Sum Length of Targeted Regions

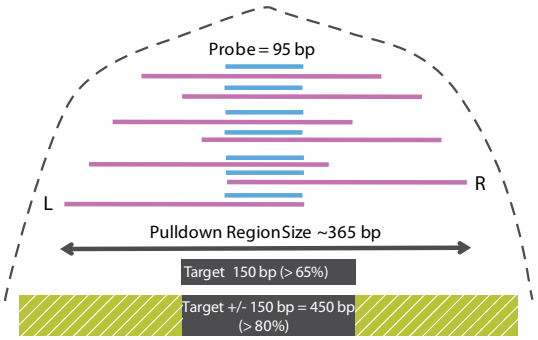
The sum length of targeted regions is equal to the total amount of genomic sequence (bp) targeted in the enrichment assay. For example, with the TruSeq Exome Enrichment Kit, the total amount of targeted sequence is 62 Mb, including 5' UTR, coding exons, 3' UTR, microRNAs, microRNA targets, and other selected and conserved regions of interest. Each 95mer probe targets 300–400 bp libraries (insert size of 180–280 bp), enriching 265–465 bases centered symmetrically on the midpoint of the probe (Table 1).

Figure 1: Probe Footprint

When a 350 bp DNA library (mean insert size = 230 bases) is enriched with the biotinylated TruSeq probes, each probe pulls down the entire DNA library.

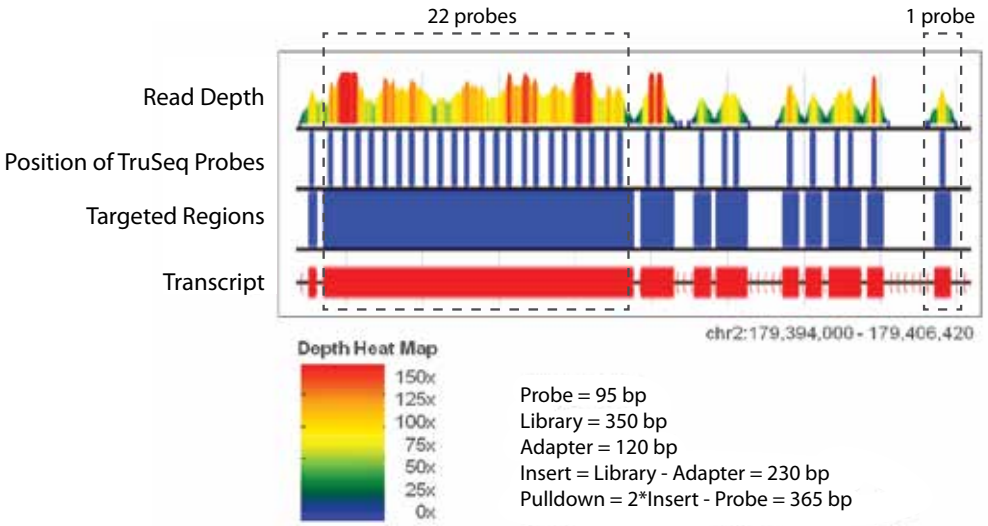
To illustrate the probe footprint, Figure 1 shows a 350 bp library with an insert of 230 bp and adapters of approximately 60 bp on either side. Each 95mer probe targets a region of interest (Figure 2). Conservatively assuming a probe requires all 95 bases to hybridize to a single 230 base insert, the leftmost position (L) of the insert would be 230 bases upstream of the right most position of the probe, and the rightmost position (R) of the insert would be 230 bases downstream of the leftmost position of the probe. The distance between the leftmost and rightmost position is the pulldown width that can be calculated (Pulldown = 2*Insert – Probe = 2*230 – 95 = 365). This is likely a conservative estimate of the pulldown width due to empirical evidence showing that up to 15 mismatches within an 80mer probe can be tolerated, meaning that the effective pulldown region may be larger.

Figure 2: Pulldown Region Size



The dotted line represents the cumulative depth of enriched inserts by position. Greater than 65% of passing filter reads that map to the reference will overlap the targets. Greater than 80% of passing filter reads will map to the reference within 150 bases of these targets. The pulldown region of 365 bases of sequence is symmetrically centered on the midpoint of the probe.

Figure 3: 95mer Probes are Optimally Spaced to Enrich Targets



The positions of the 95mer probes (thin blue lines) are evenly spaced across the targeted regions (blue blocks). The colorized read depth histogram demonstrates that the exons from this transcript subselection are uniformly covered and without gaps.

Probe Design

Each 95mer probe is designed against a TruSeq exome target by balancing GC composition with specificity. Figure 3 shows a subset of a transcript (red blocks) and illustrates the optimized interval distance between each probe (thin blue lines). The top track displays a depth histogram with colors corresponding to the 'Depth Heat Map' (linear distribution of depth ranging from 0 to 150×). On the far right, a single probe effectively pulls down a symmetrical distribution of reads that is centered around its target, with a width equal to 365 bases. The largest exon target (on the left) is enriched with 22 evenly spaced 95mer probes and exhibits uniformity of coverage with a depth ranging from 50×–150×. No gaps in coverage exist across these targeted exons.

Enrichment Efficiency

The fraction of reads 'on target' is indicative of enrichment efficiency. Calculating enrichment efficiency, or percent enrichment, is achieved by dividing the number of passing filter (PF) reads mapping to the targets by the total number of PF reads mapping to the reference per sample.

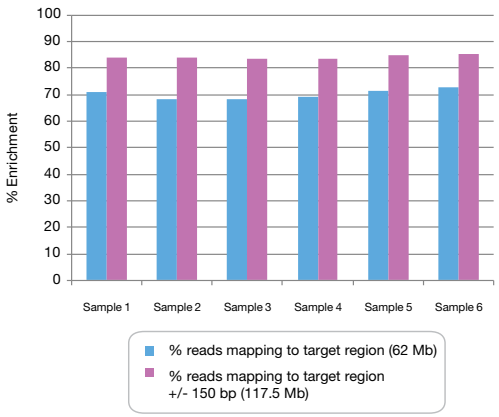
$$\text{Enrichment efficiency} = \frac{\text{Number PF reads mapping to targets}}{\text{Total number of PF reads mapping to the reference}}$$

An example TruSeq exome enrichment experiment is shown in Figure 4. For the six samples, an enrichment efficiency of ~70% was achieved when considering only the reads mapping to the 62 Mb targeted region. Due to the nature of TruSeq Enrichment technology, the padded regions up- and down-stream of the actual probe-targeted regions are also captured. An increase from > 65% enrichment (blue bars) to > 80% enrichment (purple bars) is observed when reads are expanded to include those mapping to the targeted regions +/- 150 bp (117.5 Mb). Therefore, a researcher can maximize the amount of DNA captured with the fewest number of probes, while also capturing reads adjacent to any given target, accessing biological information from regions that may be difficult to capture using competing technologies.

Distribution of Coverage Depth for Targeted Regions

Determining the distribution of coverage depth for targeted regions requires the generation of normalized coverage plots. Simply calculating the mean sequencing coverage will provide only a summary of the average read depth across the bases targeted in the enriched sample. The most commonly used methods report a given percentage of targeted bases covered at a particular depth (e.g., 90% of targeted bases covered at 10× read depth). It is possible to increase the total

Figure 4: High Target Specificity



Six pooled samples were analyzed on the HiSeq™ 2000 to demonstrate the specificity obtained in an optimized TruSeq exome enrichment experiment. The percent enrichment (y-axis) shows a high proportion of total reads mapping to the target regions (blue bars). The target region of +/- 150 bp represents percentage of total reads within 150 bases of the defined target regions (purple bars).

Figure 5: Creation of a TruSeq Normalized Mean Distribution Plot

A

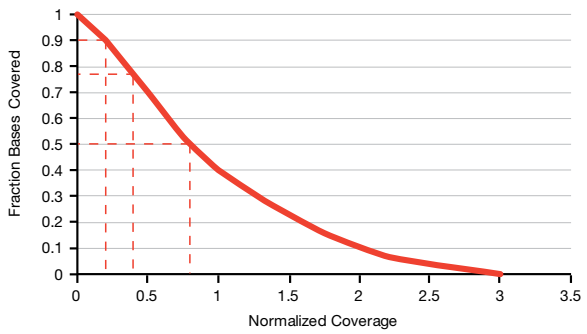
Base Position	Coverage	Normalized Coverage (Coverage divided by the Average)
1	2	0.17
2	3	0.26
3	5	0.43
4	6	0.52
5	10	0.86
6	10	0.86
7	15	1.29
8	20	1.72
9	20	1.72
10	25	2.16

11.6 = average coverage over all 10 bases

B

Normalized Coverage	Number of Bases at Given Normalized Coverage	Fraction of Bases at Given Normalized Coverage
≥ 0	10	1.0
≥ 0.2	9	0.9
≥ 0.5	7	0.7
≥ 1.0	4	0.4
≥ 2.0	1	0.1
≥ 3.0	0	0

C



D

% Bases Covered	Mean Normalized Coverage	Mean Sequencing Coverage Required (at 10× desired coverage)
90	0.2	50×
75	0.4	25×
50	0.8	12.5×

TruSeq exome analysis scripts generate normalized coverage plots as part of secondary data analysis, enabling researchers to calculate the mean sequencing coverage, i.e. how much sequencing will be required to yield a given percentage of targeted bases at a particular read depth.

sequencing coverage by over-sequencing the sample, however, this is a highly inefficient approach.

TruSeq exome analysis scripts generate mean normalized coverage plots, showing the distribution of coverage depth across all targeted bases. This enables researchers to calculate just how much sequencing will be required to yield a given percentage of targeted bases at a particular read depth. It is important to note that this normalization is a characteristic property of the Illumina TruSeq Exome Enrichment assay, averaged over a number of runs. TruSeq analysis scripts have been designed to report the distribution of depth of coverage in both an absolute and normalized manner (e.g., 90% of targeted bases covered at 0.2× of the normalized mean).

Creation of Cumulative Normalized Coverage Plots

TruSeqExome1.1 scripts generate normalized coverage plots as part of secondary data analysis (also available for download from https://icom.illumina.com/). The plots are constructed by first determining the coverage of each individual base and normalizing the coverage over all the bases.

To calculate normalized coverage, the coverage is divided by the average coverage over all 10 base positions (e.g., 11.6) (Figure 5A). Normalized coverage is then grouped into ranges (0 to ≥ 3.0), as denoted by the shading in Figures 5A and 5B, and the total number of bases that fall in those ranges are counted. Then the number of bases at a given normalized coverage is divided by the fraction of the total number of base positions (e.g., 10) to obtain the fraction of bases at a given normalized coverage (Figure 5B). A normalized mean coverage plot is constructed by determining how many bases have at least (≥) a given normalized coverage (Figure 5C).

Using the cumulative normalized coverage plot (Figure 5C), the mean normalized coverage (Figure 5D) is determined by finding the desired fraction of targeted bases on the y-axis (e.g., 0.9) and drawing a line horizontally over to the plotted line and down the x-axis to find the corresponding mean normalized coverage (e.g., 0.2). This means that 90% of bases are covered at 0.2× of the mean coverage.

The mean sequencing coverage required (Figure 5D) is calculated by dividing the desired coverage by the mean normalized coverage. If for example, 10× coverage of 90% of the bases is desired, simply divide the desired coverage by the mean normalized coverage obtained from the normailzed coverage plot (e.g., 0.2) (Step 1 in the following example).

Once you have determined the desired percentage of bases covered at a particular depth and obtained the mean normalized coverage from the normalized coverage plot, you can calculate the required total amount of sequence data per sample by simply factoring in the total amount of bases targeted in the assay and the enrichment efficiency (Step 2 in the following example).

Calculating Required Amount of Sequencing Data

Using the following steps* we can determine the amount of sequencing required to obtain a specified amount of coverage for a certain fraction of bases.

Example:

Step 1

The following equation can be used to calculate the mean sequencing coverage required:

$$\text{Mean sequencing coverage required} = \frac{\text{Desired coverage}}{\text{Mean normalized coverage}}$$

For the purpose of this example, we can calculate mean sequencing coverage for 90% of bases covered at 10× coverage. The mean normalized coverage for 90% of bases covered is 0.2 (Figure 4D).

$$\text{Mean sequencing coverage} = \frac{10\times}{0.2} = 50\times$$

Step 2

The total amount of sequencing can be calculated by using the following equation.

$$\frac{\text{(total targeted bases) (mean sequencing coverage)}}{\text{(enrichment efficiency)}} = \text{Required amount of PF mapped sequence}$$

The TruSeq Exome Enrichment Kit offers the most comprehensive exome coverage, with the total amount of targeted sequence equal to 62 Mb. For this example, the enrichment efficiency will be conservatively set at 0.65 (Figure 5), where the experiment yields ~68–72% enrichment mapping to the target regions.

Now the total amount of sequence can be calculated using our equation and the following givens:

Total targeted bases = 62 Mb
Mean sequencing coverage = 50×
Enrichment efficiency = 0.65

Inputting these into the equation, we calculate that 4.8 Gb of sequencing data will be required to obtain 10× coverage of 90% of the targeted bases.

$$\frac{\text{(62 Mb) (50×)}}{0.65} = \text{4.8 Gb of PF and mapped sequence}$$

Steps 1 and 2 calculate the amount of PF mapped sequence required to cover a specified percentage (90%) of the 62 million targeted bases to a specified depth of coverage (10×). It is further possible to calculate the amount of PF sequence required by factoring in the percentage of reads passing filter that map to the reference genome (~90%). [(4.8 Gb)(1/.90) = 5.3 Gb PF sequence]

Illumina, Inc. • 9885 Towne Centre Drive, San Diego, CA 92121 USA • 1.800.809.4566 toll-free • 1.858.202.4566 tel • techsupport@illumina.com • illumina.com

FOR RESEARCH USE ONLY

illumina®