

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**

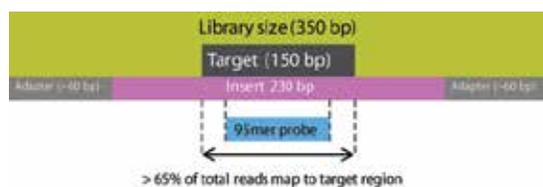
This document discusses these key parameters in detail and provides a method for precalculating the amount of sequencing and mean coverage required to fully optimize any targeted sequencing study.

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).

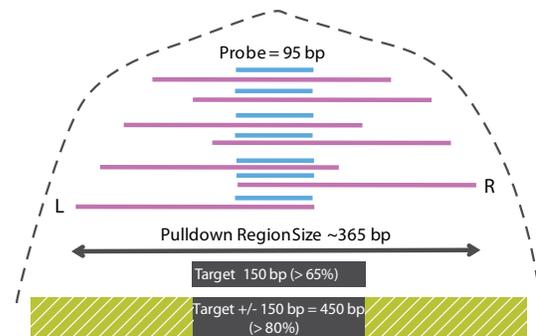
To illustrate the probe footprint, Figure 1 shows a 350 bp library with an insert of 230 bp and adapters of approximately 60 bp on either side. Each 95mer probe targets a region of interest (Figure 2). Conservatively assuming a probe requires all 95 bases to hybridize to a single 230 base insert, the 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 ($\text{Pulldown} = 2 * \text{Insert} - \text{Probe} = 2 * 230 - 95 = 365$). 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 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.

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.

