Estimating Sequencing Coverage

Before starting a sequencing experiment, you should know the depth of sequencing you want to achieve. This Technical Note helps you estimate that coverage.

Next-generation shotgun sequencing approaches require sequencing every base in a sample several times for two reasons:

- **You need multiple observations per base to come to a reliable base call.**
- **Reads are not distributed evenly over an entire genome, simply because the reads will sample the genome in a random and independent manner[1,2]. Therefore many bases will be covered by fewer reads than the average coverage, while other bases will be covered by more reads than average. You need to account for this in your planning.**

This is expressed by the coverage metric, which is the number of times a genome has been sequenced (the depth of sequencing). For applications where you aim to sequence only a defined subset of an entire genome, like targeted resequencing or RNA sequencing, coverage means the amount of times you sequence that subset. For example, for targeted resequencing, coverage means the number of times the targeted subset of the genome is sequenced.

This Technical Note provides information on how to calculate the coverage required for an experiment, and how to estimate the number of flow cells or lanes you need to use.

Coverage Requirements Depend on Application

Illumina does not have an official recommendation for sequencing coverage level.

Most users determine the necessary coverage level based on the type of study, gene expression level, size of reference genome, published literature, and best practices defined by the scientific community. For example, the level of coverage for human genome mutations/SNPs/rearrangements detection that most publications require is from 10× to 30× depth of coverage depending on the application and statistical model. For ChIP-Seq studies where reads map to only a subset of a genome, often the researchers/publications require coverage around 100×.

For RNA sequencing, determining coverage is complicated by the fact that different transcripts are expressed at different levels. This means that more reads will be captured from highly expressed genes, and few reads will be captured from genes expressed at low levels. When planning RNA sequencing experiments, researchers usually think in terms of numbers of millions of reads to be sampled. The number of reads required will depend on how sensitive the experiment needs to be for genes expressed at low levels. Detecting rarely expressed genes might require an increase in the depth of coverage.

Standards are Set by Field and Journals

The standards are ultimately set by journals and the scientific field you are in. The "Recent Publications" section on Illumina's website (http://www.illumina.com/publications/overview.ilmn) provides a resource for users to search publications for Whole Genome Resequencing, De Novo Sequencing, Targeted Resequencing, Transcriptomics and many other fields. This is recommended as a starting point for determining the target depth of coverage for a particular study. Another good resource for RNA Sequencing is provided by the ENCODE project:

http://genome.ucsc.edu/ENCODE/protocols/dataStandards/ENCODE_RNAseq_Standards_V1.0.pdf

Estimating Sequencing Runs

Coverage Equation

The Lander/Waterman equation is a method for computing coverage[1]. The general equation is:

$$C = LN / G$$

- **C stands for coverage**
- **G is the haploid genome length**
- **L is the read length**
- **N is the number of reads**

So, if we take one lane of single read human sequence with v3 chemistry, we get

$$C = (100 \text{ bp}) * (189 \times 10^6) / (3 \times 10^9 \text{ bp}) = 6.3$$

This tells us that each base in the genome will be sequenced between six and seven times on average.

Coverage Calculator

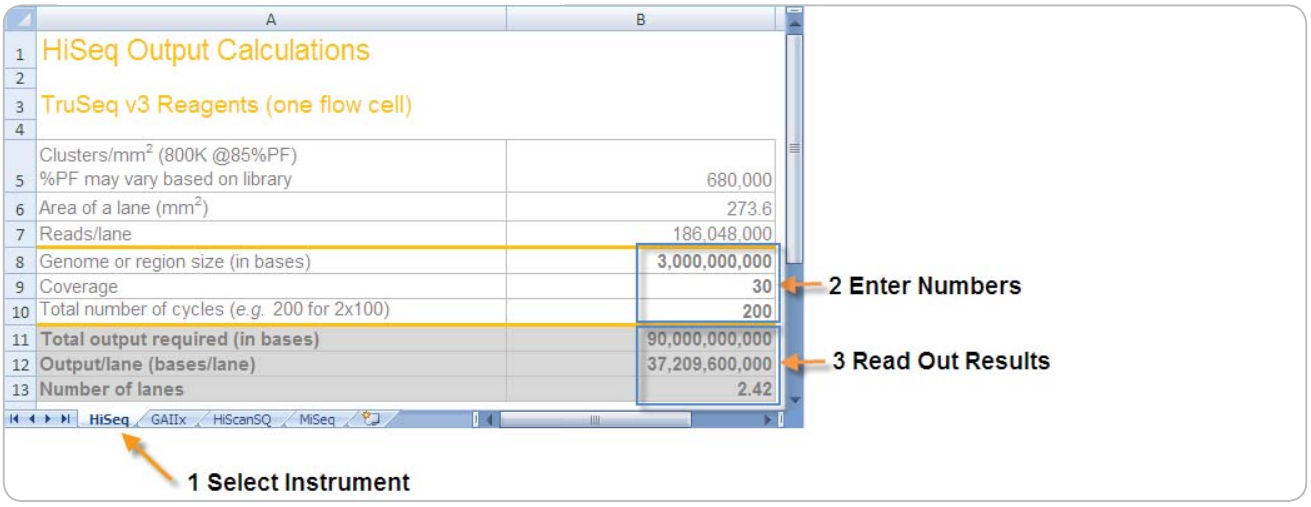
Illumina provides an Excel spreadsheet that calculates the reagents and sequencing runs needed to arrive at the desired coverage for your experiment, based on the Lander/Waterman equation. The spreadsheet can be found here:

http://www.illumina.com/CoverageCalculator

Perform the following steps to run the calculator:

1. **Click on the tab to choose your instrument (HiSeq/GAIIx/HiScanSQ/MiSeq).**
2. **Enter numbers:**

- **Target genome or region size, for example, input 3000000000 (3 Gb) for human genome.**
- **Coverage you want.**
- **Total number of cycles. For example, if you want to perform**



100 bp paired-end runs (2×100), enter 200.

3. **Read out the total output required, output per lane, and number of lanes you need to use for the desired coverage.**

For example, if you want to perform a human genome resequencing experiment on a HiSeq 2000 with TruSeq™ SBS v3 reagents at 30× coverage with 2×100 bp reads, your result would look like Figure 1.

Note that the calculator uses an estimate of reads passing filter commonly found for balanced genomes (such as PhiX or the human genome). If you plan to sequence an unbalanced genome, you may have a lower number of reads passing filter, and consequently a lower output per lane.

When to Sequence More

In Illumina sequencing experiments, it is very easy to increase the coverage or sequence depth, if you later decide you need more data. Provided you still have your original sample, you can just sequence more, and combine the sequencing output from different flow cells. There are a number of reasons to sequence more than the originally estimated coverage, these include:

- **The effects you see are not statistically significant. Sequencing more reads will generally increase the power of your assay.**
- **You are investigating events that are very rare. For example, you may want to look at transcripts that are expressed at a very low level in RNA Sequencing, or look at very low binding activities in ChIP Sequencing.**
- **Certain journals or fields may require a higher level of coverage for your particular application.**
- **Certain genomes may need more sequencing. For example, certain regions may be hard to sequence requiring more coverage, or the genome may be polyploid.**

References

1. Lander ES, Waterman MS.(1988) Genomic mapping by fingerprinting random clones: a mathematical analysis, Genomics 2(3): 231-239.
2. Estimating the number of times a base is expected to be sequenced. Lander and Waterman made two assumptions about the sequencing:
   • Reads will be distributed randomly across the genome
   • Overlap detection doesn't vary between reads.
   Based upon these two assumptions, they reached the conclusion that the number of times a base is sequenced follows a Poisson distribution. The Poisson distribution can be used to model any discrete occurrence given an average number of occurrences.The probability function is the following:
   $P(Y=y) = (C^y \times e^{-C})/y!$
   • y is the number of times a base is read
   • C stands for coverage
   We can use the Poisson distribution to compute the probability of a base being sequenced a certain number of times. We can use the coverage as the average number of occurrences and y as the exact number of times a base is sequenced, and then compute the probability that would happen:
   $P(Y=3) = (6.3^3 \times e^{-6.3})/3! = 0.077$
   Of course, this is the value for exactly 3. It probably is more interesting to see the probability the base is sequenced 3 times or less, as most SNP callers require at least four calls at a base position to call SNPs. We can determine this probability simply by summing up the probabilities for Y=2, Y=1, and Y=0:
   $P(Y<=3) = P(Y=3) + P(Y=2) + P(Y=1) + P(Y=0) =$
   $0.077 + 0.036 + 0.012 + 0.002 = 0.127$
   So we see that about 12.7% of the bases in the genome will be covered by three or fewer reads, and we will probably want to increase our coverage for this experiment.The same formula can be used in a couple other ways. For example, by simply computing the Y=0 probability, we can estimate the percentage of a genome not yet sequenced: in our example above 0.2% of the genome was not sequenced at all. By multiplying 0.2% by the genome size, we see that we would have a total gap length of about 6,000,000 bp. We can also estimate the number of gaps by multiplying the number of reads used by the percentage of the genome not covered: 0.2% * 189,000,000 gives 378,000 gaps in the 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

Pub. No.770-2011-022 Current as of 17 October 2011

illumina®