illumina®

# Sequencing Coverage Calculation Methods for Human Whole-Genome Sequencing

An overview of Illumina coverage calculation methods using BaseSpace® or third-party analysis tools.

## What is Sequencing Coverage?

While it is easy to assume that whole-genome sequencing (WGS) involves sequencing each base in a genome one time, most applications require sequencing each base multiple times to achieve high confidence base calls. Sequencing coverage (or sequencing depth), at the nucleotide level, refers to the number of times a reference base is represented within a set of sequencing reads. When describing sequencing coverage at the whole genome level, this value is expressed as an average or median of all the per base coverage values. For example, a genome with 30× coverage will have an average of 30 reads spanning any given position within the genome—some regions will have higher coverage and some will have lower coverage. Under certain assumptions, shotgun sequencing coverage follows a Poisson distribution. Therefore, if the average coverage is 30×, the data would be expected to fall to 15× or below about 0.2% of the time.[1]

## Sequencing Coverage Level for Human WGS

Sequencing at increased levels of coverage enables the generation of a high quality, highly accurate consensus sequence, providing confidence in the results and allowing accurate detection of variants or assembly of complex genomes. Human WGS studies are typically performed at 30×–40× coverage or higher. Depending on the study design and final application, more or less coverage may be required. For example, certain applications, such as tumor-normal analysis or rare variant detection, may require higher coverage levels to identify low-frequency variants. For more information on coverage level calculations for specific applications, see the "Estimating sequencing coverage" technical note.[1]

## Alignment Tools and Coverage Level Results

Various aligners and filter settings will produce slightly different mean coverage values. These differences stem from the way aligners tolerate sequencing errors, trim sequences, align chimeric reads, and handle overlapping bases within read pairs. Furthermore, the algorithms employed may lead to different analyses of all possible read matches to the reference genome. The Burrows-Wheeler Aligner[2] (BWA) covers more of the search space and generally leads to a higher alignment rate at the expense of time. In contrast, k-mer matching algorithms, such as those used by the Isaac™ aligner, are faster, albeit at the expense of memory usage.[3,4]

## How Does Illumina Calculate Human WGS Coverage?

Illumina defines sequencing coverage as "the average coverage of unique reads across the non-N portion of the human genome."* Researchers can perform coverage calculations equivalent to Illumina calculation methods using BWA WGA App v1.0[5] or Isaac WGS App v2.0.[6] Both apps are freely available on BaseSpace.[7] These workflows employ the data filtration options outlined in Table 1. When calculating mean coverage, Illumina does not filter reads based on mapping quality or base quality following the alignment stage of analysis.

* The non-N portion of the PAR-masked human genome spans 2,858,674,665 bases (from hg19).[8]

**Table 1: Data Excluded with the BWA WGA App v1.0 and the Isaac WGA App v2.0 [a]**

| | |
|---|---|
| **N-Calls** | Bases in the read or reference genome that could not be resolved to one of the four DNA nucleotides (A, T, C, or G). |
| **Duplicate Reads** | Paired end reads that have the exact same alignment starting positions for Read 1 and Read 2 as another set of paired end reads. |
| **Soft-Clipped Bases** | Bases masked by the aligner in order to map reads with higher confidence, such as adapter sequence. |
| **Overlapping Bases [b,c]** | In a read pair, reference bases covered by both Read 1 and Read 2. |

a. Data filtration options are subject to change in future software versions.
b. The Isaac WGS App v1.0 excludes overlapping bases from only one read (ie, overlapping bases are counted one time).
c. The BWA WGS App does not exclude overlapping bases from either read (ie, overlapping bases are counted twice).

## Coverage Calculation Using Third-Party Tools

All Illumina sequencing instruments produce raw sequence data in *.bcl file format. Data in the *.bcl format can be converted to FASTQ format using the free, Illumina BCL2FASTQ software.[9] FASTQ files are compatible with most third-party software tools. BWA, a popular alignment tool, can be used to produce alignments in *.bam format, which can then be analyzed with Picard,[10] a commonly used analysis suite. The Picard algorithm CollectWgsMetrics[11] reports the following metrics, which can be used to arrive at the Illumina definition of coverage, using the calculation below:

$$\frac{(\text{MEAN\_COVERAGE}) \times (1 - \text{PCT\_EXC\_DUPE} - \text{PCT\_EXC\_OVERLAP})}{(1 - \text{PCT\_EXC\_TOTAL})}$$

These metrics are defined on the Picard Metrics Definitions website.[12]

## Summary

To account for specific application requirements as well as typical shotgun sequencing characteristics, such as PCR bias and non-uniform read distribution, the genome or target sequence must be sequenced to the appropriate coverage level. Different aligners and filter settings will produce slight variations in mean coverage values. Coverage depth for human WGS can be calculated with the Illumina BWA WGA 1.0 or Issac WGA v2.0 apps or using common third-party analysis tools.

## References


1. Illumina (2011) Estimating sequencing coverage tech note. (www.illumina.com/documents/products/technotes/technote_coverage_calculation.pdf).
2. Li H and Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009;25(14): 1754-60.
3. Raczy C, Petrovski R, Saunders CT, et al. Isaac: ultra-fast whole-genome secondary analysis on Illumina sequencing platforms. *Bioinformatics*. 2013;29(16): 2041-2043.
4. Illumina (2013) Isaac genome alignment and Isaac variant caller technical note. (www.illumina.com/documents/products/whitepapers/whitepaper_isaac_workflow.pdf).
5. BWA WGA App v1.0 (basespace.illumina.com/apps/279279/BWA-Whole-Genome-Sequencing-v1-0).
6. Isaac WGS App v2.0 (basespace.illumina.com/apps/278278/Isaac-Whole-Genome-Sequencing-v2).
7. BaseSpace (basespace.illumina.com/home/sequence).
8. GRCh37, hg19, Feb. 2009 (hgdownload.cse.ucsc.edu/goldenPath/hg19/bigZips/).
9. BCL2FASTQ conversion software (support.illumina.com/downloads/bcl2fastq_conversion_software.html).
10. Picard. A set of tools for working with next-generation sequencing data in the BAM format. (broadinstitute.github.io/picard/).
11. Picard CollectWgsMetrics (broadinstitute.github.io/picard/command-line-overview.html#CollectWgsMetrics).
12. Picard Metrics Definitions (broadinstitute.github.io/picard/picard-metric-definitions.html).


**Illumina** • 1.800.809.4566 toll-free (U.S.) • +1.858.202.4566 tel • techsupport@illumina.com • www.illumina.com

FOR RESEARCH USE ONLY


Pub. No. 770-2014-042 Current as of 28 October 2014

illumina®