illumina®

HiSeq Analysis Software v2.0 Whole-Genome Sequencing Data Analysis Pipelines

Fast, high-quality analysis pipelines for HiSeq X™ Series.

Introduction

HiSeq Analysis Software v2.0 meets the speed and scalability requirements of the HiSeq X Series with fast and accurate analysis of a full range of variant types, including single nucleotide variants (SNVs), insertion-deletions (indels), structural variants (SVs), and copy number variants (CNVs). The software supports 2 pipelines for human whole-genome sequencing (WGS):

- **WGS Analysis Pipeline** (Figure 1)
- **Tumor Normal Analysis Pipeline** (Figure 2)

The pipelines leverage a suite of proven algorithms including Isaac™ Aligner,[1] Starling Small Variant Caller,[1] Manta Structural Variant Caller,[2] Canvas Copy Number Variant Caller,[3] Strelka Somatic Variant Caller,[4] and SENECA Copy Number Aberration Caller.[5] The HiSeq Analysis Software v2.0 pipelines can be easily deployed through a command-line software package on commodity hardware, lowering IT infrastructure costs. Standard output file formats include BAM, VCF, and gVCF. The pipelines also generate automated reports in PDF format displaying read statistics, coverage histograms, variant summary tables, and more (Figure 3, Figure 4).

This white paper describes a validation experiment comparing the performance of the HiSeq Analysis Software v2.0 WGS Pipeline to trusted tools such as the Burrows-Wheeler Aligner (BWA)[6] and the Genome Analysis Toolkit v3.0 (GATK).[7] Additional analysis capabilities offered by the HiSeq Analysis Software v2.0 pipelines that are not currently available with other tools are also described.

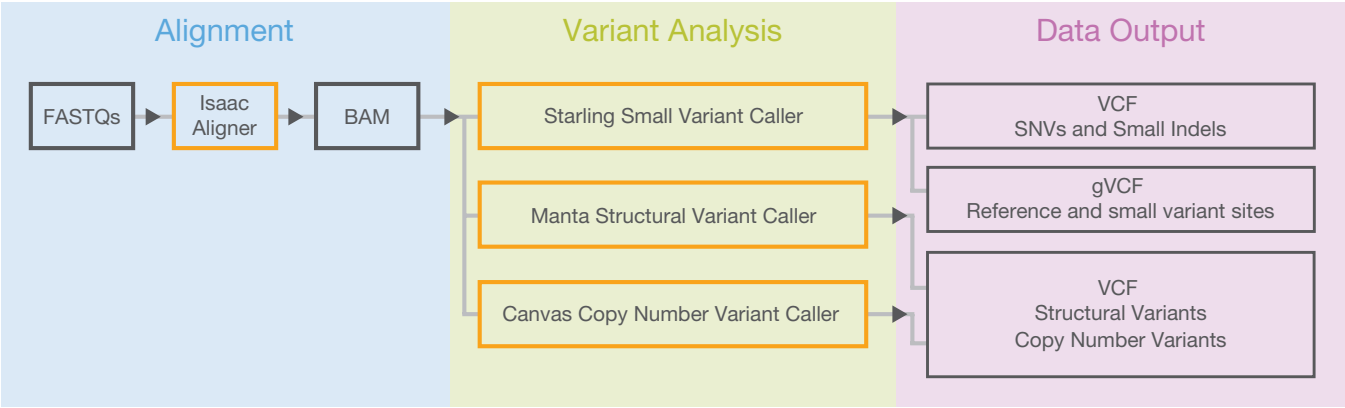


Figure 1: HiSeq Analysis Software v2.0 WGS Analysis Pipeline—With the recommended hardware configuration, the HiSeq Analysis Software v2.0 WGS Analysis Pipeline can analyze a 33× human genome in approximately 6.5 hours. Resequencing workflow (FASTQ data to variant calls) with recommended hardware configuration (Table 4). Optimized algorithms are highlighted in orange boxes.

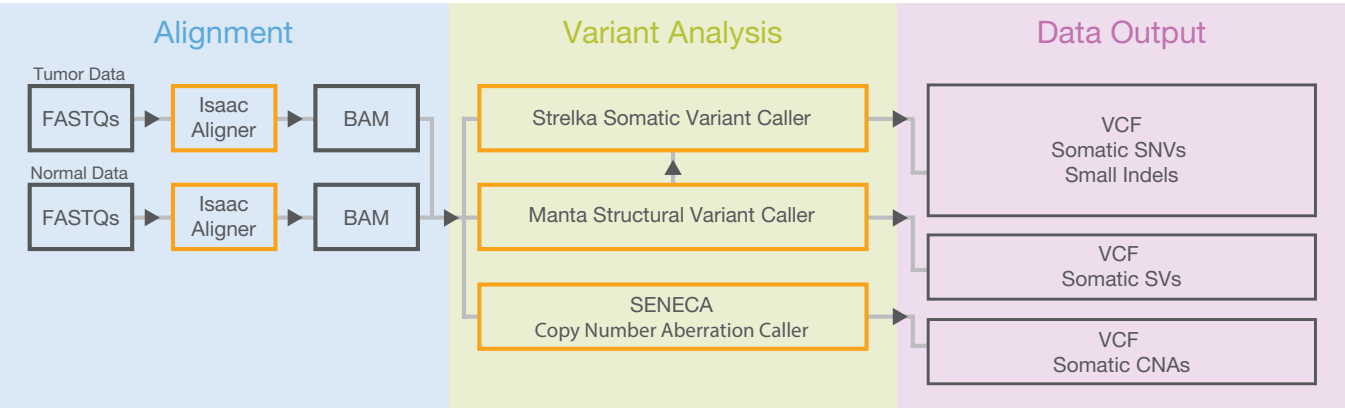


Figure 2: HiSeq Analysis Software v2.0 Tumor Normal Analysis Pipeline—The HiSeq Analysis Software v2.0 Tumor Normal Analysis Pipeline can perform somatic analysis in approximately 16.6 hours with the recommended hardware configuration (Table 4). Optimized algorithms are highlighted in orange boxes.

Small Variants Summary

	SNVs	Insertions	Deletions
Total Passing	3,564,342	280,773	292,298
Percent Found in dbSNP	99.18%	80.29%	82.87%
Het/Hom Ratio	1.62	1.91	2.21
Ts/Tv Ratio	2.08	-	-

Structural Variants Summary

SV Type	Total	Number in Genes
CNV	89	54
Insertion	686	343
Tandem Duplication	139	84
Deletion	4,309	2,290
Inversion	127	61

Coverage Histogram

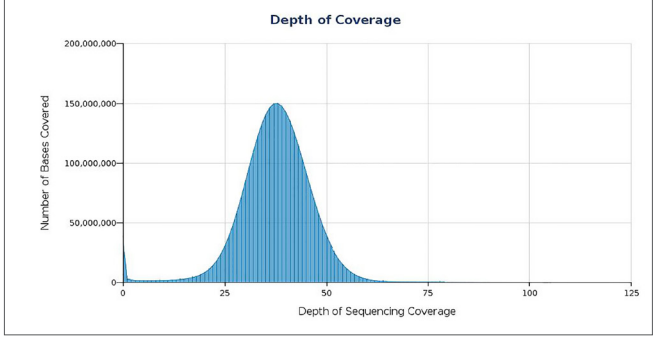


Figure 3: Resequencing Report—Variant tables and a coverage histogram from the HiSeq Analysis Software v2.0 WGS Analysis Pipeline are displayed to assess the quality of each genome analyzed.

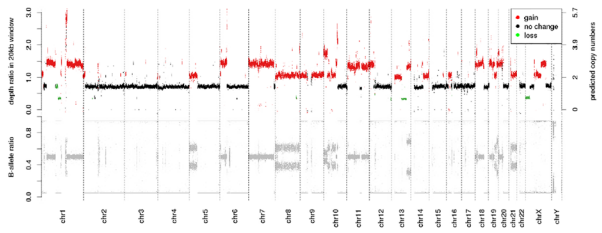
Somatic Small Variants Summary

	SNVs	Deletions	Insertions
Total	15,345	551	421
Number in Genes	6,785	246	200
Number in Exons	284	13	10
Number in Coding Regions	169	8	4
Splice Site Region	16	0	1
Stop Gained	8	0	0
Stop Lost	0	0	0
Frameshift	0	6	4
Non-synonymous	130	2	0
Synonymous	31	0	0
Mature miRNA	0	0	0
UTR Region	115	5	6
dbSNP	1,608	16	41

Somatic Structural Variants Summary

SV Type	Total	Number in Genes
CNV	169	162
Deletion	156	109
Tandem Duplication	93	85
Insertion	0	0
Inversion	19	16
Translocation Breakend	38	19

Depth/B-Allele Plot



Top: Ratio of the tumor read depth to the normal read depth after normalizing for sequencing coverage. Each point represents a 20 kb genomic region. Points are classified as either copy number gain (**red**), copy number loss (**green**), or copy number unchanged (**black**).

Bottom: Variant allele frequencies in the tumor sample at dbSNP positions where the normal sample is heterozygous.

Figure 4: Somatic Analysis Report—Tables and graphics from HiSeq Analysis Software v2.0 Tumor Normal Analysis Pipeline summarizing somatic variant analysis of a tumor-normal sample pair.

Materials and Methods

Library Preparation and Sequencing

Sequencing libraries for human WGS were prepared from sample NA12878 (Coriell Institute for Medical Research) using the TruSeq® DNA PCR Free Library Preparation Kit (Illumina, Catalog No. FC-121-3001). The libraries were sequenced to 40× on a HiSeq System, using paired end 2 × 100 read length.

For paired tumor-normal analysis, sequencing libraries were prepared from the cell line HCC2218 (University of Texas Southwestern Medical Center). For large somatic variants, paired tumor-normal samples were prepared from HCC1187 (University of Texas Southwestern Medical Center). The libraries were prepared using the TruSeq DNA PCR Free Library Preparation Kit and sequenced on a HiSeq System, using paired end 2 × 100 read length. The paired tumor-normal samples were sequenced to 80× and 40× respectively.

Whole-Genome Sequencing Pipelines

WGS data were analyzed with the **HiSeq Analysis Software v2.0 WGS Analysis Pipeline**, which includes alignment with the Isaac Aligner and identification of SNVs and small indel calls with Starling Small Variant Caller. For large variant calling, the Canvas CNV Caller predicted copy number deletions and duplications, while deletions, insertions, and tandem duplications were predicted with the Manta Structural Variant Caller. For quality comparison, WGS data was also analyzed with the BWA/GATK v3.0 pipeline.

Tumor-Normal Sequencing Pipelines

Tumor-normal sequence data from cell line HCC2218 (or cell line HCC1187 for large somatic variants) were analyzed with the **HiSeq Analysis Software v2.0 Tumor Normal Analysis Pipeline**, which includes alignment with the Issac Aligner and determination of somatic SNVs with Strelka Somatic Variant Caller. For large variant calling, the Canvas CNV Caller predicted copy number deletions and duplications, while deletions, insertions, and tandem duplications were predicted with the Manta Structural Variant Caller.

Calculation of Precision and Recall

Precision and recall statistics were generated by comparing the results of each pipeline to the Illumina Platinum Genomes v7.0[8] data set using haploCompare 0.7.0.[9] For tumor-normal large somatic variant data, results from the HiSeq Analysis Software v2.0 Tumor Normal Analysis Pipeline were compared to the known variants from the Catalogue of Somatic Mutations in Cancer (COSMIC)[10] and additional manually reviewed calls with statistics generated using haploCompare 0.7.0.

Comparison and Results

Variant Calls with the WGS Analysis Pipeline and BWA/GATK v3.0

The human WGS data set was analyzed with the WGS Analysis Pipeline and again with BWA/GATK v3.0. A comparison of the data shows that recall and precision for both SNVs and indels using both methods produce comparable results (Table 1).

Table 1: HiSeq Analysis Software v2.0 WGS Analysis Pipeline and BWA/GATK v3.0 Variant Call Statistics

	SNV Recall	SNV Precision	Indel Recall	Indel Precision.
WGS Analysis Pipeline	96%	> 99%	84%	98%
BWA/GATK v3.0	98%	> 99%	89%	99%

Large Variant Calls with the WGS Analysis Pipeline

Structural variants (SVs) are common genetic features and their discovery is essential for understanding both normal and disease-related development. However, large structural variants have been traditionally difficult to identify using current variant callers due to their multiallelic nature. An advantage of the HiSeq Analysis Software v2.0 pipeline is that it can call a broad range of variant types—including the traditionally challenging SVs and CNVs (Figure 1). The Manta Structural Variant Caller combines paired-end and split read evidence during structural variant discovery and scoring to improve performance. It can identify and score various structural variant types (50 bp–10 kb in length) including deletions, insertions, inversions, tandem duplications, and interchromosomal translocations. The Canvas CNV Caller scans diploid genomes for regions > 10 kb that have an unexpected number of short read alignments. To demonstrate the quality of large variant calls with the WGS pipeline, CNVs were predicted with the Canvas CNV Caller, and deletions, insertions, and tandem duplications were predicted with the Manta Structural Variant Caller (Table 2).

Table 2: HiSeq Analysis Software v2.0 WGS Analysis Pipeline Large Variant Calling Quality Statistics

Variant Type	Variant Size	Recall	Precision	Events
Deletion	51–99 bp	74%	86%	796
Deletion	100 bp–1 kb	78%	95%	1951
Deletion	1–10 kb	61%	98%	830
Copy Number Deletion	> 10 kb	77%	n/a	22
Insertion	51–99 bp	70%	82%	503
Insertion	100–199 bp	83%	86%	513
Insertion	200–299 bp	78%	94%	100
Insertion	> 300 bp	66%	88%	146
Tandem Duplication	100 bp–1 kb	90%	59%	21
Tandem Duplication	1–10 kb	67%	80%	18
Copy Number Duplication	> 10 kb	77%	n/a	13

WGS Analysis Pipeline and GATK v3.0 Speed Comparison

Data processing power and analysis speed are critical factors for achieving maximum cost efficiencies on the HiSeq X Series systems. The WGS pipeline takes only 6.5 hours to analyze a single sample, while the Tumor Normal Pipeline takes 16.6 hours with the recommended hardware for HAS 2.0 (Table 4). Compared to BWA/GATK v3.0, which takes approximately 38 hours to execute alignment and small variant calling on a single sample, HiSeq Analysis Software v2.0 provides comparable data in a significantly shorter amount of time. As a result, we find that the quality and speed of the HiSeq Analysis Software v2.0 is well suited for the production scale of the HiSeq X Series.

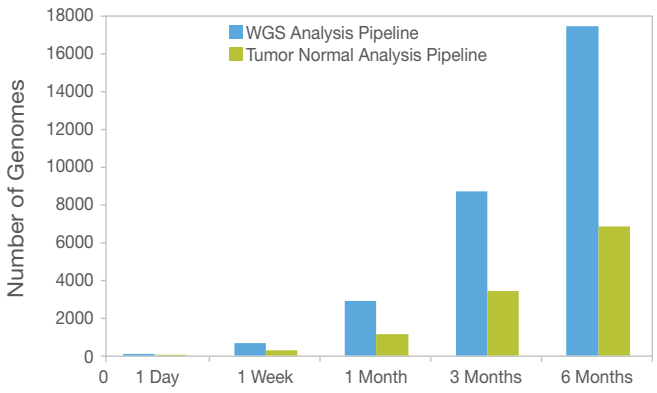


Figure 5: HiSeq Analysis Software v2.0 Analysis Speed—With the recommended hardware configuration, the WGS Analysis Pipeline can analyze over 17,000 human genomes at 30× coverage in 6 months. The Tumor Normal Analysis Pipeline can analyze over 6,000 tumor-normal paired samples at 80× and 40× respectively in 6 months.

HiSeq Analysis Software v2.0 Pipelines Provide Broad Somatic Variant Calling Capabilities

The HiSeq Analysis Software v2.0 Tumor Normal Analysis Pipeline uses the Manta Structural Variant Caller, the Strelka Somatic Variant Caller, and the SENECA Copy Number Aberration Caller to identify somatic variants. Strelka uses a unique approach with a combined rather than a subtractive method for calling somatic variants. The statistical models operate on the combined cancer and normal reads as input rather than the subtraction of variant calls.

We find that using a combined approach produces robust, and often, superior results.[4] In our testing of the HiSeq Analysis Software v2.0 pipeline, somatic SNV calls had 76% recall and 81% precision. Another feature of the Tumor Normal Analysis Pipeline includes the distribution of relevant candidate data between callers. For example, after the Strelka Somatic Variant Caller finds a set of candidate indels, the data are supplemented with candidate indels (≤ 50 bps) discovered by the Manta Structural Variant Caller, providing a more comprehensive data set.

Large Somatic Variant Calls with the Tumor Normal Analysis Pipeline

Identification of large somatic variants also presents a challenge for tumor-normal paired data. While most pipelines have the ability to call small variants, the Tumor Normal pipeline can call these challenging variants to create a more comprehensive picture of somatic variants across the genome landscape. The Tumor Normal Pipeline uses the Manta Structural Variant Caller and the SENECA Copy Number Aberration Caller to identify large somatic variants that can be difficult to call using alternative methods. To demonstrate variant call quality, precision and recall were calculated for the large somatic variants predicted with the Tumor Normal Analysis Pipeline (Table 3).

Conclusion

HiSeq Analysis Software v2.0 is a seamless, fully supported software solution that provides high-quality data comparable to the top tools currently utilized across diverse variant classes. The HiSeq Analysis Software v2.0 pipelines provide complete whole genome analysis with high-quality data for both germline and somatic variant calling, identifying even the most challenging variants with a high level of precision (up to 98%). Addressing difficult-to-call regions that likely contain biologically relevant variants, HiSeq Analysis Software v2.0 calls indels, and challenging SV and CNV calls with an average of 75% and 85% for recall and precision respectively. In addition to providing high-quality variant data across a broad range of variant classes, the HiSeq Analysis Software v2.0 pipelines are significantly faster—a critical factor for avoiding HiSeq X Series analysis bottlenecks.

Learn More

For more on the HiSeq Analysis Software v2.0, visit www.illumina.com/systems/hiseq-x-sequencing-system/software.html.

Table 3: HiSeq Analysis Software v2.0 WGS Analysis Pipeline Large Variant Calling Quality Statistics

Variant Type	Variant Size	Recall	Precision	Events
Deletion	51 bp–1 kb	77%	92%	57
Deletion	1–10 kb	79%	95%	24
Deletion	> 10 kb	76%	91%	41
Duplication	> 10 kb	99%	92%	83
Inversions	-	90%	75%	21
Translocation	-	87%	68%	30
Copy Number Aberration	10–50 kb	76%	n/a	2
Copy Number Aberration	50–500 kb	76%	n/a	44
Copy Number Aberration	> 500 kb	76%	n/a	75

Table 4: Recommended Hardware Configuration for HiSeq Analysis Software v2.0

Feature	Requirements
Speed	2.8 Ghz
Network	10 Gb network
Dual Core Processors	10
RAM Memory	128 GB 1867 MHz
Local Storage	6 TB with 500 MB
Ethernet Adaptors	10 Gbits per sec
Total Number of Servers	26

References

1. Raczy C, Petrovski R, Saunders CT, et al. Isaac: ultrafast whole-genome secondary analysis on Illumina sequencing platforms. Bioinformatics. 2013;29:2041-2043.
2. Manta Structural Variant Caller (support.illumina.com/downloads/whole_genome_sequencing_service_user_guide.html).
3. Canvas CNV Caller (support.illumina.com/downloads/whole_genome_sequencing_service_user_guide.html).
4. Saunders CT, Wong WS, Swamy S, Becq J, Murray LJ, and Cheetham RK. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. Bioinformatics. 2012;28:1811-1817.
5. SENECA Copy Number Aberration Caller (support.illumina.com/downloads/cancer_analysis_service_user_guide.html).
6. Burrows-Wheeler-Aligner (bio-bwa.sourceforge.net/).
7. Genome Analysis Toolkit (GATK) (www.broadinstitute.org/gatk/).
8. Illumina Platinum Genomes Project (www.illumina.com/platinumgenomes).
9. haploCompare 0.7.0 is an internal Illumina analysis tool.
10. Catalogue of Somatic Mutations in Cancer (COSMIC) (cancer.sanger.ac.uk/cosmic).

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

illumina®