Human Whole-Genome Sequencing with the NovaSeq™ 6000 Sequencing System

With a simple, intuitive design, the NovaSeq 6000 System delivers the highest daily throughput and exceptional data quality for human whole-genome sequencing.

Highlights

- **Scalable Output**
  Generate up to 6 Tb and 20 billion reads in dual flow cell mode with simple streamlined automated workflows
- **Ultimate Flexibility**
  Configure system to enable sequencing a trio in one day or up to 48 genomes in ~2 days
- **Exceptional Data Quality**
  Highly accurate Illumina sequencing by synthesis (SBS) chemistry delivers proven industry-leading data quality

Introduction

The NovaSeq 6000 System introduces a new era in sequencing with groundbreaking innovations to provide users with the throughput, speed, and flexibility required to complete projects faster and more economically than ever before. Combining the best features of previous Illumina platforms, the NovaSeq 6000 System incorporates additional innovations to deliver tunable output of up to 6 Tb and 20 billion reads—all in about 2 days. The NovaSeq 6000 System leverages proven Illumina sequencing by synthesis (SBS) chemistry—the most widely adopted next-generation sequencing (NGS) technology.[1] SBS chemistry delivers exceptional data accuracy, the highest yield of error-free reads, and the highest percentage of base calls above Q30 in the industry.[2,3]

This application note describes and compares human whole-genome sequencing (WGS) from a single run on a development NovaSeq 6000 instrument to human WGS data previously generated on the HiSeq™ 2500, HiSeq 4000, and HiSeq X™ Systems. We show that the data quality meets or exceeds data quality from previous runs performed on existing platforms.

Methods

Libraries were prepared from 1000 ng of Coriell NA12878 genomic DNA using the TruSeq™ DNA PCR-Free Library Preparation Kit (Illumina, Cat. No. FC-121-3001). After PCR quantitation and dilution, cluster generation, and sequencing were performed on a NovaSeq 6000 System using NovaSeq 6000 S2 Reagents (Illumina, Cat. No. 200012860). All libraries were run at 2 × 150 bp with dual indexing except the HiSeq 2500 run, which was configured at 2 × 125 bp with dual indexing (Table 1). Six samples were loaded per NovaSeq S2 flow cell. Clustering and sequencing steps, including Read 1, Index 1 read (i7), Index 2 read (i5), paired-end turn, and Read 2, proceeded automatically without user intervention.

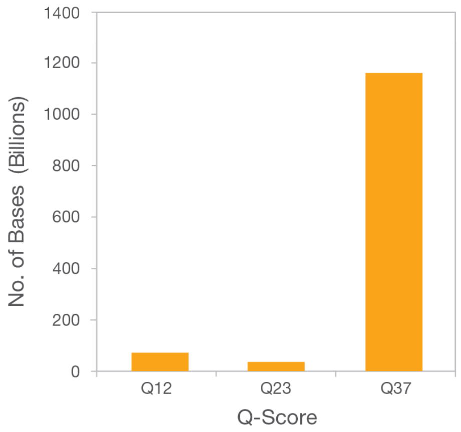

Figure 1: NovaSeq Quality Score Distribution—Quality scores for a human genome 2 × 151 base pair run on a NovaSeq 6000 System. This example shows more than 91% of bases sequenced above Q30.

Data files generated by the NovaSeq 6000 System were aligned against the human reference genome GRCh38.[4] BAM and variant call files were generated using using the Illumina Whole Genome Sequencing Analysis App v5 in BaseSpace™ Sequence Hub for both original and downsampled data.[5] Variant calls, build depth, and additional secondary analysis metrics are shown (Table 2). Raw data and variant call files from this run are available in the Public Data section of BaseSpace Sequence Hub.

Results

The NovaSeq 6000 System generated high-quality whole-genome data. The NovaSeq 6000 sequencing run generated 1268 Gb on a single NovaSeq S2 flow cell with 91% of bases above Q30 (Figure 1 and Table 1). Higher yields from the NovaSeq S2 flow cell enable each flow cell to accommodate up to eight human whole-genome samples or two WGS trio samples, depending on desired depth of coverage.

Furthermore, these runs are performed with lower cost, higher speed, and greater depth of coverage than HiSeq 2500 or HiSeq 4000 platforms.[1]

For Research Use Only. Not for use in diagnostic procedures.

Table 1: Comparison of Sequencing Run Metrics from Single Flow Cell Runs

Metric[a]	HiSeq 2500 v4	HiSeq 4000	HiSeq X	NovaSeq 6000
Run Configuration (No. of Reads × Read Length)	2 × 125	2 × 150	2 × 150	2 × 150
Reads Passing Filter (Millions)	2095	2937	3862	4037
Bases ≥ Q30[b]	87.7%	89.7%	91.9%	91.3%
Yield (Gb)	524	883	1190	1268

a. All data generated from a single run on a single flow cell
b. Average of both reads. For base calls with a quality score of 30 (Q30), one in 1000 base calls is predicted to be incorrect

Table 2: Coverage and Variant Analysis Metrics

Metric[a,b]	HiSeq 2500	HiSeq 4000	HiSeq X	NovaSeq 6000
Build Depth[c]	30×	30×	30×	30×
Total SNPs	3,628,296	3,684,413	3,706,366	3,677,388
Het:Hom Ratio	1.54	1.57	1.57	1.55
Ti:Tv Ratio	2.07	2.06	2.06	2.06
Matching Position in dbSNP	95.04%	94.64%	94.53%	94.87%
Percent Coverage ≥ 10×	97.81%	98.33%	98.44%	98.40%
SNP Precision	99.85	99.79	99.80	99.87
SNP Recall	96.28	96.97	97.23	97.08
Indel Precision	96.93	96.06	97.34	97.42
Indel Recall	87.09	90.40	94.90	95.28

a. Abbreviations: **SNPs** = single nucleotide polymorphisms, **Indel** = insertion-deletion mutation, **Het:Hom Ratio** = heterozygous/homozygous ratio, **Ti:Tv Ratio** = transition/transversion ratio, **Precision** (accuracy) = calculated as the ratio of [# of True Positive Calls/(# of True Positive Calls + # of False Positive Calls)], and **recall** (sensitivity) = calculated as the ratio of [# of True Positive Calls/(# of True Positive Calls + # of False Negative Calls)]
b. Metrics are average numbers across samples: 4 samples for HiSeq 2500, 2 samples for HiSeq 4000, 8 samples for HiSeq X, and 6 samples for NovaSeq
c. Build depth downsampled to 30×

Beyond high data yield with NovaSeq S2 reagents, percent aligned reads data showed similar performance across platforms. Ninety-five percent of reads (average of Read 1 and Read 2) aligned to the reference genome across all samples sequenced on the S2 flow cell. For variant discovery, data sets from all platforms were down sampled to 30× coverage and build metrics were analyzed across all samples. The 30× build metrics for NovaSeq S2 flow cell demonstrated equivalent performance compared to the HiSeq X flow cell (Table 1)and surpassed most variant analysis metrics on the HiSeq 2500 and HiSeq 4000 Systems (Table 2). Alignment and read statistics for this data set are available on BaseSpace Sequence Hub.

Overall, analysis of WGS data from the NovaSeq 6000 System, compared to data from existing HiSeq Systems demonstrates very high-quality data with consistent and highly concordant performance.

Conclusions

Applications requiring large amounts of data, such as human WGS can now be completed easily and in a more cost-effective manner. The simple load-and-go operation, convenient onboard cluster generation, and automatic integration with data storage and analysis tools help streamline the overall experimental workflow. The NovaSeq 6000 System expands NGS possibilities for all researchers. Whether running a single system or a large fleet, the NovaSeq System opens new avenues across a range of samples types and applications. With streamlined operations, exceptional sample scalability, and tremendous flexibility to support a range of applications, the NovaSeq 6000 System is the most flexible and powerful high-throughput Illumina sequencing system to date.

Learn More

For more information about the NovaSeq 6000, visit the NovaSeq System page.

To access the raw data and variant call files used in this application note, visit the public data section in BaseSpace Sequence Hub.

References

1. Data calculations on file. Illumina, Inc., 2015.
2. Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, et al. Characterizing and measuring bias in sequence data. *Gen Biol.* 2013;14(5):R51.
3. Liu L, Li Y, Li S, et al. Comparison of next-generation sequencing systems. *J Biomed Biotechnol.* 2012: 251364.
4. National Center for Biotechnology Information (NCBI) Human Genome Build GRCh38. Accessed July 2017.
5. Illumina BaseSpace Sequence Hub App. Whole Genome Sequencing Analysis App v5. Accessed July 2017.