Genome, Exome and Transcriptome Analysis on the HiSeq® 4000 System

Comparison of the HiSeq 4000 to the HiSeq 2500 System Performance.

Highlights

- **Unparalleled Data Output and Throughput**
  Run more samples in less time and accelerate your research
- **Exceptional Data Quality**
  SBS chemistry provides proven high-quality results
- **Flexible Platform for Multiple Applications**
  Run a broad range of applications from whole-genome sequencing to small targeted panels

Figure 1: HiSeq 3000 and HiSeq 4000 Systems—Generate up to 750 Gb per flow cell on the single flow cell HiSeq 3000 or dual flow cell HiSeq 4000 Systems.

Introduction

The HiSeq 3000 and HiSeq 4000 Systems (Figure 1) build on the proven performance of the HiSeq 2500 System. The HiSeq 3000 and 4000 instruments take high-throughput sequencing to a new level with innovative patterned flow cell technology. Patterned flow cells, first introduced on the HiSeq X® platform, contain billions of nanowells at fixed locations across the flow cell. This structured organization provides consistent spacing between adjacent clusters, with a fixed feature size, allowing accurate resolution of clusters during imaging. Patterned flow cells enable a tremendous increase in daily throughput—generating > 200 Gb per flow cell per day and allowing customers to sequence more samples, at greater depth, in less time. With the increased output on the HiSeq 3000 and 4000 Systems, users can now process up to 6 human genome samples at 30× coverage, 58 exomes, or 50 whole transcriptome samples per flow cell (Table 1).

Although the HiSeq 2500 has been the platform of choice for production-level, next-generation sequencing, the HiSeq 3000 and 4000 Systems now set a new standard in daily throughput while maintaining the same level of exceptional data quality. To demonstrate the exceptional performance of the HiSeq 3000 and 4000 Systems, this application note compares data from human whole-genome, exome, and RNA-Seq libraries run on the HiSeq 4000 and HiSeq 2500 Systems.

Table 1: HiSeq Series Sample Throughput

Sample Type (# of Samples per Dual Flow Cell Run)	HiSeq 2500	HiSeq 4000
Human Whole-Genome (30× coverage, 120 Gb)	8	12
Nextera® Rapid Capture Exome (100× coverage, 6.5 Gb)	96	116
Whole Transcriptome (50 M reads)	80	100

HiSeq Series System Applications

The HiSeq 3000 and HiSeq 4000 Systems can sequence a broad range of applications. The systems are compatible with numerous library prep methods, including but not limited to:

- TruSeq DNA PCR-Free Library Prep Kits
- TruSeq Nano DNA Library Prep Kits
- TruSeq DNA Methylation Library Prep Kit
- Nextera XT DNA Library Prep Kit
- Nextera Mate-Pair Library Prep Kit
- ChIP-Seq DNA Library Prep Kit
- TruSeq RNA Library Prep Kits
- TruSeq Stranded mRNA Library Prep Kit
- TruSeq Stranded Total RNA Library Prep Kits with RiboZero
- TruSeq Small RNA Library Prep Kit
- TruSeq Synthetic Long Read DNA Library Prep Kit

Methods

Whole-Genome Sequencing

WGS libraries were prepared from NA12878 genomic DNA (Coriell Institute for Medical Research) using the TruSeq® DNA PCR-Free Library Prep Kit (Illumina, Catalog No. FC-121-3001) with an insert size of 350 bp. Samples were run at the maximum read lengths for the given platform: 2 × 150 bp on the HiSeq 4000 and 2 × 125 bp on the HiSeq 2500 (Illumina, Catalog No. FC-401-4003 and FC-410-1003).

Normalized data were downsampled to 30× coverage. Reads were aligned and variants identified using the Isaac™ Whole-Genome Sequencing v4 BaseSpace® App.[1] Results were compared to the NIST Genome in a Bottle calls v0.2.[2] Single nucleotide variant (SNV) precision and recall along with indel precision and recall were calculated with the Variant Calling Assessment Tool (VCAT) v2.3.0 BaseSpace Labs App.[3]

Exome Analysis

Exome libraries were prepared using sample NA12878 genomic DNA (Coriell Institute for Medical Research). Exonic regions were targeted using the Nextera Rapid Capture Exome Kit (Illumina, Catalog No. FC-140-1003), which targets 37 Mb of exonic regions in the human genome. Samples were run using 2 × 100 bp read lengths and downsampled to 100× coverage depth to minimize discrepancies in the data due to differences in the coverage. Analysis was performed with the BWA Enrichment v1.0 BaseSpace App, which includes the Burrows-Wheeler Aligner (BWA) for alignment and the Genome Analysis Toolkit (GATK) for variant detection.[4] SNV precision and recall, as well as indel precision and recall were calculated by comparison to NIST Genome in a Bottle v0.2 with VCAT version 2.3.0.

RNA-Seq Analysis

Replicate mRNA-Seq and Total RNA-Seq libraries were prepared for both Human Reference Brain (HBRR Life Technologies, Catalog No. AM6050) and Universal Human Reference RNA (UHRR Agilent Technologies, Catalog No. 740200). Libraries were prepared using theTruSeq RNA Stranded mRNA Kit (Illumina, Catalog No. RS-122-2101) and the TruSeq Stranded Total RNA Kit (Illumina, Catalog No. RS-122-2201). These libraries were combined into an 8-plex pool and run on both HiSeq 4000 and HiSeq 2500 Systems. Gene-level fragments per kilobase of transcript per million mapped reads (FPKM) and differential gene expression were calculated using TopHat[5] and Cufflinks[6] applications available in BaseSpace.

Results

WGS Results

Data quality on both platforms was high in a side-by-side comparison of WGS samples run on the HiSeq 4000 and HiSeq 2500 Systems with the majority of bases having quality scores of Q30 or higher. The HiSeq 4000 generated 88% of bases with Q-scores greater than Q30, and the HiSeq 2500 generated 87.5% of bases with Q-scores greater than Q30. Secondary analysis results are also highly congruous with greater than 99% SNV precision and greater than 96% SNV recall. Indel precision was greater than 96% on both platforms, while indel recall was slightly higher on the HiSeq 4000 System with 84% indel recall on the HiSeq 4000 System compared to ~82% on the HiSeq 2500 System (Table 2). Minor differences in the data are expected as there will be run-to-run and sample-to-sample variability. Longer read lengths (2 × 150 bp) on the HiSeq 4000 can also contribute to the differences in the percentage of indels called.

Table 2: WGS Data—Variant Calling Assessment[a]

System	SNV Precision	SNV Recall	Indel Precision	Indel Recall
HiSeq 4000	99.82	97.50	96.68	84.46
HiSeq 2500	99.86	96.64	96.77	81.95

a. HiSeq 4000 System run at 2 x 150 bp, HiSeq 2500 System run at 2 x 125 bp

Exome Results

Quality scores for the exome samples were consistently high on both platforms: the HiSeq 4000 produced 93% of bases with Q-scores of Q30 or greater and the HiSeq 2500 produced 87% of bases with Q-scores of Q30 or greater. Data generated on the HiSeq 4000 and the HiSeq 2500 Systems is highly concordant; total SNV numbers vary by less than 0.01% and total indel calls differ by 0.05% (Table 3). SNV recall along with indel precision and indel recall are highly synonymous with the NIST Genome in a Bottle gold standard on both platforms. Data show greater than 99% SNV precision and 92% SNV recall and greater than 96% indel precision and 78% indel recall (Table 4). A closer look at the variant calls made on both platforms shows highly consistent results between the 2 platforms with the same calls being made on each system (Figure 2).

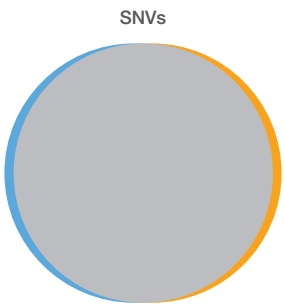
Table 3: Exome Data—Total Variant Counts

System	Total SNV Count	Total Indel Count
HiSeq 4000	33,562	2151
HiSeq 2500	33,825	2043

Table 4: Exome Data—Variant Calling Assessment[a]

System	SNV Precision	SNV Recall	Indel Precision	Indel Recall
HiSeq 4000	99.44	92.35	96.07	79.04
HiSeq 2500	99.40	92.81	96.04	78.28

a. HiSeq 4000 and HiSeq 2500 Systems were run at 2 x 100 bp and downsampled to 100x coverage depth.



Unique to HiSeq 4000	Common in both platforms	Unique to HiSeq 2500
Total count: 1,299	Total count: 32,263	Total count: 1,562
Percent: 3.87%		Percent: 4.62%
Het/Hom: 3.37	Het/Hom: 1.77	Het/Hom: 2.73
Ts/Tv: 1.81	Ts/Tv: 2.68	Ts/Tv: 2.05

Figure 2: Exome Data Variant Calling Assessment—The data show a strong concordance between the HiSeq 4000 and HiSeq 2500 platforms with greater than 32,000 SNVs being called on both platforms

RNA-Seq Results

Gene-level FPKM counts on both the HiSeq 4000 and HiSeq 2500 platforms show consistent performance with R squared values > 0.99 for approximately 16,000 genes detected at > 0.1 FPKM (Figure 3). The log2 fold-change ratio of genes between Brain and UHRR samples as calculated from data produced by the HiSeq 2500 or the HiSeq 4000 System also displays high R squared values (> 0.99).

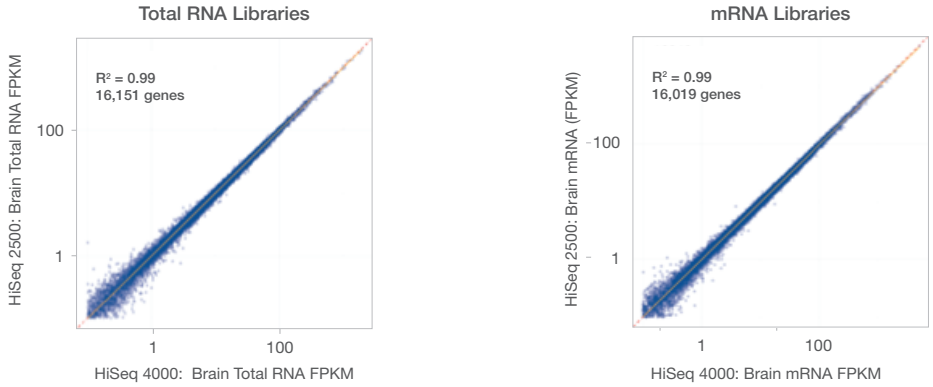


Figure 3: Gene-Level FPKM Comparisons—Gene-level expression values (in FPKM units) are shown for the same library run on the HiSeq 4000 and HiSeq 2500 Systems. In both cases, the R2 value is > 0.99 for about 16,000 genes detected at > 0.1 FPKM.

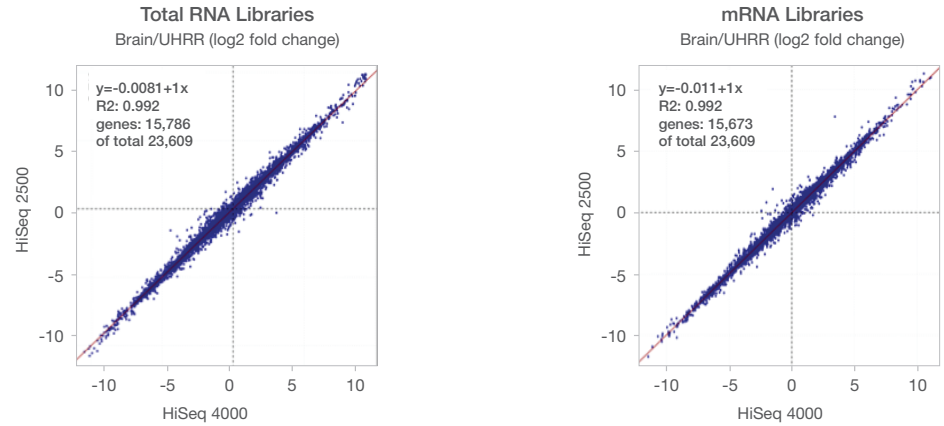


Figure 4: Concordance of Differential Expression—Log2 fold-change gene level expression from mRNA and Total RNA samples run on the HiSeq 4000 and HiSeq 2500 Systems. Data trends in a straight line for both libraries demonstrating high concordance between the platforms

Together, these data generated on the HiSeq 4000 and the HiSeq 2500 platforms, demonstrate high concordance of differential expression (Figure 4).

Conclusion

An analysis of WGS, exome, and RNA-Seq data from the HiSeq 4000 System compared to data from the HiSeq 2500 System demonstrates consistent and highly concordant performance between the HiSeq platforms. Capacity and throughput gains on the HiSeq 3000 and HiSeq 4000 platforms set a new standard for high-throughput sequencing laboratories by providing exceptional performance, significant capacity gains, and high-quality data. Building on the proven HiSeq 2500 System and harnessing innovative patterned flow cell technology, the HiSeq 3000 and 4000 Systems drive throughput to new levels, helping users achieve their research goals faster with richer, more meaningful data.

Ordering Information

Product	Catalog No.
Systems	
HiSeq 4000 Sequencing System	SY-401-4001
HiSeq 3000 Sequencing System	SY-401-3001
HiSeq 4000 System Upgrade	SY-401-4002
Reagent Kits	
HiSeq 3000/4000 SBS Kit (50 cycles)	FC-410-1001
HiSeq 3000/4000 SBS Kit (150 cycles)	FC-410-1002
HiSeq 3000/4000 SBS Kit (300 cycles)	FC-410-1003
HiSeq 3000/4000 SR Cluster Kit	GD-410-1001
HiSeq 3000/4000 PE Cluster Kit	PE-410-1001

References

1. Isaac Whole Genome Sequencing v4 BaseSpace App (www.illumina.com/informatics/research/sequencing-data-analysis-management/basespace/basespace-apps/isaac-whole-genome-sequencing-1825824.html).
2. NIST Genome in a Bottle (sites.stanford.edu/abms/giab).
3. Variant Calling Assessment Tool (VCAT) v2.3.0 BaseSpace Labs App (www.illumina.com/informatics/research/sequencing-data-analysis-management/basespace/basespace-apps/vcat-1344343.html).
4. BWA Enrichment v1.0 BaseSpace App (www.illumina.com/informatics/research/sequencing-data-analysis-management/basespace/basespace-apps/bwa-enrichment-303303.html).
5. TopHat Alignment (www.illumina.com/informatics/research/sequencing-data-analysis-management/basespace/basespace-apps/tophat-alignment-544544.html).
6. Cufflinks Assembly & DE (www.illumina.com/informatics/research/sequencing-data-analysis-management/basespace/basespace-apps/cufflinks-assembly-and-de-408408.html).

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

illumina®