TECHNICAL NOTE

Longer NGS reads and greater read depth improve metagenomic characterization of samples

High-quality metagenomic data using 600-cycle kits on NextSeq™ 1000 and NextSeq 2000 Systems

illumina®

Characterizing metagenomic samples with NGS

Next-generation sequencing (NGS) is an excellent tool for characterizing the species present in real-world microbial samples and their proportional abundance. Unlike amplicon sequencing approaches, like 16S sequencing, shotgun metagenomic sequencing captures comprehensive genomic information for every organism present in a sample.[1,2] The ability to capture full genomes means that shotgun metagenomics can identify species missed by amplicon sequencing and that the resulting data contains functional information not available from amplicon methods.[3]

This technical note demonstrates the performance of the NextSeq 1000 and NextSeq 2000 Systems using available 600-cycle kits for shotgun metagenomic studies. The shotgun metagenomics workflow integrates library preparation, proven Illumina NGS, and push-button secondary data analysis to provide a comprehensive solution for microbiome discovery (Figure 1). Specifically, we look at the effects of read length and read depth for characterizing samples.

Methods

Library preparation

To demonstrate performance on real-world samples, previously described stool samples[4] were obtained for analysis. Libraries were prepared with the TruSeq™ Nano DNA High Throughput Library Prep Kit (96 samples) (Illumina, Catalog no. 20015965) and IDT for Illumina–TruSeq DNA UD Indexes v2 (96 Indexes, 96 Samples) (Illumina, Catalog no. 20040870).

Sequencing

Prepared libraries were pooled and sequenced using a NextSeq 1000/2000 P1 Reagents Kit (600 cycles) (Illumina, Catalog no. 20075294). Sequencing was performed on the NextSeq 2000 System (Illumina, Catalog no. 20038897).

Analysis

Sequencing data from pooled libraries were demultiplexed using the BaseSpace™ Sequence Hub genomics cloud-computing platform and processed using the DRAGEN™ Metagenomics pipeline. Metagenomes were assembled using SPAdes Genome Assembler. Taxonomic classifications were conducted through the DRAGEN Metagenomics pipeline.

For read length performance comparisons, reads generated using 600 cycles were trimmed *in silico* to 300 cycles using the DRAGEN FASTQ Toolkit, available on BaseSpace Sequence Hub. To allow read depth comparisons between samples, each sample was downsampled to the same number of reads (30M, 10M, 1M) using the DRAGEN FASTQ Toolkit. Downsampling is required when only a subset of the sample can be processed by an application (eg, *de novo* assembly with memory constraints) or when the full data set is not necessary to process a sample (eg, for validating an approach at varying levels of genomic coverage).

Figure 1: Shotgun metagenomics NGS workflow on the NextSeq 1000 and NextSeq 2000 Systems.

Results

To demonstrate the effect of longer reads on metagenome assembly in complex samples, real-world stool samples were sequenced on the NextSeq 2000 System using the NextSeq 1000/2000 P1 Reagents Kit (600 cycles). To generate shorter reads for this comparison, reads from the real-world samples were trimmed from 600 cycles to 300 cycles using the DRAGEN FASTQ Toolkit app. Next, Kraken2, a k-mer–based taxonomic classifier,* was used to determine the percentage of classified reads for samples at either 600 cycles or 300 cycles (Figure 2). Data indicated that longer reads provided some improvements for k-mer–based taxonomic classification with diverse environmental samples.

Using the same trimmed data, we examined the read length impacts on sample richness (ie, the number of species detected in a sample) and the Shannon index (ie, the proportional representation of the species detected in the sample). These metrics indicate that the observed microbial diversity of the stool samples increased when the read length increased, while the proportion of detected species quantified by the Shannon index, as expected, remained relatively unchanged (Figure 3).

Greater sequencing depth improves sample characterization

Next, we examined the importance of read depth on measures of population diversity for the real-life stool samples. For this analysis, the number of reads from the NextSeq 2000 System was downsampled to 30M, 10M, and 1M reads using the DRAGEN FASTQ Toolkit and the richness and Shannon index were calculated using the DRAGEN Metagenomics app. Diversity metrics demonstrated that the microbial diversity of the stool samples increased when the sequencing depth was increased, while the proportion of the species indicated by the Shannon index remained relatively unchanged (Figure 4).

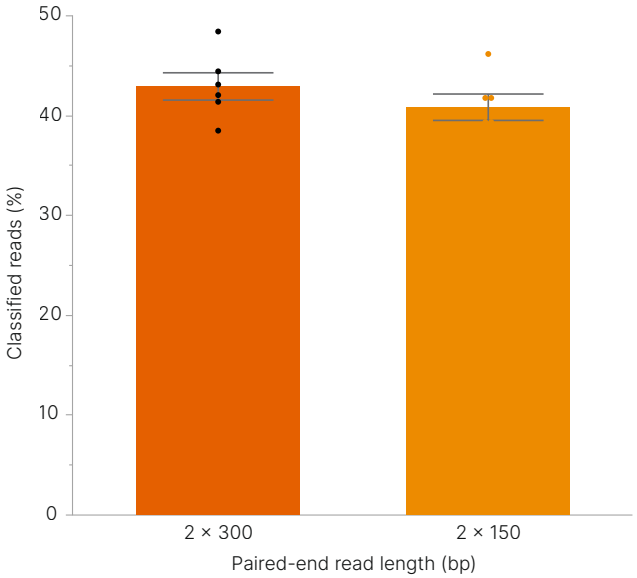

Figure 2: Longer read length improves reads classification for real-life stool samples sequenced on the NextSeq 2000 System—The DRAGEN FASTQ Toolkit app was used to trim reads from the NextSeq 2000 System from 600 cycles (2 × 300 bp) to 300 cycles (2 × 150 bp). Next, Kraken2 was used to determine the percentage of classified reads for each sample. Read depth was 30M reads for the analysis. Error bars represent one standard error from the mean.

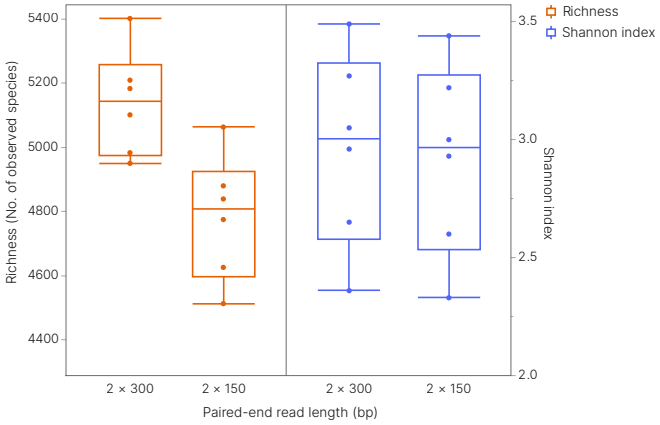

Figure 3: Microbial richness of the stool samples increases with longer read length—The DRAGEN FASTQ Toolkit App was used to trim reads from the NextSeq 2000 System from 600 cycles (2 × 300 bp) to 300 cycles (2 × 150 bp). Next, the DRAGEN Metagenomics App was used to calculate the richness and Shannon index to quantify the number of species detected and proportional diversity, respectively. Richness increased with longer reads, while the Shannon index remained relatively unchanged. Read depth was 30M reads for the analysis.

* Kraken2 Metagenomics taxonomic classification is available through the Illumina DRAGEN Metagenomics BaseSpace App.

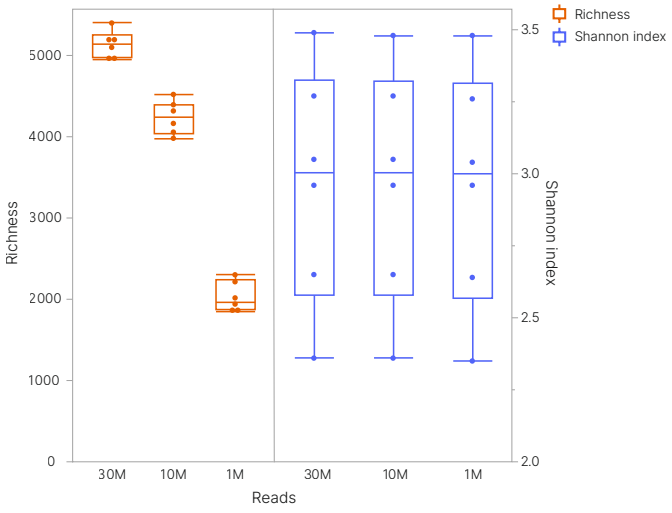

Figure 4: Microbial richness of the stool samples increases as sequencing depth increases—Richness and Shannon index were calculated using the DRAGEN Metagenomics app to measure the number of species detected and the proportional diversity observed with sequencing at 30M, 10M, and 1M reads. As expected, richness decreases in downsampled data, while the Shannon index stays approximately the same.

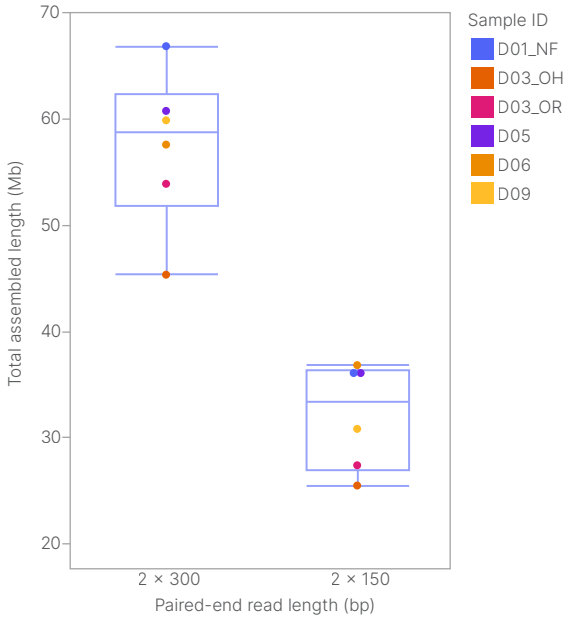

Figure 5: Longer read length supports an increased overall number of assembled bases for real-life stool samples sequenced on the NextSeq 2000 System—NextSeq 2000 System metagenomic sequencing data was trimmed to 1M reads and read lengths from 2 × 300 bp to 2 × 150 bp using the DRAGEN FASTQ Toolkit app. For comparisons, the SPAdes Genome Assembler was used to generate total contig lengths from sequencing reads set at 2 × 300 bp and 2 × 150 bp.

Longer reads improve taxonomic identification

One of the current challenges with profiling diverse environmental microbial populations is the lack of complete reference genomes for many rare and unculturable species. To examine the impact of longer reads on sample characterization, data was trimmed to 1M reads and read lengths from 2 × 300 bp to 2 × 150 bp using the DRAGEN FASTQ Toolkit App. Data showed that the number of assembled contigs in highly diverse microbial populations was greater when using longer read lengths, as demonstrated by the larger total length of assemblies (Figure 5). Shotgun metagenomic sequencing with Illumina 2 × 300 bp sequencing read lengths improved *de novo* assembly of metagenomes from environmental samples, contributing significantly to the overall completeness of each assembled metagenome.

Summary

This technical note examines the impact of using longer reads and higher read depths on analysis of metagenomic data from real-world stool samples. The 600-cycle kits available on the NextSeq 1000 and NextSeq 2000 Systems provide an excellent mix of read length and data output that enables for deep sequencing of metagenomic samples, making them an excellent choice for this application.

We demonstrate that longer reads support gains in percentage of classified reads and richness measures, while supporting better genomic assemblies in complex samples. The Shannon index was minimally affected by changes in read length or read depth, indicating that those parameters were sufficient across all comparisons in the reported analyses to generate a reasonable representation of proportional sample diversity.

Learn more

Shotgun metagenomics

NextSeq 1000 and NextSeq 2000 Sequencing Systems

TruSeq DNA Nano

References

1. Peterson D, Bonham KS, Rowland S, Pattanayak CW; RESONANCE Consortium, Klepac-Ceraj V. Comparative Analysis of 16S rRNA Gene and Metagenome Sequencing in Pediatric Gut Microbiomes. *Front Microbiol*. 2021;12:670336. Published 2021 Jul 15. doi:10.3389/fmicb.2021.670336
2. Durazzi F, Sala C, Castellani G, Manfreda G, Remondini D, De Cesare A. Comparison between 16S rRNA and shotgun sequencing data for the taxonomic characterization of the gut microbiota. *Sci Rep*. 2021;11(1):3030. Published 2021 Feb 4. doi:10.1038/s41598-021-82726-y
3. Hillmann B, Al-Ghalith GA, Shields-Cutler RR, et al. Evaluating the Information Content of Shallow Shotgun Metagenomics. *mSystems*. 2018;3(6):e00069-18. Published 2018 Nov 13. doi:10.1128/mSystems.00069-18
4. Illumina. 16S rRNA sequencing on NextSeq 1000 and NextSeq 2000 Systems. illumina.com/content/dam/illumina/gcs/assembled-assets/marketing-literature/nextseq-600c-16s-rrna-application-note-m-gl-01146/nextseq-600c-16s-rrna-application-note-m-gl-01146.pdf. Published 2023. Accessed February 6, 2024.

illumina®

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

M-GL-03217 v1.0