TECHNICAL NOTE

# A shotgun metagenomics NGS workflow for assessing microbial populations in complex samples

## 600-cycle kits on the NextSeq™ 1000 and NextSeq 2000 Systems provide accuracy and flexibility for species identification

illumina®

## Metagenomic classification of complex samples

Shotgun metagenomic sequencing is an alternative method to amplicon sequencing approaches, such as 16S and internal transcribed spacer (ITS) ribosomal RNA (rRNA) sequencing, for assessing microbial diversity in complex samples. Unlike amplicon approaches, shotgun metagenomic sequencing using next-generation sequencing (NGS) captures comprehensive genomic information for every organism present in a sample. The ability to capture full genomes means that shotgun metagenomics can identify species missed by amplicon sequencing[1] and that the resulting data contains functional information not available from amplicon methods.[2,3]

This application note demonstrates the performance similarities of the NextSeq 1000, NextSeq 2000, and MiSeq™ Systems for high-throughput shotgun metagenomic studies. Using data generated on the venerable NextSeq 550 System, we also demonstrate advantages that 600-cycle kits have over commonly used 300-cycle kits in metagenomics applications. We present data from synthetic population and real-life samples to demonstrate superior genera and species identification when using 600-cycle kits.

## Methods

The NextSeq 1000/2000 P1 Reagents (600 cycles) kit and the NextSeq 1000/2000 P2 Reagents (600 cycles) kit expand the capacity and sequencing output of the NextSeq 1000 and NextSeq 2000 Systems with specifications that are ideal for shotgun metagenomic sequencing. The NextSeq 1000 and NextSeq 2000 Systems use load-and-go reagents with no onboard fluidics, reducing the number of workflow steps and the risk of sample contamination. The shotgun metagenomics workflow integrates library preparation, proven Illumina NGS, and push-button secondary data analysis for a complete solution to microbiome discovery (Figure 1).

### Library preparation

Microbial genomic DNA samples were obtained from two sources. The first sample was the commercially available American Type culture collection ATCC 20 Strain Staggered Mix Genomic Material (ATCC, Catalog no. MSA-1003). This ATCC sample is a mock microbial community composed of a staggered distribution of prepared genomic DNA from bacterial strains selected based on attributes such as Gram stain, GC content, and sporulation attributes. A second set of real-world stool samples, described previously,[4] was also obtained for analysis.

Libraries were prepared with Illumina DNA Prep, (M) Tagmentation (24 Samples, IPB) (Illumina, Catalog no. 20060060) and IDT for Illumina DNA/RNA UD Indexes Set A, Tagmentation (96 Indexes, 96 Samples)

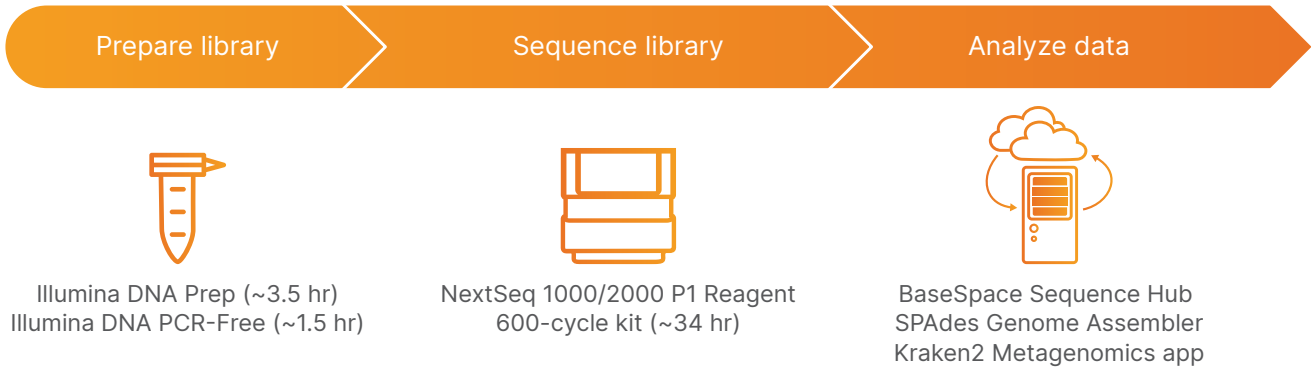


Figure 1: Shotgun metagenomics NGS workflow on the NextSeq 1000 and NextSeq 2000 Systems

(Illumina, Catalog no. 20027213). IDT for Illumina DNA/RNA UD Indexes Sets A to D allow users to generate 384 metagenomic libraries.

### Sequencing

Prepared libraries were pooled and loaded into a prefilled NextSeq 1000/2000 P1 Reagents (600 cycles) kit flow cell, a MiSeq Reagent Kit v3 (600-cycle) flow cell (Illumina, Catalog no. MS-102-3003), or a NextSeq 500/550 High Output Kit v2.5 (300 cycles) flow cell (Illumina, Catalog no. 20024908). Sequencing was performed on the NextSeq 2000 System, a MiSeq System, or a NextSeq 550 System, respectively. Representative sequencing data for all runs are available on the BaseSpace™ Sequence Hub demo data web page.

### Analysis

Pooled libraries were demultiplexed in the BaseSpace Sequence Hub genomics cloud computing platform. The DRAGEN™ Metagenomics pipeline was used to process data generated on the NextSeq 2000, MiSeq, and NextSeq 550 Systems. Metagenomes were assembled using SPAdes Genome Assembler. Taxonomic classifications were conducted through the DRAGEN Metagenomics pipeline.

To compare NextSeq 2000 data generated using a 600-cycle kit with NextSeq 500 data generated using a 300-cycle kit, NextSeq 2000 reads were trimmed using the DRAGEN FASTQ Toolkit, available on BaseSpace Sequence Hub. To allow comparisons between samples, each sample was downsampled to the same number of reads (30M, 10M, 1M) through the DRAGEN FASTQ Toolkit. Downsampling is required in cases 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).

## Results

### Improved sequence primary metrics

NextSeq 1000/2000 P1 Reagents (600 cycles) on the NextSeq 2000 System shows a higher percentage of quality scores ≥ Q30 when compared to the MiSeq Reagent Kit v3 (600-cycle) run on the MiSeq System. NextSeq 1000/2000 P1 Reagents (600 cycles) also provides up to 100M single-end reads passing filters or 200M paired-end reads passing filters. At approximately 60 Gb, NextSeq 1000/2000 P1 Reagents (600 cycles) generates four-fold more data than the MiSeq Reagent Kit v3 (600-cycle) at approximately 15 Gb. In addition, sequencing runs with NextSeq 1000/2000 P1 Reagents (600 Cycles) kit completed in about 34 hr, which is approximately 20 hr less time than a MiSeq Reagent Kit v3 (600-cycle) sequencing run (Figure 2).

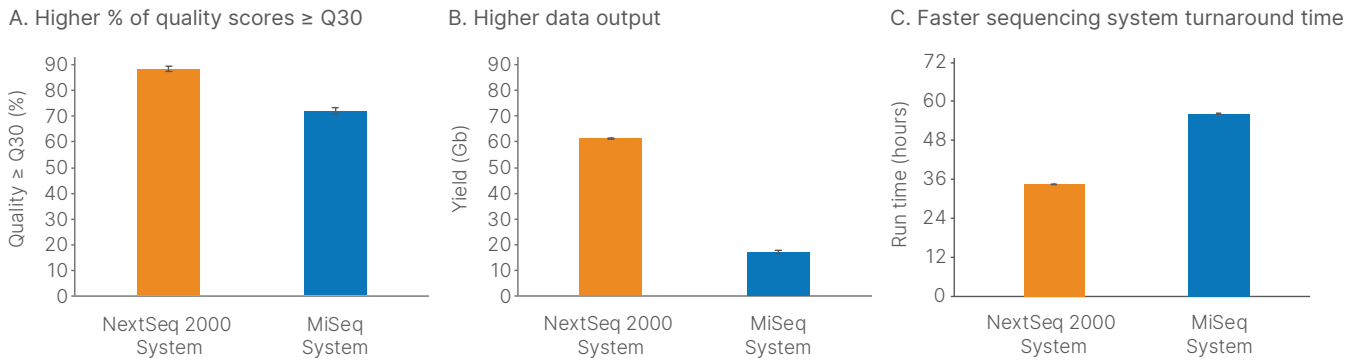


Figure 2: Primary performance metrics comparisons for NextSeq 2000 and MiSeq Systems—Compared to the MiSeq Reagent Kit v3 (600-cycle) run on the MiSeq System, NextSeq 1000/2000 P1 Reagents (600 cycles) on the NextSeq 2000 System offer (A) a higher percentage of quality scores ≥ Q30, (B) four-fold higher data output at, and (C) ~20 hr shorter instrument run time when using the NextSeq P1 flow cell.

The NextSeq 500/550 High Output Kit v2.5 (300 cycles) (Illumina, Catalog no. 20024908) is able to generate up to 120 Gb of high-quality sequencing data, at a 2 × 150 bp read length, in about 29 hours. The quality specification for this kit is > 75% of bases ≥ Q30 (data not shown).

## Metagenomic analysis comparisons

To compare performance across systems, the 20 Strain Staggered Mix Genomic Material was sequenced on the NextSeq 2000, MiSeq, and NextSeq 550 Systems. The DRAGEN Metagenomics app on BaseSpace Sequence Hub was used for downstream analysis elucidating taxonomic classifications. Metagenomic NGS analysis identified all expected members of the mock bacterial community and showed comparable results between all three sequencing systems (Figure 3).

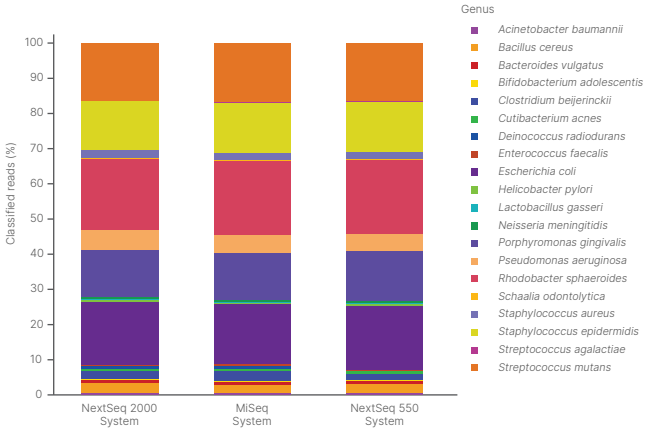


Figure 3: Comparative microbial composition analysis reveals composition for ATCC 20 Strain Staggered Mix analyzed on the NextSeq 2000, MiSeq, and NextSeq 550 Systems—DRAGEN Metagenomics app analysis of microbial composition demonstrates excellent, reproducible genera identity and distribution.

While the ATCC sample shows highly repeatable performance, a mock sample may not be indicative of performance with real-world samples. Therefore, we tested organism detection performance with real-world stool samples. The 20 highest represented genera in the stool samples were compared across the NextSeq 2000, NextSeq 550, and MiSeq Systems (Figure 4.) While the results are not as uniform with this more complex sample type, the metagenomics community profiles of all real-world stool samples were highly concordant between the

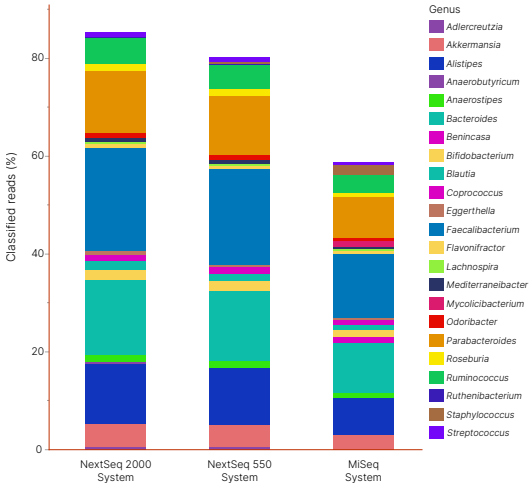


Figure 4: Metagenomic analysis reveals similar community composition for a representative real-life stool sample analyzed on multiple systems—DRAGEN Metagenomics app to analyze the microbial composition of real-life stool samples, limited to the 20 highest represented genera in the sample, sequenced on the NextSeq 2000, NextSeq 550 and MiSeq Systems. Data demonstrate similar genera coverage across platforms, even with complex samples.

NextSeq 2000, MiSeq, and NextSeq 550 Systems. This indicates that all three systems are capable of performing reliable metagenomic sequencing in various sample types.

## Longer read length improves sample characterization

While 300-cycle flow cells are able to provide meaningful metagenomic sequencing data, 600-cycle kits offer advantages when working to identify individual species in complex samples. To demonstrate the effect of longer reads on metagenome assembly, reads from the NextSeq 2000 System were trimmed from 600 cycles to 300 cycles using the DRAGEN FASTQ Toolkit app. Kraken2, a k-mer–based taxonomic classifier,* was used to determine the percentage of classified reads for each sample at either 600 cycles or 300 cycles (Figure 5). Data suggest that longer reads do provide some improvements for k-mer–based taxonomic classification with diverse environmental samples.

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

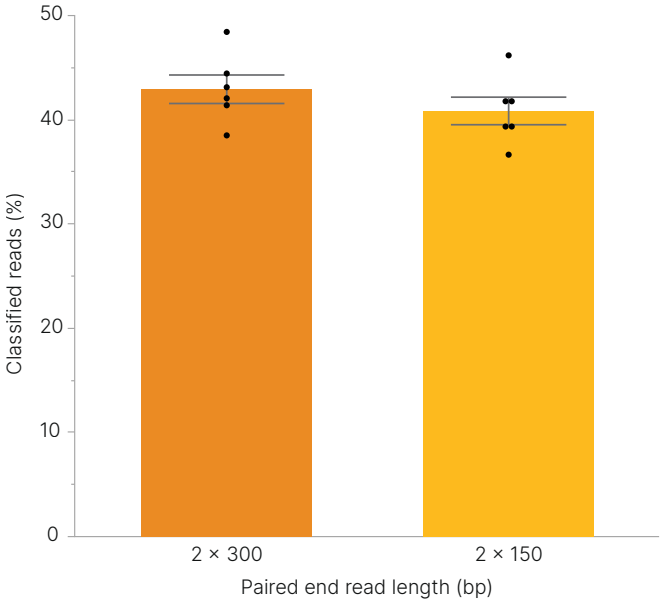


Figure 5: Longer read length improves reads classification for real-life stool samples sequenced on the NextSeq 2000—The DRAGEN FASTQ Toolkit app was used to trim reads from the NextSeq 2000 System from 600 cycles (2 × 300 bp) to 300 cycles (2 × 150 bp). Next, Kraken2 was used to determine the percentage of classified reads for each sample. Read depth was 30M reads for the analysis. Error bars represent one standard error from the mean.

Using the same trimmed data, 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) were also examined. These metrics indicate that the observed microbial diversity of the stool samples increases when the read length increases, while the proportion of detected species quantified by the Shannon index, as expected, remains relatively unchanged (Figure 6).

## 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. The number of reads from the NextSeq 2000 System were 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

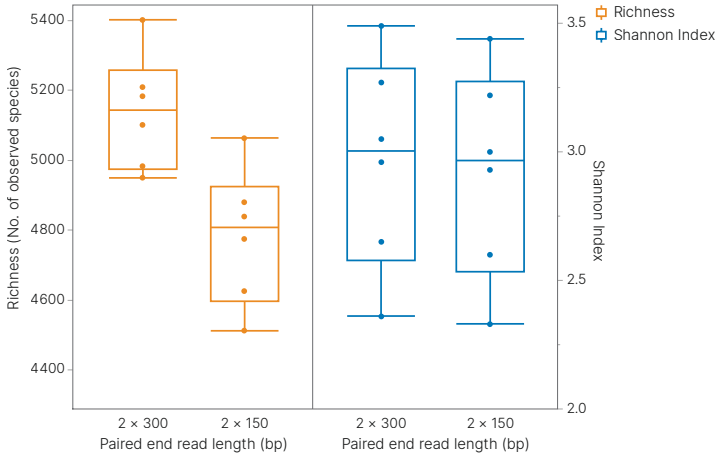


Figure 6: Microbial richness of the stool samples increases with longer read length— The DRAGEN FASTQ Toolkit App was used to trim reads from the NextSeq 2000 System from 600 cycles (2 × 300 bp) to 300 cycles (2 × 150 bp). Next, the DRAGEN Metagenomics App was used to calculate the richness and Shannon index to quantify the number of species detected and proportional diversity, respectively. Richness increases with longer reads while the Shannon index remains relatively unchanged. Read depth was 30M reads for the analysis.

the microbial diversity of the stool samples increases when the sequencing depth increases, while the proportion of the species indicated by the Shannon index remains relatively unchanged (Figure 7).

## 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. Generally, the number of assembled contigs for highly diverse microbial populations is greater for longer read lengths, as shown in the larger total length of assembly (Figure 8). Shotgun metagenomic sequencing with Illumina 2 × 301 bp sequencing read lengths generally improves *de novo* assembly of metagenomes from environmental samples, contributing significantly to the overall completeness of each assembled metagenome.

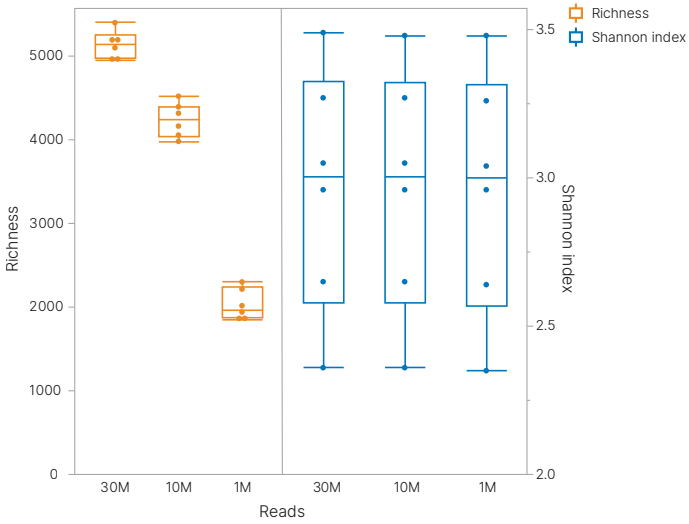


Figure 7: Microbial richness of the stool samples increases as sequencing depth increases—Richness and Shannon index were calculated using the DRAGEN Metagenomics app to measure the number of species detected and the proportional diversity observed with sequencing at 30M, 10M, and 1M reads. As expected, richness decreases in downsampled data, while the Shannon index stays approximately the same.

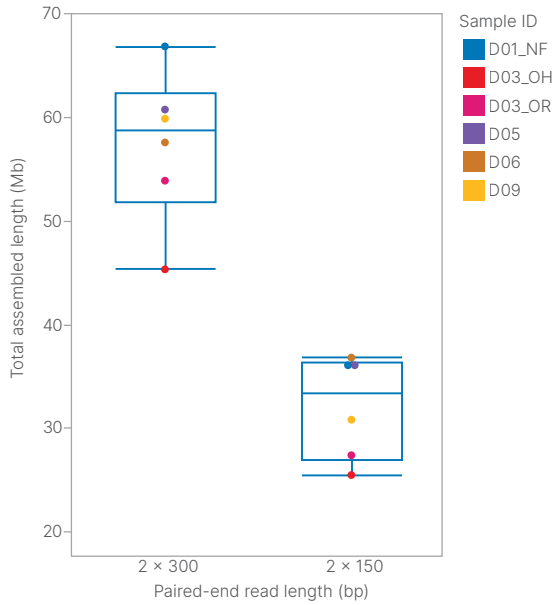


Figure 8: Longer read length supports increased overall number of assembled bases for real-life stool samples sequenced on the NextSeq 2000 System—NextSeq 2000 System metagenomic sequencing data was trimmed to 1M reads, and read lengths from 2 × 300 bp to 2 × 150 bp using the DRAGEN FASTQ Toolkit app. 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.

## Summary

The purpose of this app note is to demonstrate the similar performance of the 600-cycle kits on the NextSeq 2000 System and the MiSeq System. The NextSeq 550 System is also capable of accurate metagenomic analysis, but does not have a 600-cycle option. All three systems achieved accurate *de novo* assemblies regardless of which instrument was used. Concordant metagenomic profiles, particularly the observed proportional diversity of genera, were achieved across all three systems.

Overall, the NextSeq 2000 System offer advantages over the MiSeq and NextSeq 550 Systems in terms of resolving sample richness detail from diverse, culture-free samples due to additional read length and read depth capabilities.† These advangags are especially important for complex metagenomics samples, such as stool or environmental samples.

In summary, the NextSeq 1000/2000 P1 Reagents (600 cycles) and NextSeq 1000/2000 P2 Reagents (600 cycles) on the NextSeq 1000 and NextSeq 2000 Systems offer high-quality sequencing with a faster turnaround time than the MiSeq Reagent Kit v3 (600-cycle) on the MiSeq System. The 600-cycle kits on the NextSeq 1000 and NextSeq 2000 Systems maintain high Q30 scores and low error rates as demonstrated in this application note. In addition, the NextSeq 1000/2000 P1 (600 cycles) NextSeq 1000/2000 P2 Reagents (600 cycles) deliver flexible data output that is ideal for small- and medium-scale genome sequencing experiments. Finally, the 600-cycle kits for NextSeq 1000 and NextSeq 2000 Systems enable application expansion and operational simplicity while maintaining the data quality established on the proven MiSeq System.

† The NextSeq 1000 System is functionally similar to the NextSeq 2000 System and should achieve similar results to that instrument using the same 600-cycle kits.

## Learn more

Illumina sequencing platforms

NextSeq 1000 and NextSeq 2000 System

NextSeq 1000/2000 reagents

Illumina DNA Prep

Illumina DNA Prep crude lysate protocol for NGS

## 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. Stothart MR, McLoughlin PD, Poissant J. Shallow shotgun sequencing of the microbiome recapitulates 16S amplicon results and provides functional insights. *Mol Ecol Resour*. 2023;23(3):549-564. doi:10.1111/1755-0998.13713
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.


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


M-GL-01147 v1.0