APPLICATION NOTE

Long read lengths for shotgun metagenomics

2 × 251 bp sequencing on the NovaSeq™ 6000 System

- Explore taxonomic and functional diversity of microbial communities
- Perform efficient *de novo* assembly of metagenomes with a comprehensive workflow

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

Introduction

Shotgun metagenomics is a powerful, unbiased approach to assess microbial composition, diversity, and functional potential. This method harnesses next-generation sequencing (NGS) technology to sequence thousands of microbial genomes in parallel from culture-free human, animal, and environmental samples. Experimental objectives and sample complexity dictate the depth of coverage needed to perform a comprehensive analysis of each metagenomic sample. Currently, Illumina offers a diverse array of sequencing platforms to support metagenomics studies and each can be tailored for specific microbial applications. This application note demonstrates how the NovaSeq 6000 Sytem can be used for high-throughput shotgun metagenomic studies using Illumina long-read sequencing.[1]

In this study, a diverse array of bacterial and metagenomic samples were used to investigate the effects of library preparation kit, read length, and sequencing instrument on the quality of metagenomics analysis for taxonomic classification and *de novo* assembly (Figure 1). Shotgun metagenomics using 2 × 251 bp sequencing was performed in parallel on the NovaSeq 6000 System and the HiSeq™ 2500 System. The HiSeq 2500 System enables up to 600 million high-quality 2 × 251 bp reads in Rapid Run mode.[2] The results of this study demonstrate that the NovaSeq 6000 System delivers high-quality 2 × 251 bp reads at a much faster rate compared to the HiSeq 2500 System, while providing sample size flexibility for cost-effective and efficient metagenomic studies (Figure 2). By delivering up to 1.6 billion paired-end reads, the NovaSeq SP flow cell is optimal for large-scale studies that include highly diverse environmental samples and enables high-quality *de novo* assembly for downstream metagenomics analysis.

Methods

Isolate and mock community samples

A combination of commercially available bacterial isolates and mock community samples were evaluated (Table 1).

Stool samples

Stool samples from three donors were collected and provided by collaborators at PerkinElmer and stored at 4°C for 20 hours before DNA extraction.

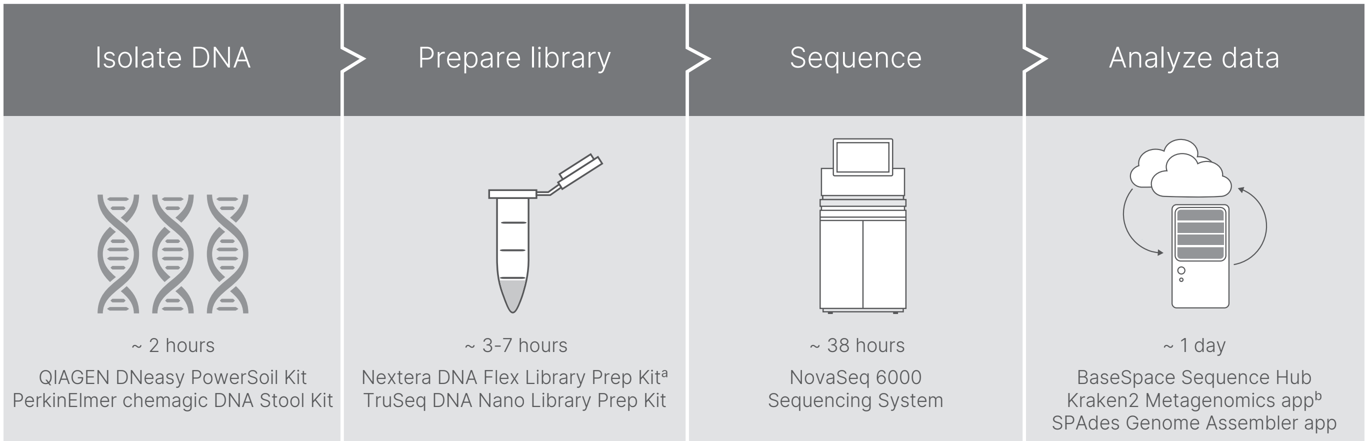
Isolate DNA
Prepare library
Sequence
Analyze data
~ 2 hours
QIAGEN DNeasy PowerSoil Kit
PerkinElmer chemagic DNA Stool Kit
~ 3-7 hours
Nextera DNA Flex Library Prep Kit[a]
TruSeq DNA Nano Library Prep Kit
~ 38 hours
NovaSeq 6000
Sequencing System
~ 1 day
BaseSpace Sequence Hub
Kraken2 Metagenomics app[b]
SPAdes Genome Assembler app

Figure 1: The shotgun metagenomics NGS workflow on the NovaSeq 6000 System—Shotgun metagenomics studies on the NovaSeq 6000 System use a comprehensive NGS workflow that includes DNA extraction, library preparation, long-read sequencing, and data analysis.

a. The Nextera DNA Flex Library Prep Kit, used for this experiment, is now known as Illumina DNA Prep

b. Kraken2 Metagenomics taxonomic classification, used for this experiment, is now available through the DRAGEN Metagenomics app

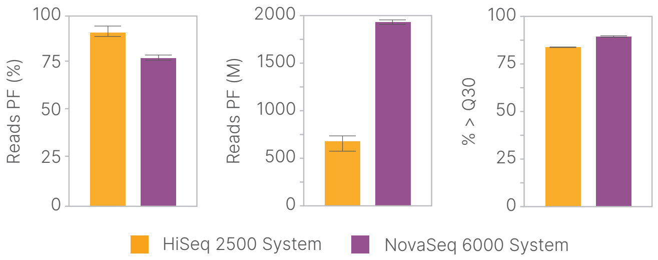

Figure 2: Primary performance metrics comparison—The NovaSeq 6000 System (purple) delivers comparable sequencing results to the HiSeq 2500 System (yellow), as measured by reads passing filter (PF) and percentage of bases > Q30. Error bars indicate averaged metric for each system for 2 × 251 bp runs.

Soil samples

Soil samples were collected from Encinitas, CA and Madison, WI. Topsoil samples were collected with a trowel and stored at 4°C in 50 ml conical tubes until DNA extraction.

DNA extraction

Automated stool genomic DNA (gDNA) extractions were performed by PerkinElmer using the chemagic DNA Stool kit (Catalog no. CMG-1076, PerkinElmer) on the chemagic 360 Instrument (Catalog no. 2024–0020, PerkinElmer).[3,4] Samples were eluted in a volume of 150 µl.

Manual soil gDNA extractions were performed using the DNeasy PowerSoil Kit (Catalog no. 12888-100, QIAGEN).[5] Each extraction was performed with 0.25 mg of soil input and eluted with a volume of 100 µl. Integrity and concentration of the DNA was assessed with the Fragment Analyzer (Catalog no. M5311AA, Agilent Technologies) using the HS NGS Fragment kit (Catalog no. DNF474-1000, Agilent Technologies).[6] These methods provide unbiased extraction of purified microbial gDNA ready for Illumina library preparation and sequencing.

Table 1: Commercially available isolate and mock community samples

Sample	Vendor	Catalog no.
E. coli	ATCC	700926
B. cereus	ATCC	10987
R. sphaeroides	ATCC	17023
ATCC 10 Strain Staggered Mix Genomic Material	ATCC	MSA-1001
ZymoBIOMICS microbial community DNA standard	Zymo Research	D6306

ATCC, American Type Culture Collection

Library preparation

Eleven libraries were prepared manually using the Nextera™ DNA Flex Library Prep Kit* (Table 2) (Catalog no. 20018705, Illumina).[7] The total DNA input (10 ng) was normalized to a volume of 30 µl with nuclease-free water prior to on-bead tagmentation and amplified using IDT for Illumina Nextera DNA UD Indexes Set A (Catalog no. 20027213, Illumina).[8]

Nineteen libraries were prepared manually using the TruSeq™ DNA Nano Low Throughput Library Prep Kit (Catalog no. 20015964, Illumina) and IDT for Illumina – TruSeq DNA UD Indexes (Catalog no. 20020590, Illumina) were utilized for ligation.[9,10] Eleven of these 19 libraries were prepared with 100 ng gDNA input and sheared with the Covaris S220 Focused-ultrasonicator (Catalog no. M020011, Covaris) to a target insert size of 350 bp. The remaining eight libraries were prepared with 200 ng gDNA input and sheared to a target insert size of 550 bp (Table 2). Standard sonication conditions described in the TruSeq DNA Nano Low Throughput Library Prep Kit manual were used.[11] The quality and concentration of PCR-amplified libraries were assessed using the Fragment Analyzer prior to pooling.

* The Nextera DNA Flex Library Prep Kit is now known as Illumina DNA Prep. The two kits have identical product performance specifications and kit configurations.

Table 2: No. of prepared libraries from metagenomics samples

Sample type	TruSeq DNA Nano 100 ng input, 350 bp insert	TruSeq DNA Nano 200 ng input 550 bp insert	Nextera DNA Flex[a] 10 ng input 350 bp insert
Isolate	3	3	3
Mock community	2	1	2
Soil	3	3	3
Stool	3	1	3
Total	11	8	11

a. The Nextera DNA Flex Library Prep Kit, used for this experiment, is now known as Illumina DNA Prep; The two kits have identical product performance specifications and kit configurations

Sequencing

Pooled libraries were sequenced on the HiSeq 2500 System (Rapid Run mode) and the NovaSeq 6000 System (SP flow cell) with a run configuration of 2 × 251 bp.

Data analysis

Pooled libraries were demultiplexed in BaseSpace™ Sequence Hub, the Illumina genomics cloud computing platform. Taxonomic classification and downsampling were conducted through the Kraken2 Metagenomics† and FASTQ Toolkit Apps, respectively.[12,13] Prior to gene identification, metagenomes were assembled using SPAdes Genome Assembler, available through BaseSpace Sequence Hub (Table 3).[14] Gene detection was performed using online comparative analysis tools available through Joint Genome Institute (JGI) (GOLD for project submission and IMG/M for gene detection and functional profiling).[15,16]

Table 3: Metagenomics sequencing BaseSpace apps

BaseSpace App	Description
Kraken2 Metagenomics[b]	Assigns taxonomic labels to short DNA sequences with high sensitivity and speed
SPAdes Genome Assembler	Assembles small genomes from standard bacterial data sets
FASTQ Toolkit	Provides a modular set of analyses for quality control checks on raw sequence data before downstream analysis

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

Results

To evaluate the performance of the NovaSeq 6000 System for shotgun metagenomics with 2 × 251 bp reads, four types of microbial samples were used for this study: bacterial isolates, microbial mock communities, stool from human donors, and soil. Kraken2, a k-mer–based taxonomic classifier,† was used to determine the percentage of classified reads for each sample (Figure 3). The number of classified reads were comparable between the HiSeq 2500 and NovaSeq 6000 Systems (Figure 3). Furthermore, this data suggests that longer reads provide slight improvements for k-mer–based taxonomic classification for diverse environmental samples (Figure 3).

Efficient *de novo* assembly of metagenomes

One of the current challenges with profiling diverse environmental microbial populations is the lack of complete reference genomes for many rare and unculturable species. Therefore, it is desirable to identify a cost-effective method for high-quality *de novo* genome assembly from culture-free samples. Shotgun metagenomic sequencing with Illumina 2 × 251 bp sequencing offers a solution for efficient *de novo* assembly of metagenomes from environmental samples (Figure 4).

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

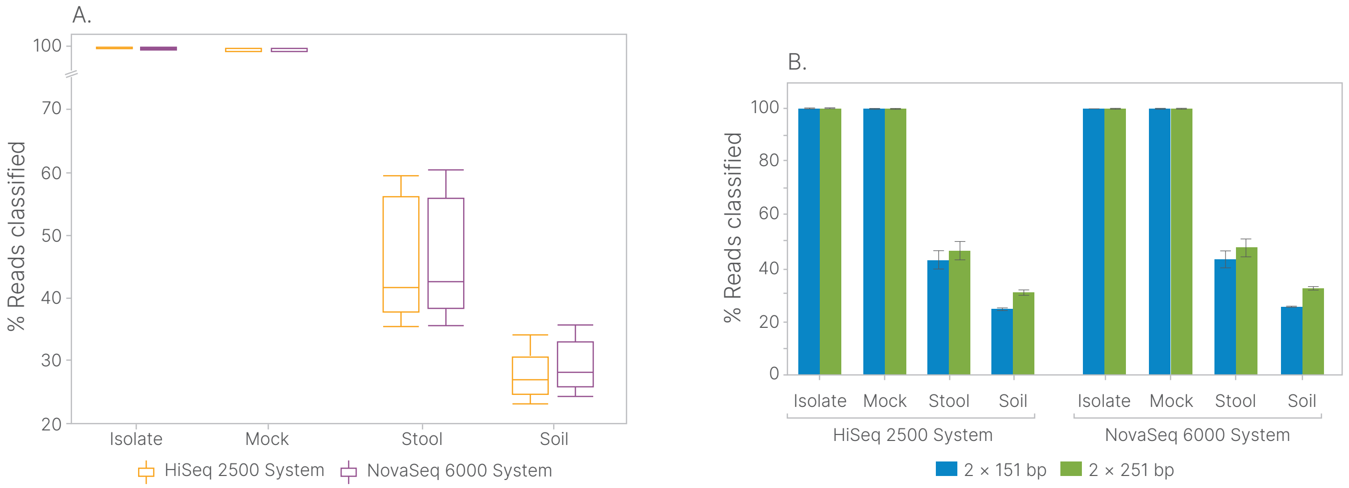

Figure 3: Comparison of classified reads by system and read length—The percentage of classified reads for four different microbial sample types were comparable between (A) the HiSeq 2500 System (yellow) and NovaSeq 6000 System (purple) and (B) 2 × 151 bp (blue) and 2 × 251 bp (green) read lengths.

Generally, the number of contigs for highly diverse microbial populations is greater for longer read lengths, contributing significantly to the overall completeness of each assembled metagenome, as shown in the larger total length of assembly. To this end, the NovaSeq 6000 System offers an advantage in terms of yield for data output to increase contig length and total assembly length, important for metagenomic assembly from diverse culture-free samples.

Longer read lengths improve genome assembly as biodiversity increases

Interestingly, for samples of low diversity (bacterial isolates and mock communities), longer read lengths do not display the same degree of *de novo* assembly advantages as seen with diverse samples, likely due to the significantly smaller genome sizes (Figure 5). With 20 million paired-end reads, complete draft genomes were assembled for bacterial isolates where the total length of each genome assembled did not vary between tested read lengths. For the mock communities, each composed of gDNA from 10 microbial species, relatively complete draft genomes were assembled for each microorganism, though the longer read lengths provided slight improvements in fraction assembled genome (Figure 5).

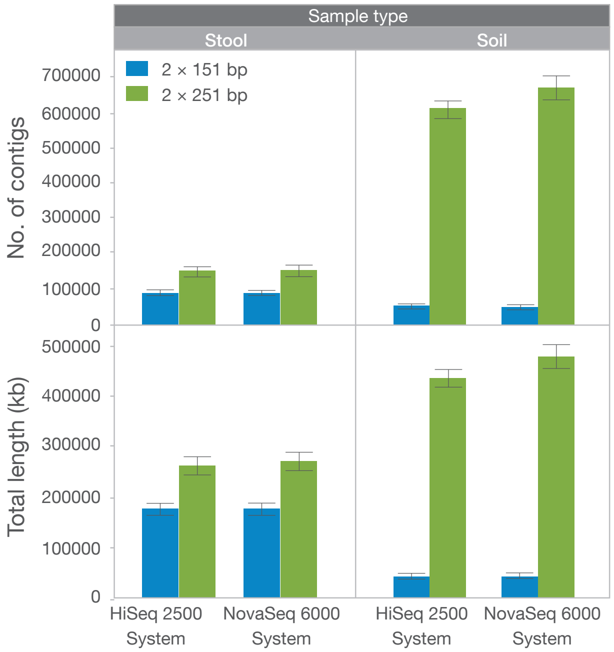

Figure 4: High-quality *de novo* assembly of metagenomes—The combination of longer read length and the NovaSeq 6000 System produces more contigs (top panel) and longer total assembly length (bottom panel), all of which are important metrics for metagenomic assembly from diverse culture-free samples. Analysis performed with SPAdes Genome Assembler; comparing read lengths of 2 × 151 bp (blue) and 2 × 251 bp (green).

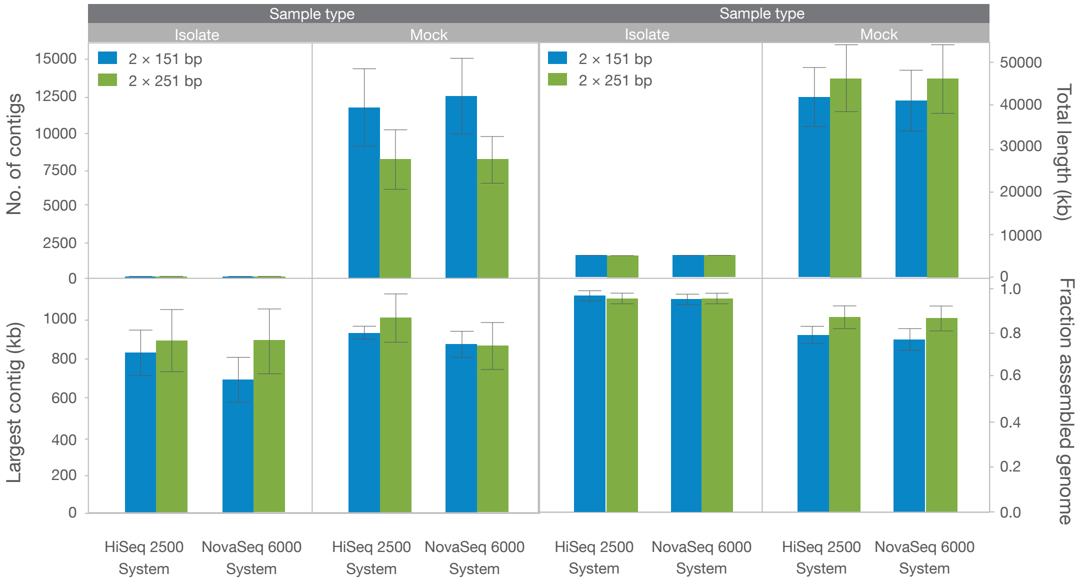

Figure 5: Effect of read length on genome assembly for controls—Low-diversity isolate and mock community samples sequenced on the HiSeq 2500 and NovaSeq 6000 Systems resulted in comparable *de novo* assembly; comparing read lengths of 2 × 151 bp (blue) and 2 × 251 bp (green). Longer read lengths provided slight improvements in the fraction assembled genome in a mock community with higher biodiversity compared to isolates.

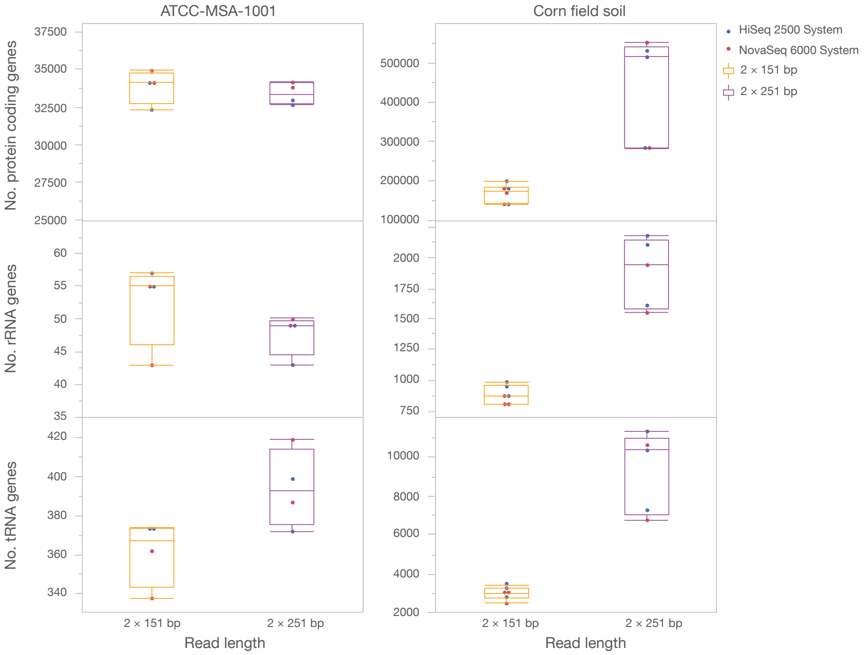

Figure 6: Improved gene annotation with longer read lengths—Longer read lengths result in improved gene annotation for functional profiling, with increases in the number of protein-coding genes (top panel), mRNA genes (middle panel), and tRNA genes (bottom panel) for a soil metagenome. ATCC-MSA-1001 is a mock community of low diversity used as a control with read length displaying minimal impact on gene annotation. Results were comparable between the HiSeq 2500 System (blue) and NovaSeq 6000 System (red). Comparing read lengths of 2 × 151 bp (yellow) and 2 × 251 bp (purple).

Longer read lengths improve gene annotation

Due to better *de novo* metagenome assembly with longer read lengths, the overall number of gene elements detected for soil metagenomes was significantly higher with 2 × 251 bp run configuration when compared to the trimmed 2 × 151 bp read lengths (Figure 6). The American Type Culture Collection (ATCC) mock community was used as a control for a metagenome with low diversity, where minimal differences in gene detection was expected between read length due to saturated coverage (Figure 6). Furthermore, there was no significant difference detected when comparing the number of annotated genes between the HiSeq 2500 and NovaSeq 6000 systems, demonstrating the exceptional data output of both sequencing systems.

Summary

The NovaSeq 6000 System offers high-quality long-range sequencing at rapid speed, while maintaining high Q30 scores and low error rates. Furthermore, this system delivers high-output data that is ideal for large-scale experiments, especially *de novo* metagenomic assembly. This application note demonstrates that robust metagenomics analysis was achieved between the HiSeq 2500 and NovaSeq 6000 Systems with the 2 × 251 bp run configuration, supporting precise taxonomic classification and metagenomic assembly. Also, no significant difference was observed between library preparation kits and insert sizes (data not shown), suggesting that longer read length is the major contributor to improved metagenomics analysis. The accessibility of analytical tools via BaseSpace Sequence Hub allows for simple and efficient processing of high-throughput data for a diverse array of metagenomic applications.

Learn more

NovaSeq 6000 System, illumina.com/novaseq

References

1. Illumina. The NovaSeq 6000 System Specification Sheet. illumina.com/content/dam/illumina/gcs/assembled-assets/marketing-literature/novaseq-6000-spec-sheet-770-2016-025/novaseq-6000-spec-sheet-770-2016-025.pdf. Published 2016. Updated 2020. Accessed June 4, 2021.
2. Illumina. HiSeq 2500 Sequencing System Specification Sheet. illumina.com/content/dam/illumina-marketing/documents/products/datasheets/datasheet-hiseq-2500.pdf. Published 2015. Accessed June 22, 2021.
3. Illumina. An automated Nextera DNA Flex library preparation workflow for high-throughput metagenomics. illumina.com/content/dam/illumina-marketing/documents/products/appnotes/nextera-dna-flex-ht-metagenomics-workflow-770-2018-007.pdf. Published 2018. Accessed June 22, 2021.
4. chemagic 360 Instrument, PerkinElmer. Compact, High Volume, High Throughput Nucleic Acid Isolation. perkinelmer.com/lab-solutions/resources/docs/BRO_chemagic-360_ROW-CT6-30-0116-01.pdf. Published 2016. Accessed June 22, 2021.
5. DNeasy PowerSoil Kit, QIAGEN. DNeasy PowerSoil Kit Handbook. qiagen.com/us/resources/resourcedetail?id=5a0517a7-711d-4085-8a28-2bb25fab828a&lang=en. Published 2017. Accessed June 22, 2021.
6. Fragment Analyzer, Agilent Technologies. NGS Analysis. aati-us.com/documents/brochures/ngs-analysis-brochure.pdf. Published 2018. Accessed December 19, 2018.
7. Illumina. Nextera DNA Flex Library Preparation Kit Data Sheet. illumina.com/content/dam/illumina-marketing/documents/products/datasheets/nextera-dna-flex-data-sheet-770-2017-011.pdf. Published 2017. Accessed June 22, 2021.
8. IDT for Illumina UD Indexes. support.illumina.com/sequencing/uniquedualindex.html. Accessed June 22, 2021.
9. Illumina. TruSeq DNA Nano Data Sheet. illumina.com/content/dam/illumina-marketing/documents/products/datasheets/datasheet_truseq_nano_dna_sample_prep_kit.pdf. Published 2017. Accessed June 22, 2021.

10. IDT for Illumina TruSeq UD Indexes. support.illumina.com/sequencing/sequencing_kits/idt-truseq-dna-rna-udi.html. Accessed June 22, 2021.
11. Illumina. TruSeq Nano DNA Library Prep Reference Guide. support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/samplepreps_truseq/truseqnanodna/truseq-nano-dna-library-prep-guide-15041110-d.pdf. Published 2017. Accessed June 22, 2021.
12. Kraken2 Metagenomics. ccb.jhu.edu/software/kraken2/. Accessed June 22, 2021.
13. FASTQ Toolkit. illumina.com/products/by-type/informatics-products/basespace-sequence-hub/apps/fastq-toolkit.html. Accessed June 4, 2021.
14. SPAdes Genome Assembler. illumina.com/products/by-type/informatics-products/basespace-sequence-hub/apps/algorithmic-biology-lab-spades-genome-assembler.html. Accessed June 22, 2021.
15. JGI Gold. gold.jgi.doe.gov. Accessed June 22, 2021.
16. Chen IA, Chu K, Palaniappan K, et al. IMG/M v.5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes. *Nucleic Acids Res.* 2019:47(D1):D666–D677.

illumina®

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