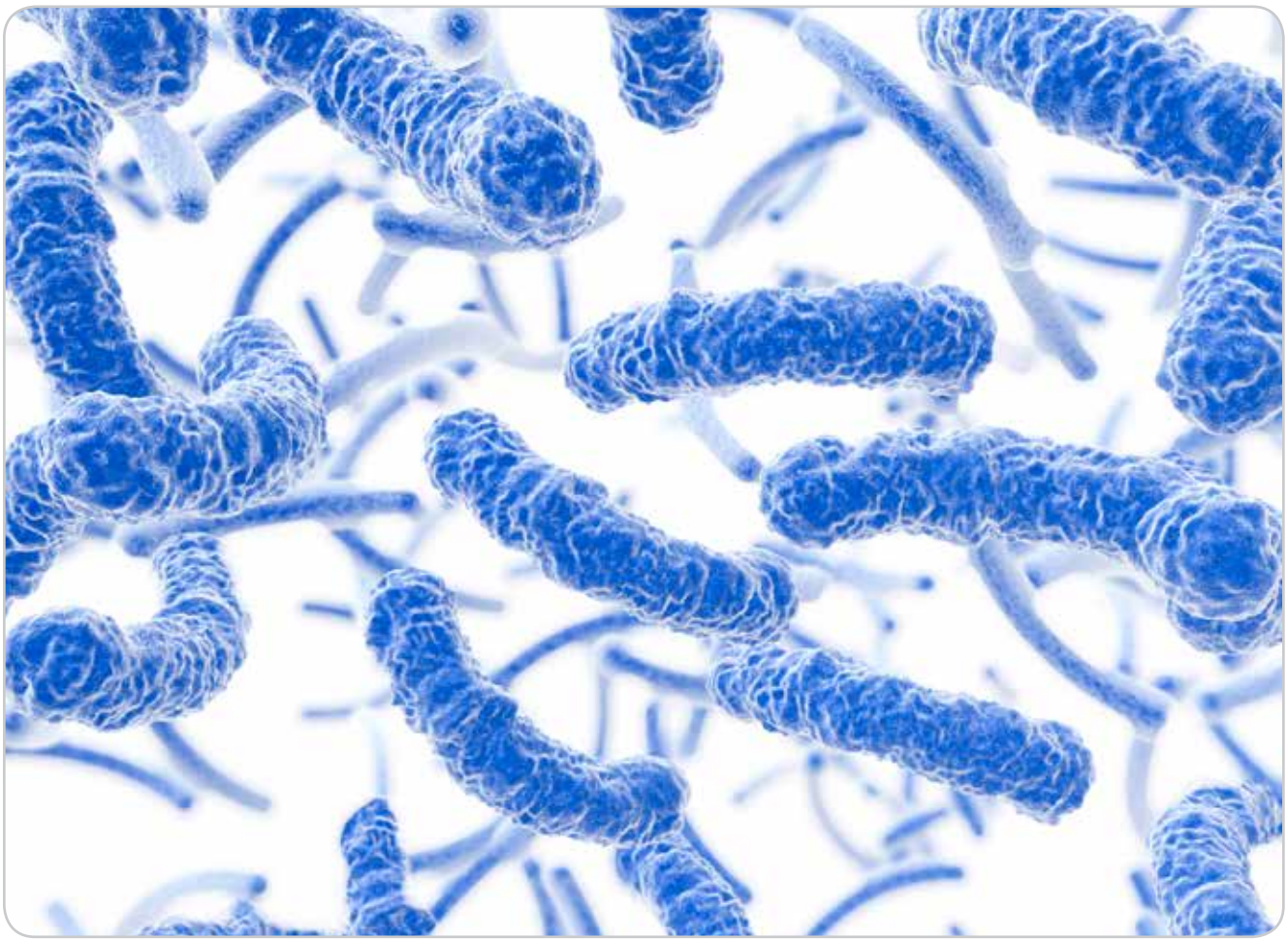
An Introduction to Illumina Next-Generation Sequencing Technology for Microbiologists

Deciphering DNA sequences is essential for virtually all branches of biological research. Capillary electrophoresis (CE)-based sequencing has enabled scientists to elucidate genetic information from almost any organism or biological system. Although this technology has become widely adopted, inherent limitations in throughput, scalability, cost, speed, and resolution can hinder scientists from obtaining essential genomic information. To overcome these barriers, an entirely new technology was developed—next-generation sequencing (NGS), a fundamentally different approach to sequencing that has triggered numerous ground-breaking discoveries. The years since the introduction of NGS have seen a major transformation in the way scientists extract genetic information from biological systems, revealing insight about the genome, transcriptome, and epigenome. This introduction will highlight the benefits of using NGS for microbiology research.



www.illumina.com/microbiology

illumina®

Welcome to Next-Generation Sequencing

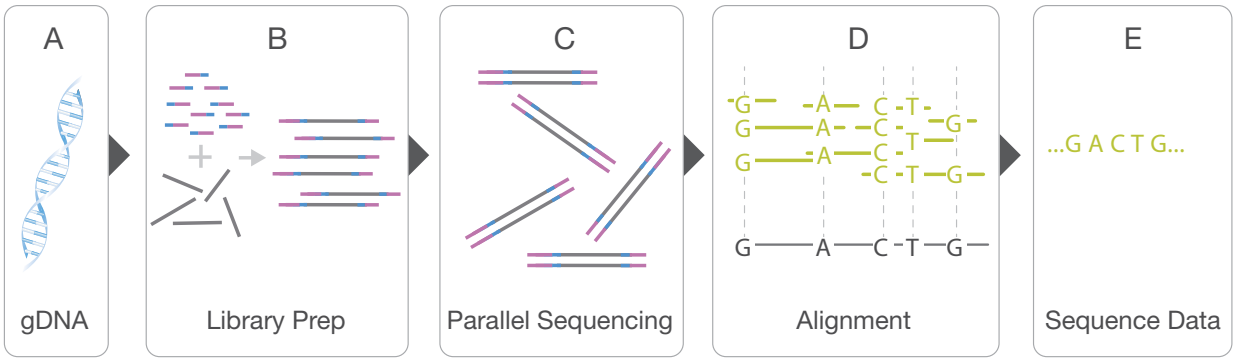
Advances in next-generation sequencing have been instrumental in advancing scientific fields from human disease research to environmental and evolutionary science. NGS lends itself particularly well to the microbial laboratory, where the genomes are small and the data analysis is relatively simple. One major advantage of using NGS over other genome interrogation methods is the ability to measure changes anywhere in the genome without prior knowledge, such as for unculturable organisms. Single-base resolution allows tracking of microbial adaptation over short periods of time, both in the laboratory and in the environment.

As evidence of the uptake of this technology, NGS data output has increased at a phenomenal rate, more than doubling each year since it was introduced. In 2007, a single sequencing run could produce about one gigabase (Gb) of data. By 2012, that rate had jumped 1000×, to one terabase (Tb) of data in a single sequencing run. With this tremendous increase in output has come a 10^5-fold decrease in the cost of determining the genome of a microorganism. In 1995, sequencing the 1.8 megabase (Mb) genome of *Haemophilus influenzae* with CE technology cost ~1 million US dollars, taking over one year. Today, sequencing the 5 Mb genome of *Escherichia coli* with Illumina NGS can be done in one day at a fraction of the cost.

Basic Concepts of NGS

In principle, NGS is similar to Sanger-based, or CE sequencing. The bases of a small fragment of DNA are sequentially identified from signals emitted as each fragment is re-synthesized from a DNA template strand. NGS extends this process across millions of reactions in a massively parallel fashion, rather than being limited to a single or a few DNA fragments. This advance enables rapid sequencing of large stretches of DNA, with the latest instruments capable of producing hundreds of gigabases of data in a single sequencing run. To illustrate how this process works, consider a single genomic DNA (gDNA) sample. The gDNA is first fragmented into a library of small segments and sequenced. The newly identified strings of bases, called reads, are then reassembled using a known reference genome as a scaffold (resequencing), or assembled together using advanced computational techniques if no reference genome is available (de novo sequencing). The full set of aligned reads reveals the entire genomic sequence of the sample (Figure 1). Once the sample library is prepared, all of the sequencing steps through data analysis can be performed on a single instrument, facilitating rapid turnaround with minimal hands-on time.

Figure 1: Concepts of Next-Generation Sequencing



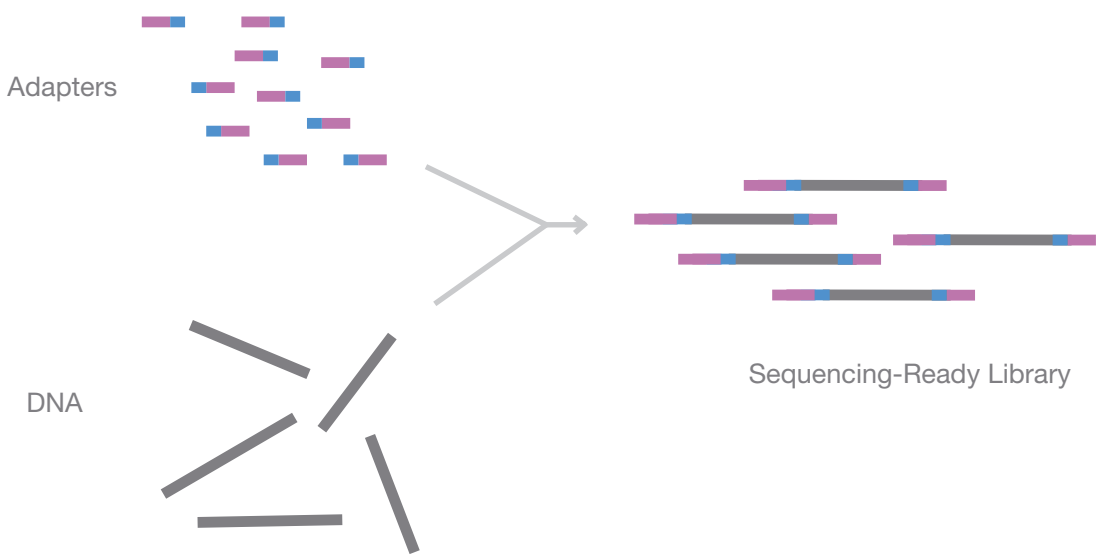
A. Extracted gDNA.
B. Sample preparation fragments genomic DNA and adds adapters to generate a library.
C. DNA fragments within the library are each sequenced in parallel.
D. Individual sequence reads are reassembled by aligning to a reference genome.
E. The whole genome sequence is derived from the consensus of aligned reads.

Sample Preparation

How NGS is used experimentally is largely dictated by the way sequencing libraries are prepared and the way the data is analyzed, with the actual sequencing steps remaining fundamentally unchanged. A growing number of library preparation kits provide complete reagent sets and protocols for sequencing whole genomes, small genomes, mRNA, targeted regions such as whole exomes, custom-selected regions, protein-binding regions, and more. To address specific research objectives, researchers have developed many novel protocols to isolate specific regions of the genome associated with a given biological function.

Sample preparation protocols for NGS are generally more rapid and straightforward than those for CE-based Sanger sequencing. With NGS, researchers can start directly from a gDNA or cDNA library. The DNA fragments are then ligated to specific oligonucleotide adapters needed to perform the sequencing biochemistry, requiring as little as 90 minutes with Illumina's Nextera® technology (Figure 2). In contrast, CE-based Sanger sequencing requires genomic DNA to be fragmented first and cloned into either bacterial artificial chromosomes (BACs) or yeast artificial chromosomes (YACs). Then, each BAC/YAC must be further subcloned into a sequencing vector and transformed into the appropriate microbial host. Template DNA is then purified from individual colonies or plaques prior to sequencing. This process can take days or even weeks to complete.

Figure 2: NGS Sample Preparation Overview



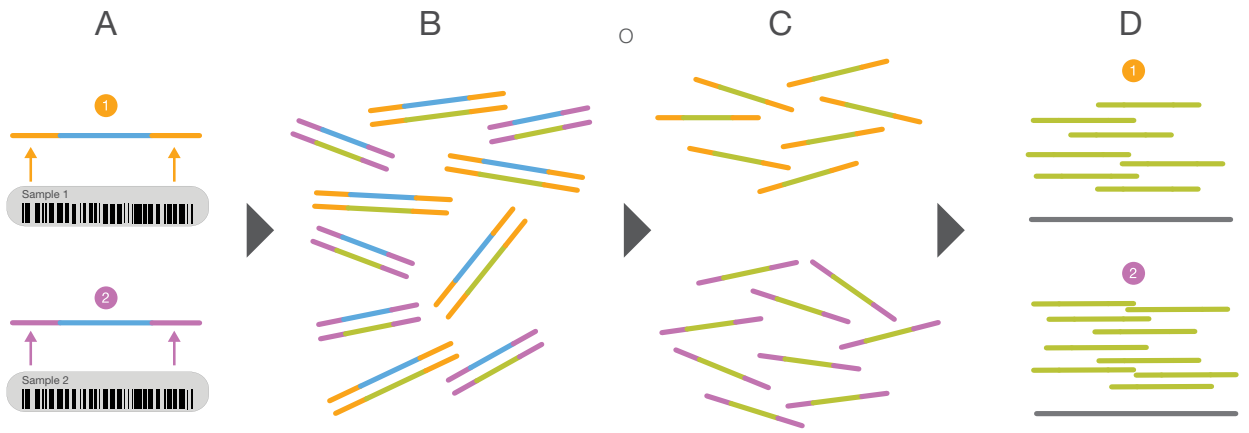
NGS library preparation starts directly from fragmented genomic DNA. Platform-specific oligonucleotide adapters, needed to perform the sequencing biochemistry and index samples, are ligated to each end of the fragments to yield the sequencing-ready library.

Scalable Studies Enabled by Multiplexing

For sequencing small bacterial/viral genomes, a researcher can choose to use a lower output instrument and process a smaller number of samples per run, or can opt to process a large number of samples. Multiplexing enables large numbers of samples to be simultaneously sequenced during a single experiment (Figure 3). To accomplish this, individual "barcode" sequences are added to each sample so they can be differentiated during the data analysis.

With multiplexing, NGS dramatically reduces the time to data for large numbers of samples. Processing hundreds of amplicons using CE technology generally requires several weeks or months. The same number of samples can now be sequenced in a matter of hours and fully analyzed within two days using NGS. With highly automated, easy-to-use protocols, researchers can go from experiment to data to publication faster and easier than ever before.

Figure 3: Sample Multiplexing Overview



A. Two representative DNA fragments from two unique samples, each attached to a specific barcode sequence that identifies the sample from which it originated.
B. Libraries for each sample are pooled and sequenced in parallel. Each new read contains both the fragment sequence and its sample-identifying barcode.
C. Barcode sequences are used to de-multiplex, or differentiate reads from each sample.
D. Each set of reads is aligned to the reference sequence.

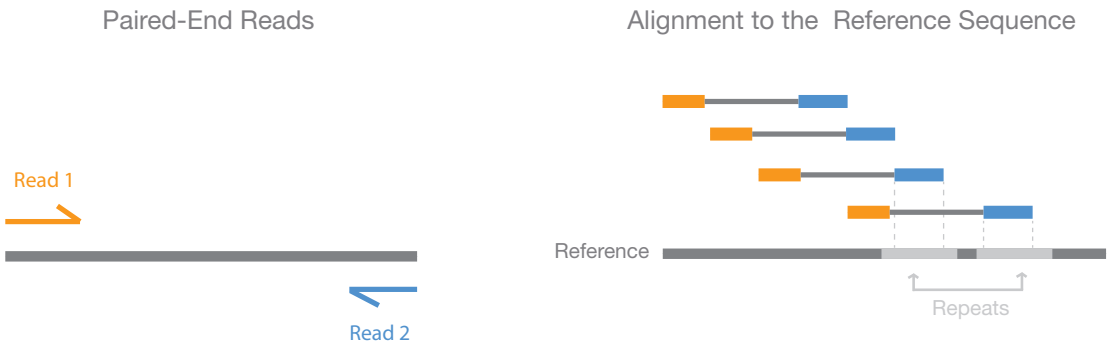
Paired-End Sequencing

Paired-end (PE) sequencing, where both ends of a DNA fragment are sequenced (Figure 4) allows long range positioning of the DNA fragment. Because the distance between each paired read is known, alignment algorithms can use this information to precisely map the reads, resulting in superior alignment across difficult-to-sequence or repetitive genome regions. Illumina NGS offers the flexibility of variable insert sizes and read lengths (35–300 bp), allowing high resolution characterization of any genome.

Analyze, Store, and Share in Illumina's BaseSpace® Cloud

Data analysis is an important factor to consider for sequencing applications. One of the biggest challenges with NGS systems has been the requirement for a high-performance computing infrastructure, enterprise-level storage, and highly skilled bioinformatics and IT staff. While complex, primary data processing including alignment and variant calling happens seamlessly and behind the scenes as the sequencing run progresses. Depending on the application, most subsequent analyses can be run directly on optimized software installed on the sequencer's internal computer, or in BaseSpace, Illumina's unique cloud computing environment. Essentially, push-button informatics solutions simplify the analysis, allowing researchers to focus on the biology. By storing and analyzing data in the cloud, BaseSpace users can instantly share data with collaborators across the hallway or across the globe. The BaseSpace Application Store will provide seamless access to a wide variety of commercial software tools, as well a collection of well-known and open-source algorithms from academic institutions. These tools will provide biological interpretation and insights to further your research.

Figure 4: Paired-End Sequencing and Alignment



Paired-end sequencing enables both ends of the DNA fragment to be sequenced. Because the distance between each paired read is known, alignment algorithms can use this information to map the reads over repetitive regions more precisely. This results in much better alignment of the reads, especially across difficult-to-sequence, repetitive regions of the genome.

End-to-End Solution

Only Illumina NGS provides a fully supported solution from DNA to results, with specialized sample prep choices for the application you are working on, to robust and proven sequencing reagents, and a wide range of simple data analysis tools (Figure 5).

Figure 5: Illumina's End-to-End NGS Workflow on the MiSeq Benchtop Sequencer

NGS workflow on the MiSeq system includes sample preparation, massively parallel sequencing, automated data analysis, and cloud-enabled data analysis, storage, and sharing.

Microbiology Applications

Genomic sequencing can further any microbial identification study, including those based on known organisms, and those with incomplete or no information, as in microbiome studies or environmental surveillance. High-resolution genome data can be instrumental for examining pathogenesis, horizontal gene transfer, pangenomes, and co-evolution of hosts and symbionts/parasites. The wealth of information enabled by NGS is beneficial for mutational studies of all kinds, including directed evolution strategies, lab adaptation analyses, mutagenesis screens, or studying the temporal and spatial dynamics of epidemics and transmission.

Another important advantage of NGS is the abundance of sequence information. Deep sequencing makes it possible to detect very low abundant members of complex populations. As a result, the ability to detect low abundance populations can profoundly impact the interpretation of microbiological changes. Sequencing of microbial genomes has become routine and individual cells have been sequenced. This accumulation of sequence information has greatly expanded our appreciation of the dynamic nature of microbial populations and their impact on the environment and human health.

Whole-Genome Sequencing of Small Genomes

Until recently, sequencing an entire genome was a major endeavor. While NGS is commonly associated with sequencing large genomes, the scalability of the technology makes it just as useful for small viral or bacterial genomes. By contrast, whole-genome sequencing using CE-based Sanger technology requires significant time and resources, even for compact genomes. The ability of NGS platforms to produce a large volume of data in a short period of time makes it a powerful tool for whole-genome sequencing. The power and speed of NGS was demonstrated during the 2011 enteroaggregative *E. coli* outbreak in Europe, which prompted a rapid scientific response. Using NGS data, researchers were able to quickly generate a high-quality, whole-genome sequence of the bacterial strain, enabling them to better understand the genetic mutations conferring the increased virulence.

Table 1: Illumina NGS and CE-Based Sanger Sequencing for Small Genomes

Parameter	MiSeq System	Sanger Sequencing
Strains to be sequenced	4*	4
Genome size	3 Mb	3 Mb
Time for sample prep	1.5 hours	weeks
Sequencing time	2 days	243 days**
Price per genome†	$249 USD	$84,000 USD
Project price†	< $995 USD	> $336,000 USD
Coverage depth per genome	> 583 x	7 x
On instrument data analysis?	Yes	No

* Can sequence up to 75 strains per run at 30x coverage
**Assumes single 3730xl instrument running 24 hr/day
† Excluding sample prep

Example: *De novo* Sequencing

One challenge associated with sequencing small genomes is the lack of reference genomes available for most species. This means that whole-genome sequencing must often be done *de novo*, where the reads are assembled without aligning to a reference sequence. Paired-end reads and increasing reads lengths up to 300 bp result in good alignment across regions containing repetitive sequences and produce longer contigs for *de novo* sequencing by filling gaps in the consensus sequence, resulting in more complete coverage. Compared to CE sequencing, NGS enables researchers to simultaneously analyze many strains in one experiment, at significant time and cost savings (Table 1).

Targeted Sequencing

With targeted sequencing, only a subset of genes or defined regions in a genome are sequenced, allowing researchers to focus time, expenses, and data storage resources on the regions of the genome in which they are most interested. Amplicon sequencing refers to sequencing selected regions of the genome spanning hundreds of base pairs. The latest NGS amplicon library preparation kits allow researchers to perform rapid in-solution amplification of custom-targeted regions from genomic DNA. Using this approach, thousands of amplicons spanning multiple samples can be simultaneously prepared and indexed in a matter of hours. With the ability to process numerous amplicons and samples on a single run, NGS enables researchers to simultaneously analyze all genomic content of interest in one experiment, at fraction of the time and cost of conventional CE sequencing.

Table 2. Illumina NGS and CE-Based Sanger Sequencing for Targeted Applications

Parameter	MiSeq System	Sanger Sequencing
Samples in project	96	96
Number of amplicons	12	12
Target panel size	~5 kb	~5 kb
Time for sample prep	< 3 hours	< 3 hours
Sequencing time	1 day	6 days
Price per amplicon*	$1 USD	$4 USD
Project price*	< $2000 USD	> $4500 USD
Coverage depth per amplicon	> 13,000 x	2 x**
On instrument data analysis?	Yes	No

* Excluding PCR amplification.
** Including bidirectional sequencing.

Example: 16S Metagenomic Sequencing

A common amplicon sequencing application is comparing the bacterial 16S rRNA gene, a widely used method for studying phylogeny and taxonomy. This method has been used to evaluate bacterial diversity in many environments, allowing researchers to characterize microbiomes from samples that are otherwise difficult or impossible to study. NGS, with its ability to sequence thousands of organisms in parallel, is uniquely suited to this application. The ability to pool samples and obtain high sequence coverage during a single run allows NGS to identify rarer variants that are missed, or too expensive to identify, using CE-based sequencing approaches (Table 2). It is worth mentioning that the example in Table 2 is based on only 96 samples for the purpose of comparison. True metagenome sequencing studies, comprising hundreds or thousands of possible genomes, are cost- and labor-prohibitive with CE sequencing, and only possible with high-throughput NGS systems such as Illumina's HiSeq and NextSeq™ systems.

Take Your Research to the Next Level

The advent of NGS has enabled researchers to study biological systems at a level never before possible. With clear benefits over Sanger-based CE sequencing, next-generation sequencing can transform your microbiology research, opening new avenues to explore. To identify the sequencing platform that is optimal for your research needs, visit www.illumina.com.

From Innovation to Publication

As NGS technology continues to evolve, researchers are making fascinating discoveries in a number of biological fields, unlocking answers never before possible in all fields of research. As a result, there has been an explosion in the number of peer-reviewed scientific publications, including over 4,500 featuring Illumina sequencing technology. Selected recent examples relevant to microbiology are listed below.

Whole-Genome Sequencing

1. Toprak E, Veres A, Michel JB, Chait R, Hartl DL, et al. (2011) Evolutionary paths to antibiotic resistance under dynamically sustained drug selection. Nat Genet 44: 101–105.
2. Chua, KYL, Seemann T, Harrison PF, Monagle S, Korman TM, et al. (2011) The dominant Australian community-acquired methicillin-resistant *Staphylococcus aureus* clone ST93-IV [2B] is highly virulent and genetically distinct. PLoS ONE 6:
3. Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, et al. (2012) Performance comparison of benchtop high-throughput sequencing platforms. Nat Biotechnol 30(5): 434–9.

De novo Sequencing

4. Chitsaz H, Yee-Greenbaum JL, Tesler G, Lombardo MJ, Dupont CL, et al. (2011) Efficient de novo assembly of single-cell bacterial genomes from short-read data sets. Nat Biotechnol 29: 915–921.
5. Rodrigue S, Malmstrom R, Berlin A, Birren B, Henn M, et al. (2009) Whole genome amplification and de novo assembly of single bacterial cells PLoS One (4)9 e6864.

Metagenomics

6. Caporaso JG, Lauber CL, Walkers, WA, Berg-Lyons D, Lozupone CA et al. (2011) Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. Proc Natl Acad Sci USA 108:4516–22.
7. Mackelprang, R, Waldrop MP, DeAngelis KM, David MM, Chavarria KL, et al. (2011) Metagenomic analysis of a permafrost microbial community reveals a rapid response to thaw. Nature 480: 368–371.

FOR RESEARCH USE ONLY

Pub No. 770-2012-023 Current as of 20 December 2013

illumina®