An Introduction to Illumina Next-Generation Sequencing Technology for Agriculture

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 agricultural research.

www.illumina.com/agriculture

illumina®

Welcome to Next-Generation Sequencing

Advances in NGS have been instrumental in advancing scientific fields from human disease research to environmental and evolutionary science. NGS lends itself particularly well to the agricultural laboratory, where genomes can be complex and *a priori* knowledge scarce. NGS allows identification of novel SNPs accounting for valuable phenotypic traits or the sequencing of unculturable organisms. For veterinary diagnostic applications, single-base resolution allows tracking of microbial or viral adaptation over short periods of time, both in the laboratory and in the environment.

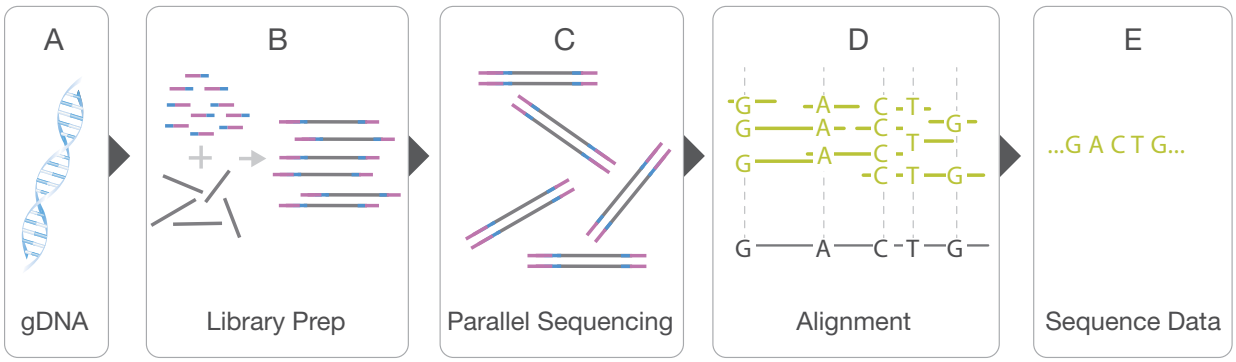
As evidence of the uptake of this technology, NGS data output has increased at a phenomenal rate, more than doubling each year since it was introduced. In 2007, a single sequencing run could produce about one gigabase (Gb) of data. By 2012, that rate had jumped 1000× to one terabase (Tb) of data in a single sequencing run. With this tremendous increase in output has come a 105-fold decrease in the cost of bacterial whole-genome sequencing. 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. This advance enables rapid sequencing of many large stretches of DNA, with the latest instruments capable of producing 1 Tb 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 bioinformatically using a known reference genome (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: Next-Generation Sequencing Steps



A. Extracted gDNA.
B. Library 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.

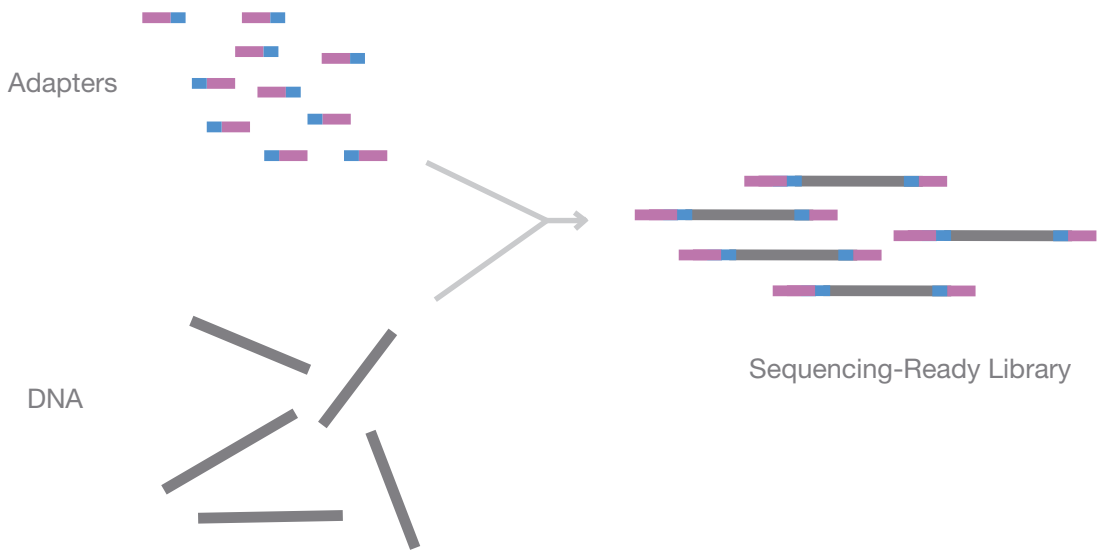
Library Preparation

How NGS is used experimentally is largely dictated by the way sequencing libraries are prepared and the way the data are analyzed, with the actual sequencing steps remaining fundamentally unchanged. A growing number of library preparation kits are currently available to provide complete reagents and protocols for sequencing whole genomes, small genomes, mRNA, targeted regions such as whole exomes, targeted or enriched 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.

Library 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 whole-genome 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 Library Preparation Overview



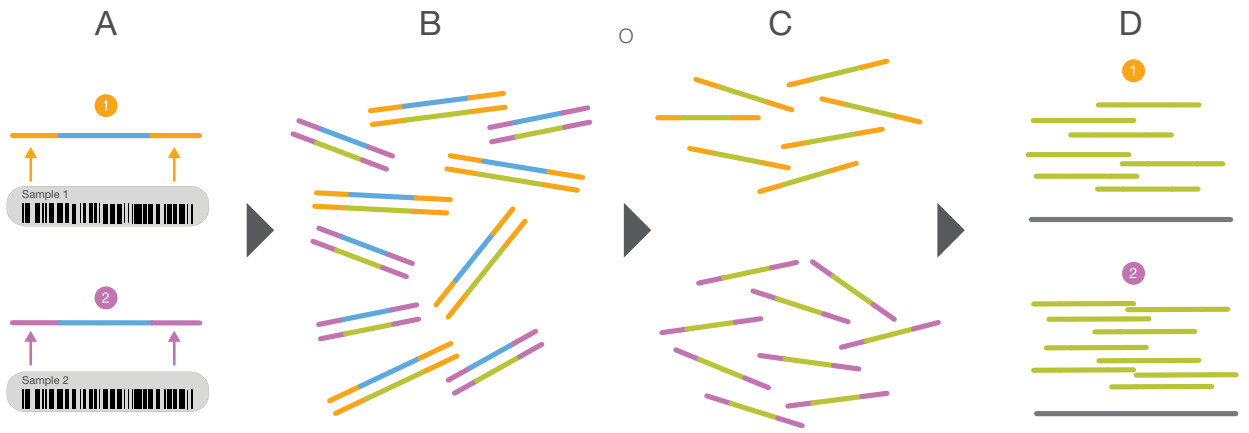
NGS library preparation starts directly from fragmented genomic DNA. Platform-specific oligonucleotide adapters, needed to perform the sequencing biochemistry and index samples, are ligated to each end of the fragments to yield the sequencing-ready library.

Scalable Studies Enabled by Multiplexing

For veterinary diagnostic applications such as sequencing small bacterial or viral genomes, researchers can choose to use a lower output instrument and process a smaller number of samples per run, or opt for a higher output instrument 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 on hundreds of samples 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, more easily, and more cost-effectively than ever before.

Figure 3: Sample Multiplexing Overview

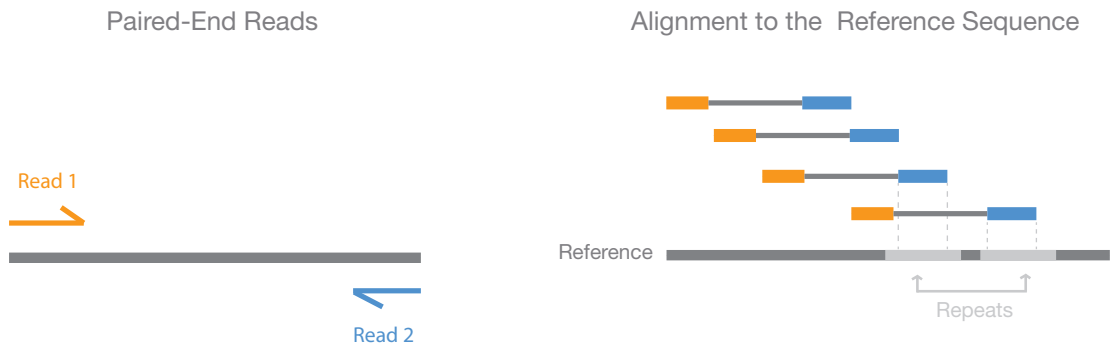


A. Two representative DNA fragments from two unique samples, each attached to a specific barcode sequence that identifies the sample from which it originated.
B. Libraries for each sample are pooled and sequenced in parallel. Each new read contains both the fragment sequence and its sample-identifying barcode.
C. Barcode data 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) and Mate Pair (MP or long-insert, paired-end) sequencing, where both ends of a DNA fragment are sequenced (Figure 4), allow 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 regions or junctions near repetitive genome regions, such as those found in many plant (wheat and maize) genomes. Illumina NGS offers the flexibility of variable insert sizes (typically 400 bp to 5 kb) and read lengths (35–300 bp), allowing high resolution characterization of any genome.

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 or repetitive regions of the genome.

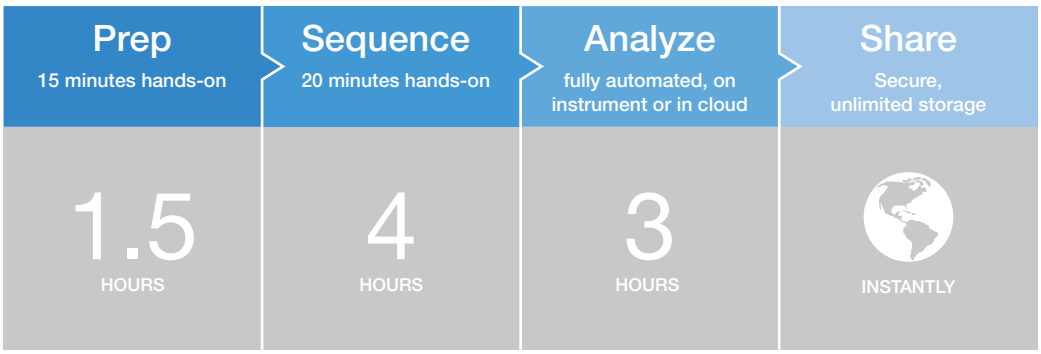
Analyze, Store, and Share in the Cloud

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. 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. 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 as collection of well-known and open-source algorithms from academic institutions.

End-to-End Solution

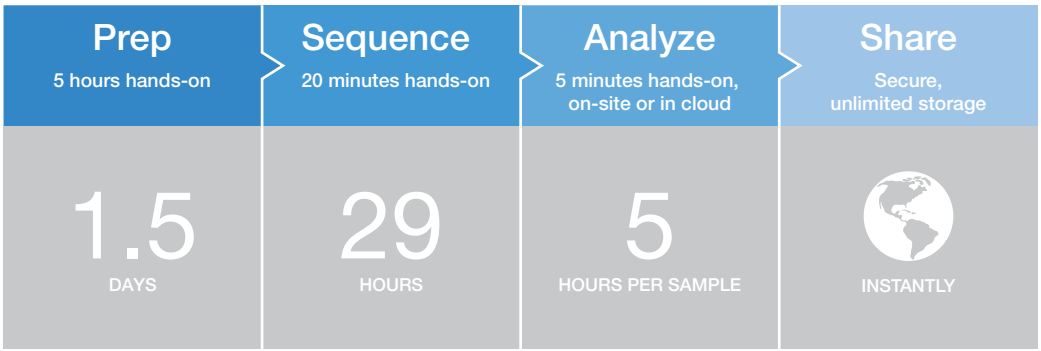
Only Illumina NGS provides a fully supported solution from DNA to results, with specialized library prep choices for the applications you are working on, robust and proven sequencing reagents, and a wide range of simple data analysis tools (Figures 5 and 6).

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



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

Figure 6: Illumina's End-to-End NGS Workflow on the NextSeq 500 System



The NextSeq 500 system workflow delivers highly accurate sequencing data. Data analysis includes alignment and variant calling.

Agricultural Applications

Until recently, sequencing an entire genome was a major endeavor. The ability of NGS to produce a large volume of data in a short period of time makes it a powerful tool for whole-genome sequencing. Its speed and cost-effectiveness also make it an ideal method for targeted sequencing, which focuses on defined regions of the genome.

To accelerate the return on investment in breeding practices, there's an increasing drive towards lower cost methods that will enable the implementation of SNP genotyping tools for routine screening in more agriculture crops. The recent evolution of sequence-based genotyping methods is a direct result of the pressing need for more cost-effective solutions.

Whole-Genome Sequencing

Whether a research project is focused on a novel species or just one that has never been investigated before using genetic tools, *de novo* sequencing is a first step toward understanding the genetic underpinnings of a plant or animal's functions and its interaction with the environment. NGS is fully scalable, capable of sequencing large (wheat and oats), medium (bovine, porcine, and ovine), small (rice), and tiny genomes (viral and bacterial). By contrast, whole-genome sequencing using CE-based Sanger technology requires significant time and resources, even for small genomes. 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.

NGS has also transformed metagenomics, enabling the study of large microbial communities directly in their natural environment without prior culturing. These studies can yield important information about the complex and diverse populations of microbes associated with animal and plant development, from rumen flora that enhance animal digestion to root-associated bacteria involved in nitrogen fixation.

Example: *De novo* Sequencing

One challenge associated with sequencing several plant and animal genomes is the lack of reference genomes available for many 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 help produce longer contigs for *de novo* sequencing by filling gaps in the consensus sequence, resulting in a more complete assembly. The resulting assembled genome can be used to assign map positions and stack diverse breed information, enabling subsequent resequencing to discover SNPs and other genetic variations.

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 most impactful regions of the genome for their research. Amplicon sequencing refers to sequencing selected regions of the genome spanning hundreds of base pairs up to several kilobases. For bovine studies, the Illumina TruSeq® Custom Amplicon Library Preparation kit allows 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.

Example: 16S Metagenomic Sequencing

A common use of amplicon sequencing involves comparing the bacterial 16S rRNA gene, a widely used method for studying phylogeny and taxonomy in prokaryotes. 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 1 illustrates the comparison of 96 samples done on the MiSeq versus CE sequencing for the purpose of comparison.

Table 1: 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 library 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.

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 the Illumina HiSeq®, NextSeq™, and MiSeq® systems.

Next-Generation Genotyping (NGG)

Implementing SNP genotyping tools for routine screening in many agriculture crops relies upon low cost genetic screening methods. The lower the cost, the more researchers can accelerate the return on investment in breeding practice for genetic mapping, screening backcross lines, purity testing, constructing haplotype maps, and performing association and genomic evaluation for plant agrigenomics studies. We use NGG to refer to genotyping using NGS sequencing methods. NGG encompasses methods to discover novel SNPs and genotype them, often simultaneously in many individuals or specimens. For small genomes such as *Drosophila* or high profile research species like *Arabidopsis*, most NGG methods can be completed with standard library preparations and manufacturer multiplexing protocols. Target enrichment or reduction of genome complexity can be employed for larger genomes, requiring homebrew library preparation and optimized multiplexing to

ensure low cost, while offering sufficient overlap in sequencing coverage to call SNPs reliably. NGG benefits for agricultural research vary slightly by method and offer lower cost, reduced ascertainment bias, and in some cases, an ability to perform comparative analysis across samples in the absence of a reference genome. Proven NGG methods include PCR-based approaches (amplicon or long range PCR), restriction enzyme methods (RAD-Seq or published GBS methods by Cornell), and targeted methods of enrichment to capture desired regions of the genome, such as microsatellites, exome, and methylation sensitive sequencing.

Example: RAD-Seq

Restriction site-associated DNA sequencing (RAD-Seq) is a genome complexity reduction method, designed to reliably interrogate a fraction of a target genome instead of the entire genome sequence. This method involves use of a tailored restriction enzyme protocol optimized to characterize SNPs and sequencing depth to meet a given project goal. The short fragments (or RAD tags) of genomic DNA that flank the enzyme recognition site are screened for the presence of genetic variation. NGS compatible RAD fragment libraries enable massively parallel and multiplexed sample sequencing of RAD-tag libraries, facilitating the rapid discovery and genotyping of tens to hundreds of thousands of identical SNPs across large populations.

Discover. Develop. Deploy.

The advent of NGS has enabled agricultural researchers to study livestock, crops, and biological systems at a level never before possible. With clear benefits over Sanger-based CE sequencing, next-generation sequencing can accelerate and enhance agricultural research, advance the development of high-value trait screening methods, and enable the swift deployment of these applications in the real word. 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 agriculture are listed below.

Whole-Genome Sequencing

1. Larkina DM, Daetwylerb HD, Hernandezc AG, Wrightc CL, Hetrickc LA, et al. (2012) Whole-genome resequencing of two elite sires for the detection of haplotypes under selection in dairy cattle. Proc Natl Acad Sci U S A. 109:7693–7698.
2. The International Barley Genome Sequencing Consortium (2012) A physical, genetic and functional sequence assembly of the barley genome. Nature: 10.1038/nature11543. [Epub ahead of print].
3. Bannantine JP, Wu C, Hsu C, Zhou S, Schwartz DC, et al. (2012) Genome sequencing of ovine isolates of *Mycobacterium avium* subspecies paratuberculosis offers insights into host association. BMC Genomics 13:89.
4. Huang X, Feng Q, Qian Q, Zhao Q, Wang L, et al. (2009) High-throughput genotyping by whole-genome resequencing. Genome Res 19: 1068–1076.
5. Metzger ML (2010) Sequencing technologies—the next generation. Nat Rev Genet. 11: 31–46.

De novo Sequencing

6. Liu M, Qiao G, Jiang J, Yang H, Xie L, et al. (2012) Transcriptome sequencing and *de novo* analysis for ma bamboo (*Dendrocalamus latiflorus Munro*) using the Illumina platform. PLoS ONE 7: e46766.
7. Lulin H, Xiao Y, Pei S, Wen T, Shangqin H (2012) The first Illumina-based *de novo* transcriptome sequencing and snalysis of safflower flowers. PLoS ONE 7: e38653.
8. Dejun Li, Zhi Deng, Bi Qin, Xianghong Liu. and Zhonghua Men (2012) *De novo* assembly and characterization of bark transcriptome using Illumina sequencing and development of EST-SSR markers in rubber tree (*Hevea brasiliensis Muell. Arg.*).BMC Genomics 13:192.

Metagenomics

9. Kava AB, Sassona G, Jamia E, Doron-Faigenboima A, Benharb I, et al. (2012) Insights into the bovine rumen plasmidome. Proc Natl Acad Sci U S A.109:5452–5457.
10. Avrani S, Wurtzel O, Sharon I, Sorek R and Lindell D (2011) Genomic island variability facilitates *Prochlorococcus*-virus coexistence. Nature 474: 604–608.

Targeted Sequencing

11. Cronn R, Knaus BJ, Liston A, Maughan PJ, Parks M, et al. (2012) Targeted enrichment strategies for next-generation plant biology. Am J Bot. 99: 291–311.

Next-Generation Genotyping

12. Baird N, Etter P, Atwood T, Currey M, Shiver A, et al. (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. PloS One 3:e3376.
13. Poland JA, Brown PJ, Sorrells ME, Jannink J-L (2012) Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. PLoS ONE 7: e32253
14. Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, et al. (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species.PloS One 6: e19379.
15. Kirst M, Resende M, Munoz P, and Neves L. (2011) Capturing and genotyping the genome-wide genetic diversity of trees for association mapping and genomic selection. BMC Proceedings 5: I7.
16. Peterson BK, Weber JN, Kay EH, Fisher HS, and Hoekstra HE (2012) Double digest RADseq: an inexpensive method for *de novo* SNP discovery and genotyping in model and non-model species. PLoS One 7: e37135.

FOR RESEARCH USE ONLY

Pub No. 770-2012-048 Current as of 11 November 2014

illumina®