Genomic Sequencing

Illumina Sequencing offers an unparalleled combination of read lengths, read depth, and paired-end insert size ranges. These attributes support sequencing of complex genomes and comprehensive characterization of the widest range of structural variants.

Highlights Of Illumina Genomic Sequencing

- **Flexible Platform:**
  Powerful combination of read length and paired read options tailored to individual applications
- **Wide Range of Applications:**
  SNP and structural variant detection, de novo assembly, transcript sequencing, methylation profiling
- **High-Quality Sequence Data:**
  Accurate base-calling for tens of billions of bases per flow cell
- **Efficient Sample Prep:**
  Fast automated workflow and low input requirements to generate data in less than a week

Figure 1: Genome Analyzer And Paired-End Sequencing Module

The Genome Analyzer (right) is the fluidics and imaging device for conducting Illumina sequencing. The Paired-End Module (left) is an optional device that automates paired-end sequencing.

Introduction

Scientists are using the Illumina Sequencing to explore the fullest extent of genetic diversity across various populations[1–3]. By being flexible and easy to use, the Genome Analyzer has made a wide range of genomescale applications routine and is the most widely adopted next-generation sequencing platform (Figure 1).

The Illumina Sequencing reliably generates tens of billions of bases of data per week. The flexible workflow supports any organism and a wide array of application areas. Simple sample preparation methods and robust chemistry support single reads or paired reads with a range of separation distances from 200 bp to 5 kb. Empowered by these capabilities, even the smallest labs are able to greatly expand upon what is known about individual genomes and make breakthrough discoveries from studies in any species.

Illumina sequencing technology is truly versatile, with the potential to transform all genomic research applications. With its powerful combination of features, you can quickly generate the most meaningful data and go where the biology takes you (Table 1).

Read Length Flexibility

The robust chemistry supports a wide range of read lengths. This flexibility allows researchers to tailor each run to their requirements while benefiting from the platform's high raw read accuracy. Highly accurate 75 bp paired reads are more likely to align to a reference or generate larger continuous contigs, so the quality of the entire data set improves.

For applications such as *de novo* sequencing, metagenomics, transcriptomics, and targeted resequencing, longer reads provide increased ability to read through highly repetitive and homologous regions. This effectively increases the proportion of the genome that is mappable, increases the confidence of genomic assembly, and generates greater overall sequencing yields.

Counting applications, such as ChIP-Seq and tag-based expression profiling, can leverage shorter reads to achieve quick turnaround times. For example, 18-cycle tag sequencing runs at high depth can be completed in a single day.

With an unmatched combination of longer reads, high read depth, and flexible read pair spacing, the Genome Analyzer is the ideal platform for a multitude of applications, providing simplified alignment, improved variant detection, and increased genomic coverage.

Coverage Through Depth And Breadth

Maximal sequencing efficiency is achieved as a result of both depth of coverage and uniform read distribution. Illumina sequencing technology draws on its unique combination of hardware, chemistry, and sample preparation techniques to deliver the most useful data in the shortest amount of time. The robust chemistry inherent to Illumina's cluster generation yields a wealth of unique reads, which provide uniform coverage. Using the impressive throughput of the Genome Analyzer, an *E. coli* genome can be sequenced at 430× coverage using only one lane of an eight-lane flow cell.

Table 1: Flexible Paired Sequencing Provides Optimal Detection Of Any Variant

Variant	Single Read	Short Insert Paired-Ends (200–500 bp)	Long Insert Mate Pairs (2–5 kb)	Paired-End And Mate Pair Combined
SNP	++	++++	++	++++
Small indels	++	++++	++	++++
Insertion	+	+++	+++	++++
Amplification	++	+++	+++	++++
Deletion	+	+++	++	++++
Inversion	+	+++	++	++++
Complex rearrangement	+	+++	++	++++
Large rearrangement	+	++	+++	++++

Only by combing short and long inserts can researchers be certain to find all different sizes and types of variants. In particular, short inserts are essential to identifying small indels and mate pairs are essential for identifying the largest rearrangements.

Paired Read Flexibility

Many sequencing applications require a level of overall genomic coverage that can only be achieved through both a high number of alignable reads and the ability to access difficult-to-sequence regions of the genome. For these applications, high-diversity paired reads generate greater genomic information than single reads (e.g., 2×75 bp paired reads rather than 150 bp single reads). Illumina offers a flexible range of paired read separation distances from 200 bp to 5 kb, enabling researchers to optimize each run to their specific goals.

Standard paired-end libraries (200–500 bp) can be used to detect large and small insertions, deletions (indels), inversions, and other rearrangements. Paired-end sequencing also provides greater ability to overcome the obstacles of characterizing repetitive sequence elements by filling in gaps of consensus sequence to achieve complete overall coverage (Figure 2). Moreover, these short-insert paired-end reads are essential for reliable detection of high complexity structural rearrangements, such as inversions within a deleted region (Figure 3) and small indels that would otherwise be undetectable because they lie within the noise of any long insert approach.

Illumina's streamlined long-insert mate pair approach, unlike other protocols, provides the highest fragment diversity relative to starting input material, yielding more uniform sequencing coverage. Mate pair libraries can be generated with insert sizes ranging from 2 to 5 kb, optimal for *de novo* assembly applications including both genome scaffold generation and genome finishing. Long insert libraries generated using Illumina's mate pair protocol also provide an efficient method for identifying large structural variants via sequencing.

Combining short insert paired-end and long-insert mate pair sequencing is the most powerful approach for maximal coverage across the genome. The combination of insert sizes enables detection of the widest range of structural variant types and is essential for accurately identifying more complex rearrangements (Table 1, Figure 3). With its combination of high throughput capability and a broad range of supported insert sizes, Illumina's sequencing technology provides all the tools necessary to sequence even the most difficult-to-access regions of the genome, making it the most versatile sequencing platform available.

Broadest Applications Flexibility

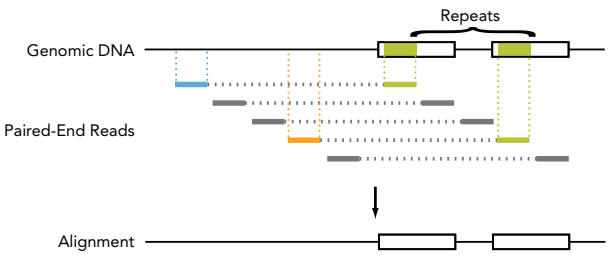
Researchers using the Genome Analyzer can identify an expansive range of genetic variants with a single technology and overcome the obstacles of characterizing many repetitive sequence elements. In addition to accurate and efficient *de novo* assembly and resequencing, the suite of Illumina sequencing sample preparation methods opens the door to many diverse applications. For transcriptome sequencing, paired reads maximize the sensitivity for discovery of polymorphisms, splice variants, alternative promoters and termination sites, and novel genes, as well as accurate transcript quantitation. With minor modifications to the sample preparation procedure, Illumina paired-end sequencing can be extended to BAC end sequencing, di-tag sequencing of cDNA or genomic fragments, bisulfite sequencing, and a wide range of other applications.

High Quality Data

Illumina sequencing provides high-throughput sequence information with industry leading accuracy. Rigorous functional testing ensures robust and reproducible performance. With 2×75 bp paired-end sequencing, the Genome Analyzer consistently generates 12–15 Gb of mappable data, with more than 70% of base calls having Q30 or greater quality scores.

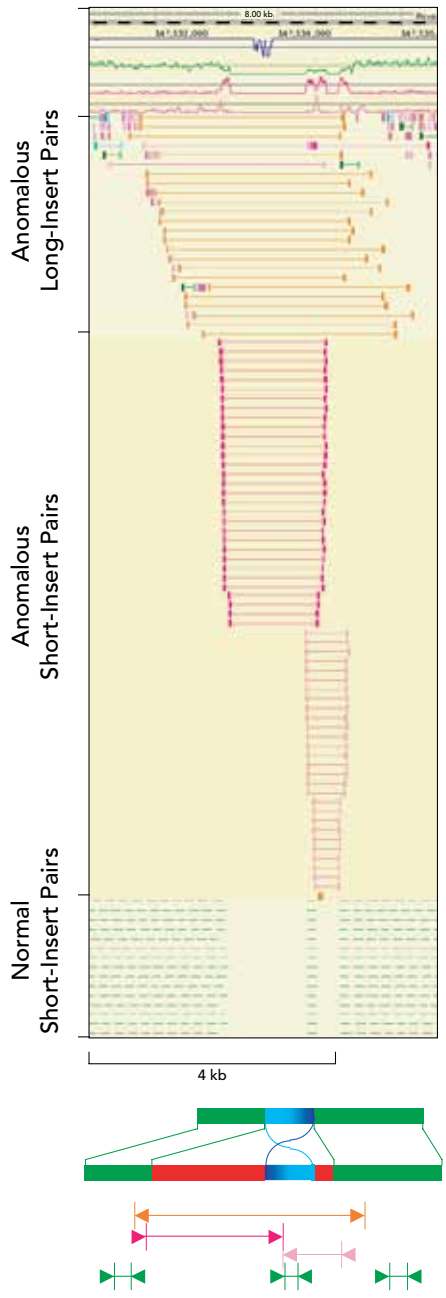
For paired-end sequencing, templates are regenerated between reads to provide equivalent sequencing fidelity across both reads. The end result is consistently high data quality for the entire multi-gigabase data set (Figure 5).

Figure 2: Unique Alignment Of Paired Reads In Repeats



Reads in repeats (green) can be unambiguously aligned in complex genomes. Each read is associated with a paired read (blue or orange) and the separation between read pairs is known from the fragment size of the input DNA.

Figure 3: Read Diversity Enables Discovery Of Complex Rearrangements



This complex rearrangement involves an inversion of 369 bp (blue bar in bottom schematic) flanked by deletions (red bars) of 1,206 and 164 bp, respectively, at the left- and right-hand breakpoints1. Pairs of reads are indicated by color-coded blocks, and DNA fragment inserts are indicated by lines.
The schematic diagram at bottom depicts the arrangement of normal and anomalous read pairs relative to the rearrangement. Top line, structure of NA18507; second line, structure of reference sequence.

Illumina Sequencing Technology

Illumina sequencing technology uses a unique process to generate high-density, massively parallel sequencing runs with reads from one or both ends of tens to hundreds of millions of templates per flow cell. The fully automated Illumina Cluster Station isothermally amplifies DNA on a flow cell surface to create clusters, each containing 500–1000 clonal copies of a single template molecule. The resulting high-density array of templates on the flow cell surface is sequenced with the fully automated Genome Analyzer. Templates undergo sequencing by synthesis in parallel using proprietary fluorescently labeled reversible terminator nucleotides. For paired-end reads, after completion of the first read, the clusters are modified *in situ* to regenerate the template for the paired read. The same clusters are then sequenced using a second sequencing primer to generate the second read (Figures 6B–C).

Simple and Flexible Workflow

Illumina sequencing technology is amenable to a wide range of insert sizes and read lengths. With user-friendly products and streamlined workflows, sample preparation is fast and easy, contributing to customers' rapid successes and Illumina's position at the forefront of next-gen sequencing.

Illumina sample preparation kits are straightforward and use standard molecular biology techniques (described in Figures 6A–C). Paired-end sample preparation methods do not use restriction enzymes to prepare fragments and thus avoid constraints on read length or fragment size, maximizing yield and utility of the data. Single or paired-end read sample preparation can be completed in less than a day by one person and uses minimal starting DNA (one microgram or less). For long-insert paired reads, Illumina offers the simplest mate pair library generation approach with an optimized protocol requiring limited hands-on time for efficient generation of highly diverse libraries.

Sequencing runs are also streamlined and are fully automated. In less than a week, tens of billions of bases of high-quality sequence information can be obtained from hundreds of millions of paired reads in a single run. Researchers can progress from DNA collection to data analysis in as few as three days with Illumina sequencing for the fastest path to discoveries and publication[20].

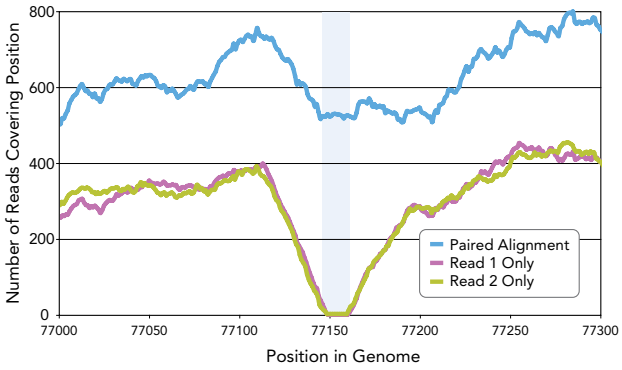
Data Processing and Analysis

The Genome Analyzer system includes a robust analysis software suite. Images from the Genome Analyzer are processed in real time, minimizing the time to results and the need to archive primary data. Illumina's Genome Analyzer Pipeline software produces reads and assigns quality values to each called base. These reads are aligned to a chosen reference for downstream genetic analysis. The open architecture of Illumina's software allows users to customize analysis workflows and to take advantage of a broad array of analysis tools.

Detection of Genetic Variation

Illumina's ELAND alignment algorithm is designed to be fast and is optimized for downstream detection of SNPs. ELAND can match reads to the transcriptome, in addition to the genome, allowing for the identification of splice junctions and novel RNA isoforms in RNA sequencing experiments. Confidence scores are determined for all alignments, and aligned reads from one or many lanes can be imported into the

Figure 4: Paired-End Reads Fill In Sequence Gaps

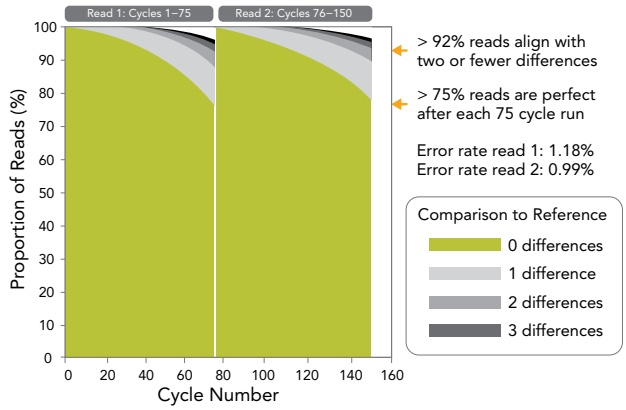


The coverage plot shows that paired reads are aligned across the entire region (blue). If the read-pair information is omitted from the analysis and the same data set is treated as single reads (purple and green), the coverage of aligned reads dips to zero in the plot at the location of a short repeat (blue shaded region).

CASAVA (Consensus Assessment of Sequence and Variation) software package. CASAVA performs secondary analyses (including SNP allele calls from DNA samples or counts of exons, genes, and splice junctions from RNA samples) and exports genomic builds that can be imported into GenomeStudio™ Software or other software packages.

Using Illumina's paired-end technology enables powerful identification of structural variants. In this case, ELAND is used to identify perfectly aligning fragments with aberrant pair separation distances, which is critical to identify insertions, deletions, and more complex rearrangements (Figure 3).

Figure 5: High Accuracy Paired-End Reads



The Genome Analyzer provides a powerful combination of high output quantity and quality. This graph depicts the high per base accuracy profile from a 14.1 Gb run with 2×75 bp paired-end sequencing. Both reads show equivalently high rates of perfect reads (> 75%) and reads with two or fewer differences (> 92%). Results were internally generated using the current Genome AnalyzerII System.

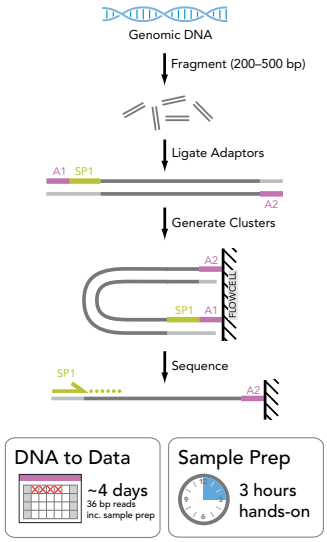
Illumina Sequencing Supports The Broadest Range Of Applications

- **Discover all types of genetic variation:** SNPs, insertions, deletions, copy number variants, and rearrangements[1,5–7]
- **Use targeted sequencing of association or linkage peaks to identify variants that cause disease**
- **Characterize new bacterial isolates by de novo sequencing and re-sequencing[10–12]**
- **Resequence a collection of samples from any population or species[8–10]**
- **Profile DNA methylation status across the entire genome[13–15]**
- **Define somatic variations in cancer[2]**
- **Characterize complex RNA populations for new genes and transcript structures[16,17]**
- **Create new applications enabled by massively parallel sequencing[18,19]**

GenomeStudio Data Analysis Software

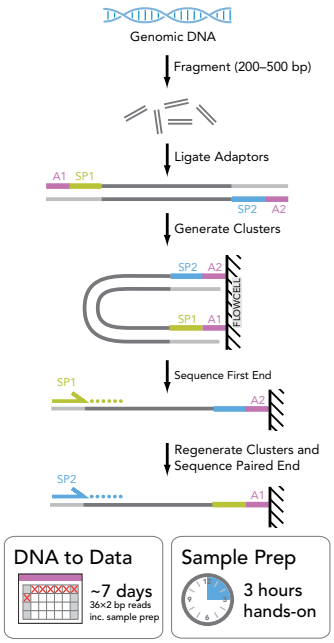
GenomeStudio Software provides integrated data visualization and results analysis for all Illumina assay platforms, including DNA sequencing. Data generated using the Genome Analyzer and Pipeline Software tools can be analyzed to discover and confirm SNPs and chromosomal breakpoint regions. Visualization tools display consensus reads in the reassembled genome and graphically indicate SNPs (Figure 7). Newly discovered SNPs can be exported to use for designing customized iSelect® genotyping arrays.

Figure 6A: Single-Read Sequencing



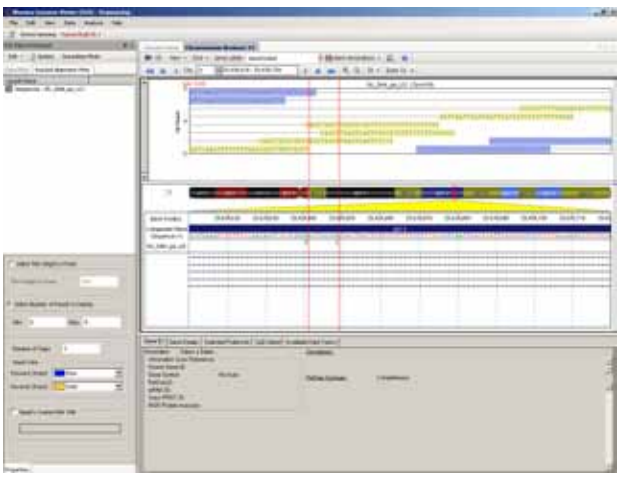
Fragmented sample DNA is size-selected and adaptors are ligated to the ends. Adaptors (A1 and A2) are used to attach fragments to the flow cell, and A1 includes the sequencing primer site (SP1). Libraries are deposited on a flow cell and clusters are generated in the Illumina Cluster Station. Flow cells prepared with template clusters are sequenced in the Genome Analyzer.

Figure 6B: Paired-End Sequencing



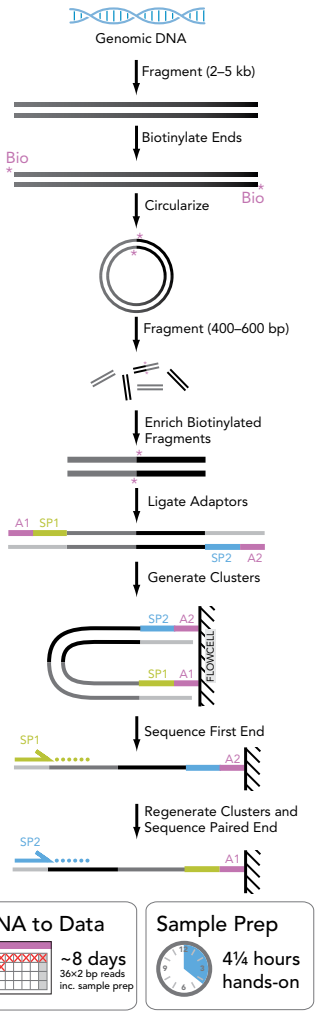
Adapters containing attachment sequences (A1 & A2) and sequencing primer sites (SP1 & SP2) are ligated onto DNA fragments (e.g., genomic DNA). The resulting library of single molecules is attached to a flow cell. Each end of every template is read sequentially.

Figure 7: SNPS Identified From Aligned Reads Displayed In GenomeStudio Software



Aligned sequencing reads (yellow and purple blocks) are stacked on a reference genome in the Illumina Chromosome Browser (ICB). SNPs are identified with red characters and indicated with a ruler (red line) indicating the called SNPs from the stacked aligned reads relative to the reference genome.

Figure 6C: Mate Pair Sequencing



Mate pair library preparation is designed to generate short fragments consisting of two segments that originally had a separation of several kilobases in the genome. Fragments of sample genomic DNA is end-biotinylated to tag the eventual mate pair segments. Self-circularization and refragmentation of these large fragments generates a population of small fragments, some of which contain both mate pair segments with no intervening sequence. These mate pair fragments are enriched using their biotin tag. Mate pairs are sequenced using a similar two-adaptor strategy as described for paired-end sequencing.

The Illumina Informatics Community

As a result of the broad and rapid adoption of the Genome Analyzer, many third-party tools are also available for data analysis and data management. *De novo* assembly of Genome Analyzer sequence reads is supported by a number of software tools, such as Velvet, Abyss, and EULER[20]. The Genome Analyzer Pipeline software output files are used as direct input to these packages or others, including CLCBio Genomics Workbench, Partek, and GenoLogics Geneus.

Go Where The Biology Takes You

Illumina sequencing provides the ideal combination of tunable sequencing read lengths, a broad range of paired-end insert sizes, and balanced genomic coverage. The technology's flexibility and data quality enable the best and broadest range of solutions for studying complex genomes.

Straightforward procedures and convenient reagent kits make sample preparation easy, fast, and reproducible. Template amplification and sequencing are automated and require minimal intervention. Paired-end and mate pair sequencing approaches increase assembly robustness through highly repetitive regions, and enable de novo sequencing of complex genomes, detection of SNPs and indels, and characterization of structural variants.

Be confident in your results using the Genome Analyzer because it provides the most complete range of applications, superior data quality and the throughput required to get your studies published in record time. Leverage the platform's unmatched flexibility to go where the biology takes you.

References

1. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, et al. (2008) Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456: 53–59.
2. Ley TJ, Mardis ER, Ding L, Fulton B, McLellan MD, et al. (2008) DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. Nature 456: 66–72.
3. Wang J, Wang W, Li R, Li Y, Tian G, et al. (2008) The diploid genome sequence of an Asian individual. Nature 456: 60–65.
4. Specifications are based on Genome AnalyzerII available at the time of publication.
5. Campbell PJ, Stephens PJ, Pleasance ED, O'Meara S, Li H, et al. (2008) Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. Nat Genet 40: 722–729.
6. Chen W, Kalscheuer V, Tzschach A, Menzel C, Ullmann R, et al. (2008) Mapping translocation breakpoints by next-generation sequencing. Genome Res 18: 1143–1149.
7. Hillier LW, Marth GT, Quinlan AR, Dooling D, Fewell G, et al. (2008) Whole-genome sequencing and variant discovery in C. elegans. Nat Methods 5: 183–188.
8. Holt KE, Parkhill J, Mazzoni CJ, Roumagnac P, Weill FX, et al. (2008) High-throughput sequencing provides insights into genome variation and evolution in Salmonella Typhi. Nat Genet 40: 987–993.
9. Kistler AL, Gancz A, Clubb S, Skewes-Cox P, Fischer K, et al. (2008) Recovery of divergent avian bornaviruses from cases of proventricular dilatation disease: identification of a candidate etiologic agent. Virol J 5: 88.
10. Salzberg SL, Sommer DD, Puiu D, Lee VT (2008) Gene-boosted assembly of a novel bacterial genome from very short reads. PLoS Comput Biol 4: e1000186.
11. Srivatsan A, Han Y, Peng J, Tehranchi AK, Gibbs R, et al. (2008) High-precision, whole-genome sequencing of laboratory strains facilitates genetic studies. PLoS Genet 4: e1000139.
12. Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18: 821–829.
13. Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, et al. (2008) Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. Nature 452: 215–219.
14. Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, et al. (2008) Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell 133: 523–536.
15. Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, et al. (2008) Genome-scale DNA methylation maps of pluripotent and differentiated cells. Nature 454: 766–770.
16. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet.
17. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, et al. (2008) Alternative isoform regulation in human tissue transcriptomes. Nature 456: 470–476.
18. Craig DW, Pearson JV, Szelinger S, Sekar A, Redman M, et al. (2008) Identification of genetic variants using bar-coded multiplexed sequencing. Nat Methods 5: 887–893.
19. Cronn R, Liston A, Parks M, Gernandt DS, Shen R, et al. (2008) Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. Nucleic Acids Res 36: e122.
20. De Novo Assembly with the Genome Analyzer. http://www.illumina.com/downloads/DeNovoAssembly.pdf
21. http://www.illumina.com/publications.

Ordering Information

Product	Catalog No.
Standard Genomic DNA Sample Prep Kit	FC-102-1001
Standard Cluster Generation Kit (for GA$_{II}$)	GD-203-1001
Paired-End DNA Sample Prep Kit	PE-102-1001
Mate Pair Library Prep Kit	PE-112-1002
Paired-End Cluster Generation Kit (for GA$_{II}$)	PE-203-1001

Illumina, Inc. • 9885 Towne Centre Drive, San Diego, CA 92121 USA • 1.800.809.4566 toll-free • 1.858.202.4566 tel • techsupport@illumina.com • illumina.com

FOR RESEARCH USE ONLY

illumina®

AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
TCAACGTACCGTAACGAACGTATCATTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTT
CGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACC
AACGTACCATTAAGAGCTACCGTGCAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAA
AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
GATTACTTGATCCACTGATTCAACGTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAGCAACGAGC
CGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGAAAAGAATGATAACAGTA