Protein Cell 2010, 1(6): 520–536
DOI 10.1007/s13238-010-0065-3

Protein & Cell

REVIEW

The next-generation sequencing technology and application

Xiaoguang Zhou[1,3] ✉, Lufeng Ren[1,3], Qingshu Meng[1], Yuntao Li[2,3], Yude Yu[2,3], Jun Yu[1,3] ✉

[1] Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100029, China
[2] Institute of Semiconductor, Chinese Academy of Sciences, Beijing 100083, China
[3] The Joint Laboratory of Bioinformation Acquisition and Sensing Technology, Beijing 100029, China
✉ Correspondence: junyu@big.ac.cn (J. Yu), joezhou@big.ac.cn (X. Zhou)
Received May 20, 2010 Accepted May 29, 2010

ABSTRACT

As one of the key technologies in biomedical research, DNA sequencing has not only improved its productivity with an exponential growth rate but also been applied to new areas of application over the past few years. This is largely due to the advent of newer generations of sequencing platforms, offering ever-faster and cheaper ways to analyze sequences. In our previous review, we looked into technical characteristics of the next-generation sequencers and provided prospective insights into their future development. In this article, we present a brief overview of the advantages and short-comings of key commercially available platforms with a focus on their suitability for a broad range of applications.

KEYWORDS next-generation sequencing technology, RNA-seq, ChIP-seq, metagenome, transcriptome, epigenome

INTRODUCTION

DNA sequencing technology has played a pivotal role in the advancement of molecular biology (Gilbert, 1980). Over the past decade, we have witnessed tremendous transformation in this field. The neck-breaking speed of this change has been comparable with the rapid evolution of semiconductor industry under the Moore's law (Moore, 1965; Shendure et al., 2004). The advancement of disruptive sequencing technologies afford us unprecedented high-throughput and low-cost sequencing platforms. The fast and low-cost sequencing approaches not only change the landscape of genomes sequencing projects but also usher in new opportunities for sequencing in various applications—the innovative ways of applying these technologies being created. Over the past few years, we have seen next-generation technologies being applied in a variety of contexts, including *de novo* whole-genome sequencing, resequencing of genomes for variations, profiling mRNAs and other small and non-coding RNAs, assessing DNA binding proteins and chromatin structures, and detecting methylation patterns. Traditional and established techniques such as microarray in functional genomics applications have been increasingly challenged by the high-throughput next-generation sequencing techniques. With so many applications and instrument platforms available on the market, a common question would be: which platform is the best choice for a given biologic experiment? As a follow-up to our previous review article on generations of sequencing technologies (Zhou et al., 2010), here we describe a few widely-used, commercially available next-generation instrument platforms and devote most of the following space to highlight their transforming potential and suitability to various applications. We believe that the synergetic relationship between sequencing technologies and its applications will ensure the continuing push toward faster, cheaper and more reliable approaches of producing sequences into the foreseeable future.

CURRENT NEXT-GENERATION SEQUENCING PLATFORMS

Even though the capillary sequencer that is based on Sanger's chain-termination chemistry and developed in the mid-1970s (Sanger and Coulson, 1975), such as the ABI 3730 (Applied Biosystems), is still viable and good for sequencing PCR products and other small-scale sequencing projects (Mardis, 2008), most sequencing operations today are performed on next-generation instruments. The three dominate commercial platforms currently on the market are the Roche 454 Genome Sequencer, the Illumina Genome Analyzer, and the Life Technologies SOLiD System. To help readers understand the advantages and limitations of these instruments in relating to wide range of applications, a brief overview of their inner workings is in order.

All three platforms were developed at the end of 1990s and commercialized around 2005. They all adopted conceptually similar work flow as outlined in our previous article (Zhou et al., 2010), from template library preparation, template amplification, to parallel sequencing by chain-extension, with variations in array formation, cluster generation and enzyme-based sequencing biochemistry. New breed of sequencing-by-synthesis instruments with single molecule detection is also on the horizon. Recently, Helicos Biosciences has introduced its version of single-molecule sequencing (tSMS), the Helicos Genetic Analysis System (http://www.helicosbio.com/). Pacific Bioscience also introduced its Zero-Mode Waveguide based SMRT technology (http://www.pacificbiosciences.com). Comparison of the next-generation sequencing platforms is summarized in Table 1.

The Roche 454 Genome Sequencer FLX system

The GS FLX system based on sequencing-by-synthesis with pyrophosphate chemistry, was developed by 454 Life Sciences and was the first next-generation sequencing platform available on the market (Margulies et al., 2005). In this system, the DNA sample is first sheared into fragments. Two short adaptors, an A-adaptor and a B-adaptor are then ligated to the fragments. The adaptors provide priming sites for amplification and sequencing, as well as a special key sequence. The B-adaptor also contains a 5′-biotin tag that enables the immobilization of library fragments onto streptavidin-coated magnetic beads. The double-stranded products bound to the beads are then denatured to release the complementary non-biotinylated strands containing both an A- adaptor sequence and a B-adaptor sequence. These denatured strands form the single-stranded template DNA library (Fig. 1A). For DNA amplification, the Genome Sequencer FLX system employs emulsion-based clonal amplification, called emPCR (Dressman, 2003). The single-stranded DNA library is immobilized by hybridization onto primer-coated capture beads. The process is optimized to produce beads where a single library fragment is bound to each bead. The bead-bound library is emulsified along with the amplification reagents in a water-in-oil mixture. Each bead with a single library fragment is captured within its own emulsion microreactor, where the independent clonal amplification takes place. After amplification, the microreactors are broken, releasing the DNA-positive beads for further enrichment (Fig. 1B). For sequencing, the DNA beads are layered onto a PicoTiterPlate device, depositing the beads into the wells, followed by enzyme beads and packing beads. The enzyme beads contain sulfurylase and luciferase, which are key components of the sequencing reaction, while the packing beads ensure that the DNA beads remain positioned in the wells during that sequencing reaction (Fig. 1C). The fluidics sub-system delivers sequencing reagents that contain buffers and nucleotides by flowing them across the wells of the plate. Nucleotides are flowed sequentially in a specific order over the PicoTiterPlate device. When a nucleotide is complementary to the next base of the template strand, it is incorporated into the growing DNA strand by the polymerase. The incorporation of a nucleotide releases a pyrophosphate moiety. The sulfurylase enzyme converts the pyrophosphate molecule into ATP using adenosine phosphosulfate. The ATP is hydrolyzed by the luciferase enzyme using luciferin to produce oxyluciferin and give off light. The light emission is detected by a CCD camera, which is coupled to the PicoTiterPlate device. The intensity of light from a particular well indicates the incorporation of nucleotides (Fig. 1D). Across multiple cycles, the pattern of detected incorporation events reveals the sequence of templates represented by individual beads. The sequencing is ‘asynchronous’ in that some features may get ahead or behind other features depending on their sequence relative to the order of base addition. Raw reads processed by the 454 platform are screened by various quality filters to remove poor-quality sequences, mixed sequences (more than one initial DNA fragment per bead), and sequences without the initiating key sequence. For downstream analysis, three different bioinformatic tools are available: GS *De Novo* Assembler, GS Reference Mapper, and GS Amplicon Variant Analyzer (http://454.com/products-solutions/analysis-tools/index.asp). Using these graphical analysis tools, researchers can quickly obtain biologically informative results from sequence data.

A major limitation of the 454 technology relates to resolution of homopolymer-containing DNA segments, such as AAA and GGG (Rothberg and Leamon, 2008). Because there is no terminating moiety preventing multiple consecutive incorporations at a given cycle, pyrosequencing relies on the magnitude of light emitted to determine the number of repetitive bases. This is prone to a greater error rate than the discrimination of incorporation versus nonincorporation. As a consequence, the dominant error type for the 454 platform is insertion-deletion, rather than substitution. Another disadvantage of 454 sequencing platform is that the per-base cost of sequencing is much higher than that of other next-generation platforms, e.g., SOLiD and Solexa (Rothberg and Leamon, 2008). It is therefore unsuitable for sequencing targeted fragments from small numbers of DNA samples, such as those for phylogenetic analysis. Comparing to other next-generation platforms, the key advantage of the 454 platform is its read length (Metzker, 2010). The 454 System can generate more than 1,000,000 individual reads with improved Q20 read length of 400 bases per 10 h instrument

Table 1 Comparison of next-generation sequencing platforms (Metzker, 2010)

platform	library	sequencing principle	read length (bases)	run time (days)	Gb per run	machine cost ($)	pros	cons	biological applications
Roche 454	fragment/emulsion PCR	pyrosequencing	400	0.4	0.5	500,000	longer reads improve mapping in repetitive regions, fast run times	high reagent cost, high error rates in homopolymer repeats	bacterial and insect genome *de novo* assemblies, medium scale (< 3 Mb) exome capture, 16S in metagenomics
Illumina GAII	fragment/polony	sequencing by synthesis	75	$4^a/9^b$	$18^a/35^b$	540,000	currently the most widely used platform in the field	low multiplexing capability of samples	variant discovery by whole-genome resequencing or whole-exome capture, gene discovery in metagenomics
Life Technologies SOLiD3	fragment/emulsion PCR	sequencing by ligation	50	$7^a/14^b$	$30^a/50^b$	595,000	two-base encoding provides inherent error correction	long run times	variant discovery by whole-genome resequencing or whole-exome capture, gene discovery in metagenomics
Helicos Biosciences Heliscope	fragment/single molecule	sequencing by synthesis	32	8	37	999,000	non-bias representation of templates for genome and seq-based applications	high error rates compared with other reversible terminator chemistries	seq-based methods
Pacific Bioscience RS system	fragment/single molecule	sequencing by synthesis/real time	1100	N/A	13	N/A	has the greatest potential for reads exceeding 1 kb	highest error rates compared with other NGS chemistries	full-length transcriptome sequencing, complements other resequencing efforts in discovering large structural variants and haplotype blocks

[a] Fragment run.
[b] Mate-pair run.

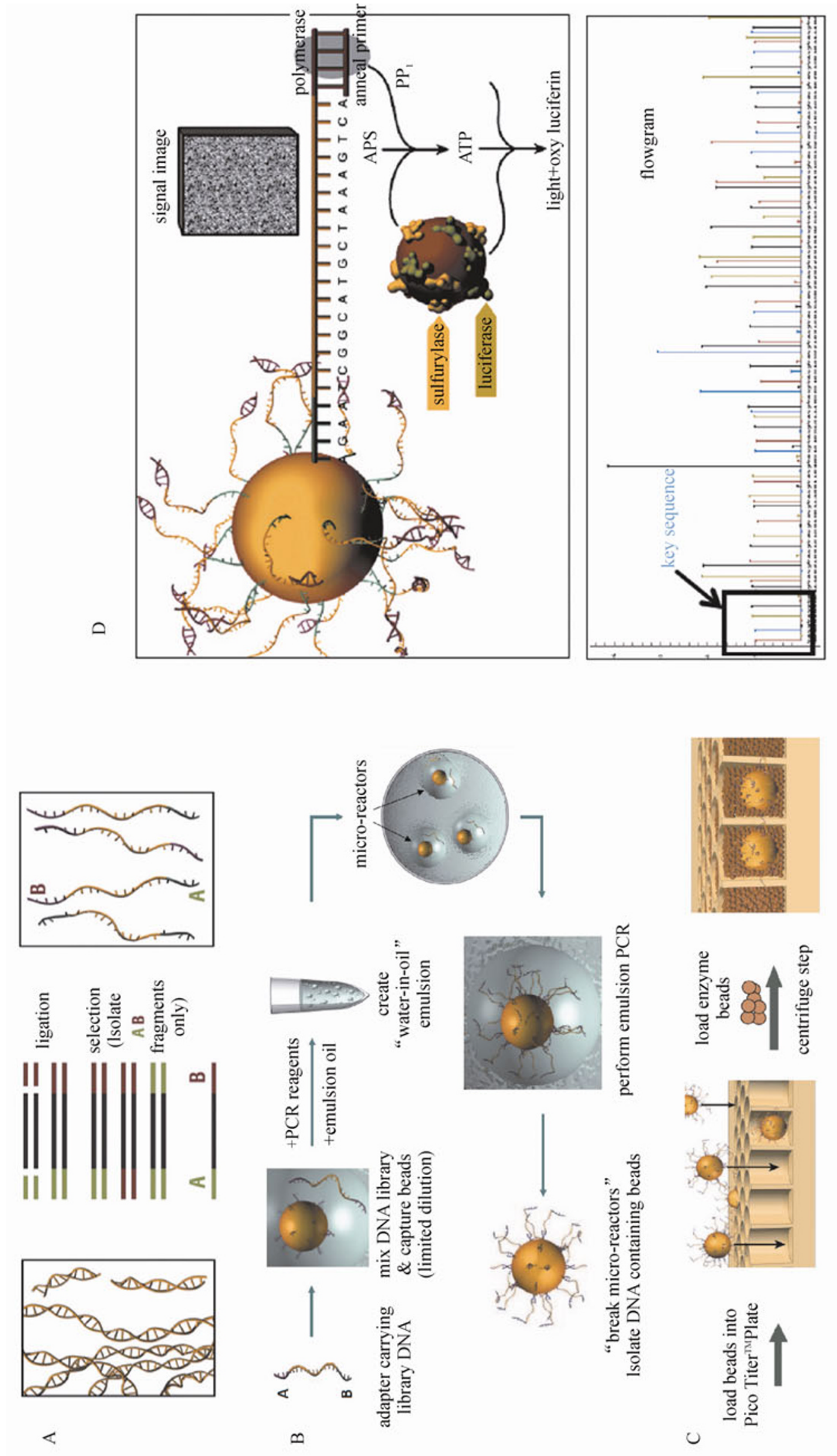

Figure 1. The GS FLX system working principle. (A) Prepare adapter ligated ssDNA library (A-[insert]-B). (B) Emulsion based clonal amplification. (C) Depositing DNA beads into the PicoTiter™ plate. (D) Sequencing and base calling. (http://www.454.com)

run. It may be a best choice for certain applications where long read-lengths are critical, such as *de novo* assembly and metagenomics. The company is also projected to place a low-throughput version (1/10 of the GS throughput) of the instrument GS Junior, into the market later this year, which is ideal for sequencing bacterial genomes and yields about 20 × coverage of a typical bacterial genome, ~5 Mb in size (http://www.gsjunior.com/).

The Illumina (Solexa) Genome Analyzer

The Solexa sequencing platform was commercialized in 2006. The working principle (Fig. 2) is sequencing-by-synthesis chemistry. Input DNA is fragmented by hydrodynamic shearing to generate < 800 bp fragments. The fragments are blunt ended and phosphorylated, and a single 'A' nucleotide is added to the 3′-ends of the fragments. Then DNA fragments are ligated at both ends to adapters that have a single-base 'T' overhang. After denaturation, DNA fragments are immobilized at one end on a solid support-flow cell. The surface of the flow cell is coated densely with the adapters and the complementary adapters. Each single-stranded fragment that is immobilized at one end on the surface creates a 'bridge' structure by hybridizing with its free end to the complementary adapter on the surface of the flow cell. The adapters on the surface also act as primers for the following PCR amplification. Adding mixtures containing the PCR amplification reagents to the flow cell surface, the DNA fragments are amplified by "bridge PCR" (Adessi, 2000; Fedurco et al., 2006). After several PCR cycles, about 1000 copies of single-stranded DNA fragments are created on the surface, forming a surface-bound colony (the cluster). The reaction mixture for the sequencing chemistry and DNA synthesis is supplied onto the surface, which contains four reversible terminator nucleotides, each labeled with a different fluorescent dye. After incorporation into the DNA strand, the terminator nucleotide as well as its position on the support surface are detected and identified via its fluorescent dye by the CCD camera. The terminator group at the 3′-end of the base and the fluorescent dye are then removed from the base and the synthesis cycle is repeated. This series of steps continues for a specific number of cycles, as determined by user-defined instrument settings. A base-calling algorithm assigns sequences and associated quality values to each read and a quality checking pipeline evaluates the Illumina data from each run, removing poor-quality sequences.

In 2008, Illumina introduced an upgrade, the Genome Analyzer II, to its predecessor, which offered a powerful combination of the cBot and Paired-End Module (http://www.illumina.com/systems/genome_analyzer.ilmn). cBot is a revolutionary automated system that creates clonal clusters from single molecule DNA templates, preparing them for sequencing by synthesis on the Genome Analyzer. The Paired-End Module is a fluidics station that attaches to the Genome Analyzer. After completion of the first read, the templates can be regenerated *in situ* to prepare for the second round of sequencing from the opposite end of the fragments. First, the newly sequenced strands are stripped off and the complementary strands are bridge amplified to form clusters. Once the original templates are cleaved and removed, the reverse strands undergo sequencing by synthesis. The Paired-End Module enables paired-end sequencing up to 2 × 100 bp for fragments ranging from 200 bp to 5 kb. For Genome Analyzer II, the run time is highly decreased and the output per paired-end run can reach 45–50 Gb (gigabasepairs). Compared to Sanger sequencing, the Illumina system is able to produce more data at a reduced time and cost; however, error rates are higher (often resulting in false-positive when identifying sequence variations) and reads are shorter (Metzker, 2010). Usually error rate can be overcomed by coverage but contiguity is rather limited by the read length as the Lander-Waterman Curve describes (Lander and Waterman, 1988). Illumina also offers a newer version of its next-generation sequencer, HiSeq2000, which has a dual flowcell system installed to increase the efficiency of data generation (http://www.illumina.com/systems/hiseq_2000.ilmn).

The Life Technologies SOLiD system

The Life Technologies SOLiD system is based on a sequencing-by-ligation technology. This platform has its origins in the system described by Shendure et al. (2005) and in work by McKernan and colleagues (2006) at Agencourt Personal Genomics (acquired by Applied Biosystems in 2006). The generation of a DNA fragment library and the sequencing process by subsequent ligation steps are shown in Fig. 3. In this technology, two types of libraries—fragment or mate-paired library can be constructed depending on the researchers' purposes. Then DNA fragments are ligated to adapters and bound to beads. DNA fragments on the beads are amplified by the emulsion PCR, after which the templates are denatured and bead enrichment is performed to select beads with extended templates. The template on the selected beads undergoes a 3′ modification to allow covalent bonding to the glass slide. The sequencing methodology is based on sequential ligation with dye-labeled oligonucleotides (Housby and Southern, 1998). In the first step, a primer is hybridized to the adapter sequence within the library template. Next, a set of four fluorescently labeled oligonucleotide octamers compete for ligation to the sequencing primer. In these octamers, the first and second di-base are characterized by one of four fluorescent labels at the end of the octamer. After the detection of the fluorescence from the label, bases 1 and 2 in the sequence are thus determined. The ligated octamer oligonucleotides are cleaved off after the fifth base, removing the fluorescent label, then hybridization and ligation cycles are repeated, this time determining bases 6 and 7 in the

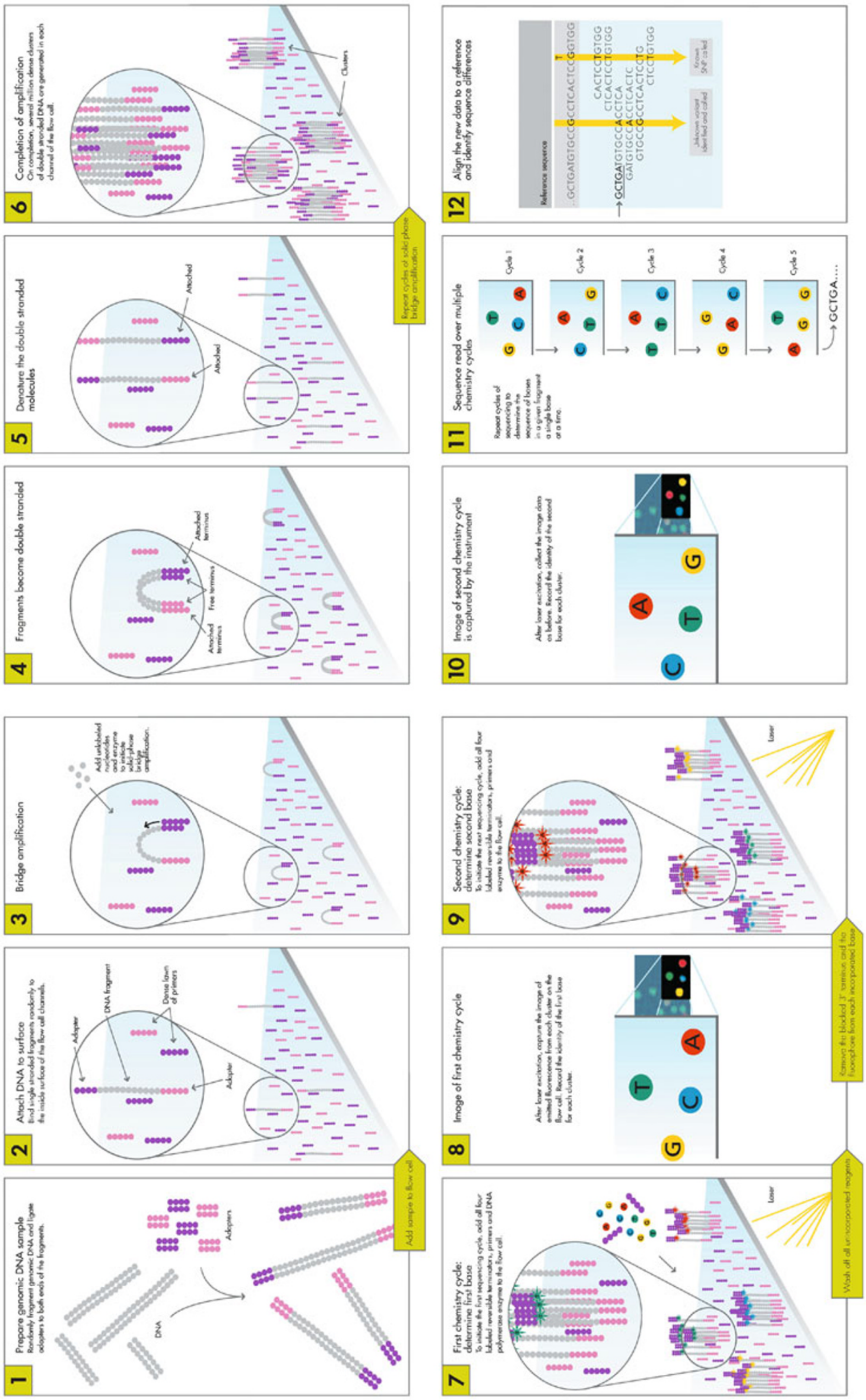

Figure 2. The working principle of the Solexa sequencing platform. (http://www.illumina.com)

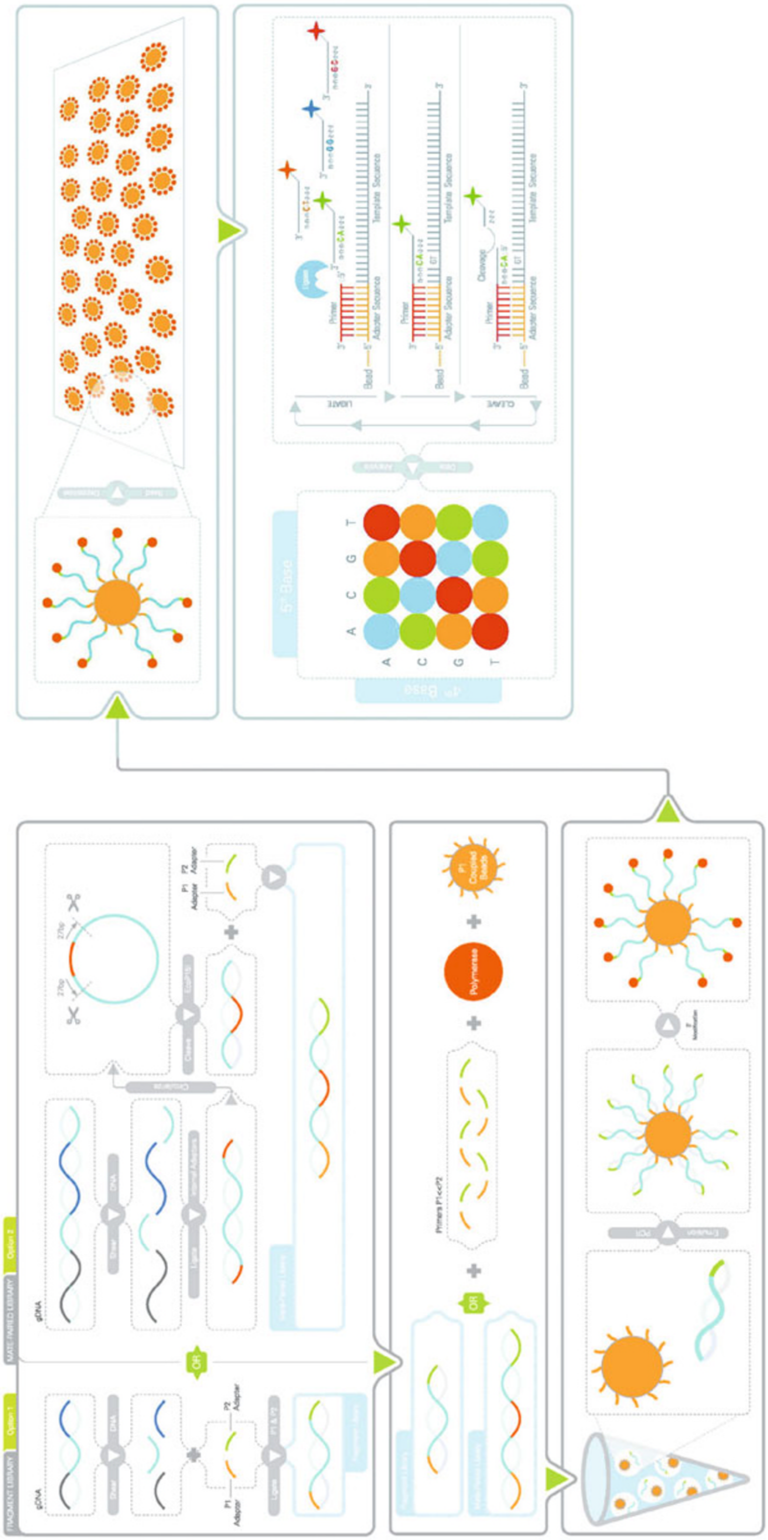

Figure 3. The generation of a DNA fragment library and the sequencing process by subsequent ligation steps of SOLiD. (http://www.appliedbiosystems.com)

sequence; in the subsequent cycle bases 11 and 12 are determined, etc. Progressive rounds of octamer ligation enable sequencing of every five bases. Following a series of ligation cycles, the extension product is removed and the template is reset with another octamer complementary to the n-1 position for a second round of ligation cycles. After five rounds of primer reset completed for each sequence tag, each base is interrogated in two independent ligation reactions by two different primers. This method is called 'Two Base Encoding' (Mckernan et al., 2006). Two Base Encoding is a unique and powerful approach designed to clearly discriminate measurement errors. The combination of ligase enzymology, primer reset and two-base encoding all contribute to the low error rate and reduced systemic noise.

In December 2009, Applied Biosystems has updated the platform to SOLiD™ 3 Plus (http://www3.appliedbiosystems.com/AB_Home/applicationstechnologies/SOLiD-System-Sequencing-B/index.htm). The SOLiD™ 3 Plus System can generate more than 60 Gb of mappable sequence or greater than 1 billion reads per run. The SOLiD™ 4 system is claimed to produce comparable amount of raw data as its competitors claimed. The cost of the instrument is substantially lower than that of other next-generation sequencing platforms. The current read length, however, significantly limits its applications (Metzker, 2010).

Single molecule sequencing

The HeliScope Genetic Analysis System by Helicos Biosciences, based on the work from Quake's group (Braslavsky et al. 2003), is the first single molecule sequencing system and recently available on market. It utilizes sequencing by synthesis on single molecule. Constructed single-stranded DNA library is disorderly arrayed on a flat substrate without any amplification. DNA polymerase and one of four fluorescently labeled nucleotides are flowed into the system at each sequencing cycle. Strands in the array that have undergone template-directed base extension are lightened up by fluorescent label, which are recorded with a CCD camera. After washing, fluorescent labels on the extended strands are chemically removed, and another cycle of base extension repeats. Like in pyrosequencing with 454, each iterative cycle is asynchronous—some strands in the array may pull ahead, fall behind, or completely fail to extend all together, leading to similar problem with homopolymer run. However, unlike Roche 454 platform, single molecule affords us mitigation by playing trick with enzyme kinetics to slow down the rate of chain extension so to reduce the chance of two consecutive base incorporations before being washed away (Harris et al. 2008).

One of the key challenges this technology faces is the raw sequencing accuracy due to the difficulty with detecting single molecule event. Therefore, the dominant error type with this instrument is deletion. A two-pass strategy can somewhat mitigate the shortcoming. Single molecule sequencing means that we can now reset the tethered template DNA to its original state by lifting off the newly extended strand after one sequencing run. Another sequencing pass can then be performed in opposite direction from distal adaptor, yielding a second sequence from the same template. The duplicated sequences can be used to average out detection errors, and thus, give rise to much higher accuracy than otherwise.

Pacific Biosciences is another company currently developing the single molecule sequencing technology, the SMRT (Single Molecule Real Time) technology (Eid et al., 2009). The SMART technology relies on a nano-structure, the Zero Mode Waveguide (ZMW), for real-time observation of DNA polymerization (Levene et al., 2003). ZMW chip consists of thousands upon thousands of sub-wavelength holes, tens of nanometers in diameter, fabricated by perforating a thin metal film supported by transparent substrate. When illuminated from the side of glass, light is not able to penetrate through the hole but leaves an exponentially-decayed evanescent wave at very bottom of each hole, and thus, creating a very small volume of fluorescence detection. Furthermore, the DNA polymerase is planted to the bottom of each waveguide. During a sequencing assay, each time a fluorescently labeled base is grabbed on by the polymerase, it brings fluorophore to the detection volume, creating a burst of fluorescent light. If the nucleotide is complementary to the template strand, it will go through a time-consuming synthesis process, and therefore, stay in the detection volume longer until the fluorescent moiety is released as part of pyrophosphate. The color-coded fluorescence burst and its duration reveal the identity of the complementary base on template DNA. By continuously following the bursts of fluorescence at each waveguide in real-time, sequences of template DNA can be rapidly determined.

This technology has the great potential to achieve high speed with long read length. However, error stemmed from real-time single-molecule detection might put a damper on its raw accuracy as with other SMS-based sequencing platforms. Current CCD technology has also limited the maximum ZMW chip area that can be simultaneously observed. Low yield ratios (~30%) of polymerase-occupied waveguide further limit the number of useable wells on the chip (Korlach et al., 2008). Even with this limitation, the first version of the instrument when introduced has promised a read length of no less than 1500 bp, at a speed of 15 min per run, and with reagent cost no more than $60 per run. It is anticipated that future version of this platform, after technical issues are resolved, could churn out 100 Gb of data per day with read length up to 100,000 bp.

APPLICATION OF NEXT-GENERATION SEQUENCING TECHNOLOGIES

The production of low cost reads by next-generation sequencing technologies makes them useful in a variety of areas (Table 2). Important applications include: (1) *de novo* genome

Table 2 Applications of next-generation sequencing technologies

category	examples of applications	references
genome	*de novo* sequencing: the initial generation of large eukaryotic genomes	Velasco et al., 2007 Diguistini et al., 2009 Huang et al., 2009 Li et al., 2010
	whole-genome resequencing: comprehensive SNP, indels, copy number and structual variations in individual human genomes	Bentley, 2006 Ossowski et al., 2008 Denver et al., 2009 Xia et al., 2009
	targeted resequencing: targeted polymorphism and mutation discovery	Hodges et al., 2007 Porreca et al., 2007 Harismendy et al., 2009
trancriptome	quantification of gene expression and alternative splicing; transcript annotation; discovery of transcribed SNPs or somatic mutations	Axtell et al., 2006 Sultan et al., 2008 Sugarbaker et al., 2008 Jacquier, 2009
	small RNA profiling	Berezikov et al., 2006 Houwing et al., 2007
epigenome	transcription factor with its direct targets	Johnson et al., 2007 Robertson et al., 2007
	genomic profiles of histone modifications	Impey et al., 2004 Mikkelsen et al., 2007
	DNA methylation	Cokus et al., 2008 Costello et al., 2009
	genomic profiles of nucleosome positions	Fierer et. al., 2006 Johnson et al., 2006
metagenome	environmental	Edwards et al., 2007 Hubert et al., 2007
	human microbiome	Turnbaugh et al., 2007 Qin et al., 2010

sequencing, whole-genome resequencing or more targeted sequencing for discovery of mutations or polymorphisms; (2) transcriptome analysis and cataloguing, where shotgun libraries derived from mRNA or small RNAs are deeply sequenced; (3) large-scale analysis of DNA methylation, by deep sequencing of bisulfite-treated DNA; (http://454.com/products-solutions/analysis-tools/index.asp) genome-wide mapping of DNA-protein interactions, by deep sequencing of DNA fragments pulled down by chromatin immunoprecipitation (ChIP-Seq); (http://454.com/products-solutions/analysis-tools/index.asp) species classification and/or gene discovery by metagenomics and pangenomics. As mentioned previously, there are different advantages and limitations among the next-generation platforms in respect to specific applications.

De novo sequencing and assembly

De novo sequencing is the initial generation of the primary genomic sequence of a particular organism. A detailed genetic analysis of any organism is possible only after *de novo* sequencing has been performed (Goldberg et al., 2006; Durfee et al., 2008; Reinhardt et al., 2009), i.e., reference sequence produced. Until March 1, 2010, there have been 740 eukaryotic genome sequencing projects submitted to NCBI (Table 3), while only 23 genomes are completed, and most of them are in draft assemblies or work-in-progress (http://www.illumina.com/systems/genome_analyzer.ilmn). Four of these organisms are sequenced using next-generation sequencing technologies independently or in combination with the traditional Sanger method (Velasco et al., 2007; Diguistini et al., 2009; Huang et al., 2009; Li et al.,

Table 3 Statistics of eukaryotic genome sequencing projects (data tabulated on March 1, 2010)

organism group	complete	assembly	in progress	sum
animals	4	137	146	287
plants	3	23	85	111
fungi	10	120	93	223
protists	6	49	64	119
total	23	329	388	740

2010) (Table 4). A hybrid Sanger/pyrosequencing approach resolved a complex heterozygous grape genome, where consensus sequence of the genome and a set of mapped marker loci were generated. This is the first project that utilizes both the long Sanger and short SBS reads to assemble the genome sequence of a large eukaryotic genome (Velasco et al., 2007). A draft sequence of the giant panda genome was successfully generated and assembled based on next-generation sequencing technology alone, taking the advantage of excellent colinerarity of the mammalian genomes. The assembled contigs (2.25 Gb) cover approximately 94% of the whole genome and the remaining gaps (0.05 Gb) seem to contain carnivore-specific repeats and tandem repeats (Li et al., 2010). When taking on large genomes, i.e., over 1 Gb in total length, one should be more cautious in designing a sequencing experiment since the effort could be severely hampered by polyploidy and large repetitive fraction. Nevertheless, successful sequencing projects demonstrate the feasibility of using next-generation sequencing technologies for accurate, cost-effective, and rapid *de novo* assembly of large eukaryotic genomes (Imelfort and Edwards, 2009; Turner et al., 2009a).

With its long read lengths and high accuracy, capillary electrophoresis-based sequencing has been the gold standard for *de novo* genome sequencing projects in the past decades. However, the throughput of these systems makes *de novo* assembly of most organisms a lengthy and costly endeavor. Next-generation sequencing technologies hold great promise in reducing the time and cost. Compared to just a few years ago, it is now much easier and cheaper to sequence entire genomes, and a wide variety of species are being studied using these advanced tools every day.

Whole-genome or targeted resequencing

By far, the most common use of next-generation sequencing platform has been resequencing (Davies, 2007). To identify single nucleotide polymorphisms, indels, copy number and structural variations, multiple individuals or strains, or a population-based sampling of a species have to be resequenced (Bentley, 2006; Ossowski et al., 2008; Denver et al., 2009; Xia et al., 2009; Pleasance et al., 2010). In humans, such an endeavor has already commenced with the publication of several complete genomes (Table 5), with the list

Table 4 *De novo* eukaryotic genomes sequencing using next-generation technologies

organism	group	genome size (Mb)	Chr	status	method	platform	depth	references
Ailuropoda melanoleuca	animals	2460	21	assembly	WGS[a]	Solexa	56 ×	Li et al., 2010
Grosmannia clavigera	fungi	32.5		assembly	WGS	Sanger, 454 and Solexa	50 ×	Diguistini et al., 2009
Vitis vinifera	plants	500	19	assembly	WGS	Sanger and 454	11 ×	Velasco et al., 2007
Cucumis sativus	plants	367	7	assembly	W&C[b]	Sanger and Solexa	72.2 ×	Huang et al., 2009

[a] Whole genome shotgun.
[b] Whole genome shotgun combined with clone-based method.

Table 5 Sequencing statistics of six individual human genomes

platform	individual	No. of reads (millions)	read length (bases)	read coverage	genome coverage (%)	SNPs (millions)	No. of runs	estimated cost (US$)	references
Sanger	J. Craig Venter	31.9	800	7.5 ×	N/A	3.21	>340,000	70,000,000	Levy et al., 2007
Roche 454	James D. Watson	93.2	250	7.4 ×	95	3.32	234	1,000,000	Wheeler et al., 2008
SOLiD	James R. Lupski	238	35	29.6 ×	99.8	3.42	3	75,000	Lupski et al., 2010
Illumina Solexa	Yoruba male (NA18507)	3681	35	40.6 ×	99.9	4	40	250,000	Pushkarev 2009
	Han Chinese male (YH)	2950	35	36 ×	99.9	3.07	35	500,000	Wang et al., 2008
	Korean male (SJK)	1647	35, 74	29.0 ×	99.9	3.44	15	250,000	Ahn et al., 2009
	Korean male (AK1)	1910	36, 88, 106	27.8 ×	99.8	3.45	30	200,000	Kim et al., 2009
Helicos	Stephen R. Quake	2725	32	28 ×	90	2.81	4	48,000	Pushkarev et al., 2009

growing by the day. The first is from J. Craig Venter and achieved using traditional Sanger sequencing methods (Levy et al., 2007) as part of the Human Genome Project and the second is from James D. Watson, which was sequenced using the Roche 454 technology to 7.5 × genome coverage. The reads were aligned to the NCBI reference sequence using a combination of the BLAT and Smith-Waterman algorithms. The sequence differs from the reference at 3.32 Mb, of which 2.7 Mb are known differences (Wheeler et al., 2008). The next four human genome sequences are from a Chinese (Wang et al., 2008), an African (Pushkarev et al., 2009), and two Korean individuals (Ahn et al., 2009; Kim et al., 2009); all were done using the Illumina Genome Analyzer and sequenced to around 20 × haploid genome coverage with the exception of the African male's genome which was also resequenced on ABI SOLiD system (McKernan et al. 2009). For all four genomes, reads covered more than 99% of the NCBI human reference genome, revealing approximately 3 million SNPs. More recently, James Lupski's genome was sequenced to 30 × base coverage using ABI's SOLiD System (Lupski et al., 2010). Resequencing of human genome was not limited to the 2nd-generation platforms. Steven Quake's genome, for example, was sequenced to 90% genome coverage on Helicos' single-molecule sequencing platform (Pushkarev et al., 2009).

Target-region resequencing refers to sequencing a targeted region of a species' genome from multiple individuals; it enables scientists to investigate variations of interested genomic regions or genes with high coverage and lower cost (Harismendy et al., 2009). Two methods of target-region resequencing are widely used: PCR-based candidate gene (Dracatos et al., 2009; Goossens et al., 2009; Harismendy and Frazer, 2009; Tewhey et al., 2009) and whole exome approaches (Hodges et al., 2007; Porreca et al., 2007; Choi et al., 2009; Turner et al., 2009b). Ji and colleagues developed a procedure for massive parallel resequencing of multiple human genes. It combines a highly multiplexed and target-specific amplification process with a parallel sequencing technology (Dahl et al., 2007). They demonstrated parallel resequencing of 10 cancer genes covering 177 exons with average sequence coverage per sample of 93%. Through exome sequencing, Bamshad et al. discovered the gene for a rare mendelian disorder of unknown cause, the Miller syndrome (Ng et al., 2010), which demonstrates that exome sequencing of a small number of unrelated affected individuals is a powerful, efficient strategy for identifying the genes underlying rare mendelian disorders and will likely transform the genetic analysis of monogenic traits.

Whole transcriptome shotgun sequencing: RNA-Seq

The transcriptome is the complete set of transcripts in a cell, and their quantity, for a specific developmental stage or physiologic condition (Jacquier, 2009). Understanding the transcriptome is essential for interpreting the functional elements of the genome and revealing the molecular constituents of cells and tissues, and also for understanding development and disease. The specific aims of transcriptomics are: (1) to catalog all transcripts in a context of cell types for a species, including mRNAs, non-coding RNAs and small RNAs; (2) to determine the transcriptional structure of genes, in terms of their start sites, 5′- and 3′-ends, splicing patterns and other post-transcriptional modifications; and (3) to quantify the expression levels of each transcript during development or under different physiologic and pathological conditions. With the availability of faster and cheaper next-generation sequencing platforms, more transcriptomic analyses are performed using a recently-developed deep sequencing approach, RNA-Seq (Wang et al., 2009). Studies using this method have already altered our view of the extent and complexity of eukaryotic transcriptomes (Cloonan et al., 2008; Mortazavi et al., 2008; Sugarbaker et al., 2008; Sultan et al., 2008; Tang et al.,2010).

The current gold standard for protein-coding gene annotation is EST or full-length cDNA sequencing followed by alignment to a reference genome, but it has been estimated that most EST studies using Sanger sequencing detect only about 60% of transcripts in the cell, which fails to cover the poorly expressed or long transcripts (Brent, 2008). This information gap can be addressed using the next-generation sequencing technologies, which have been used to generate transcriptomes for many species and tissues (Mortazavi et al., 2008; Nagalakshmi et al., 2008; Sultan et al., 2008). For instance, a study used the 454 technology to generate 391,157 EST reads from the brain transcriptome of the wasp *P. metricus* (Toth et al., 2007). The reads were then aligned to the genome sequence and EST resources from the honeybee, *Apis mellifera*, to annotate *P. metricus* transcripts. Interestingly, the study found wasp EST matches to 39% of the honeybee mRNAs and observed a strong correlation between the expression levels of the corresponding transcripts from the two species.

The short reads produced by high-throughput next-generation technologies, particularly Illumina and SOLiD, are arguably suitable for gene expression profiling based on tens of millions of short reads rather than tens of thousands of based on the Sanger method. RNA-Seq has been used to accurately monitor gene expression during yeast vegetative growth (Nagalakshmi et al., 2008), yeast meiosis (Wilhelm et al., 2008) and mouse embryonic stem-cell differentiation (Cloonan et al., 2008), to track gene expression changes during development, and to provide a 'digital measurement' of gene expression difference among different tissues.

Before the advent of transcriptome shotgun sequencing, the starts and ends of most transcripts had not been precisely resolved and the extent of spliced heterogeneity remained poorly understood. RNA-Seq, with its high resolution and sensitivity, has revealed many novel transcribed regions and splicing isoforms of known genes. It also helps to map 5′- and 3′-boundaries of many genes. Using RNA-seq method, the 5′- and 3′-boundaries of 80% and 85% of all annotated genes, respectively, were mapped in *S. cerevisiae* (Nagalakshmi et al., 2008). Similarly, in *S. pombe* (Wilhelm et al., 2008) many boundaries were defined by RNA-Seq data in combination with tiling array data. In humans, 31,618 known splicing events were confirmed (11% of all known splicing events) and 379 novel splicing events were discovered (Morin et al., 2008a). In mice, extensive alternative splicing was observed for 3462 genes (Mortazavi et al., 2008). In addition, results from RNA-Seq suggest the existence of a large number of novel transcribed regions in every genome surveyed, including those of *A. thaliana* (Lister et al., 2008), mouse (Cloonan et al., 2008; Mortazavi et al., 2008), human (Morin et al., 2008a), *S. cerevisiae* (Nagalakshmi et al., 2008) and *S. pombe* (Wilhelm et al., 2008). These novel transcribed regions, combined with many undiscovered novel splicing variants, suggest that there is considerably more transcriptomic complexity than previously appreciated.

Small RNA analysis

A related application of next-generation sequencing technologies to the analysis of transcriptomes is small RNA discovery and profiling. High-throughput sequencing offers a greater potential for the identification of novel small RNAs as well as profiling of known and novel small RNA genes. Small RNA profiling with 454 pyrosequencing technology has been widely reported, which include studies in the moss *Physcomitrella patens*(Axtell et al., 2006), *A. thaliana* (Henderson et al., 2006; Lu et al., 2006; Rajagopalan et al., 2006), *Triticum aestivum* (Yao et al., 2007), the basal eudicot species *Eschscholzia californica* (Barakat et al., 2007), the lycopod *Selaginella moellendorffii* (Axtell et al., 2006), the unicellular alga *Chlamydomonas reinhardtii* (Zhao et al., 2007), Marek disease virus (Burnside et al., 2006), and some primates (Berezikov et al., 2006). More importantly, these studies contributed to the discovery of a novel class of small RNAs, termed Piwi-interacting RNAs. They are expressed in mammalian testes and are presumably required for germ cell development in mammals and other species (Girard et al., 2006; Lau et al., 2006; Houwing et al., 2007).

The higher throughput of Illumina and SOLiD technologies enables the generation of deeper small RNA libraries. Using Illumina sequencing, Morin et al. (2008b) identified 334 known plus 104 novel miRNA genes expressed in human embryonic stem cells, while Glazov et al. (2008) detected 449 novel and all known chicken miRNAs in the chicken embryo. In addition, small RNA profilings in locust, *Xenopus tropicalis* (Armisen et al., 2009), *C. elegans* embryos (Stoeckius et al., 2009) and *Gossypium hirsutum* L (Pang et al., 2009) have also been reported.

Epigenomic Analysis

Epigenetics is the study of heritable gene regulation that does not involve the DNA sequence itself but its modifications and higher-order structures. The next-generation sequencing technologies offer the potential to substantially accelerate epigenomic research. To date, these technologies have been applied in several epigenomic areas, including the characterization of DNA methylation patterns, posttranslational modifications of histones, the interaction between transcription factors and their direct targets, and nucleosome positioning on a genome-wide scale. These areas are summarized into the following two major sections.

Methylome

DNA cytosine methylation is a central epigenetic modification that plays essential roles in cellular processes including genome regulation, development and disease. Single-base resolution analysis of DNA methylation sites can be achieved by sodium bisulfite (BS) treatment of genomic DNA, which converts cytosines, but not methylcytosines, to uracil. Subsequent sequencing of PCR-amplified bisulfite-converted DNA allows determination of the methylation state of the cytosines in the sequenced region of the genome, as methylcytosine will be sequenced as cytosine, and unmethylated cytosine as thymine. Taylor et al. (2007) improved the bisulfite DNA sequencing procedure by combining with the 454 sequencing technology. The approach was applied to analyze methylation patterns in 25 gene-related CpG-rich regions from > 40 cases of primary cells. The study generated > 1600 individual sequence far beyond the few clones (< 20) typically analyzed by traditional bisulfite sequencing. Using the Illumina Genome Analyzer, Cokus et al. (2008) and Lister et al. (2008) generated 2–3 gigabases of uniquely aligned bisulfite sequence to comprehensively identify sites of DNA methylation throughout the *Arabidopsis* genome at a single-base resolution, including previously unidentified sites of cytosine methylation, and local sequence motifs associated with DNA methylation. The approach of bisulfite DNA sequencing is widely used for DNA methylation profiling in various organisms now (Costello et al., 2009; Smith et al., 2009; Bormann Chung et al., 2010).

DNA-protein interactions: ChIP-Seq

ChIP-seq is a recently developed technique for genome-wide profiling of DNA binding proteins, histone modifications and nucleosomes. The association between DNA and proteins is a fundamental biologic interaction that plays a key part in regulating gene expression and controlling the availability of DNA for transcription, replication and other biologic processes. These interactions can be studied using a technique called chromatin immunoprecipitation, and traditionally followed by microarray analysis (Aparicio et al., 2004). With the advent of high-throughput next-generation sequencing technologies, a more powerful approach based on chromatin immunoprecipitation followed by sequencing (ChIP-seq) has emerged. The precedent-setting paper for ChIP-seq was published by Johnson and colleagues in 2006, who used *Caenorhabditis elegans* and the Roche 454 platform to elucidate nucleosome positioning on genomic DNA (Johnson et al., 2006). This study established that sequencing the nuclease-derived digestion products of genomic DNA was sufficient to generate a genome-wide, highly precise positional profile of chromatin. Subsequent studies utilized a ChIP-based approach and the Illumina platform to provide insights into transcription factor binding sites in the human genome such as neuron-restrictive silencer factor (NRSF) (Johnson et al., 2007) and signal transducer and activator of transcription 1 (STAT1) (Robertson et al., 2007). The first applications of ChIP-seq to profile histone modifications were done in CD4+ T cells (Impey et al., 2004) and mouse embryonic stem (ES) cells (Mikkelsen et al., 2007). In a landmark study, Mikkelsen and coworkers (2007) explored the connection between chromatin packaging of DNA and differential gene expression using mouse embryonic stem cells and lineage-committed mouse cells (neural progenitor cells and embryonic fibroblasts). Their work provided a next-generation sequencing-based framework for using genome-wide chromatin profiling to characterize cell populations.

In comparison to array-based predecessor, i.e., ChIP-chip technology, ChIP-seq offers higher resolution, lower noise and better coverage. With the ever-decreasing cost of sequencing, ChIP-seq has become an indispensable tool for studying gene regulation and epigenetic mechanisms.

Metagenomic sequencing

Metagenomics involves the genomic analysis of microorganisms by direct extraction of DNA from uncultured ensemble of microbial communities. It is not until recent years that scientists are able to unravel a wider range of microorganisms, thanks largely to advances in DNA-sequencing technology. Robert Edwards and colleagues published the first sequences of environmental samples generated with next generation sequencing technique with Roche's 454 pyrosequencing instrument (Edwards et al., 2006). Since then, a wide range of metagenomes has been studied with this technique, including some real large metagenomic project such as the Human Microbiome Project (HMP) (Turnbaugh et al., 2007). The aim of HMP is to lay bare the microbial communities associated with various parts of the human body, including the gut. In a recent publication, an international team of researchers have cataloged human gut microbial genes by metagenomic sequencing (Qin et al., 2010). They generated over 570 Gb of sequence data from 124 individuals, assembled and characterized 3.3 million non-redundant microbial genes. This helped scientists, for the first time, to define the minimal human gut metagenome and the minimal gut bacterial genomes. Even though it has been widely accepted that the 454 system is the most promising next-generation sequencing technology for metagenomic analysis, due to its long read length, this recent work on human gut microbial was done with Illumina Genomic Analyzer.

CONCLUSIONS

The field of sequencing technology and application development is a fast-moving area of biomedical research. As we can see from the introduction above, the next-generation sequencing technologies have extended to an impressive array of applications beyond just genomic sequencing and its large-scale operations. As we described, new innovations are being developed each day. Novel generations of sequencing technologies, such as single-molecule sequencing and nanostructure-based sequencing, holds greater promise to achieve ever-faster, cheaper, more accurate and reliable ways to produce sequence data. Shortcomings of today's next-generation sequencing platforms, e.g., short-read and less base accuracy, will be overcomed with the development of new technologies. This, indeed, makes this an exciting area for genomic studies.

ACKNOWLEDGEMENTS

This work was supported by the Chinese Academy of Sciences Scientific Research Equipments (Grant No. YZ200823) and the Institutional Director's Initiative Fund awarded to Jun Yu.

REFERENCES

Adessi, C., Matton, G., Ayala, G., Turcatti, G., Mermod, J.J., Mayer, P., and Kawashima, E. (2000). Solid phase DNA amplification: characterisation of primer attachment and amplification mechanisms. Nucleic Acids Res 28, E87.

Ahn, S.M., Kim, T.H., Lee, S., Kim, D., Ghang, H., Kim, D.S., Kim, B.C., Kim, S.Y., Kim, W.Y., Kim, C., *et al*. (2009). The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. Genome Res 19, 1622–1629.

Ansorge, W.J. (2009). Next-generation DNA sequencing techniques. New Biotechnol 25, 195–203.

Aparicio, O., Geisberg, J.V., and Struhl, K. (2004). Chromatin immunoprecipitation for determining the association of proteins with specific genomic sequences in vivo. Curr Protoc Cell Biol Chapter 17, Unit 17 17.

Armisen, J., Gilchrist, M.J., Wilczynska, A., Standart, N., and Miska, E.A. (2009). Abundant and dynamically expressed miRNAs, piRNAs, and other small RNAs in the vertebrate Xenopus tropicalis. Genome Res 19, 1766–1775.

Axtell, M.J., Jan, C., Rajagopalan, R., and Bartel, D.P. (2006). A two-hit trigger for siRNA biogenesis in plants. Cell 127, 565–577.

Barakat, A., Wall, K., Leebens-Mack, J., Wang, Y.J., Carlson, J.E., and Depamphilis, C.W. (2007). Large-scale identification of microRNAs from a basal eudicot (Eschscholzia californica) and conservation in flowering plants. Plant J 51, 991–1003.

Bentley, D.R. (2006). Whole-genome re-sequencing. Curr Opin Genet Dev 16, 545–552.

Berezikov, E., Thuemmler, F., van Laake, L.W., Kondova, I., Bontrop, R., Cuppen, E., and Plasterk, R.H. (2006). Diversity of microRNAs in human and chimpanzee brain. Nat Genet 38, 1375–1377.

Blow, N. (2008). DNA sequencing: generation next-next. Nat Methods 5, 267–274.

Bormann Chung, C.A., Boyd, V.L., McKernan, K.J., Fu, Y.T., Monighetti, C., Peckham, H.E., Barker, M., and Khanin, R. (2010). Whole methylome analysis by ultra-deep sequencing using two-base encoding. PLoS ONE 5, e9320.

Braslavsky, I., Hebert, B., Kartalov, E., and Quake, S.R. (2003). Sequence information can be obtained from single DNA molecules. Proc Natl Acad Sci U S A 100, 3960–3964.

Brent, M.R. (2008). Steady progress and recent breakthroughs in the accuracy of automated genome annotation. Nat Rev Genet 9, 62–73.

Burnside, J., Bernberg, E., Anderson, A., Lu, C., Meyers, B.C., Green, P.J., Jain, N., Isaacs, G., and Morgan, R.W. (2006). Marek's disease virus encodes MicroRNAs that map to meq and the latency-associated transcript. J Virol 80, 8778–8786.

Choi, M., Scholl, U.I., Ji, W.Z., Liu, T.W., Tikhonova, I.R., Zumbo, P., Nayir, A., Bakkaloğlu, A., Ozen, S., Sanjad, S., *et al*. (2009). Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. Proc Natl Acad Sci U S A 106, 19096–19101.

Cloonan, N., Forrest, A.R.R., Kolle, G., Gardiner, B.B.A., Faulkner, G. J., Brown, M.K., Taylor, D.F., Steptoe, A.L., Wani, S., Bethel, G., *et al*. (2008). Stem cell transcriptome profiling via massive-scale mRNA sequencing. Nat Methods 5, 613–619.

Cokus, S.J., Feng, S.H., Zhang, X.Y., Chen, Z.G., Merriman, B., Haudenschild, C.D., Pradhan, S., Nelson, S.F., Pellegrini, M., and Jacobsen, S.E. (2008). Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. Nature 452, 215–219.

Costello, J.F., Krzywinski, M., and Marra, M.A. (2009). A first look at entire human methylomes. Nat Biotechnol 27, 1130–1132.

Dahl, F., Stenberg, J., Fredriksson, S., Welch, K., Zhang, M., Nilsson, M., Bicknell, D., Bodmer, W.F., Davis, R.W., and Ji, H.L. (2007). Multigene amplification and massively parallel sequencing for cancer mutation discovery. Proc Natl Acad Sci U S A 104, 9387–9392.

Davies, K. (2007). Next-Generation Sequencing: Scientific and Commercial Implications of the $1000 Genome (Insight Pharma Reports)

Denver, D.R., Dolan, P.C., Wilhelm, L.J., Sung, W., Lucas-Lledó, J.I., Howe, D.K., Lewis, S.C., Okamoto, K., Thomas, W.K., Lynch, M., *et al*. (2009). A genome-wide view of Caenorhabditis elegans base-substitution mutation processes. Proc Natl Acad Sci U S A 106, 16310–16314.

Diguistini, S., Liao, N.Y., Platt, D., Robertson, G., Seidel, M., Chan, S. K., Docking, T.R., Birol, I., Holt, R.A., Hirst, M., *et al*. (2009). *De novo* genome sequence assembly of a filamentous fungus using Sanger, 454 and Illumina sequence data. Genome Biol 10, R94.

Dracatos, P.M., Cogan, N.O.I., Sawbridge, T.I., Gendall, A.R., Smith, K.F., Spangenberg, G.C., and Forster, J.W. (2009). Molecular characterisation and genetic mapping of candidate genes for qualitative disease resistance in perennial ryegrass (Lolium perenne L.). BMC Plant Biol 9, 62.

Dressman, D., Yan, H., Traverso, G., Kinzler, K.W., and Vogelstein, B. (2003). Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. Proc Natl Acad Sci U S A 100, 8817–8822.

Durfee, T., Nelson, R., Baldwin, S., Plunkett, G. 3rd, Burland, V., Mau, B., Petrosino, J.F., Qin, X., Muzny, D.M., Ayele, M., *et al*. (2008). The complete genome sequence of Escherichia coli DH10B: insights into the biology of a laboratory workhorse. J Bacteriol 190, 2597–2606.

Edwards, R.A., Rodriguez-Brito, B., Wegley, L., Haynes, M., Breitbart, M., Peterson, D.M., Saar, M.O., Alexander, S., Alexander, E.C. Jr, and Rohwer, F. (2006). Using pyrosequencing to shed light on deep mine microbial ecology. BMC Genomics 7, 57.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., *et al*. (2009). Real-time DNA sequencing from single polymerase molecules. Science 323, 133–138.

Fedurco, M., Romieu, A., Williams, S., Lawrence, I., and Turcatti, G. (2006). BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. Nucleic Acids Res 34, e22.

Fierer, N., Breitbart, M., Nulton, J., Salamon, P., Lozupone, C., Jones, R., Robeson, M., Edwards, R.A., Felts, B., Rayhawk, S., *et al*. (2007). Metagenomic and small-subunit rRNA analyses reveal the genetic diversity of bacteria, archaea, fungi, and viruses in soil. Appl Environ Microbiol 73, 7059–7066.

Gilbert, W. (1981). DNA sequencing and gene structure Nobel lecture, 8 December 1980. Biosci Rep 1, 353–375.

Girard, A., Sachidanandam, R., Hannon, G.J., and Carmell, M.A. (2006). A germline-specific class of small RNAs binds mammalian Piwi proteins. Nature 442, 199–202.

Glazov, E.A., Cottee, P.A., Barris, W.C., Moore, R.J., Dalrymple, B.P., and Tizard, M.L. (2008). A microRNA catalog of the developing chicken embryo identified by a deep sequencing approach. Genome Res 18, 957–964.

Goldberg, S.M.D., Johnson, J., Busam, D., Feldblyum, T., Ferriera, S., Friedman, R., Halpern, A., Khouri, H., Kravitz, S.A., Lauro, F.M., *et al*. (2006). A Sanger/pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes. Proc Natl Acad Sci U S A 103, 11240–11245.

Goossens, D., Moens, L.N., Nelis, E., Lenaerts, A.S., Glassee, W., Kalbe, A., Frey, B., Kopal, G., De Jonghe, P., De Rijk, P., *et al*. (2009). Simultaneous mutation and copy number variation (CNV) detection by multiplex PCR-based GS-FLX sequencing. Hum Mutat 30, 472–476.

Harismendy, O., and Frazer, K.A. (2009). Method for improving sequence coverage uniformity of targeted genomic intervals amplified by LR-PCR using Illumina GA sequencing-by-synthesis technology. Biotechniques 46, 229–231.

Harismendy, O., Ng, P.C., Strausberg, R.L., Wang, X.Y., Stockwell, T.B., Beeson, K.Y., Schork, N.J., Murray, S.S., Topol, E.J., Levy, S., *et al*. (2009). Evaluation of next generation sequencing platforms for population targeted sequencing studies. Genome Biol 10, R32.

Harris, T.D., Buzby, P.R., Babcock, H., Beer, E., Bowers, J., Braslavsky, I., Causey, M., Colonell, J., Dimeo, J., Efcavitch, J. W., *et al*. (2008). Single-molecule DNA sequencing of a viral genome. Science 320, 106–109.

Henderson, I.R., Zhang, X.Y., Lu, C., Johnson, L., Meyers, B.C., Green, P.J., and Jacobsen, S.E. (2006). Dissecting Arabidopsis thaliana DICER function in small RNA processing, gene silencing and DNA methylation patterning. Nat Genet 38, 721–725.

Hodges, E., Xuan, Z., Balija, V., Kramer, M., Molla, M.N., Smith, S.W., Middle, C.M., Rodesch, M.J., Albert, T.J., Hannon, G.J., *et al*. (2007). Genome-wide in situ exon capture for selective resequencing. Nat Genet 39, 1522–1527.

Housby, J.N., and Southern, E.M. (1998). Fidelity of DNA ligation: a novel experimental approach based on the polymerisation of libraries of oligonucleotides. Nucleic Acids Res 26, 4259–4266.

Houwing, S., Kamminga, L.M., Berezikov, E., Cronembold, D., Girard, A., van den Elst, H., Filippov, D.V., Blaser, H., Raz, E., Moens, C.B., *et al*. (2007). A role for Piwi and piRNAs in germ cell maintenance and transposon silencing in Zebrafish. Cell 129, 69–82.

Huang, S.W., Li, R.Q., Zhang, Z.H., Li, L., Gu, X.F., Fan, W., Lucas, W.J., Wang, X.W., Xie, B.Y., Ni, P.X., *et al*. (2009). The genome of the cucumber, Cucumis sativus L. Nat Genet 41, 1275–1281.

Huber, J.A., Mark Welch, D.B., Morrison, H.G., Huse, S.M., Neal, P. R., Butterfield, D.A., and Sogin, M.L. (2007). Microbial population structures in the deep marine biosphere. Science 318, 97–100.

Imelfort, M., and Edwards, D. (2009). *De novo* sequencing of plant genomes using second-generation technologies. Brief Bioinform 10, 609–618.

Impey, S., McCorkle, S.R., Cha-Molstad, H., Dwyer, J.M., Yochum, G. S., Boss, J.M., McWeeney, S., Dunn, J.J., Mandel, G., and Goodman, R.H. (2004). Defining the CREB regulon: a genome-wide analysis of transcription factor regulatory regions. Cell 119, 1041–1054.

Jacquier, A. (2009). The complex eukaryotic transcriptome: unexpected pervasive transcription and novel small RNAs. Nat Rev Genet 10, 833–844.

Johnson, D.S., Mortazavi, A., Myers, R.M., and Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. Science 316, 1497–1502.

Johnson, S.M., Tan, F.J., McCullough, H.L., Riordan, D.P., and Fire, A.Z. (2006). Flexibility and constraint in the nucleosome core landscape of Caenorhabditis elegans chromatin. Genome Res 16, 1505–1516.

Kim, J.I., Ju, Y.S., Park, H., Kim, S., Lee, S., Yi, J.H., Mudge, J., Miller, N.A., Hong, D., Bell, C.J., *et al*. (2009). A highly annotated whole-genome sequence of a Korean individual. Nature 460, 1011–1015.

Korlach, J., Marks, P.J., Cicero, R.L., Gray, J.J., Murphy, D.L., Roitman, D.B., Pham, T.T., Otto, G.A., Foquet, M., and Turner, S. W. (2008). Selective aluminum passivation for targeted immobilization of single DNA polymerase molecules in zero-mode waveguide nanostructures. Proc Natl Acad Sci U S A 105, 1176–1181.

Lander, E.S., and Waterman, M.S. (1988). Genomic mapping by fingerprinting random clones: a mathematical analysis. Genomics 2, 231–239.

Lau, N.C., Seto, A.G., Kim, J., Kuramochi-Miyagawa, S., Nakano, T., Bartel, D.P., and Kingston, R.E. (2006). Characterization of the piRNA complex from rat testes. Science 313, 363–367.

Levene, M.J., Korlach, J., Turner, S.W., Foquet, M., Craighead, H.G., and Webb, W.W. (2003). Zero-mode waveguides for single-molecule analysis at high concentrations. Science 299, 682–686.

Levy, S., Sutton, G., Ng, P.C., Feuk, L., Halpern, A.L., Walenz, B.P., Axelrod, N., Huang, J., Kirkness, E.F., Denisov, G., *et al*. (2007). The diploid genome sequence of an individual human. PLoS Biol 5, e254.

Li, R.Q., Fan, W., Tian, G., Zhu, H.M., He, L., Cai, J., Huang, Q.F., Cai, Q.L., Li, B., Bai, Y.Q., *et al*. (2010). The sequence and *de novo* assembly of the giant panda genome. Nature 463, 311–317.

Lister, R., O'Malley, R.C., Tonti-Filippini, J., Gregory, B.D., Berry, C.C., Millar, A.H., and Ecker, J.R. (2008). Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell 133, 523–536.

Lu, C., Kulkarni, K., Souret, F.F., MuthuValliappan, R., Tej, S.S., Poethig, R.S., Henderson, I.R., Jacobsen, S.E., Wang, W., Green, P.J., *et al*. (2006). MicroRNAs and other small RNAs enriched in the Arabidopsis RNA-dependent RNA polymerase-2 mutant. Genome Res 16, 1276–1288.

Lupski, J.R., Reid, J.G., Gonzaga-Jauregui, C., Rio Deiros, D., Chen, D.C.Y., Nazareth, L., Bainbridge, M., Dinh, H., Jing, C., Wheeler, D. A., *et al*. (2010). Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. N Engl J Med 362, 1181–1191.

Mardis, E.R. (2008). Next-generation DNA sequencing methods. Annu Rev Genom Hum G 9, 387–402.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z.T., *et al*. (2005). Genome sequencing in microfabricated high-density picolitre reactors. Nature 437, 376–380.

Mckernan, K., Blanchard, A., Kotler, L., and Costa, G. (2006). Reagents, methods, and libraries for bead-based sequencing. US patent application 20080003571.

McKernan, K.J., Peckham, H.E., Costa, G.L., McLaughlin, S.F., Fu, Y. T., Tsung, E.F., Clouser, C.R., Duncan, C., Ichikawa, J.K., Lee, C. C., *et al*. (2009). Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. Genome Res 19, 1527–1541.

Metzker, M.L. (2010). Sequencing technologies — the next generation. Nat Rev Genet 11, 31–46.

Mikkelsen, T.S., Ku, M.C., Jaffe, D.B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T.K., Koche, R.P., *et al* (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature 448, 553–560.

Moore, G.E. (1998). Cramming more components onto integrated circuits (Reprinted from Electronics, pg 114–117, April 19, 1965). P Ieee 86, 82–85.

Morin, R.D., Bainbridge, M., Fejes, A., Hirst, M., Krzywinski, M., Pugh, T.J., McDonald, H., Varhol, R., Jones, S.J.M., and Marra, M.A. (2008a). Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. Biotechniques 45, 81–94.

Morin, R.D., O'Connor, M.D., Griffith, M., Kuchenbauer, F., Delaney, A., Prabhu, A.L., Zhao, Y., McDonald, H., Zeng, T., Hirst, M., *et al*. (2008b). Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. Genome Res 18, 610–621.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B.

(2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5, 621–628.

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., and Snyder, M. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. Science 320, 1344–1349.

Ng, S.B., Buckingham, K.J., Lee, C., Bigham, A.W., Tabor, H.K., Dent, K.M., Huff, C.D., Shannon, P.T., Jabs, E.W., Nickerson, D.A., *et al*. (2010). Exome sequencing identifies the cause of a mendelian disorder. Nat Genet 42, 30–35.

Ossowski, S., Schneeberger, K., Clark, R.M., Lanz, C., Warthmann, N., and Weigel, D. (2008). Sequencing of natural strains of Arabidopsis thaliana with short reads. Genome Res 18, 2024–2033.

Pang, M.X., Woodward, A.W., Agarwal, V., Guan, X.Y., Ha, M., Ramachandran, V., Chen, X.M., Triplett, B.A., Stelly, D.M., and Chen, Z.J. (2009). Genome-wide analysis reveals rapid and dynamic changes in miRNA and siRNA sequence and expression during ovule and fiber development in allotetraploid cotton (Gossypium hirsutum L.). Genome Biol 10, R122.

Pleasance, E.D., Stephens, P.J., O'Meara, S., McBride, D.J., Meynert, A., Jones, D., Lin, M.L., Beare, D., Lau, K.W., Greenman, C., *et al*. (2010). A small-cell lung cancer genome with complex signatures of tobacco exposure. Nature 463, 184–190.

Porreca, G.J., Zhang, K., Li, J.B., Xie, B., Austin, D., Vassallo, S.L., LeProust, E.M., Peck, B.J., Emig, C.J., Dahl, F., *et al*. (2007). Multiplex amplification of large sets of human exons. Nat Methods 4, 931–936.

Pushkarev, D., Neff, N.F., and Quake, S.R. (2009). Single-molecule sequencing of an individual human genome. Nat Biotechnol 27, 847–852.

Qin, J.J., Li, R.Q., Raes, J., Arumugam, M., Burgdorf, K.S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., *et al*, and the MetaHIT Consortium. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. Nature 464, 59–65.

Rajagopalan, R., Vaucheret, H., Trejo, J., and Bartel, D.P. (2006). A diverse and evolutionarily fluid set of microRNAs in Arabidopsis thaliana. Genes Dev 20, 3407–3425.

Reinhardt, J.A., Baltrus, D.A., Nishimura, M.T., Jeck, W.R., Jones, C. D., and Dangl, J.L. (2009). *De novo* assembly using low-coverage short read sequence data from the rice pathogen Pseudomonas syringae pv. oryzae. Genome Res 19, 294–305.

Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M., Zhao, Y.J., Zeng, T., Euskirchen, G., Bernier, B., Varhol, R., Delaney, A., *et al*. (2007). Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat Methods 4, 651–657.

Rothberg, J.M., and Leamon, J.H. (2008). The development and impact of 454 sequencing. Nat Biotechnol 26, 1117–1124.

Sanger, F., and Coulson, A.R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J Mol Biol 94, 441–448.

Shendure, J., and Ji, H.L. (2008). Next-generation DNA sequencing. Nat Biotechnol 26, 1135–1145.

Shendure, J., Mitra, R.D., Varma, C., and Church, G.M. (2004). Advanced sequencing technologies: methods and goals. Nat Rev Genet 5, 335–344.

Shendure, J., Porreca, G.J., Reppas, N.B., Lin, X.X., McCutcheon, J. P., Rosenbaum, A.M., Wang, M.D., Zhang, K., Mitra, R.D., and Church, G.M. (2005). Accurate multiplex polony sequencing of an evolved bacterial genome. Science 309, 1728–1732.

Smith, Z.D., Gu, H.C., Bock, C., Gnirke, A., and Meissner, A. (2009). High-throughput bisulfite sequencing in mammalian genomes. Methods 48, 226–232.

Stoeckius, M., Maaskola, J., Colombo, T., Rahn, H.P., Friedländer, M. R., Li, N., Chen, W., Piano, F., and Rajewsky, N. (2009). Large-scale sorting of C. elegans embryos reveals the dynamics of small RNA expression. Nat Methods 6, 745–751.

Sugarbaker, D.J., Richards, W.G., Gordon, G.J., Dong, L., De Rienzo, A., Maulik, G., Glickman, J.N., Chirieac, L.R., Hartman, M.L., Taillon, B.E., *et al*. (2008). Transcriptome sequencing of malignant pleural mesothelioma tumors. Proc Natl Acad Sci U S A 105, 3521–3526.

Sultan, M., Schulz, M.H., Richard, H., Magen, A., Klingenhoff, A., Scherf, M., Seifert, M., Borodina, T., Soldatov, A., Parkhomchuk, D., *et al*. (2008). A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science 321, 956–960.

Tang, F.C., Barbacioru, C., Nordman, E., Li, B., Xu, N.L., Bashkirov, V. I., Lao, K.Q., and Surani, M.A. (2010). RNA-Seq analysis to capture the transcriptome landscape of a single cell. Nat Protoc 5, 516–535.

Taylor, K.H., Kramer, R.S., Davis, J.W., Guo, J., Duff, D.J., Xu, D., Caldwell, C.W., and Shi, H. (2007). Ultradeep bisulfite sequencing analysis of DNA methylation patterns in multiple gene promoters by 454 sequencing. Cancer Res 67, 8511–8518.

Tettelin, H., and Feldblyum, T. (2009). Bacterial genome sequencing. Methods Mol Biol 551, 231–247.

Tewhey, R., Warner, J.B., Nakano, M., Libby, B., Medkova, M., David, P.H., Kotsopoulos, S.K., Samuels, M.L., Hutchison, J.B., Larson, J. W., *et al*. (2009). Microdroplet-based PCR enrichment for large-scale targeted sequencing. Nat Biotechnol 27, 1025–1031.

Toth, A.L., Varala, K., Newman, T.C., Miguez, F.E., Hutchison, S.K., Willoughby, D.A., Simons, J.F., Egholm, M., Hunt, J.H., Hudson, M. E., *et al*. (2007). Wasp gene expression supports an evolutionary link between maternal behavior and eusociality. Science 318, 441–444.

Turnbaugh, P.J., Ley, R.E., Hamady, M., Fraser-Liggett, C.M., Knight, R., and Gordon, J.I. (2007). The human microbiome project. Nature 449, 804–810.

Turner, D.J., Keane, T.M., Sudbery, I., and Adams, D.J. (2009a). Next-generation sequencing of vertebrate experimental organisms. Mamm Genome 20, 327–338.

Turner, E.H., Lee, C.L., Ng, S.B., Nickerson, D.A., and Shendure, J. (2009b). Massively parallel exon capture and library-free resequencing across 16 genomes. Nat Methods 6, 315–316.

Velasco, R., Zharkikh, A., Troggio, M., Cartwright, D.A., Cestaro, A., Pruss, D., Pindo, M., Fitzgerald, L.M., Vezzulli, S., Reid, J., *et al*. (2007). A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. PLoS ONE 2, e1326.

Wang, J., Wang, W., Li, R.Q., Li, Y.R., Tian, G., Goodman, L., Fan, W., Zhang, J.Q., Li, J., Zhang, J.B., *et al*. (2008). The diploid genome sequence of an Asian individual. Nature 456, 60–65.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10, 57–63.

Wheeler, D.A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y.J., Makhijani, V., Roth, G.T., *et al* (2008). The complete genome of an individual by massively parallel DNA sequencing. Nature 452, 872–876.

Wilhelm, B.T., Marguerat, S., Watt, S., Schubert, F., Wood, V., Goodhead, I., Penkett, C.J., Rogers, J., and Bähler, J. (2008). Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. Nature 453, 1239–1243.

Xia, Q.Y., Guo, Y.R., Zhang, Z., Li, D., Xuan, Z.L., Li, Z., Dai, F.Y., Li, Y. R., Cheng, D.J., Li, R.Q., *et al*. (2009). Complete resequencing of 40 genomes reveals domestication events and genes in silkworm (Bombyx). Science 326, 433–436.

Yao, Y.Y., Guo, G.G., Ni, Z.F., Sunkar, R., Du, J.K., Zhu, J.K., and Sun, Q.X. (2007). Cloning and characterization of microRNAs from wheat (Triticum aestivum L.). Genome Biol 8, R96.

Zhao, T., Li, G.L., Mi, S.J., Li, S., Hannon, G.J., Wang, X.J., and Qi, Y. J. (2007). A complex system of small RNAs in the unicellular green alga Chlamydomonas reinhardtii. Genes Dev 21, 1190–1203.

Zhou, X.G., Ren, L.F., Li, Y.T., Zhang, M., Yu, Y.D., and Yu, J. (2010). Next-generation sequencing technology: A technology review and future perspective. Sci China C Life Sci 53, 44–57.

