Genomics 92 (2008) 255–264

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Review

Applications of next-generation sequencing technologies in functional genomics

Olena Morozova, Marco A. Marra *

BC Cancer Agency Genome Sciences Centre, Suite 100, 570 West 7th Avenue, Vancouver, BC V5Z 4S6, Canada

ARTICLE INFO

Article history:
Received 3 April 2008
Accepted 9 July 2008
Available online 24 August 2008

Keywords:
Illumina/Solexa
ABI/SOLiD
454/Roche
Functional genomics
Next-generation sequencing technology
Transcriptome
Epigenome
Deep sequencing
Sequencing by synthesis
Sequencing by ligation

ABSTRACT

A new generation of sequencing technologies, from Illumina/Solexa, ABI/SOLiD, 454/Roche, and Helicos, has provided unprecedented opportunities for high-throughput functional genomic research. To date, these technologies have been applied in a variety of contexts, including whole-genome sequencing, targeted resequencing, discovery of transcription factor binding sites, and noncoding RNA expression profiling. This review discusses applications of next-generation sequencing technologies in functional genomics research and highlights the transforming potential these technologies offer.

Contents

* Corresponding author. Fax: +1 604 675 8178.
E-mail address: mmarra@bcgsc.ca (M.A. Marra).

0888-7543/$ – see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2008.07.001

Since first introduced to the market in 2005, next-generation sequencing technologies have had a tremendous impact on genomic research. The next-generation technologies have been used for standard sequencing applications, such as genome sequencing and resequencing, and for novel applications previously unexplored by Sanger sequencing. In this review we first describe the three commercially available next-generation sequencing technologies in comparison to a state-of-the-art Sanger sequencer, and follow this with a discussion of the novel kind of data produced by next-generation sequencers and the issues associated with it. We then turn our attention to the application of next-generation sequencing technologies to functional genomics research, particularly focusing on transcriptomics and epigenomics. We end with a discussion of future prospects that next-generation technologies hold for functional genomics research. This review does not address genome sequencing and resequencing applications of next-generation sequencers [1] or the huge impact these technologies have had in metagenomics (reviewed in [2]).

Advances in DNA sequencing technologies

The landmark publications of the late 1970 s by Sanger's and Gilbert's groups [3,4] and notably the development of the chain termination method by Sanger and colleagues [5] established the groundwork for decades of sequence-driven research that followed. The chain-termination method published in 1977 [5], also commonly referred to as Sanger or dideoxy sequencing, has remained the most commonly used DNA sequencing technique to date and was used to complete human genome sequencing initiatives led by the International Human Genome Sequencing Consortium and Celera Genomics [6–8]. Very recently, the Sanger method has been partially supplanted by several "next-generation" sequencing technologies that offer dramatic increases in cost-effective sequence throughput, albeit at the expense of read lengths. The next-generation technologies commercially available today include the 454 GS20 pyrosequencing-based instrument (Roche Applied Science), the Solexa 1G analyzer (Illumina, Inc.), the SOLiD instrument from Applied Biosystems, and the Heliscope from Helicos, Inc. As of this writing, information on the performance of the Heliscope in functional genomics applications is lacking, and so we have restricted our comments to the 454, SOLiD, and 1G sequencing platforms. These new technologies as well as the current state-of the-art Sanger sequencing platform are summarized in Table 1 and discussed in some detail below. For a review of the history of DNA sequencing the reader is referred to [9]; for a more comprehensive review of emerging sequencing technologies see [10].

Table 1
Advances in DNA sequencing technologies

Technology	Approach	Read length	Bp per run	Company name and Web site
Automated Sanger sequencer ABI3730xl	Synthesis in the presence of dye terminators	Up to 900 bp	96 kb	Applied Biosystems www.appliedbyosystems.com
454/Roche FLX system	Pyrosequencing on solid support	200–300 bp	80–120 Mb	Roche Applied Science www.roche-applied-science.com
Illumina/ Solexa	Sequencing by synthesis with reversible terminators	30–40 bp	1 Gb	Illumina, Inc. http://www.illumina.com/
ABI/SOLiD	Massively parallel sequencing by ligation	35 bp	1–3 Gb	Applied Biosystems www.appliedbyosystems.com

Sanger sequencing

Since its initial report in 1977, the Sanger sequencing method has remained conceptually unchanged. The method is based on the DNA polymerase-dependent synthesis of a complementary DNA strand in the presence of natural 2′-deoxynucleotides (dNTPs) and 2′,3′-dideoxynucleotides (ddNTPs) that serve as nonreversible synthesis terminators [5]. The DNA synthesis reaction is randomly terminated whenever a ddNTP is added to the growing oligonucleotide chain, resulting in truncated products of varying lengths with an appropriate ddNTP at their 3′ terminus. The products are separated by size using polyacrylamide gel electrophoresis and the terminal ddNTPs are used to reveal the DNA sequence of the template strand.

Originally, four different reactions were required per template, each reaction containing a different ddNTP terminator, ddATP, ddCTP, ddTTP, or ddGTP. However, advances in fluorescence detection have allowed for combining the four terminators into one reaction by having them labeled with fluorescent dyes of different colors [11,12]. Subsequent advances have replaced the original slab gel electrophoresis with capillary gel electrophoresis, thereby enabling much higher electric fields to be applied to the separation matrix. One effect of this advance was to enhance the rate at which fragments could be separated [13]. The overall throughput of capillary electrophoresis was further increased by the advent of capillary arrays whereby many samples could be analyzed in parallel [14]. In addition, breakthroughs in polymer biochemistry, including the development of linear polyacrylamide [15] and polydimethylacrylamide [16] have allowed the reuse of capillaries in multiple electrophoretic runs, thus further increasing sequencing efficiency. For further reading on improvements in Sanger sequencing research the reader is referred to [10] and [17].

These and many other advances in sequencing technology contributed to the relatively low error rate, long read length, and robust characteristics of modern Sanger sequencers. For instance, a commonly used automated high-throughput Sanger sequencing instrument from Applied Biosystems, the ABI 3730xl, has a 96-capillary array format and is capable of producing 900 or more PHRED 20 [18] bp per read for a total of up to 96 kb for a 3-h run (Table 1). However, despite the many advances in chemistries and the robust performance of instruments like the 3730xl, the application of relatively expensive Sanger sequencing to large sequencing projects has remained beyond the means of the typical grant-funded investigator. This is a limitation that has been apparently successfully addressed, to varying degrees, by all of the latest technology offerings.

454 sequencing technology: pyrosequencing in high-density picoliter reactors

An inherent limitation of Sanger sequencing is the requirement of in vivo amplification of DNA fragments that are to be sequenced, which is usually achieved by cloning into bacterial hosts. The cloning step is prone to host-related biases, is lengthy, and is quite labor intensive [2]. The 454 technology [19], the first next-generation sequencing technology released to the market, circumvents the cloning requirement by taking advantage of a highly efficient in vitro DNA amplification method known as emulsion PCR [20]. In emulsion PCR, individual DNA fragment-carrying streptavidin beads, obtained through shearing the DNA and attaching the fragments to the beads using adapters, are captured into separate emulsion droplets. The droplets act as individual amplification reactors, producing ~10^7 clonal copies of a unique DNA template per bead [19]. Each template-containing bead is subsequently transferred into a well of a picotiter plate and the clonally related templates are analyzed using a pyrosequencing reaction. The use of the picotiter plate allows hundreds of thousands of pyrosequencing reactions to be carried out in parallel, massively increasing the sequencing throughput [19]. The

pyrosequencing approach [21,22] is a sequencing-by-synthesis technique that measures the release of inorganic pyrophosphate (PP_i) by chemiluminescence. The template DNA is immobilized, and solutions of dNTPs are added one at a time; the release of PP_i, whenever the complementary nucleotide is incorporated, is detectable by light produced by a chemiluminescent enzyme present in the reaction mix. The sequence of DNA template is determined from a "pyrogram," which corresponds to the order of correct nucleotides that had been incorporated. Since chemiluminescent signal intensity is proportional to the amount of pyrophosphate released and hence the number of bases incorporated, the pyrosequencing approach is prone to errors that result from incorrectly estimating the length of homopolymeric sequence stretches (i.e., indels).

The current state-of-the-art 454 platform marketed by Roche Applied Science is capable of generating 80–120 Mb of sequence in 200- to 300-bp reads in a 4-h run. The 454 technology has been the most widely published next-generation technology, having so far been featured in more than 100 research publications (Roche Applied Sciences).

Illumina: sequencing by synthesis of single-molecule arrays with reversible terminators

The Illumina/Solexa approach [23–25] achieves cloning-free DNA amplification by attaching single-stranded DNA fragments to a solid surface known as a single-molecule array, or flow cell, and conducting solid-phase bridge amplification of single-molecule DNA templates (Illumina, Inc.). In this process, one end of single DNA molecule is attached to a solid surface using an adapter; the molecules subsequently bend over and hybridize to complementary adapters (creating the "bridge"), thereby forming the template for the synthesis of their complementary strands. After the amplification step, a flow cell with more than 40 million clusters is produced, wherein each cluster is composed of approximately 1000 clonal copies of a single template molecule. The templates are sequenced in a massively parallel fashion using a DNA sequencing-by-synthesis approach that employs reversible terminators with removable fluorescent moieties and special DNA polymerases that can incorporate these terminators into growing oligonucleotide chains. The terminators are labeled with fluors of four different colors to distinguish among the different bases at the given sequence position and the template sequence of each cluster is deduced by reading off the color at each successive nucleotide addition step. Although the Illumina approach is more effective at sequencing homopolymeric stretches than pyrosequencing, it produces shorter sequence reads [25] and hence cannot resolve short sequence repeats. In addition, due to the use of modified DNA polymerases and reversible terminators, substitution errors have been noted in Illumina sequencing data [9]. Typically, the 1G genome analyzer from Illumina, Inc., is capable of generating 35-bp reads and producing at least 1 Gb of sequence per run in 2–3 days.

ABI/SOLiD: massively parallel sequencing by ligation

Massively parallel sequencing by hybridization–ligation, implemented in the supported oligonucleotide ligation and detection system (SOLiD) from Applied Biosystems, has recently become available. The ligation chemistry used in SOLiD is based on the polony sequencing technique that was published in the same year as the 454 method [26]. Construction of sequencing libraries for analysis on the SOLiD instrument begins with an emulsion PCR single-molecule amplification step similar to that used in the 454 technique. The amplification products are transferred onto a glass surface where sequencing occurs by sequential rounds of hybridization and ligation with 16 dinucleotide combinations labeled by four different fluorescent dyes (each dye used to label four dinucleotides). Using the four-dye encoding scheme, each position is effectively probed twice, and the identity of the nucleotide is determined by analyzing the color that results from two successive ligation reactions. Significantly, the two-base encoding scheme enables the distinction between a sequencing error and a sequence polymorphism: an error would be detected in only one particular ligation reaction, whereas a polymorphism would be detected in both. The newly released SOLiD instrument is capable of producing 1–3 Gb of sequence data in 35-bp reads per an 8-day run.

Making use of next-generation sequencer data format: pains and gains of plentiful short reads

Sequence census applications

The read lengths currently achievable by 454 technology are approaching 300 bp (Roche Applied Sciences), yet are still shorter than the 700–900 bp achieved by Sanger sequencing. The Illumina and ABI/SOLiD instruments generate even shorter ~35-bp reads. The large numbers of short reads produced by next-generation sequencers provide opportunities for the development of new applications that benefit from the particular data format. For instance, next-generation technologies have been widely applied in contexts whereby sequencing of only a portion of the molecule is sufficient (referred to as sequence census applications [27]).

The sequence census approach (Fig. 1) uses short reads (or "tags") to assign the site of origin of the read instead of determining the entire sequence of the original DNA molecule. By mapping the sequence read to its molecule of origin, the presence of the molecule is established. Significantly, the number of reads that map to a particular nucleic acid species correlates with the abundance of the species in the cell [28–30]. The sequence census approach is conceptually similar to the serial analysis of gene expression (SAGE) method initially developed with Sanger sequencing [31]. In SAGE, the abundance of a particular mRNA species is estimated from the count of sequence fragments (tags) derived from its 3′ end [31]. Since the advent of next-generation sequencers that reduce the cost

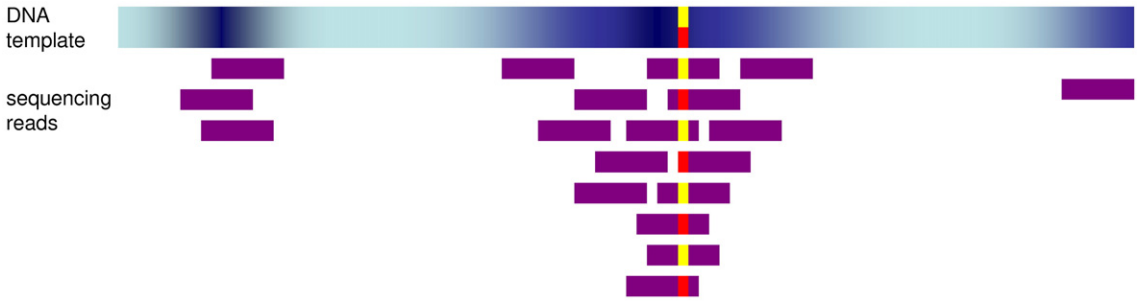

Fig. 1. Sequence census approach. In the sequence census approach used in next-generation sequencing, short reads are mapped to the template molecule to provide three types of information. Sequence data are used to reveal sequence polymorphisms in the template, e.g., a SNP (red and yellow), the abundance of reads is used as a quantitative measure of the abundance of the template, and the particular areas of the template covered by reads reveal the internal structure of the template, e.g., the presence of exons and introns.

of sequencing and massively increase throughput, the applicability of such sequence census methods has been expanded to include many research areas [27]. To date, sequence census methods have been most commonly used for the analysis of transcribed portions of the genome, such as gene expression and noncoding RNA profiling [28–30]. The applications of these methods for studying transcriptomes are discussed in more detail in the transcriptome section of this review.

A novel use of the sequence census approach is the identification of protein binding sites on the DNA using chromatin immunoprecipitation followed by next-generation sequencing (ChIP-Seq) [32–37]. This technique couples the commonly used chromatin immunoprecipitation procedure, in which DNA–protein complexes are cross-linked and precipitated using an antibody [38], to next-generation sequencing of DNA fragments bound to the precipitated protein [36]. To date, ChIP-Seq has been applied to the identification of transcription factor binding sites as well as histone modifications on a genome-wide scale [32,34,36,37]. The application of this technology for studying epigenomes is described in detail in the epigenome section of this review.

Read pairs and read accuracy issues

A limitation of short-read sequence data is the difficulty in de novo sequence assembly. This shortcoming is particularly an issue in sequencing new genomes and in sequencing highly rearranged genome segments, such as one might discover in cancer genomes [39] or in regions of structural variation [40]. Paired-end sequencing approaches, in which both ends of a fragment of defined size are sequenced to provide more information about the fragment, have the potential to improve the utility of short reads for sequencing rearranged genomic segments and for de novo sequence assembly [2,41]. Although widely adopted for Sanger sequencing, in which paired-end reads are obtained by sequencing both ends of a clone insert, paired-end approaches are currently in their infancy relative to most next-generation technologies [2]. A few reports of paired-end approaches for next-generation sequencers included the paired-end polony sequencing approach, applied to resequence the genome of an evolved strain of *Escherichia coli* [26], and the multiplex sequencing of paired-end ditags (MS-PET) method for 454 sequencing [42]. More recently, a paired-end mapping (PEM) procedure has been developed for the 454 technology and used to map structural rearrangements in two previously studied human genomes [43]. The results of PEM were in concordance with those obtained from previous investigations, including the HapMap project [43]. Another paired-end mapping approach involving paired-end ditags has been described for the detection of gene fusions and transcribed retrotransposons [44]. At the time of writing, no paired-end studies have been reported using Illumina or SOLiD technologies, although both platforms are developing or have implemented paired-end approaches.

Given the quantity of reads and their short length, read accuracy becomes critical for mapping the reads to a reference sequence and for detecting sequence polymorphisms. The base accuracy, and the PHRED method [18] for evaluating the quality of Sanger-sequenced bases, is well established. This is currently not the case for the next-generation technologies ([2], but see [45]). To compensate for the uncertainty related to sequence quality and base accuracy, a general reliance on redundancy of sequence coverage is commonly invoked in next-generation sequencing. Multiple overlapping reads are thus used to confirm the accuracy of the base calls in applications in which accuracy is paramount, such as in the reliable detection of mutations or sequence polymorphisms [46]. Increasing the accuracy of individual base calls will ultimately lead to reductions in the high levels of redundancy currently invoked for confident base assignment and thus will presumably decrease sequencing costs.

Transcriptome sequencing by next-generation technologies

The sequencing of cDNA rather than genomic DNA focuses analysis on the transcribed portion of the genome. This focus reduces the size of the sequencing target space, which can be viewed as desirable given the fact that, even with next-generation sequencers, sequencing an entire vertebrate genome is still an expensive undertaking. Transcriptome sequencing has been used for applications ranging from gene expression profiling, genome annotation, and rearrangement detection to noncoding RNA discovery and quantification. A unique feature of high-throughput transcriptome sequencing studies is the versatility of the data, which can simultaneously be analyzed to provide insight into the level of gene expression, the structure of genomic loci, and sequence variation present at loci (e.g., SNPs). To date, the 454 technology has dominated next-generation applications in transcriptomics; but at least one recent paper describes the use of the Illumina sequencer for profiling microRNAs [30].

Gene expression profiling using novel and revisited sequence census methods

The identification and quantification of mRNA species under different conditions or in different cell types have long been of interest to biologists. Two conceptually different approaches to high-throughput gene expression profiling have emerged in the past decade to allow the simultaneous interrogation of gene expression levels on a genome-wide scale [47]. One group of methods is based on microarrays, in which cDNA is hybridized to an array of complementary oligonucleotide probes corresponding to genes of interest, and the abundance of a particular mRNA species is estimated from its hybridization intensity to the relevant probe [48]. A variety of microarray-based platforms and techniques have been developed in recent years; see for review [49].

A conceptually different group of methods uses sequencing of cDNA fragments followed by counting the number of times a particular fragment has been observed. This group of methods includes the well-known SAGE method [31] and the more recent massively parallel signature sequencing (MPSS) [50]. In SAGE, restriction enzymes are used to obtain short sequence fragments (tags) of 14–17 bp, usually derived from the 3′ end of an mRNA; the tags are concatenated and sequenced to determine the expression profiles of their corresponding mRNAs [31]. The MPSS method also generates small fragment signatures of each mRNA species; however, it uses a different protocol that does not involve propagation in bacteria and a different non-gel-based sequencing method [50]. SAGE and MPSS are often termed “clone-and-count” techniques as they provide a digital overview of gene expression profiles in a cell [47]. Advantages of such digital readouts include statistical robustness and less stringent standardization and replication requirements than those used for microarrays [50,51]. Some disadvantages that have hindered the use of SAGE and MPSS up until recently included the cost of sequencing and the biases introduced by the necessary cloning step. Furthermore, the MPSS technology has been restricted to only a few specialized laboratories [52].

Despite its excellent performance at detecting highly abundant transcripts, SAGE as commonly employed (i.e., sequencing to depths of 30,000–200,000 tags) involves relatively limited sequencing that does not robustly detect rare mRNAs [53]. This is due to the costs incurred with extensive Sanger sequencing of SAGE libraries. Because of these costs, no conventionally sequenced SAGE library exhibits saturating tag number kinetics that would suggest complete representation of the cellular transcriptome [53]. In contrast, next-generation technologies offer substantial cost-effective increases in sequencing throughput, such that millions of sequences can be obtained for a few thousands of dollars or less. In addition, the short read lengths are compatible with the short tags generated using SAGE-like library

construction techniques. A recent study by Nielsen et al. [52] has described an extension of the SAGE method based on the LongSAGE protocol [54], for generation of longer tags of 17 bp versus 14 bp in the original SAGE, and the 454 sequencing technology. The next-generation sequencing-based SAGE method, termed DeepSAGE, greatly simplifies the sample preparation procedure by removing the cloning step and replacing it with emulsion PCR-based amplification; the sequencing is conducted by the 454 protocol that allows multiple samples to be sequenced in a single run at a high depth [52]. In particular, the authors estimate that a typical DeepSAGE experiment would generate 300,000 tags with less effort than a typical LongSAGE experiment generating 50,000 tags [52]. Nielsen et al. [52] applied the DeepSAGE protocol to the analysis of the transcriptome of the potato and showed that it was efficient at detecting rare transcripts (gauged by examining the expression of potato transcription factors) and due to the much increased depth provides more robust expression level estimates than LongSAGE [52].

A novel sequence census technique for surveying mRNA levels using 5′-end sequence fragments has been developed and termed rapid analysis of 5′ transcript ends (5′-RATE) [55]. The technique involves three steps, 5′ oligocapping of mRNA; ditag formation using RL-SAGE [56], a modification of the LongSAGE protocol; and 454 sequencing of tags. The technique was applied to the analysis of maize transcripts and was shown to provide an effective means for surveying the transcriptome. Some key features of 5′-RATE are the tag length (~80 bp), which is longer than that of LongSAGE and MPSS, and the ability to generate tags from the 5′ end, facilitating the identification of transcription start sites. In addition, similar to DeepSAGE, 5′-RATE is a relatively simple, fast, and productive procedure that does not involve cloning.

Other sequencing-based methods such as full-length cDNA sequencing [57] and the generation of expressed sequence tags (ESTs), which are single sequencing reads derived from one end of a cDNA clone [58,59], have been used to characterize cellular mRNA profiles. However, primarily due to the cost of sequencing, these methods had been even less effective than SAGE at providing a quantitative and comprehensive representation of cellular transcripts or transcript variability [60]. With the development of next-generation sequencing technologies, EST sequencing has gained potential as one of the sequence census methods for studying mRNA profiles on a genome-wide scale. A number of studies have been successful at constructing EST libraries using the 454 technology; so far EST libraries have been constructed from plants, including the mustard weed *Arabidopsis thaliana* [61,62], the model legume *Medicago truncatula* [63], and maize, *Zea mays* [64], as well as the insects *Drosophila melanogaster* [28] and wasp, *Polistes metricus* [65]. A study conducted at our Genome Centre used 454 sequencing to generate ESTs from a human prostate cancer cell line, LNCaP [29]. The 454 technology is well suited to EST sequencing, as it is currently capable of generating ~400,000 reads per run [29,63] and provides an unbiased representation of all regions of a transcript independent of length or expression level [61]. Importantly, a single 454 run has been shown to provide a representative view of the mRNA population in the cell [61,63]. Further, unlike the shorter reads generated using the Illumina or SOLiD sequencers, the length of the 454 reads allows for interpretation of sequences generated from species lacking a genome sequence or extensive transcriptome sequences for comparison (e.g., [66]).

Another novel sequence census approach based on the 454 technology focuses on sequencing unique fragments found at 3′ untranslated regions (3′-UTRs) of genes [67]. This approach is particularly useful for distinguishing closely related transcripts, such as those resulting from paralogs, and for studying allele-specific expression [67]. In addition, Eveland et al. [67] estimated that 3′-UTR sequencing is superior to EST sequencing at identifying individual transcripts owing to the decreased sequencing redundancy achieved by restricting reads to the 3′-UTR regions of genes [67]. In particular, the authors identified 47,299 distinct mRNAs compared to 17,500 identified by a similar EST study in *A. thaliana* [61]. Significantly, the method does not rely on a complete genome sequence and has been shown to be successful at surveying transcription in *Z. mays*, whose genome is currently being sequenced (Maize Genome Sequencing Consortium).

Small noncoding RNA profiling and the discovery of novel small RNA genes

A related application of next-generation sequencing technologies to the analysis of transcriptomes is small noncoding (ncRNA) discovery and profiling. ncRNAs are RNA molecules that are not translated into a protein product. This class of RNAs includes transfer RNA (tRNA), ribosomal RNA (rRNA), small nuclear and small nucleolar RNA, and microRNA and small interfering RNA (miRNA and siRNA). Recent research has implicated microRNAs, approximately 21-nucleotide-long RNA molecules, as crucial posttranscriptional regulators of gene expression in both animals and plants [68]. A related class of noncoding RNAs, siRNAs, 21–24 nt in length, is the predominant class of small RNA molecules in plants [69]. Definite evidence for the presence of endogenous siRNAs in animals is lacking [70]. While miRNAs and siRNAs are similar in size and are both involved in posttranscriptional regulation of gene expression, their biogenesis and exact functions are different [71].

MicroRNAs were first identified in *Caenorhabditis elegans* [72] and since then have emerged as crucial regulators of gene expression in many organisms, including humans [73]. Historically, novel miRNAs have been identified by cloning and sequencing of individual miRNAs, which involved separating them on gels and successively ligating adapters at the 5′-end monophosphate and 3′ hydroxyl groups [70,73]. However, using this approach it was difficult to distinguish miRNAs from degradation products of other ncRNAs in the cell, such as rRNA or tRNA [70]. More recently, microarray-based approaches have been developed for high-throughput miRNA profiling; however, these approaches are not suitable for the detection of novel miRNAs [70].

High-throughput sequencing of small RNAs provides great potential for the identification of novel small RNAs as well as profiling of known and novel small RNA genes. The MPSS technology has been applied to the sequencing of size-fractionated small RNAs from *A. thaliana* [69]. This approach involved the generation of 17-nt fragments corresponding to parts of mature small RNA molecules and using bioinformatic analysis to identify the corresponding small RNA genes in the genome [69]. Several disadvantages of this approach are the high complexity and cost of the MPSS technology, which involves a cloning step, and the short read lengths corresponding to only a portion of small RNA molecules [70].

Next-generation sequencing technologies do not involve cloning and produce read lengths compatible with the length of mature miRNAs and siRNAs. This provides several important advantages for ncRNA sequencing studies over MPSS; the advantages include the decreased procedural complexity and cost and the dramatically increased throughput and depth of coverage. Small RNA profiling studies with next-generation sequencing technologies currently include gel-based separation of small RNAs and construction of cDNA libraries followed by sequencing on a next-generation platform.

To date, small RNA profiling studies involving the 454 technology have been reported. These include studies in the moss Physcomitrella patens [74]; *A. thaliana* [74–80]; wheat, *Triticum aestivum* [81]; the basal eudicot species *Eschscholzia californica* [82]; the lycopod *Selaginella moellendorffii* [74]; the unicellular alga *Chlamydomonas reinhardtii* [83]; Marek disease virus [84]; and primates [85]. Importantly, small RNA sequencing studies with the 454 technology contributed to the discovery of a novel class of small RNAs, termed Piwi-interacting RNAs, that are expressed in mammalian testes and are presumably

required for germ cell development in mammals and other species [86–88].

The Illumina and SOLiD platforms allow for the generation of several millions of short 35-nt reads in comparison to up to 500,000 reads generated by the 454 technology [74], potentially providing even deeper coverage of small RNAs than the 454 platform. Our laboratory has recently used the Illumina technology to sequence small RNA libraries from human embryonic stem cells before and after their differentiation into embryonic bodies [30]. This study generated more than 6 million short sequence reads from each library and identified 334 known and 104 novel miRNA genes in one of the most comprehensive miRNA profiling exercises to date [30]. High-throughput sequencing-based approaches to small ncRNA profiling, hugely enabled by next-generation technologies, provide several advantages over microarray methods, including the ability to discover novel miRNAs and the potential to detect variations in the mature miRNA length and miRNA editing [30]. We envision that such approaches will gain even more popularity in the near future as the Illumina and SOLiD platforms are exploited in more laboratories.

Protein coding gene annotation using transcriptome sequence data

Despite the explosion of genome sequence data from multiple species fueled by advances in sequencing technologies, genome annotation for most multicellular eukaryotic species is still at its rudimentary stages. In particular, recent annotation efforts have focused on the discovery of novel noncoding RNA genes and regulatory elements that determine temporal or spatial gene expression; however, the annotation of protein-coding genes involving the elucidation of their correct exon–intron structures largely has lagged behind [89].

The current gold standard for protein-coding gene annotation is EST or full-length cDNA sequencing followed by alignment to a reference genome assembly [89]. The cDNA sequences can be aligned either to the locus from which they had been derived (*cis*-alignments) or to a homologous locus from the source genome or the genome of a related organism (*trans*-alignments). It has been estimated that most EST sequencing projects fail to cover 20–40% of transcripts, which usually include rare or very long transcripts as well as transcripts with highly specific expression patterns [89]. Another challenge of EST-driven gene annotation is alternative splicing and the complex structure of many loci from multicellular eukaryotes, resulting in a substantial number of incomplete annotations. Next-generation sequencing technologies have the potential for providing much deeper coverage of EST libraries; however, the short reads may be problematic when annotating alternative splice variants and the complete accurate structures of protein coding loci [89].

A recent study used laser capture microdissection [90] to isolate transcripts from the shoot apical meristem of *Z. mays* followed by cDNA library construction and 454 sequencing of ESTs [91]. The study used a *cis*-alignment method to annotate more than 25,000 genomic sequences from maize and detect transcription from 400 orphan genes, most of which had not been detected using other approaches [91]. Another study used the 454 technology to generate 391,157 EST reads from the brain transcriptome of the wasp *P. metricus;* the reads were then *trans*-aligned to the genome sequence and EST resources from the honeybee, *Apis mellifera,* to annotate *P. metricus* transcripts [65]. 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. Significantly, many gene expression profiling studies that use high-throughput sequencing can also provide annotation information, such as the presence of novel genes, exons, or splice events. For instance, our own study involving the generation of ESTs from the prostate cancer cell line LNCaP characterized 25 novel splicing events [29]. Another example is the 5′-RATE method [55] described above, which is particularly useful for providing annotation information about transcriptional start sites and other molecular events involving the 5′ end of transcripts. In addition, 454 transcriptome sequencing data can be useful for identifying SNPs in coding regions [92]. However, as mentioned earlier, error rates associated with next-generation sequencers require a relatively high fold coverage to call a sequence polymorphism reliably, particularly in a heterozygous sample [46].

The much deeper sequencing capacity of next-generation sequencing comes at the cost of shorter reads, which create additional challenges for gene annotation (e.g., difficulties in resolving splice isoforms). Paired-end sequencing approaches, such as the newly developed 454 sequencing-based MS-PET strategy, may facilitate annotation studies by providing mate pair information from large DNA fragments [42].

Detection of aberrant transcription events

Genome rearrangements resulting in aberrant transcriptional events are hallmarks of human cancers [93]. Techniques for detecting such genome rearrangements include cytogenetic and PCR methods as well as high-throughput array-based approaches, most notably array comparative genomic hybridization and sequencing-based techniques. Sequencing methods for genome rearrangement detection offer several advantages over array methods, such as the ability to detect multiple types of rearrangements, including previously unknown ones; the detection of absolute rather than relative changes in sequence copy numbers; and the potential for single-nucleotide resolution [39]. Large-scale transcriptome sequencing studies provide a novel means for detecting genome rearrangements in the transcribed portion of the genome. However, due to the short-read-length issue, single-end transcriptome sequencing studies using next-generation technologies, including the discussed EST studies, would be of limited use for identifying rearrangements [44].

An elegant gene identification signature analysis using paired-end ditag transcriptome sequencing methodology has been developed for the detection of gene fusions and other aberrant transcripts in cancers [44]. The approach involves generation of 18-nt-long tags from both ends of a transcript, which are then concatenated and sequenced by the 454 technology. This strategy is particularly useful for detecting fusion events in cancers, as well as actively transcribed pseudogenes that are readily distinguishable from their source genomic loci [44]. Related technologies involving other next-generation sequencing platforms are currently being developed by several laboratories.

Applications of next-generation sequencing for the analysis of epigenetic modifications of histones and DNA

Epigenetics is the study of heritable gene regulation that does not involve the DNA sequence. The two major types of epigenetic modifications regulating gene expression are DNA methylation by covalent modification of cytosine-5′ and posttranslational modifications of histone tails [94]. Regulatory RNAs provide another means of epigenetic regulation of gene expression; however, the focus of this section is on applications of next-generation sequencing to the analysis of covalent modifications of DNA and chromatin. Recent research has implicated such epigenetic modifications of prime importance in oncogenesis and development, setting the grounds for the Human Epigenome Project (HEP) initiative, which aims to catalog DNA methylation patterns on a genome-wide scale [95]. The next-generation sequencing technologies offer the potential to accelerate epigenomic research substantially. To date, these technologies have been applied in several epigenomic areas, including the characterization of DNA methylation patterns, posttranslational modifications of histones, and nucleosome positioning on a genome-wide scale.

DNA methylation profiling by bisulfite DNA sequencing

Cataloging genome-wide DNA methylation patterns, the most commonly studied epigenetic modification, is the primary goal of the HEP [95]. As part of the project, methylation profiles have been generated for chromosomes 6, 20, and 22 in 12 different tissues using bisulfite DNA sequencing on a Sanger instrument [96]. There are three main approaches to detecting DNA methylation on a large scale, including restriction endonuclease digestion coupled to microarray technology, bisulfite sequencing, and immunoprecipitation of 5′-methylcytosine to separate methylated from unmethylated DNA (for details see review by Callinan and Feinberg [94]). Bisulfite sequencing, the approach used in the HEP, is based on the chemical property of bisulfite to induce the conversion of cytosine residues to uracils while leaving 5′-methylcytosines intact. Therefore, sequencing of bisulfite-treated DNA will reveal the positions of methylated cytosines (those positions that remained cytosines following the treatment). A recent study by Taylor et al. [97] improved upon the bisulfite DNA sequencing procedure by using the 454 technology to sequence bisulfite-treated PCR amplicons corresponding to gene-related CpG-rich regions. The method, termed ultradeep bisulfite sequencing, was applied to examine methylation patterns at 25 gene-related CpG-rich regions in several hematopoietic tumors. The study generated >1600 individual sequences from each amplicon in contrast to the approximately 20 clones typically generated by conventional bisulfite sequencing, providing a superior robust alternative that does not involve cloning and allows for the simultaneous analysis of multiple genes and multiple samples [97]. A similar study using the Illumina technology has been recently reported [98].

Sequence census applications for mapping histone modifications and the locations of DNA-binding proteins

Posttranslational covalent modifications of histone tails, which include methylation, acetylation, phosphorylation, and ADP-ribosylation, are thought to control gene expression by regulating the strength of DNA–histone interactions determining the accessibility of DNA to transcriptional regulators [99]. Historically, histone modifications have been identified by chromatin immunoprecipitation (ChIP) which, in brief, involves cross-linking proteins to DNA, followed by immunoprecipitation of a protein of interest with a specific antibody, and characterization of the bound DNA by hybridization [38] or PCR amplification [100]. The genome-wide development of the ChIP method using microarrays, known as ChIP–chip, combined the ChIP procedure with hybridization to a microarray to reveal the genome-wide distribution of the protein of interest [101].

Sequence census methods have been recently coupled to the basic ChIP protocol to provide an alternative method for surveying histone modifications on a genome-wide scale. Roh et al. [102] used ChIP followed by a Sanger sequencing-based SAGE procedure (also referred to as the genome-wide mapping technique) to study the distribution of acetylated histones H3 and H4 in the yeast genome [102]. Bhinge et al. [103] replaced Sanger sequencing with the 454 sequencing technology and termed the method sequence tag analysis of genomic enrichment (STAGE). STAGE was successfully applied to identifying the genome-wide binding locations of the STAT1 transcription factor [103]; however, as in the case of the ChIP–SAGE protocol, it can also be used for the detection of histone modifications. Importantly, in these two methods the sequence reads are derived from the areas of ChIP DNA that are next to a restriction endonuclease site used in SAGE.

The introduction of next-generation sequencing to the field has brought about the development of a new sequencing-based method, named ChIP–Seq, for detecting histone modifications on a genome-wide scale. In this method, immunoprecipitated DNA is used to construct sequencing libraries for analysis on a next-generation sequencer to generate short sequence reads that, in contrast to ChIP–SAGE or STAGE, could be derived from either end of a ChIP DNA fragment regardless of the presence of a restriction site [35,36]. The number of reads that map to a particular genomic area can be used to quantify the strength of binding of the protein of interest in this area (or the amount of the assayed histone modification found at the site). To date, Illumina technology has been most commonly used for the ChIP–Seq application. The 25- to 30-bp read length obtained on an Illumina sequencer suffices to map a typical 150- to 200-bp ChIP DNA fragment that may be sequenced from both ends. ChIP–Seq has been applied to the identification of histone modifications on a genome-wide scale in the human genome [32,37]. In addition, it has been also used to reveal the genome-wide locations of transcription factor binding sites of STAT1 and NRSF [34,36].

Of the genome-wide extensions of the ChIP protocol, ChIP–Seq has the potential for the highest resolution as its resolution depends only on the size of the input chromatin fragments and the depth of sequencing. On the other hand, the resolution of ChIP–SAGE (STAGE) also depends on the distribution of the restriction enzyme sites in the input ChIP DNA. The resolution of ChIP–chip depends on the resolution of probes used for the microarray. Both ChIP–SAGE and ChIP–Seq require less PCR amplification than ChIP–chip and, therefore, may provide improved accuracy for quantifying the binding signal [99].

Applications of next-generation sequencers to the study of DNA accessibility and chromatin structure

Next-generation sequencing technologies have been applied to mapping out the positions of nucleosomes and other determinants of DNA accessibility. Nucleosomes are important factors affecting gene regulation and are usually associated with decreased accessibility of DNA to regulatory proteins. Most commonly, nucleosomes are identified by preferential cleavage of linker DNA by micrococcal nuclease (MNase) [99]. The identity of MNase digestion products, revealed by hybridization or sequencing, marks the locations of nucleosomes. Two recent studies used MNase digestion followed by sequencing with the 454 technology to map genome-wide locations of nucleosomes H2A.Z in yeast [104] and nucleosome cores in *C. elegans* [105]. Albert et al. [104] used MNase digestion followed by immunoprecipitation with an anti-H2A.Z antibody to isolate preferentially nucleosomes associated with the particular histone, while Johnson et al. [105] directly sequenced fragments liberated by digestion.

ChIP–Seq data from genome-wide histone modification profiling experiments can also be used to infer nucleosomal positions on a genome-wide scale. For instance, Schmid and Bucher [106] used ChIP–Seq data obtained by Barski et al. [37] to map the positions of two types of nucleosomes, as well as RNA PolII transcription preinitiation complexes in human $CD4^+$ T cells. They achieved this by separately analyzing sequence tags from two DNA strands and assuming that the tags mapping to the sense and antisense strands defined the 5′ and 3′ boundaries, respectively, of protein–DNA complexes [106]. The results obtained from this analysis correlated well with similar findings by Albert et al. [104], who determined the distribution of H2A.Z nucleosomes in yeast and found strong phasing of this type of nucleosome downstream of transcription start sites [104]. The use of ChIP–Seq data to map the nucleosome positions has three main limitations. First, only nucleosomes associated with a specific histone modification can be mapped in this manner; second, nucleosome positioning is regulated by certain histone modifications which can, for instance, mark a given nucleosome for removal [107]; and third, the method is not quantitative, as the abundance of reads mapping to a particular region is correlated with the abundance of the histone modification, which is not necessarily correlated with the abundance of the nucleosome [37].

Other uses of the pyrosequencing technology in epigenomics have included identifying DNase I-hypersensitive sites to help infer the role

of 3′-*BCL11B* in leukemic *cis*-activation [108] and in the development of the chromosome conformation capture (3C) method to detect higher order chromosomal structures or physical interactions between genomic loci in a high-throughput manner [109]. The original 3C method uses formaldehyde cross-linking followed by restriction enzyme digestion and intramolecular ligation to detect physically interacting genomic loci that are presumably important for regulating gene expression. The abundance of a particular ligation product (detected by quantitative PCR) is a measure of the frequency with which the particular loci interact in the nucleus. The sequencing development of 3C, termed chromosome conformation capture carbon copy (5C), replaces the quantitative PCR detection by 454 sequencing, enabling the use of the 3C approach on a genome-wide scale. The 5C approach successfully identified known and novel looping interactions involving the β-globin locus [109].

Concluding remarks

Due to their much improved cost effectiveness, compared to Sanger sequencing, and their many different uses, next-generation sequencing approaches are poised to emerge as the dominant genomics technology. Perhaps most significantly, these new sequencers have provided genome-scale sequencing capacity to individual laboratories in addition to larger genome centers. Compared to Sanger sequencing, advantages of the next-generation technologies mentioned thus far, including 454/Roche [19], Illumina/Solexa [24], and ABI/SOLiD [26], alleviate the need for in vivo cloning by clonal amplification of spatially separated single molecules using either emulsion PCR (454/Roche and ABI/SOLiD) or bridge amplification on solid surface (Illumina/Solexa). In addition to providing a means for cloning-free amplification, these methods use single-molecule templates allowing for the detection of heterogeneity in a DNA sample (e.g., identifying mutations present only in a subpopulation of cells), which is a significant advantage over Sanger sequencing [25,46].

The short read structure of next-generation sequencers provides potential problems for sequence assembly particularly in areas associated with sequence repeats. However, it has found broad applicability in sequence census studies, wherein determining the sequence of the whole DNA molecule is not essential [27]. The short read length also necessitates the development of paired-end sequencing approaches for improved mapping efficiency [41]. To date, such approaches have been reported for the 454 technology (e.g., Korbel et al. [43]) and are being made available for the Illumina and SOLiD platforms. The accuracy of next-generation sequencers is improving, but users generally rely on relatively high redundancy of sequence coverage to determine reliably the sequence of a region, particularly of that containing a polymorphism [46]. Addressing the accuracy issue by improving the reaction chemistry has the potential of further decreasing the current sequencing cost associated with next-generation sequencers.

Next-generation sequencing technologies have found broad applicability in functional genomics research. Their applications in the field have included gene expression profiling, genome annotation, small ncRNA discovery and profiling, and detection of aberrant transcription, which are areas that have been previously dominated by microarrays. Significantly, several studies found that 454 sequencing correlated well with the established gene expression profiling technologies such as microarray results (correlation coefficients of 0.83–0.91) [28] and moderately with SAGE data (correlation coefficients of 0.45) [29]. While the transcriptome sequencing studies discussed here predominantly used the 454 technology, Illumina and SOLiD technologies also offer significant potential for such applications. Another major functional genomic application is determining DNA sequences associated with epigenetic modifications of histones and DNA. Next-generation sequencing approaches have been used in this field to profile DNA methylations, posttranslational modifications of histones, and nucleosome positions on a genome-wide scale. While these areas have been previously addressed by Sanger sequencing, next-generation technologies have improved upon the throughput, the depth of coverage, and the resolution of Sanger sequencing studies.

Despite the recent exciting research advances involving next-generation sequencers, it should be noted that method development is still in its infancy. Efficient data analysis pipelines are required for many applications before they become routine, and more studies are needed to address the robustness of these techniques as well as the correspondence of results with those obtained by previous methods. Although next-generation sequencers are already being widely used, there are other sequencing methods, such as nanopore sequencing [110], whose scalability is being explored to decrease the sequencing cost and enhance throughput even further.

Acknowledgments

Support for this work was received from the National Cancer Institute of Canada, Genome Canada, Genome British Columbia, and British Columbia Cancer Foundation. O.M. is a junior trainee of the Michael Smith Foundation for Health Research and a graduate trainee of the Michael Smith Foundation for Health Research/Canadian Institutes of Health Research Bioinformatics Training Program and is supported by a Natural Sciences and Engineering Research Council of Canada fellowship. M.A.M. is a senior scholar of the Michael Smith Foundation for Health Research.

References

[1] R.E. Green, et al., Analysis of one million base pairs of Neanderthal DNA, Nature 444 (2006) 330–336.
[2] N. Hall, Advanced sequencing technologies and their wider impact in microbiology, J. Exp. Biol. 210 (2007) 1518–1525.
[3] F. Sanger, A.R. Coulson, A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase, J. Mol. Biol. 94 (1975) 441–448.
[4] A.M. Maxam, W. Gilbert, A new method for sequencing DNA, Proc. Natl. Acad. Sci. U. S. A. 74 (1977) 560–564.
[5] F. Sanger, S. Nicklen, A.R. Coulson, DNA sequencing with chain-terminating inhibitors, Proc. Natl. Acad. Sci. USA 74 (1977) 5463–5467.
[6] E.S. Lander, et al., Initial sequencing and analysis of the human genome, Nature 409 (2001) 860–921.
[7] International Human Genome Sequencing Consortium, Finishing the euchromatic sequence of the human genome, Nature 431 (2004) 931–945.
[8] J.C. Venter, et al., The sequence of the human genome, Science 291 (2001) 1304–1351.
[9] C.A. Hutchison III, DNA sequencing: bench to bedside and beyond, Nucleic Acids Res. 35 (2007) 6227–6237.
[10] M.L. Metzker, Emerging technologies in DNA sequencing, Genome Res. 15 (2005) 1767–1776.
[11] L.M. Smith, et al., Fluorescence detection in automated DNA sequence analysis, Nature 321 (1986) 674–679.
[12] J.M. Prober, et al., A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides, Science 238 (1987) 336–341.
[13] A.S. Cohen, D.R. Najarian, A. Paulus, A. Guttman, J.A. Smith, B.L. Karger, Rapid separation and purification of oligonucleotides by high-performance capillary gel electrophoresis, Proc. Natl. Acad. Sci. U. S. A. 85 (1988) 9660–9663.
[14] X.C. Huang, M.A. Quesada, R.A. Mathies, DNA sequencing using capillary array electrophoresis, Anal. Chem. 64 (1992) 2149–2154.
[15] M.C. Ruiz-Martinez, J. Berka, A. Belenkii, F. Foret, A.W. Miller, B.L. Karger, DNA sequencing by capillary electrophoresis with replaceable linear polyacrylamide and laser-induced fluorescence detection, Anal. Chem. 65 (1993) 2851–2858.
[16] R.S. Madabhushi, Separation of 4-color DNA sequencing extension products in noncovalently coated capillaries using low viscosity polymer solutions, Electrophoresis 19 (1998) 224–230.
[17] E. Carrilho, DNA sequencing by capillary array electrophoresis and microfabricated array systems, Electrophoresis 21 (2000) 55–65.
[18] B. Ewing, P. Green, Base-calling of automated sequencer traces using phred. II. Error probabilities, Genome Res. 8 (1998) 186–194.
[19] M. Margulies, et al., Genome sequencing in microfabricated high-density picolitre reactors, Nature 437 (2005) 376–380.
[20] D.S. Tawfik, A.D. Griffiths, Man-made cell-like compartments for molecular evolution, Nat. Biotechnol. 16 (1998) 652–656.
[21] P. Nyren, B. Pettersson, M. Uhlen, Solid phase DNA minisequencing by an enzymatic luminometric inorganic pyrophosphate detection assay, Anal. Biochem. 208 (1993) 171–175.

[22] M. Ronaghi, S. Karamohamed, B. Pettersson, M. Uhlen, P. Nyren, Real-time DNA sequencing using detection of pyrophosphate release, Anal. Biochem. 242 (1996) 84–89.
[23] S. Bennett, Solexa Ltd, Pharmacogenomics 5 (2004) 433–438.
[24] S.T. Bennett, C. Barnes, A. Cox, L. Davies, C. Brown, Toward the 1,000 dollars human genome, Pharmacogenomics 6 (2005) 373–382.
[25] D.R. Bentley, Whole-genome re-sequencing, Curr. Opin. Genet. Dev. 16 (2006) 545–552.
[26] J. Shendure, et al., Accurate multiplex polony sequencing of an evolved bacterial genome, Science 309 (2005) 1728–1732.
[27] B. Wold, R.M. Myers, Sequence census methods for functional genomics, Nat. Methods 5 (2008) 19–21.
[28] T.T. Torres, M. Metta, B. Ottenwalder, C. Schlotterer, Gene expression profiling by massively parallel sequencing, Genome Res. 18 (2008) 172–177.
[29] M.N. Bainbridge, et al., Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach, BMC Genomics 7 (2006) 246.
[30] R.D. Morin, et al., Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells, Genome Res. 18 (2008) 610–621.
[31] V.E. Velculescu, L. Zhang, B. Vogelstein, K.W. Kinzler, Serial analysis of gene expression, Science 270 (1995) 484–487.
[32] T.S. Mikkelsen, et al., Genome-wide maps of chromatin state in pluripotent and lineage-committed cells, Nature 448 (2007) 553–560.
[33] E.R. Mardis, ChIP-seq: welcome to the new frontier, Nat. Methods 4 (2007) 613–614.
[34] G. Robertson, et al., Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing, Nat. Methods 4 (2007) 651–657.
[35] S. Fields, Molecular biology: site-seeing by sequencing, Science 316 (2007) 1441–1442.
[36] D.S. Johnson, A. Mortazavi, R.M. Myers, B. Wold, Genome-wide mapping of in vivo protein–DNA interactions, Science 316 (2007) 1497–1502.
[37] A. Barski, et al., High-resolution profiling of histone methylations in the human genome, Cell 129 (2007) 823–837.
[38] D.S. Gilmour, J.T. Lis, Detecting protein–DNA interactions in vivo: distribution of RNA polymerase on specific bacterial genes, Proc. Natl. Acad. Sci. U. S. A. 81 (1984) 4275–4279.
[39] O. Morozova, M.A. Marra, From cytogenetics to next-generation sequencing technologies: advances in the detection of genome rearrangements in tumors, Biochem. Cell Biol. 86 (2008) 81–91.
[40] E. Tuzun, et al., Fine-scale structural variation of the human genome, Nat. Genet. 37 (2005) 727–732.
[41] E.R. Mardis, Anticipating the 1,000 dollar genome, Genome Biol. 7 (2006) 112.
[42] P. Ng, et al., Multiplex sequencing of paired-end ditags (MS-PET): a strategy for the ultra-high-throughput analysis of transcriptomes and genomes, Nucleic Acids Res. 34 (2006) e84.
[43] J.O. Korbel, et al., Paired-end mapping reveals extensive structural variation in the human genome, Science 318 (2007) 420–426.
[44] Y. Ruan, et al., Fusion transcripts and transcribed retrotransposed loci discovered through comprehensive transcriptome analysis using paired-end ditags (PETs), Genome Res. 17 (2007) 828–838.
[45] W. Brockman, et al., Quality scores and SNP detection in sequencing-by-synthesis systems, Genome Res. 18 (2008) 763–770.
[46] R.K. Thomas, et al., Sensitive mutation detection in heterogeneous cancer specimens by massively parallel picoliter reactor sequencing, Nat. Med. 12 (2006) 852–855.
[47] S. Tyagi, Taking a census of mRNA populations with microbeads, Nat. Biotechnol. 18 (2000) 597–598.
[48] M. Schena, D. Shalon, R.W. Davis, P.O. Brown, Quantitative monitoring of gene expression patterns with a complementary DNA microarray, Science 270 (1995) 467–470.
[49] S.A. Ness, Microarray analysis: basic strategies for successful experiments, Mol. Biotechnol. 36 (2007) 205–219.
[50] S. Brenner, et al., Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays, Nat. Biotechnol. 18 (2000) 630–634.
[51] S. Audic, J.M. Claverie, The significance of digital gene expression profiles, Genome Res. 7 (1997) 986–995.
[52] K.L. Nielsen, A.L. Hogh, J. Emmersen, DeepSAGE—digital transcriptomics with high sensitivity, simple experimental protocol and multiplexing of samples, Nucleic Acids Res. 34 (2006) e133.
[53] S.M. Wang, Understanding SAGE data, Trends Genet. 23 (2007) 42–50.
[54] S. Saha, et al., Using the transcriptome to annotate the genome, Nat. Biotechnol. 20 (2002) 508–512.
[55] M. Gowda, H. Li, J. Alessi, F. Chen, R. Pratt, G.L. Wang, Robust analysis of 5¢-transcript ends (5¢-RATE): a novel technique for transcriptome analysis and genome annotation, Nucleic Acids Res. 34 (2006) e126.
[56] M. Gowda, C. Jantasuriyarat, R.A. Dean, G.L. Wang, Robust-LongSAGE (RL-SAGE): a substantially improved LongSAGE method for gene discovery and transcriptome analysis, Plant Physiol. 134 (2004) 890–897.
[57] M. Seki, et al., Functional annotation of a full-length Arabidopsis cDNA collection, Science 296 (2002) 141–145.
[58] L.D. Hillier, et al., Generation and analysis of 280,000 human expressed sequence tags, Genome Res. 6 (1996) 807–828.
[59] W.R. McCombie, et al., Caenorhabditis elegans expressed sequence tags identify gene families and potential disease gene homologues, Nat. Genet. 1 (1992) 124–131.
[60] M. Sun, G. Zhou, S. Lee, J. Chen, R.Z. Shi, S.M. Wang, SAGE is far more sensitive than EST for detecting low-abundance transcripts, BMC Genomics 5 (2004) 1.
[61] A.P. Weber, K.L. Weber, K. Carr, C. Wilkerson, J.B. Ohlrogge, Sampling the Arabidopsis transcriptome with massively parallel pyrosequencing, Plant Physiol. 144 (2007) 32–42.
[62] M.W. Jones-Rhoades, J.O. Borevitz, D. Preuss, Genome-wide expression profiling of the Arabidopsis female gametophyte identifies families of small, secreted proteins, PLoS Genet. 3 (2007) 1848–1861.
[63] F. Cheung, B.J. Haas, S.M. Goldberg, G.D. May, Y. Xiao, C.D. Town, Sequencing Medicago truncatula expressed sequenced tags using 454 Life Sciences technology, BMC Genomics 7 (2006) 272.
[64] K. Ohtsu, et al., Global gene expression analysis of the shoot apical meristem of maize (Zea mays L.), Plant J. 52 (2007) 391–404.
[65] A.L. Toth, et al., Wasp gene expression supports an evolutionary link between maternal behavior and eusociality, Science 318 (2007) 441–444.
[66] J.C. Vera, et al., Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing, Mol. Ecol. 17 (2008) 1636–1647.
[67] A.L. Eveland, D.R. McCarty, K.E. Koch, Transcript profiling by 3¢-untranslated region sequencing resolves expression of gene families, Plant Physiol. 146 (2008) 32–44.
[68] W. Filipowicz, S.N. Bhattacharyya, N. Sonenberg, Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? Nat. Rev. Genet. 9 (2008) 102–114.
[69] C. Lu, S.S. Tej, S. Luo, C.D. Haudenschild, B.C. Meyers, P.J. Green, Elucidation of the small RNA component of the transcriptome, Science 309 (2005) 1567–1569.
[70] B.C. Meyers, F.F. Souret, C. Lu, P.J. Green, Sweating the small stuff: microRNA discovery in plants, Curr. Opin. Biotechnol. 17 (2006) 139–146.
[71] Z. Xie, et al., Genetic and functional diversification of small RNA pathways in plants, PLoS Biol. 2 (2004) E104.
[72] R.C. Lee, R.L. Feinbaum, V. Ambros, The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14, Cell 75 (1993) 843–854.
[73] E. Berezikov, E. Cuppen, R.H. Plasterk, Approaches to microRNA discovery, Nat. Genet. 38 Suppl. (2006) S2–S7.
[74] M.J. Axtell, C. Jan, R. Rajagopalan, D.P. Bartel, A two-hit trigger for siRNA biogenesis in plants, Cell 127 (2006) 565–577.
[75] I.R. Henderson, et al., Dissecting Arabidopsis thaliana DICER function in small RNA processing, gene silencing and DNA methylation patterning, Nat. Genet. 38 (2006) 721–725.
[76] C. Lu, et al., MicroRNAs and other small RNAs enriched in the Arabidopsis RNA-dependent RNA polymerase-2 mutant, Genome Res. 16 (2006) 1276–1288.
[77] R. Rajagopalan, H. Vaucheret, J. Trejo, D.P. Bartel, A diverse and evolutionarily fluid set of microRNAs in Arabidopsis thaliana, Genes Dev. 20 (2006) 3407–3425.
[78] N. Fahlgren, et al., High-throughput sequencing of Arabidopsis microRNAs: evidence for frequent birth and death of MIRNA genes, PLoS One 2 (2007) e219.
[79] K.D. Kasschau, et al., Genome-wide profiling and analysis of Arabidopsis siRNAs, PLoS Biol. 5 (2007) e57.
[80] M.D. Howell, et al., Genome-wide analysis of the RNA-dependent RNA polymerase 6/dicer-like 4 pathway in Arabidopsis reveals dependency on miRNA- and tasiRNA-directed targeting, Plant Cell 19 (2007) 926–942.
[81] Y. Yao, et al., Cloning and characterization of microRNAs from wheat (Triticum aestivum L.), Genome Biol. 8 (2007) R96.
[82] A. Barakat, K. Wall, J. Leebens-Mack, Y.J. Wang, J.E. Carlson, C.W. Depamphilis, Large-scale identification of microRNAs from a basal eudicot (Eschscholzia californica) and conservation in flowering plants, Plant J. 51 (2007) 991–1003.
[83] T. Zhao, et al., A complex system of small RNAs in the unicellular green alga Chlamydomonas reinhardtii, Genes Dev. 21 (2007) 1190–1203.
[84] J. Burnside, et al., Marek's disease virus encodes microRNAs that map to meq and the latency-associated transcript, J. Virol. 80 (2006) 8778–8786.
[85] E. Berezikov, et al., Diversity of microRNAs in human and chimpanzee brain, Nat. Genet. 38 (2006) 1375–1377.
[86] N.C. Lau, et al., Characterization of the piRNA complex from rat testes, Science 313 (2006) 363–367.
[87] A. Girard, R. Sachidanandam, G.J. Hannon, M.A. Carmell, A germline-specific class of small RNAs binds mammalian Piwi proteins, Nature 442 (2006) 199–202.
[88] S. Houwing, et al., A role for Piwi and piRNAs in germ cell maintenance and transposon silencing in zebrafish, Cell 129 (2007) 69–82.
[89] M.R. Brent, Steady progress and recent breakthroughs in the accuracy of automated genome annotation, Nat. Rev. Genet. 9 (2008) 62–73.
[90] M.R. Emmert-Buck, et al., Laser capture microdissection, Science 274 (1996) 998–1001.
[91] S.J. Emrich, W.B. Barbazuk, L. Li, P.S. Schnable, Gene discovery and annotation using LCM-454 transcriptome sequencing, Genome Res. 17 (2007) 69–73.
[92] W.B. Barbazuk, S.J. Emrich, H.D. Chen, L. Li, P.S. Schnable, SNP discovery via 454 transcriptome sequencing, Plant J. 51 (2007) 910–918.
[93] D. Hanahan, R.A. Weinberg, The hallmarks of cancer, Cell 100 (2000) 57–70.
[94] P.A. Callinan, A.P. Feinberg, The emerging science of epigenomics, Hum. Mol. Genet. 15 Spec. No. 1 (2006) R95–R101.
[95] M. Esteller, The necessity of a human epigenome project, Carcinogenesis 27 (2006) 1121–1125.
[96] F. Eckhardt, et al., DNA methylation profiling of human chromosomes 6, 20 and 22, Nat. Genet. 38 (2006) 1378–1385.
[97] K.H. Taylor, et al., Ultradeep bisulfite sequencing analysis of DNA methylation patterns in multiple gene promoters by 454 sequencing, Cancer Res. 67 (2007) 8511–8518.

[98] S.J. Cokus, et al., Shotgun bisulfite sequencing of the Arabidopsis genome reveals DNA methylation patterning, Nature 452 (2008) 215–219.
[99] D.E. Schones, K. Zhao, Genome-wide approaches to studying chromatin modifications, Nat. Rev. Genet. 9 (2008) 179–191.
[100] M.H. Kuo, C.D. Allis, In vivo cross-linking and immunoprecipitation for studying dynamic protein:DNA associations in a chromatin environment, Methods 19 (1999) 425–433.
[101] B. Ren, et al., Genome-wide location and function of DNA binding proteins, Science 290 (2000) 2306–2309.
[102] T.Y. Roh, W.C. Ngau, K. Cui, D. Landsman, K. Zhao, High-resolution genome-wide mapping of histone modifications, Nat. Biotechnol. 22 (2004) 1013–1016.
[103] A.A. Bhinge, J. Kim, G.M. Euskirchen, M. Snyder, V.R. Iyer, Mapping the chromosomal targets of STAT1 by sequence tag analysis of genomic enrichment (STAGE), Genome Res. 17 (2007) 910–916.
[104] I. Albert, et al., Translational and rotational settings of H2A.Z nucleosomes across the Saccharomyces cerevisiae genome, Nature 446 (2007) 572–576.
[105] S.M. Johnson, F.J. Tan, H.L. McCullough, D.P. Riordan, A.Z. Fire, Flexibility and constraint in the nucleosome core landscape of Caenorhabditis elegans chromatin, Genome Res. 16 (2006) 1505–1516.
[106] C.D. Schmid, P. Bucher, ChIP-Seq data reveal nucleosome architecture of human promoters, Cell 131 (2007) 831–832 author reply 832-833.
[107] H. Boeger, J. Griesenbeck, J.S. Strattan, R.D. Kornberg, Nucleosomes unfold completely at a transcriptionally active promoter, Mol. Cell 11 (2003) 1587–1598.
[108] S. Nagel, et al., Activation of TLX3 and NKX2-5 in t(5;14)(q35;q32) T-cell acute lymphoblastic leukemia by remote 3¢-BCL11B enhancers and coregulation by PU.1 and HMGA1, Cancer Res. 67 (2007) 1461–1471.
[109] J. Dostie, et al., Chromosome conformation capture carbon copy (5C): a massively parallel solution for mapping interactions between genomic elements, Genome Res. 16 (2006) 1299–1309.
[110] D.W. Deamer, M. Akeson, Nanopores and nucleic acids: prospects for ultrarapid sequencing, Trends Biotechnol. 18 (2000) 147–151.