The impact of next-generation sequencing technology on genetics

Elaine R. Mardis

Washington University School of Medicine, Genome Sequencing Center, St. Louis, MO 63108, USA

If one accepts that the fundamental pursuit of genetics is to determine the genotypes that explain phenotypes, the meteoric increase of DNA sequence information applied toward that pursuit has nowhere to go but up. The recent introduction of instruments capable of producing millions of DNA sequence reads in a single run is rapidly changing the landscape of genetics, providing the ability to answer questions with heretofore unimaginable speed. These technologies will provide an inexpensive, genome-wide sequence readout as an endpoint to applications ranging from chromatin immunoprecipitation, mutation mapping and polymorphism discovery to noncoding RNA discovery. Here I survey next-generation sequencing technologies and consider how they can provide a more complete picture of how the genome shapes the organism.

First described by Sanger *et al.* in 1977 [1], dideoxynucleotide sequencing of DNA has undergone a steady metamorphosis from a cottage industry into a large-scale production enterprise that requires a specialized and devoted infrastructure of robotics, bioinformatics, computer databases and instrumentation. In the process of its metamorphosis, the cost per reaction of DNA sequencing has fallen with a Moore's Law [2] (ftp://download.intel.com/research/silicon/moorespaper.pdf) precision (Moore's Law describes the trend in the history of computer hardware, whereby the number of transistors that can be placed on an integrated circuit increases exponentially, doubling approximately every 2 years). This phenomenon has been especially true in the last 5 years, largely because of efforts necessary to sequence the human genome. Although still conducted at a subsistence level in the single investigator, departmental or university core facility setting, high-throughput DNA sequencing is so specialized that it is performed in a handful of sites (i.e. http://genome.wustl.edu, http://www.broad.mit.edu/, http://www.hgsc.bcm.tmc.edu/, http://www.sanger.ac.uk/). However, just as state-of-the-art high-throughput DNA sequencing seemed to be reaching its zenith at these sequencing centers, several new sequencing instruments (so-called 'next generation' or 'massively parallel') are becoming available and already are transforming the field. Their impact on genomics is in turn causing a revolution in genetics that, because of a variety of factors, will fundamentally change the nature of genetic experimentation. When coupled with the appropriate computational algorithms, our ability to answer questions about the mutational spectrum of an organism, from single base to large copy number polymorphisms, on a genome-wide scale, is likely to radically alter our understanding of model organisms and ultimately of ourselves. Here, I review a subset of the studies that have been enabled by next-generation sequencing platforms, to gain an appreciation of the breadth and depth of their potential.

Overview of next-generation instruments

What is it that sets next-generation sequencers apart from conventional capillary-based sequencing? Namely, the ability to process millions of sequence reads in parallel rather than 96 at a time. This massively parallel throughput may require only one or two instrument runs to complete an experiment. Also, next generation sequence reads are produced from fragment 'libraries' that have not been subject to the conventional vector-based cloning and *Escherichia coli*–based amplification stages used in capillary sequencing. As such, some of the cloning bias issues that impact genome representation in sequencing projects may be avoided, although each sequencing platform may have its own associated biases. The workflow to produce next-generation sequence-ready libraries is straightforward; DNA fragments that may originate from a variety of front-end processes (described below) are prepared for sequencing by ligating specific adaptor oligos to both ends of each DNA fragment. Importantly, relatively little input DNA (a few micrograms at most) is needed to produce a library. These platforms also have the ability to sequence the paired ends of a given fragment, using a slightly modified library process. This approach can be used if a *de novo* genome sequence is to be assembled from the next-generation data, for example. Finally, next-generation sequencers produce shorter read lengths (35–250 bp, depending on the platform) than capillary sequencers (650–800 bp), which also can impact the utility of the data for various applications such as *de novo* assembly and genome resequencing (Box 1). Because they are so new, the accuracy of their sequencing reads and associated quality values are not yet well understood, although many laboratories have efforts underway to benchmark them relative to capillary electrophoresis. Aside from these generally shared features, the three commercially available sequencers differ significantly (see Table 1 for a comparison of current specifications) and are described below.

Roche (454) GS FLX sequencer

First commercially introduced in 2004, this sequencer works on the principle of 'pyrosequencing', which uses the pyrophosphate molecule released on nucleotide incorp-

Corresponding author: Mardis, E.R. (emardis@watson.wustl.edu).

0168-9525/$ – see front matter © 2007 Elsevier Ltd. All rights reserved. doi:10.1016/j.tig.2007.12.007 Available online 11 February 2008

Box 1. The '$1000' genome – toward personalized genomics?

Next-generation sequencing technologies, by enabling vast data generation, will provide a comprehensive picture of normal human genome variation in the next few years. This will set the baseline by which genome variation in a genetic disease cohort can be evaluated. Efforts to couple the discovered variations to the disease biology will provide functional annotations for gene variants that predict disease susceptibility and genetic risk factors and provide pharmacokinetic profiles. Targeted treatments might also be suggested that selectively block the impact of certain variants. At this point, inexpensive (e.g. $1000) genome sequencing as a clinical assay or a point of entry to health insurance or medical care becomes meaningful. However, the current cost of resequencing a human genome is high, even with next-generation technology. One reason is that short reads (e.g. 35–50 bases) likely will require $\sim$25- to 30-fold oversampling, or 'coverage' of the genome, to ensure that both chromosomal pairs (haplotypes) are sampled sufficiently to capture all the genetic information. At 3Gb per run, 25–30 instrument runs would be required (the human genome is $\sim$3 Gb) to provide this coverage, costing $\sim$$700 000 per genome. In addition to requiring more coverage, short reads also are limited in the power to detect sequence variation in the genome, based on their uniqueness. For example, if a given 32-base sequence is found more than once in the reference sequence, that sequence is eliminated as a potential site of variant detection because of uncertainty in aligning a 32-base sequence read (even one that contains one or more potentially variant bases) at the correct genomic location.

oration by DNA polymerase to fuel a downstream set of reactions that ultimately produces light from the cleavage of oxyluciferin by luciferase [3] (Figure 1). Instead of sequencing in discrete tubes or in microtiter plate wells, the DNA strands of the library are amplified *en masse* by emulsion PCR [4] on the surfaces hundreds of thousands of agarose beads. The surfaces of these beads have millions of oligomers attached to them, each of which is complementary to the adaptor sequences that were ligated to the fragment ends during library construction (as described above). Emulsion PCR uses a vigorously mixed oil and aqueous mixture to isolate individual agarose beads, each having a single unique DNA fragment hybridized to the oligo-decorated surface, in aqueous micelles that also contain the PCR reactants. By pipetting these tiny micelles into the wells of a conventional microtiter plate and performing temperature cycling, one can produce >1 000 000 sequence-ready 454 beads in a matter of hours. Each agarose bead surface contains up to 1 000 000 copies of the original annealed DNA fragment to produce detectable signal from the sequencing reaction. Several hundred thousand such beads (each containing a unique amplified fragment) are added to the surface of the 454 picotiter plate (PTP), which consists of single wells in the tips of fused fiber optic strands that hold each bead. Subsequently, much smaller magnetic and latex beads of 1 μm diameter, which are attached to the active enzymes needed for pyrosequencing, are added to surround the DNA-containing agarose beads in the PTP. Because the PTP also becomes the cell through which the pyrosequencing reactants flow, it is placed in the sequencer, and nucleotide and reagent solutions are delivered into it in a sequential fashion. Imaging of the light flashes from luciferase activity records which templates are adding that particular nucleotide (see http://www.454.com/enabling-technology/the-technology.asp for an explanation of the workflow and technology), and the light emitted is directly proportional to the amount of a particular nucleotide incorporated (up to the level of detector saturation). Hence, for runs of multiple nucleotides (homopolymers), the linearity of response can exceed the detector sensitivity, at which indel errors can occur in those reads. That said, the sequential flow of nucleotides almost entirely precludes the occurrence of substitution errors in the sequences. Placing a defined single nucleotide pattern in the adaptor sequence that matches the sequence of the first four nucleotide flows enables the 454 analysis software to calibrate the level of light emitted from a single nucleotide incorporation, for purposes of downstream base-calling analysis that occurs after the sequencer run is completed. Here, the signals recorded during the run for each reporting bead position on the PTP are translated into a sequence read, and several quality-checking steps remove poor quality sequences. The current 454 instrument, the GS-FLX, produces an average read length of $\sim$250 bp per sample (per bead), with a combined throughput of $\sim$100 Mb of sequence data per 7-h run. By contrast, a single ABI 3730 programmed to sequence 24 × 96-well plates per day produces $\sim$440 kb of sequence data in 7 h, with an average read length of 650 bp per sample.

Illumina genome analyzer

Introduced in 2006, the Illumina Genome Analyzer is based on the concept of 'sequencing by synthesis' (SBS) to produce sequence reads of $\sim$32–40 bp from tens of millions of surface-amplified DNA fragments simultaneously (Figure 2). Starting from a mixture of single-stranded, adaptor oligo-ligated DNA fragments, the Illumina process involves using a microfluidic cluster station to add these fragments to the surface of a glass flow cell. Each flow cell is divided into eight separate lanes, and the interior surfaces have covalently

Table 1. Comparing metrics and performance of next-generation DNA sequencers

	Platform		
	Roche(454)	**Illumina**	**SOLiD**
Sequencing chemistry	Pyrosequencing	Polymerase-based sequencing-by-synthesis	Ligation-based sequencing
Amplification approach	Emulsion PCR	Bridge amplification	Emulsion PCR
Paired ends/separation	Yes/3 kb	yes/200 bp	Yes/3 kb
Mb/run	100 Mb	1300 Mb	3000 Mb
Time/run (paired ends)	7 h	4 days	5 days
Read length	250 bp	32–40 bp	35 bp
Cost per run (total direct[a])	$8439	$8950	$17 447
Cost per Mb	$84.39	$5.97	$5.81

[a] Total direct costs include the reagents and consumables, the labor, instrument amortization cost and the disc storage space required for data storage/access.

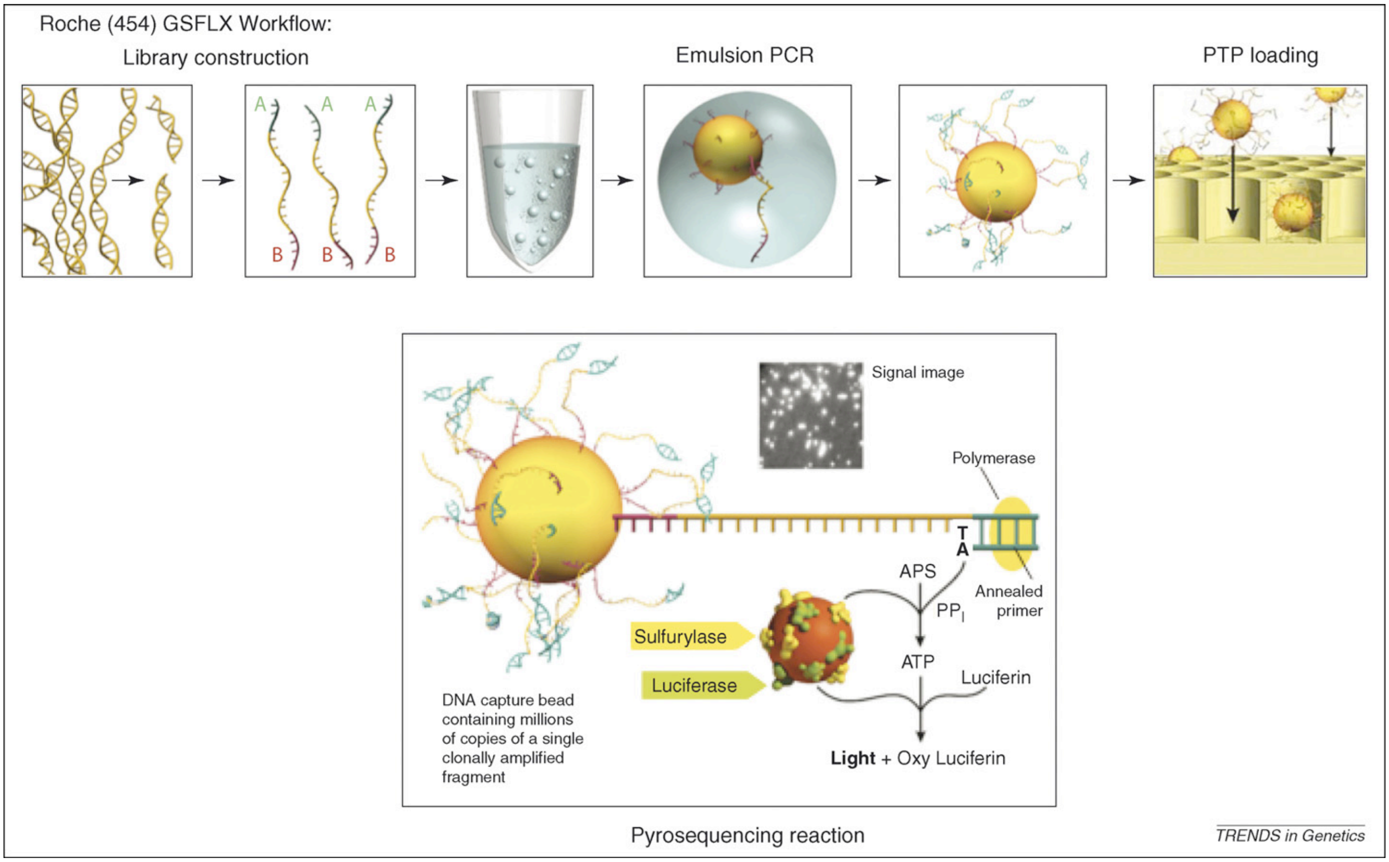

Figure 1. 454 Workflow: library construction ligates 454-specific adapters to DNA fragments and couples amplification beads with DNA in an emulsion PCR to amplify fragments before sequencing. The beads are loaded into the picotiter plate (PTP). The bottom panel illustrates the pyrosequencing reaction that occurs on nucleotide incorporation to report sequencing by synthesis.

attached oligos complementary to the specific adapters that are ligated onto the library fragments. Hybridization of these DNAs to the oligos on the flow cell occurs by an active heating and cooling step, followed by a subsequent incubation with reactants and an isothermal polymerase that amplifies the fragments in a discrete area or 'cluster' on the flow cell surfaces (see http://www.illumina.com/pages.ilmn?ID=203 for an animation of this process). The flow cell is placed into a fluidics cassette within the sequencer, where each cluster is supplied with polymerase and four differentially labeled fluorescent nucleotides that have their 3′-OH chemically inactivated to ensure that only a single base is incorporated per cycle. Each base incorporation cycle is followed by an imaging step to identify the incorporated nucleotide at each cluster and by a chemical step that removes the fluorescent group and deblocks the 3′ end for the next base incorporation cycle. At the end of the sequencing run (~4 days), the sequence of each cluster is computed and subjected to quality filtering to eliminate low-quality reads of between 32 and 40 bp (as specified by the user). A typical run yields ~40–50 million such sequences.

Applied Biosystems SOLiD sequencer

This instrument, which achieved commercial release in October 2007, uses a unique sequencing process catalyzed by DNA ligase. Each SOLiD (Sequencing by Oligo Ligation and Detection) run requires ~5 days and produces 3–4 Gb of sequence data with an average read length of 25–35 bp. The specific process couples oligo adaptor-linked DNA fragments with 1-μm magnetic beads that are decorated with complementary oligos and amplifies each bead–DNA complex by emulsion PCR. After amplification, the beads are covalently attached to the surface of a specially treated glass slide that is placed into a fluidics cassette within sequencer. In the SOLiD system, two slides are processed per run; one slide receives sequencing reactants as the second slide is being imaged. The ligation-based sequencing process starts with the annealing of a universal sequencing primer that is complementary to the SOLiD-specific adapters on the library fragments. The addition of a limited set of semi-degenerate 8mer oligonucleotides and DNA ligase is automated by the instrument. When a matching 8mer hybridizes to the DNA fragment sequence adjacent to the universal primer 3′ end, DNA ligase seals the phosphate backbone. After the ligation step, a fluorescent readout identifies the fixed base of the 8mer, which corresponds to either the fifth position or the second position, depending on the cycle number (see Table 2 for details). A subsequent chemical cleavage step removes the sixth through eighth base of the ligated 8mer by attacking the linkage between bases 5 and 6, thereby removing the fluorescent group and enabling a subsequent round of ligation. The process occurs in steps that identify the sequence of each fragment at five nucleotide intervals (Table 2), and the synthesized fragments that end at base 25 (or 35 if more cycles are performed) are removed by denaturation and washed away. A second round of sequencing initiates with the hybridization of an n-1 posi-

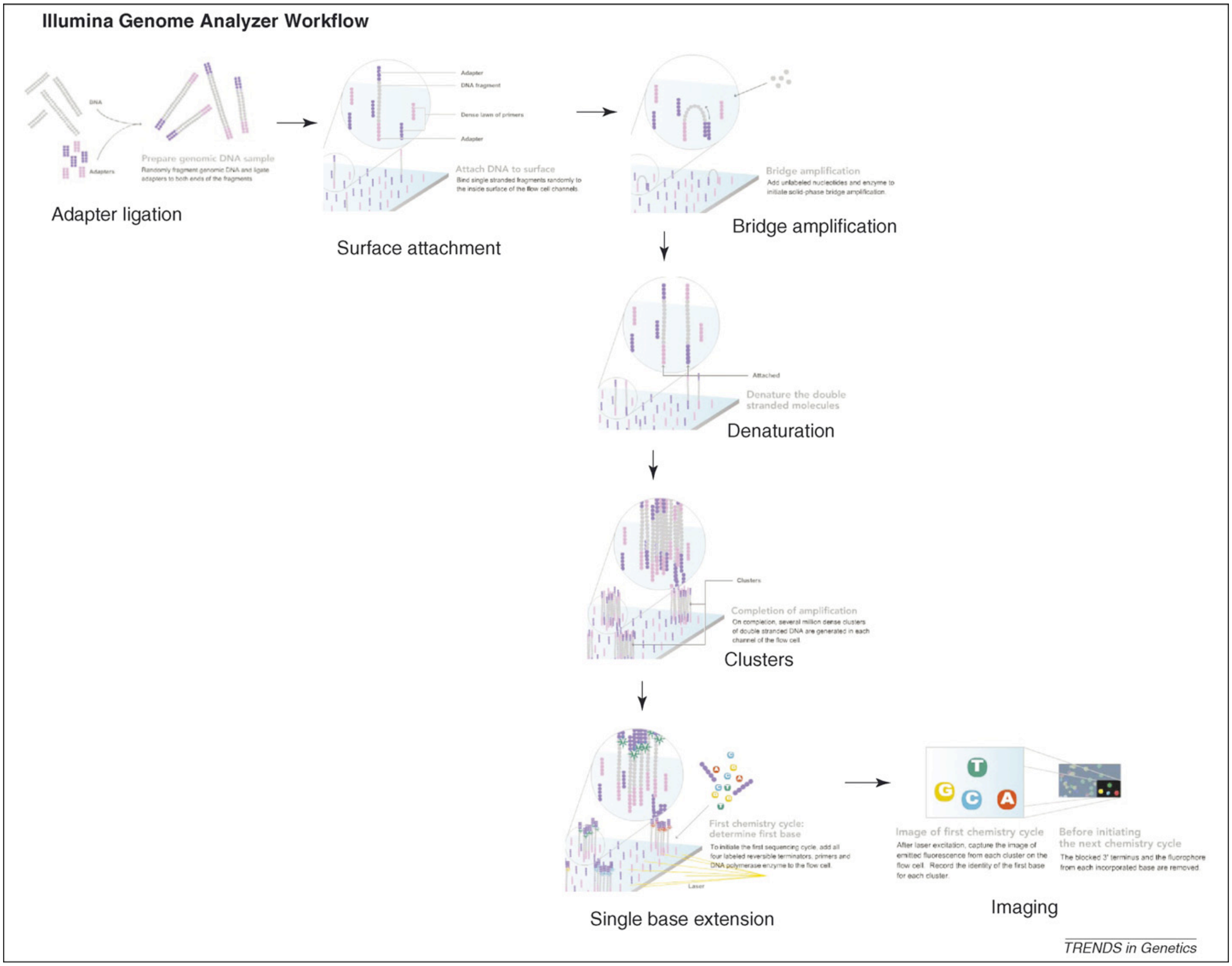
Illumina Genome Analyzer Workflow
DNA
Adapters
Prepare genomic DNA sample
Randomly fragment genomic DNA and ligate adapters to both ends of the fragments
Adapter ligation
Adapter
DNA fragment
Dense lawn of primers
Adapter
Attach DNA to surface
Bind single stranded fragments randomly to the inside surface of the flow cell channels
Surface attachment
Bridge amplification
Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification
Bridge amplification
Attached
Denature the double stranded molecules
Denaturation
Clusters
Completion of amplification
On completion, several million dense clusters of double stranded DNA are generated in each channel of the flow cell
Clusters
First chemistry cycle: determine first base
To initiate the first sequencing cycle, add all four labeled reversible terminators, primers and DNA polymerase enzyme to the flow cell
Laser
Single base extension
T
G
C
A
Image of first chemistry cycle
After laser excitation, capture the image of emitted fluorescence from each cluster on the flow cell. Record the identity of the first base for each cluster
Before initiating the next chemistry cycle
The blocked 3′ terminus and the fluorophore from each incorporated base are removed
Imaging
TRENDS in Genetics

Figure 2. Illumina workflow. Starting from similar fragmentation and adapter ligation steps, the library is added to a flow cell for bridge amplification (an isothermal process that amplifies each fragment into a cluster). The cluster fragments are denatured, annealed with a sequencing primer and subjected to sequencing by synthesis using 3′ blocked labeled nucleotides.

tioned universal primer, and subsequent rounds of ligation-mediated sequencing, and so on. An overview of the SOLiD workflow is presented at http://marketing.appliedbiosystems.com/images/Product/Solid_Knowledge/flash/102207/solid.html. The unique attribute of this ligation-based approach and the 8mer labeling is that an extra quality check of read accuracy is enabled, so-called '2 base encoding'. This essentially relies on the known fixed nucleotide identities in the 8mer sequences to identify miscalls from true nucleotide differences during the data analysis step (see Figure 3 for details).

Given the amount of data that can be produced from a single run and the cost per run ($5–$85 per Mb of sequence; see Table 2 for comparisons) for these instruments, one can envision a diverse range of applications. The following sections profile recently published studies that used these next-generation sequencers for applications including mutation discovery, metagenomic characterization, noncoding RNA and DNA–protein interaction discovery.

Mutation discovery

The discovery of mutations that determine phenotypes is a fundamental premise of genetic research and will be tremendously facilitated by next-generation sequencing approaches, both for focused and genome-wide discovery. Conventional approaches to focused mutation discovery have used directed PCR to amplify selected genomic

Table 2. AB SOLiD cycle number descriptions

Cycle number	Universal primer position	Base positions identified	Probe set[a]	Positions interrogated
1	n	4,5	NNNAA^NNN-fl	5,10,15,20,25
2	n-1	4,5	NNNAT^NNN-fl	4,9,14,19,24
3	n-2	4,5	NNNAC^NNN-fl	3,8,13,18,23
4	n	1,2	AANNN^NNN-fl	2,7,12,17,22
5	n-1	1,2	ATNNN^NNN-fl	1,6,11,16,21

[a] ^, position of cleavage on each 8mer, whereas fl indicates the position of the fluorescent group on the 8mer.

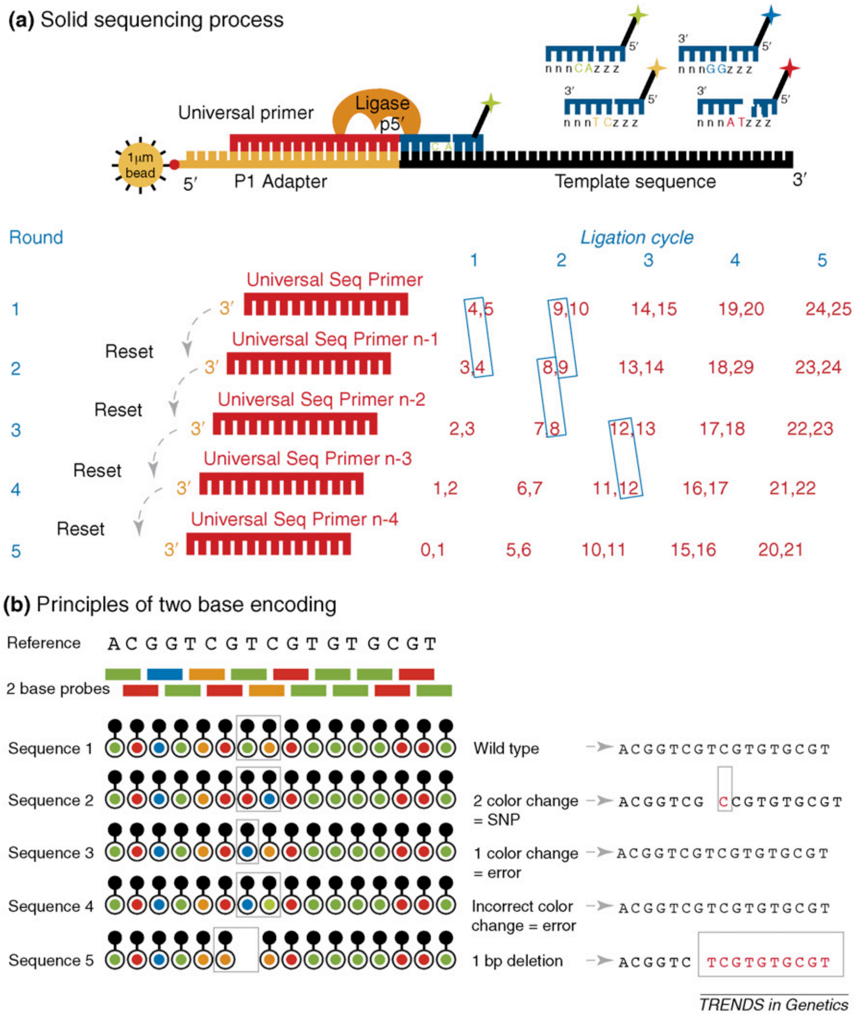

Figure 3. AB SOLiD sequencing. **(a)** AB SOLiD sequencing by ligation first anneals a universal sequencing primer then goes through subsequent ligation of the appropriate labeled 8mer, followed by detection at each cycle. **(b)** Two base encoding of the AB SOLiD data greatly facilitates the discrimination of base calling errors from true polymorphisms or indel events. Figures related to the SOLiD(tm) System are reproduced with permission from Applied Biosystems. (c) 2007 Applied Biosystems. All rights reserved.

regions from individual samples, followed by capillary sequencing, alignment of the resulting sequence traces and algorithmic detection of sequence variants [5–9]. The PCR products can instead be sequenced directly using the Roche (454) sequencer as published by Thomas *et al.* [10], who showed the high sensitivity of this platform to detect rare variants and to alleviate noisy capillary sequence data resulting from contaminating normal cells in tumor samples. A similar study by Dahl *et al.* [11] emphasized the value of this approach for highly sensitive variant detection. Recently, Porreca *et al.* [12] published an extreme example of this approach by multiplexing the amplification of 10 000 human exons using primers released from a programmable microarray and sequencing them using a massively parallel approach. An interesting variation of the PCR-directed approaches is exemplified in recent work that selects regions from the genome by microarray-based capture technology [13,14]. Once the target sequences are reclaimed from the array by denaturation and amplified, they can be directly sequenced for mutation detection. Although such approaches pose great potential for mutation discovery, they are limited for complex and repetitive genomes (such as human) because of the inability to design specific primers or capture probes.

Aside from a directed focus on select regions of a genome of interest, whole genome resequencing for variant discovery is significantly faster and less expensive using next-generation sequencers than with conventional approaches. Although there are some limitations imposed on this approach by the short read lengths these technologies deliver (Box 1), one can readily discover mutations genome-wide with a single instrument run or a portion thereof (depending on genome size). For example, recent work in our group to discover single nucleotide polymorphisms and small (1–2 bp) indels in a *Caenorhabditis elegans* strain (CB 4858) required only a single run of the Illumina sequencer (Hillier *et al.*, unpublished data). The following section expands on mutation discovery efforts using next-generation sequencing in bacterial and viral isolates.

Sequencing clinical isolates in strain-to-reference comparisons

Complete genome sequences are available for many disease-causing bacteria and viruses or for their laboratory strain equivalents (many of which are nonvirulent). Because the nature of such pathogens is to evolve continually by mutation and by exchanging sequences with one another, sequencing clinical isolates is of interest, especially if rapid data about antibiotic susceptibility and/or resistance and other virulence markers can be obtained. One clear benefit of all next-generation platforms for strain-to-reference sequencing is that each DNA sequence in a library is obtained from a single genomic fragment, such that if there are rare variants in the clinical strain population, these can be detected by virtue of the depth of sampling obtained. By contrast, this is not possible when sequencing PCR products directly obtained from a clinical sample, as is commonly done in a clinical diagnostic setting, because the low signal strength from variant nucleotides would not be detectable on a capillary sequencer. Another benefit of obviating the conventional bacterial cloning intermediate is that the cloning bias often introduced during passage of foreign sequences through a bacterial host is eliminated.

A typical project of this type involves culturing or otherwise isolating the microbe of interest and performing massively parallel sequencing of the isolate using one the approaches described above, followed by a bioinformatics-based approach to (i) align the sequence reads back to the reference genome(s); (ii) evaluate them for single nucleotide and/or indel variants and detect the presence of antibiotic resistance genes or pathogenicity islands by comparison of novel sequences to those in the public databases and (iii) evaluate any discovered variation in a functional and a biological context. Because the 454 platform has a read length appropriate for sequence assembly and a library construction approach that alleviates cloning bias, several studies using this approach have been published. In particular, two groups have produced increasingly sophisticated applications of HIV clinical isolate sequencing that identify rare members of the viral population [15,16], and that identify HIV integration sites in the host genome [17]. Further examples of the strain-to-reference application include the work of Francois *et al.* [18], who sequenced and analyzed the genomes of methicillin-resistant *Staphylococcus aureus* clinical isolates using the 454 platform, and of Poly *et al.* [19], who sequenced a clinical isolate of *Campylobacter jejuni*. Indeed, Denno *et al.* [20] have suggested that the study of the etiology of unexplained diarrhea should include the application of massively parallel sequencing to identify bacterial and viral species as potential causative agents. A remarkable study by an international consortium used 454 sequencing of *Mycobacterium tuberculosis* to identify drug targets of a diarylquinoline drug that potently inhibited both drug-sensitive and -resistant strains of the pathogen [21]. Based on these early reports, it is quite likely that our understanding of the spectrum of genome variation within clinical isolates will be greatly enhanced in the near future. This knowledge, in turn, should lead to improved diagnostics, monitoring and treatments.

Enabling metagenomics

Although sequencing a single clinical or cultured isolate of a microbe or virus seems a straightforward application of next-generation platforms, the throughput capability of these instruments has had a significant impact on a related and powerful field of endeavor known as 'metagenomics'. Metagenomics essentially entails brute force sequencing of DNA fragments obtained from an uncultured, unpurified microbial and/or viral population, followed by bioinformatics-based analyses that attempt to answer the question 'Who's there?' by comparing the metagenomic sequences obtained with all other sequenced species and isolates. Because these sequence reads are present in rough proportion to the population frequency of each microbe, inferences about relative abundance can be made. Although metagenomic studies have been accomplished with sequencing data from conventional capillary platforms, the associated cost was (and remains) prohibitive to sequence deeply into highly complex populations. Furthermore, the need for cloning before sequencing eliminates the metagenomic signatures from certain microbial and bacteriophage sequences that are not carried stably by *E. coli*. As such, the most definitive early metagenomic studies were of restricted populations that came from hostile environments [22].

Humans live in symbiosis with billions of microbial species that inhabit both the outer and inner surfaces of our bodies (including the skin, nasal/oral cavity, vagina and lower intestine). As such, many researchers posit that these symbiotic microbes provide an extension of the human genome and hence contribute to its genetic potential. This so-called 'human microbiome' represents just one of many complex metagenomic populations that are now possible to characterize using next-generation sequencing technology. Here, the combination of relatively easy, cloning-free library preparation of an uncultured sample and the sampling depth that can be obtained through inexpensive data production have ushered in a wave of metagenomic studies. These include characterizations of the microbial census of the human and mouse lower intestinal flora [23–25] and the oral cavity microbiome [26] that often combine conventional 16S typing with metagenomic sequencing of the isolates. Aside from the human body, metagenomic studies are ongoing in several important ecosystems, including global soil [27,28], deep mine [29] and the ocean [30,31]. Predictably, metagenomic sequencing can be used to uncover unknown etiologic agents, as was done in the investigation of honey bee colony collapse disorder [32]. These studies have used the Roche (454) pyrosequencing technology along with highly specialized comparative bioinformatics pipelines. It remains to be seen whether the shorter read length platforms will be used for microbiome sequencing. The human microbiome has been added to the NIH 'Roadmap' for medical research (http://nihroadmap.nih.gov/hmp/); thus, metagenomics will soon reveal whether changes in the human microbiome correlate with changes in human health and will begin to determine how the microbial census in a given body site may provide additional genetic potential, in terms of protective immunity, added enzymatic capability and so on.

Defining DNA–protein interactions

The nuclear interactions between DNA and proteins that control DNA packaging into histones or DNA transcription into mRNA have traditionally been studied in a locus-specific fashion. These processes are poorly understood in complex organisms in terms of how the genome instructs protein binding in a sequence- or structure-dependent context (or both). Moreover, the cognate binding sites are difficult to predict accurately with *in silico* methods.

Regulatory protein binding

At low throughput, chromatin immunoprecipitation (ChIP) [33] has enabled regulatory DNA–protein binding interactions to be elucidated. ChIP requires an antibody that is highly specific for the DNA-binding protein of interest. Cells are grown to the appropriate stage and treated with formaldehyde or another protein–DNA crosslinking agent, such that any protein in close association with DNA becomes linked. The cells are lysed, the DNA is fragmented and the specific antibody is used to precipitate the protein of interest along with any associated DNA fragment. These DNA pieces are subsequently released by reversing the crosslinking and identified by Southern blotting or by qPCR [33] using a probe to infer the DNA binding site sequence of the protein being studied. An intermediate version of this approach that enabled genome-wide evaluation of DNA–protein binding sites uses hybridization of fluorescently labeled DNA fragments to an appropriate microarray, after the ChIP and crosslinking reversal (ChIP-chip) [34]. Although this approach greatly facilitated binding site discovery genome-wide, there are limitations to the technique that were quickly addressed by replacing the microarray-based readout with DNA sequencing of the released fragment population using a next-generation sequencing instrument [35]; however, a recent comparative study showed that the two approaches can be complementary in this application [36]. In ChIP-seq, one can simply make an adaptor-ligated library of the released immunoprecipitated fragments and sequence them *en masse*. Follow-on bioinformatic analysis enables the genome-wide identification of the binding sites of that protein with exquisite specificity. Shorter reads are an advantage for specific definition of the binding site because they provide higher resolution, and the high throughput of these sequencers is such that even a single instrument run can provide enough data for all sites in the human genome (although replicates are typically performed to add statistical power). This approach, called ‘ChIP-seq’ was used to discover binding sites for the neuron-restrictive silencer factor (NRSF) transcription factor [37] and for the signal transducer and activator of transcription 1 (STAT1) binding protein [35] in human using the Illumina sequencer. In both studies, previously known binding sites were identified by ChIP-seq analysis, validating the approach and the analytical methods. Both studies also compared their findings with those previously discovered by ChIP-chip, determining that the resolution of ChIP-seq was superior and that fewer replicates and therefore less input DNA were required for these assays. It is likely that ChIP-seq will significantly contribute to our understanding of how protein binding sites are regulated in a genome-wide fashion in model organisms and humans. This will lead to a substantial improvement in the annotation of binding sites in these genomes and will shed light on the co-regulated genes that participate in different cellular pathways (the so-called ‘interactome’).

Exploring chromatin packaging

Chromatin packaging – how genomic DNA is packaged into histones – largely determines the availability of genes for transcription; therefore, understanding how DNA is ‘packaged’ is of great interest. Recent studies have shed light on chromatin packaging in several species and on the relationship between histone binding and gene expression. An initial 454-based study of genomic DNA packaging into nucleosomes was described for the *C. elegans* genome [38] by sequencing the DNA isolated from nucleosome cores after micrococcal nuclease digestion and mapping them to the reference genome sequence. Barski *et al.* [39] provided a comprehensive demonstration of the relationship between differential histone binding and gene expression in humans. Briefly, they used ChIP-seq with Solexa technology to compare histone methylations at promoter regions with the corresponding gene expression levels in T cells. They examined the genome-wide binding locations for 20 histones (possessing different lysine and arginine methylation states), one histone variant (H2A.Z), RNA polymerase II, and the insulator binding protein, CTCF. Mikkelsen *et al.* [40] used the Illumina platform to demonstrate the connection between chromatin packaging and gene expression in several different cell types. They used ChIP enrichment techniques followed by massively parallel Solexa sequencing to create chromatin state maps of both pluripotent and lineage-committed mouse cells. This led to a definition of three broad categories of promoters based on their chromatin state in embryonic stem cells. Moreover, when these maps were compared with gene expression patterns, Mikkelsen *et al.* [40] found that changes in chromatin state at specific promoters reflect changes in gene expression for the genes they control. Together these studies established a powerful technological paradigm that ties together data from different genome-wide approaches (e.g. histone-specific ChIP, gene expression) to enable a more comprehensive understanding of the biology of a given cell state and how it transitions into that state.

Discovering noncoding RNAs

Noncoding RNA (ncRNA) describes a broad class of regulatory RNA molecules whose myriad functions continue to be characterized in a variety of model organism systems and in human diseases, such as cancer [41]. Discovering the sequences and hence the genomic locations of ncRNAs is complicated—certain classes of ncRNAs are poorly conserved over evolutionary time; therefore, predicting precursor and processed ncRNA sequences in a genome sequence using *in silico*–based approaches is of limited use (although several programs exist). Hence, ncRNAs are most readily discovered by sequencing small RNA fragments. This is followed by extensive bioinformatics-based characterizations that examine the potential for secondary structure formation, putative genomic identification and the downstream functional assignment as regulatory molecules for specific genes. Fortunately, the lengths of ncRNAs make next-generation sequencing instruments ideal for genome-wide ncRNA discovery [42,43]. Most studies to date have used 454 technology, because of its early availability, to discover new and different ncRNA classes in several species ranging from land plants [44], *Chlamydomonas* [45], zebrafish [46], *Drosophila* [47,48], *Arabidopsis* [49], *C. elegans* [50] to human and chimpanzee brain [51]. The *C. elegans* study used RNA isolated from mixed stage worms to sequence and characterize ~400 000 small RNAs [50]. Analysis methods subsequently identified 18 novel microRNA genes, bringing the total of known microRNAs for *C. elegans* to 112, and discovered a new class of small RNAs called '21-U-RNAs'. The latter arise from >5700 loci that are found on chromosome IV and exist between and within protein-coding genes; they share a large upstream motif, yet the 21-U-RNA sequences are not conserved between *C. elegans* and *C. briggsae* (a close relative whose genome has been sequenced).

Future challenges: defining variability in many human genomes

The sequencing of the human genome [52] and the HapMap project [53,54] have impacted the study of human disease in a significant way and are enabling many genome-wide association studies that aim to elucidate the genetic component of complex diseases. Although we understand diversity among humans reasonably well at the level of common variants (present at 5% or greater in the populations studied by the HapMap project), we know little about the amount of variation below that level of allele frequency. Because common variants have not yet completely explained complex disease genetics, it is undoubtedly the case that less common and rare alleles also contribute. It will soon be possible, using next-generation platforms, to begin resequencing many 'normal' human genomes to better capture the spectrum of variability and to establish an important baseline for complex disease studies. Not only will rare alleles be discovered at the single base level, but structural variants, as well as large and small insertions and deletions, also will be captured by such an effort. A recent workshop in Cambridge, UK, began planning a pilot project toward additional human genome sequences, which will be pursued in the next 2 years and should produce an inflection point to accelerating biomedical research. This project (http://www.1000genomes.org/index.html), and the other applications of next-generation sequencing described in this article, pose significant informatic and bioinformatic challenges to the proper management and interpretation of large datasets. These are further described in the companion review article by Pop and Salzberg [55].

As our knowledge of human genome variability increases, so too will our annotation of the genome and its functional sequences increase based on the results of genome-wide inquiries such as ChIP-seq and ncRNA discovery, described earlier. These annotations, in turn, will add value to data that elucidate genome variation, making the results much more readily interpreted in terms of their potential impact(s) on function and ultimately, biology. Taken together, this integrated knowledge will significantly enhance our ability to interpret the genome in the context of observed human phenotypes, ultimately ushering in the era of personalized genomics (Box 1).

Concluding remarks

The sequence-based characterization of genomes is a relatively young pursuit in the biological sciences that has to date primarily enhanced model organism and human genetics, providing a substrate to discover genes and to understand genetics. This fundamental knowledge is now being enhanced by the ability to gather genome-wide sequence information more rapidly that can inform a higher-level appreciation of the functional genome, using front-end techniques combined with massively parallel throughput sequencing data production. Several early studies have shown the power of this new paradigm, and only time and ingenuity will determine its boundaries and its consequent ability to transform genetics.

Conflict of Interest Statement

The author has served as a Director of Applera Corporation since October 2007.

Acknowledgements

The author wishes to thank Vince Magrini for his critical reading of the manuscript and for contributing the information for Table 2.

References

1 Sanger, F. *et al.* (1977) DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.* 74, 5463–5467

2 Moore, G. (1965) Cramming more components onto integrated circuits. *Electronics* 38

3 Margulies, M. *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437, 376–380

4 Dressman, D. *et al.* (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

5 Nickerson, D.A. *et al.* (2001) Sequence-based detection of single nucleotide polymorphisms. *Methods Mol. Biol.* 175, 29–35

6 Pao, W. *et al.* (2004) EGF receptor gene mutations are common in lung cancers from 'never smokers' and are associated with sensitivity of tumors to gefitinib and erlotinib. *Proc. Natl. Acad. Sci. U. S. A.* 101, 13306–13311

7 Wilson, R.K. *et al.* (2003) Mutational profiling in the human genome. *Cold Spring Harb. Symp. Quant. Biol.* 68, 23–29

8 Wood, L.D. *et al.* (2007) The genomic landscapes of human breast and colorectal cancers. *Science* 318, 1108–1113

9 Paez, J.G. *et al.* (2004) EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. *Science* 304, 1497–1500

10 Thomas, R.K. *et al.* (2006) Sensitive mutation detection in heterogeneous cancer specimens by massively parallel picoliter reactor sequencing. *Nat. Med.* 12, 852–855

11 Dahl, F. *et al.* (2007) Multigene amplification and massively parallel sequencing for cancer mutation discovery. *Proc. Natl. Acad. Sci. U. S. A.* 104, 9387–9392

12 Porreca, G.J. *et al.* (2007) Multiplex amplification of large sets of human exons. *Nat. Methods* 4, 931–936

13 Albert, T.J. *et al.* (2007) Direct selection of human genomic loci by microarray hybridization. *Nat. Methods* 4, 903–905

14 Hodges, E. *et al.* (2007) Genome-wide *in situ* exon capture for selective resequencing. *Nat. Genet.* 39, 1522–1527

15 Hoffmann, C. *et al.* (2007) DNA bar coding and pyrosequencing to identify rare HIV drug resistance mutations. *Nucleic Acids Res.* 35, e91

16 Wang, C. *et al.* (2007) Characterization of mutation spectra with ultra-deep pyrosequencing: application to HIV-1 drug resistance. *Genome Res.* 17, 1195–1201

17 Wang, G.P. *et al.* (2007) HIV integration site selection: analysis by massively parallel pyrosequencing reveals association with epigenetic modifications. *Genome Res.* 17, 1186–1194

18 Francois, P. *et al.* (2007) Genome content determination in methicillin-resistant *Staphylococcus aureus*. *Future Microbiol.* 2, 187–198

19 Poly, F. *et al.* (2007) Genome sequence of a clinical isolate of *Campylobacter jejuni* from Thailand. *Infect. Immun.* 75, 3425–3433

20 Denno, D.M. *et al.* (2007) Explaining unexplained diarrhea and associating risks and infections. *Anim. Health Res. Rev.* 8, 69–80

21 Andries, K. *et al.* (2005) A diarylquinoline drug active on the ATP synthase of Mycobacterium tuberculosis. *Science* 307, 223–227

22 Tyson, G.W. *et al.* (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* 428, 37–43

23 Turnbaugh, P.J. *et al.* (2006) An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* 444, 1027–1031

24 Gill, S.R. *et al.* (2006) Metagenomic analysis of the human distal gut microbiome. *Science* 312, 1355–1359

25 Zhang, T. *et al.* (2006) RNA viral community in human feces: prevalence of plant pathogenic viruses. *PLoS Biol.* 4, e3

26 Marcy, Y. *et al.* (2007) Dissecting biological 'dark matter' with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth. *Proc. Natl. Acad. Sci. U. S. A.* 104, 11889–11894

27 Fierer, N. *et al.* (2007) Metagenomic and small-subunit rRNA analyses of the genetic diversity of bacteria, archaea, fungi, and viruses in soil. *Appl. Environ. Microbiol.* 73, 7059–7066

28 Leininger, S. *et al.* (2006) Archaea predominate among ammonia-oxidizing prokaryotes in soils. *Nature* 442, 806–809

29 Edwards, R.A. *et al.* (2006) Using pyrosequencing to shed light on deep mine microbial ecology. *BMC Genomics* 7, 57

30 Angly, F.E. *et al.* (2006) The marine viromes of four oceanic regions. *PLoS Biol.* 4, e368

31 Sogin, M.L. *et al.* (2006) Microbial diversity in the deep sea and the underexplored 'rare biosphere'. *Proc. Natl. Acad. Sci. U. S. A.* 103, 12115–12120

32 Cox-Foster, D.L. *et al.* (2007) A metagenomic survey of microbes in honey bee colony collapse disorder. *Science* 318, 283–287

33 Solomon, M.J. *et al.* (1988) Mapping protein-DNA interactions *in vivo* with formaldehyde: evidence that histone H4 is retained on a highly transcribed gene. *Cell* 53, 937–947

34 Ren, B. *et al.* (2000) Genome-wide location and function of DNA binding proteins. *Science* 290, 2306–2309

35 Robertson, G. *et al.* (2007) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods* 4, 651–657

36 Euskirchen, G.M. *et al.* (2007) Mapping of transcription factor binding regions in mammalian cells by ChIP: comparison of array- and sequencing-based technologies. *Genome Res.* 17, 898–909

37 Johnson, D.S. *et al.* (2007) Genome-wide mapping of *in vivo* protein-DNA interactions. *Science* 316, 1497–1502

38 Johnson, S.M. *et al.* (2006) Flexibility and constraint in the nucleosome core landscape of *Caenorhabditis elegans* chromatin. *Genome Res.* 16, 1505–1516

39 Barski, A. *et al.* (2007) High-resolution profiling of histone methylations in the human genome. *Cell* 129, 823–837

40 Mikkelsen, T.S. *et al.* (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* 448, 553–560

41 Stahlhut Espinosa, C.E. and Slack, F.J. (2006) The role of MicroRNAs in cancer. *Yale J. Biol. Med.* 79, 131–140

42 Lu, C. *et al.* (2007) Construction of small RNA cDNA libraries for deep sequencing. *Methods* 43, 110–117

43 Kiriakidou, M. *et al.* (2004) A combined computational-experimental approach predicts human microRNA targets. *Genes Dev.* 18, 1165–1178

44 Axtell, M.J. *et al.* (2007) Common functions for diverse small RNAs of land plants. *Plant Cell* 19, 1750–1769

45 Zhao, T. *et al.* (2007) A complex system of small RNAs in the unicellular green alga Chlamydomonas reinhardtii. *Genes Dev.* 21, 1190–1203

46 Houwing, S. *et al.* (2007) A role for Piwi and piRNAs in germ cell maintenance and transposon silencing in Zebrafish. *Cell* 129, 69–82

47 Brennecke, J. *et al.* (2007) Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila. *Cell* 128, 1089–1103

48 Stark, A. *et al.* (2007) Systematic discovery and characterization of fly microRNAs using 12 Drosophila genomes. *Genome Res.* 17, 1865–1879

49 Kasschau, K.D. *et al.* (2007) Genome-wide profiling and analysis of *Arabidopsis* siRNAs. *PLoS Biol.* 5, e57

50 Ruby, J.G. *et al.* (2006) Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in *C. elegans*. *Cell* 127, 1193–1207

51 Berezikov, E. *et al.* (2006) Diversity of microRNAs in human and chimpanzee brain. *Nat. Genet.* 38, 1375–1377

52 Lander, E.S. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature* 409, 860–921

53 Internation HapMap Consortium (2005) A haplotype map of the human genome. *Nature* 437, 1299–1320

54 Frazer, K.A. *et al.* (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449, 851–861

55 Pop, M. and Salzburg, S. (2008) Bioinformatics challenges of the new technology. *Trends Genet.* 24, 142–149