Emerging DNA sequencing technologies for human genomic medicine

Robert L. Strausberg, Samuel Levy and Yu-Hui Rogers

J. Craig Venter Institute, 9704 Medical Center Drive, Rockville, MD 20850, United States

The completion of draft sequences of the human genome represented a remarkable achievement for automated DNA sequencing based on Sanger technology. However, the future requires substantial leaps in sequencing technology such that whole genome sequencing will become a standard component of biomedical research and patient care. In this review we describe current advances that are in early stages of development, but that point toward technology that will enable the onset of genomic medicine encompasses strategies for preventative medicine and intervention based on complete knowledge of an individual's genome.

The opportunity

The publication of composite human genomes [1,2] several years ago established a conceptual framework for an era of medicine based on the knowledge of the human genetic repertoire. However, at present, even with available resources such as HapMap [3], we are still limited in the knowledge of human variation and how variation relates to disease predisposition and onset. The recent sequencing of the genomes of two individuals, J. Craig Venter [4] and James D. Watson provides the first glimpses of the precise extent of variation between individual humans. This is a necessary step to interface individual genomics and medical practice. Further advances will require the sequences of thousands of human genomes, interfaced with carefully annotated phenotypic and clinical datasets.

Automated Sanger-sequencing technology has served as the most widely used platform for DNA sequencing over the past two decades, from ESTs to microbial genomes to the human genome. However, while the sequence quality and long read lengths achieved with Sanger technology are well understood, the cost of this current generation of sequencing is prohibitive to the routine sequencing of human genomes. In addition, because the Sanger-sequencing readout is essentially analog (i.e. each trace peak represents a composite of many DNA molecules), there are limitations in detecting minority alleles among genomes in a heterogeneous population, such as in tumors. Therefore, rare but important mutations can be missed by Sanger sequencing alone.

The realization that the completion of drafts of the human genome was just a start of the DNA sequencing era has resulted in vigorous efforts to invent and develop new approaches to DNA sequencing. These approaches will address our anticipated future needs of throughput and cost, in a manner that allows for the multitude of current and future applications. In addition to laboratory advancements a new generation of informatics tools is required to accommodate these new approaches.

Here we describe the state-of-the-art for current generation sequencing based on Sanger technology, as well as a multitude of new approaches that will form are currently being implemented to achieve major advances in cost, throughput, as well as improving sensitivity of variant detection.

Sanger-sequencing technology

The development of enzymatic [5] and chemical-based [6] DNA sequencing technologies in the 1970s resulted in revolutionary changes in biological and biomedical science. Based on its relative suitability for automation, the Sanger dideoxy enzymatic sequencing method [5] became the gold standard for DNA sequence determination. The technology and instrumentation for the electrophoresis-based separation of Sanger-sequenced DNA fragments and the detection of individual bases have advanced substantially in the past two decades. The invention and deployment of multicolor probes for sequencing chemistries in 1986 [7] allowed scien-

Corresponding author: Strausberg, R.L. (RLS@jcvi.org)

1359-6446/06/$ - see front matter © 2008 Elsevier Ltd. All rights reserved. doi:10.1016/j.drudis.2008.03.025

tists to move away from laborious and hazardous radioisotope labeling techniques. These developments toward fluorescence-based detection were highly amenable to automated detection, computational recording and signal processing. In the mid-1990s, capillary-based electrophoresis and multi-color fluorescent detection [8–10] were successfully integrated. This enabled the creation of high-throughput sequencing machines with enhanced sample tracking, thereby enabling paired-end sequencing, a key component of whole genome shotgun sequencing. Technology development in areas such as fluorescently tagged chain-terminators [11], cycle-sequencing protocols [12] and fluorescence resonance energy transfer (FRET) dyes [13,14] have greatly enhanced the throughput, accuracy, robustness and detection sensitivity of the Sanger-sequencing approach. Moreover, the accuracy, quality and utility of the data generated by these sequencing machines also improved significantly because of advancements in instrumentation and informatics analysis tools. These multiple generations of technology development, and subsequent refinements, resulted in the development of Sanger-based instruments that served to sequence the first human genomes, and a multitude of bacterial, fungal, plant, parasitic and animal genomes including numerous mammals.

Today, state-of-the-art capillary-based sequencers that employ Sanger chemistry (ABI3730XL) have the ability to generate at least 1–2 million base pairs (bp) per 24-hour period with long read lengths (an average of 550–800 bp), and with very high accuracy [15].

Strategy development with current technology

Complementary to the development of state-of-the-art sequencing chemistries and instrumentation has been the invention of new strategies that improve sequencing efficiency dramatically. These strategies have greatly enhanced the capability of Sanger sequencing to generate genomic sequences and to gain insight to the complexities of cellular transcriptomes. For example, genomic sequencing was originally envisioned as a process in which individual clones from a set of minimally overlapping tiling path clones were completely sequenced and assembled. The last stage involving the final assembly of each fully sequenced clone into a tiling path spanning the genome. While this approach was successfully employed for sequencing eukaryotic genomes including the human, it became apparent that an alternate approach termed whole genome shotgun (WGS) sequencing enabled the production of prokaryotic and eukaryotic genomic sequencing much more cost-effectively. In the WGS approach, genomic DNA is cloned into a series of vectors that capture different size fragments from 2 to 4 kb (small insert library) to greater than 100 kb fragments (large insert library), and both clone ends (paired ends) are sequenced. Assembly algorithms such as the Celera Assembler [16] utilize information including cloned insert sizes, and the paired-end sequences to assemble the complete genome into contigs. Based on specific needs, the genome can be retained as a draft, or the gaps separating contigs can be closed by additional directed sequencing of selected clones.

Sequencing strategies evolved rapidly to obtain transcript information efficiently, thereby facilitating the identification of genes and their relative expression levels. Adams *et al.* [17] described the Expressed Sequence Tag (EST) approach to identify gene sequences rapidly. In this approach, 3′ and/or 5′ end sequence reads of cDNAs are generated. The end sequences are sometimes utilized to reveal a potentially full-length cDNA, which can then be completely sequenced by shotgun or directed strategies such as primer-walking. While the EST strategy has become the most widely employed approach for transcript identification, additional transcript tagging strategies, such as SAGE [18] and CAGE [19], were developed to generate quantitative analyses of transcripts even more cost effectively. In SAGE, for example, very short 3′ tags (up to 24 bp) are specifically generated from each transcript to generate a sequence that is unique to a transcript. These short molecules are linked together such that a single sequence lane generates tags from multiple transcripts. Similarly, tagging strategies such as digital karyotyping [20] have been applied to genomic DNA to measure events such as amplification of particular chromosomal regions in tumors and epigenetic modifications.

Current limitations of the Sanger-sequencing technology

Certain limitations of the current Sanger-sequencing technology have stimulated the development of new technological approaches. First, the expense of Sanger sequencing prohibits its use in the context of future scientific goals that include the sequencing of many thousands of human genomes. For example, an accurate high-quality mammalian genome sequence still costs several million dollars to produce even at a draft level. Moreover, for some applications, the analog nature of the sequence output results in failure to detect the variable contribution of minority genomes within a population of cells. For example, genomes within solid tumors can be very heterogeneous, and the ABI3730XL may be unable to detect rare mutations within a tumor, and yet important for the progression of a tumor and/or its response to treatment.

To address these opportunities, several promising new sequencing technologies [21–24] are being developed and made accessible in commercially available formats. These new technologies have several characteristics in common. First, several of these technologies employ the sequencing-by-synthesis principle in a massively parallel format, though different chemistries are used to extend, tag and detect the sequences. Unlike Sanger sequencing, each technology provides data derived from amplification of individual DNA molecules, thereby enabling a quantitative (digital) readout of a population of DNAs. This characteristic is important for detecting rare molecules in a population, such as in the aforementioned case of cancer. Second, in their current iterations, each of these technologies affords a much higher sequence throughput per instrument run compared with the ABI3730XL, and at substantially lower cost.

Along with many positive aspects of these technologies there are several challenges in their current implementation. The read lengths currently achieved with these technologies are substantially shorter than those achieved with the ABI3730 XL (ranging from about 35 to 250 bp), presenting new challenges for sequence assembly and mapping to the genome. In addition, front-end processes, such as for the generation of paired-end sequences in an efficient manner, are still at an early stage of development. Finally, each particular technology has its own biases in sequencing accuracy and these are actively being defined and rectified.

Despite these caveats, it is also clear that these technologies offer exciting new opportunities to revolutionize our understanding of biology and human genomics in particular. It is also apparent that, as described above for Sanger technology, evolutionary advances for these technologies are happening rapidly, such that throughput and accuracy will increase, and cost will decrease even further.

Emerging technologies

454 Pyrosequencing technology

Initial descriptions of a new approach to DNA sequencing based on sequencing-by-synthesis, termed pyrosequencing, were first reported a decade ago [25,26]. In this approach DNA synthesis is performed within a complex reaction that includes enzymes (ATP sulfurylase and luciferase) and substrates (adenosine 5′ phosphosulfate and luciferin) such that the pyrophosphate group released upon addition of a nucleotide results in the production of detectable light. Therefore, when a new nucleotide is incorporated in a growing DNA chain through the activity of DNA polymerase, pyrophosphate is generated in a stoichiometric manner resulting in the production of ATP. The ATP produced drives the enzymatic conversion of luciferin with the associated emission of photons. Subsequent to the clean-up of the reaction compounds, a new cycle of reactants are introduced and thus the incorporation of specific nucleotides is measured in a sequential manner.

Rothberg and colleagues at 454 Life Sciences (454) have developed a novel and highly parallel system based on the pyrosequencing chemistry [27] (Fig. 1). In this system, sample preparation involves fragmentation of genomic DNA followed by the ligation of adaptor sequences and clonal amplification of the target DNA on micron-sized beads using an emulsion-based PCR method (emPCR) [28]. The sample preparation process is much simplified compared with Sanger sequencing, contributing to cost effectiveness (Fig. 2) and a significant improvement in throughput. The sequencing-by-synthesis reactions are performed directly on the template-carrying beads that are preloaded into a microfabricated glass plate containing 1.6 million picoliter reactor wells. To determine DNA sequence, the four nucleotides are sequentially delivered through the plate while a CCD camera detects the wells emitting light, which is generated when the incorporation of a nucleotide occurs.

The current 454 Life Sciences commercial system (FLX) is capable of sequencing 100 million bases in a seven-hour period, and has a raw base accuracy of 98% and an average of approximately 250 bp read length (http://www.454.com/enabling-technology/the-system.asp). Compared to the current state-of-the-art Sanger-based capillary electrophoresis platform, the 454 system generates ‘raw’ sequence data with approximately 100 times higher throughput (see Table 1). Current limitations of this technology

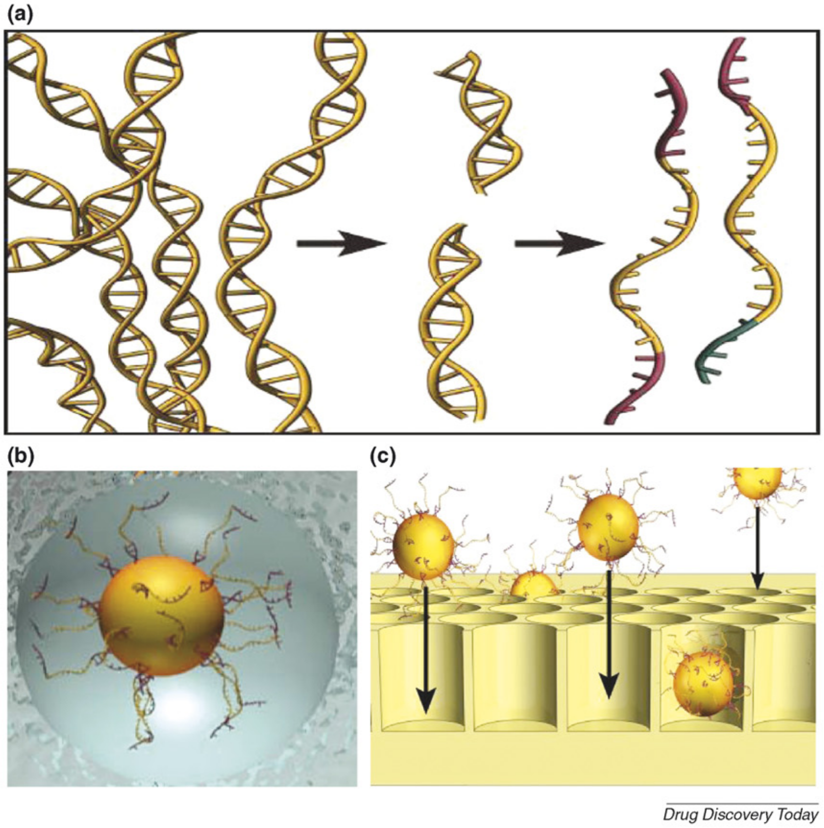
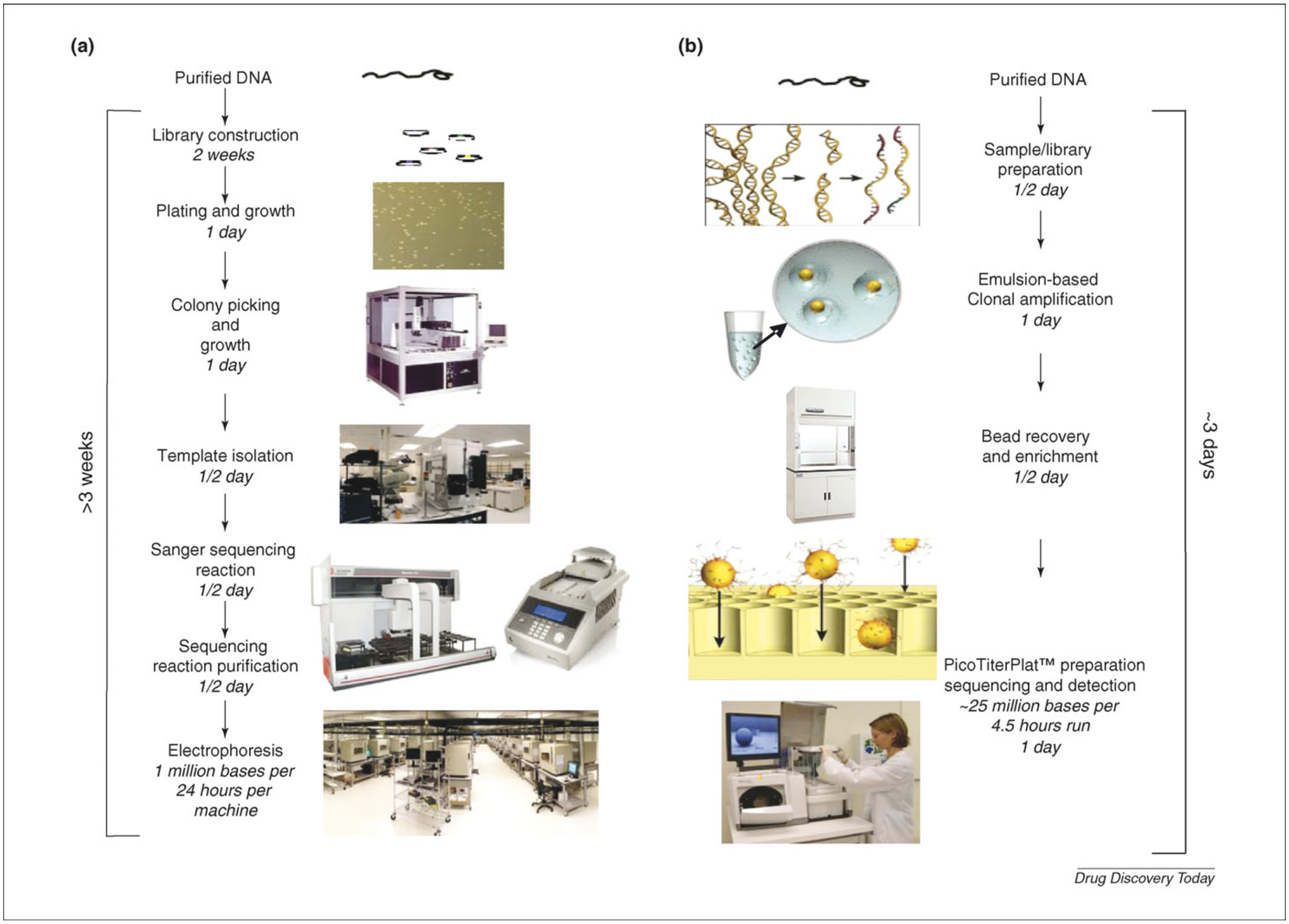


FIGURE 1

454 Pyrosequencing process overview. Following random fragmentation of genomic DNA **(a)**, adaptor sequences are ligated to both fragment ends. The target DNA is then clonally amplified on micron-sized beads using an emulsion-based PCR method (emPCR) **(b)**. Pyrosequencing-based sequencing by synthesis is then performed on the beads **(c)** and nucleotide incorporation is measured by emitted light detection (figure adapted from http://www.454.com).



Drug Discovery Today

FIGURE 2

454 versus current capillary-based sequencing. 454 Pyrosequencing process compared to the Sanger electrophoresis based process. The traditional Sanger electrophoresis approach requires a longer processing time per production cycle, substantially more support equipment, a larger facility and more labor than the 454 approach.

include relatively short read lengths (~1/3 of Sanger sequencing read length) and low base calling accuracies for some genomic regions; especially in those containing homopolymers. As a result, insertion and deletion errors are sometimes introduced in the sequence reads. However, read lengths and sequence accuracy of the 454 system have improved substantially and further advances are anticipated. With respect to sequence accuracy, it is important to develop measures for each of the new [29–31] technologies such that sequence quality can be understood much in the way that phred [32] scores have served as a currency for Sanger sequencing. Toward that end, tools are being developed that will enable the community to understand sequence quality in a context-dependent and informed manner.

Since its entry to the marketplace, the 454 pyrosequencing technology has been adapted to a wide variety of genomic applications including whole genome and transcriptome [33] sequencing [21], PCR-directed gene sequencing [34], as well as analysis of single-nucleotide polymorphisms (SNPs) [35–41], copy number variation [42], structural variation [43], and epigenetic effects [44–48]. These early studies demonstrate the potential for the technology not just with respect to cost reduction for existing applications, but also to invigorate the development of new sequencing opportunities.

TABLE 1

Comparison of sequencing platform performance

	Throughput (million bp/day)	Throughput (million bp/run)	Average read length (bp)
3730*xl*	1–2	0.08	>800
454 FLX	200	100	250
SOLiD	200–300	1000–3000	25–35
Solexa	≥200	1000–1500	25–35

Many groups are currently exploring the application of the 454 technology for genomic sequencing. Margulies *et al.* [21] reported on the capability of the GS20 instrument (the first generation 454 instrument, which generated about 100,000 reads of 100 bp each per run) to generate, from a single instrument run, the *de novo* assembly of the relatively small and previously sequenced *Mycoplasma genitalium* genome at 96% genome coverage and greater than 99.9% accuracy. Issues that are currently being addressed that will improve whole genome shotgun sequencing with 454 technology include the use of paired-end strategies, as well as improvements in read length and accuracy. These developments will be essential for the accurate *de novo* assembly of complex genomes, especially those rich in repetitive sequences. To circumvent the issues that limit *de novo* genome assembly (short reads, absence of paired ends) and systematic limitations in sequencing accuracy in specific regions (homopolymers), hybrid strategies have been employed that incorporate 454-generated sequences in addition to Sanger reads [49]. In general such strategies seek to incorporate the best features of each technology to generate an accurate sequence assembly in a cost-effective manner. Typically, this involves the use of 454 reads to provide sequence coverage while medium to long insert paired-end libraries sequenced by Sanger methods allow long range ordering of the genome assembly. The study of Goldberg *et al.* [49] demonstrated the utility of 454 reads in closing gaps in regions with clone gaps as well as structural or other features in the DNA incompatible with Sanger sequencing. It is likely that hybrid strategies will be of increasing importance given the different and often complementary characteristics of the various new sequencing methods.

The 454 technology has already been useful for identifying rare alleles within a sample owing to the digital and clonal nature of the sequence output. For example, solid tumors are known to be very heterogeneous with respect to cell types (epithelial, stromal, immune system, etc.) and there is also genome heterogeneity within the cancerous epithelial cells. Therefore mutations that could be causative of the cancer phenotype are sometimes in a minority of the genomes in a tumor. Because targeted gene sequencing by Sanger technology results in sequencing reads composed of a multitude of DNAs, variants may represent a small fraction of the population that cannot be discerned above background. Thomas *et al.* demonstrated the detection of epidermal growth factor receptor (EGFR) mutations in [34] lung cancer samples, and that these mutations appeared to explain observed resistance of these tumors to the EGFR inhibitor erlotinib. Our group has observed similar results in glioblastoma samples with respect to mutational change in the fibroblast growth factor receptor (FGFR)1 gene (present in about 10% of the genomes in the tumor) that was missed through Sanger sequencing but observed by 454 sequencing (R. Strausberg, unpublished results).

Transcriptome analysis is another particularly enabling application of the 454 technology. Previous studies of eukaryotic transcriptomes have focused on relatively light sampling of cellular transcripts in isolation of sequencing of full-length cDNAs. Various approaches for increasing the depth of transcript surveys have been pursued with expressed sequence tags (ESTs) and tagging strategies that provide a very short tag (as in SAGE) that often can be used to identify a particular transcript, but with little information about alternate processing in alternative splicing and polyadenylation. The tagging strategies are advantageous in that depth of coverage can be cost effectively increased to more than 100,000 transcripts per assay. Because the FLX format supports about 400,000 reads per instrument run with read length of about 250 bp, the technology enables a deeper view of the transcript population within a cell, at the same time providing the opportunity to identify alternative splicing and identification of polymorphisms and mutations. For example, this capability has been demonstrated with human cancer cell lines [50], as well as in the plant *Medicago truncatula* [51].

Illumina's Solexa sequencing technology

In comparison with the current 454 technology, the Illumina/Solexa sequencing-by-synthesis (SBS) technology (Fig. 3) can generate DNA sequence data at substantially higher throughput and lower cost. This sequencing system starts with a library preparation strategy similar to that utilized in the 454 approach. However, the subsequent template capturing and amplification process is quite different. It relies on the solid-phase bridge PCR technique [52] for amplification of the targeted DNA templates which form random arrays of 'clusters' on the surface of a flow cell [24].

Currently, the flow cell can retain greater than ten million clusters, each with ~1000 template copies, all in a square centimeter. The sequences of the resulting templates are detected using a four-color fluorescence format via repeated cycles of polymerase-mediated primer extension reactions using reversible dye terminators. This system currently has a throughput of approximately 1 billion bases of data in a single run with a read length of about 35 bp [53].

Although both the 454 and Illumina/Solexa systems employ the sequencing by synthesis principle for sequence determination, the potential to obtain extended read length and error characteristics of the sequence reads generated by the two systems appears to be very different. The 454 system has a higher potential to provide longer read lengths since the pyrosequencing technique employed uses non-modified deoxynucleotides. However, the non-terminating nature of the 454 chemistry results in challenges in sequencing homopolymer regions. The Illumina/Solexa system utilizes a fluorescent-based solid-phase dye terminator chemistry that requires a re-engineered DNA polymerase and modified dye labeled nucleotides as the substrates. The terminator chemistry should minimize the deletion and insertion errors such as those encountered by the 454 system. However, the need to use modified substrates and polymerase could potentially compromise the efficiency and fidelity of base incorporation during the SBS reactions. Therefore, the potential for substitution errors, as well as limited read length, appears to be the major limitation of this technology.

Although the Illumina/Solexa technology has only been recently introduced to the commercial marketplace, there are already demonstrations of its potential utility [53–58]. For example, Barski *et al.* [54] demonstrated the effectiveness of this technology in generating genome-wide mapping toward gaining new perspective of the relationship between chromatin organization and histone methylation. Their analysis of histone variants, RNA polymerase II and insulator-binding protein CTCF revealed the power of this technology in providing depth of resolution needed to reveal associations of functional regulatory elements such as the

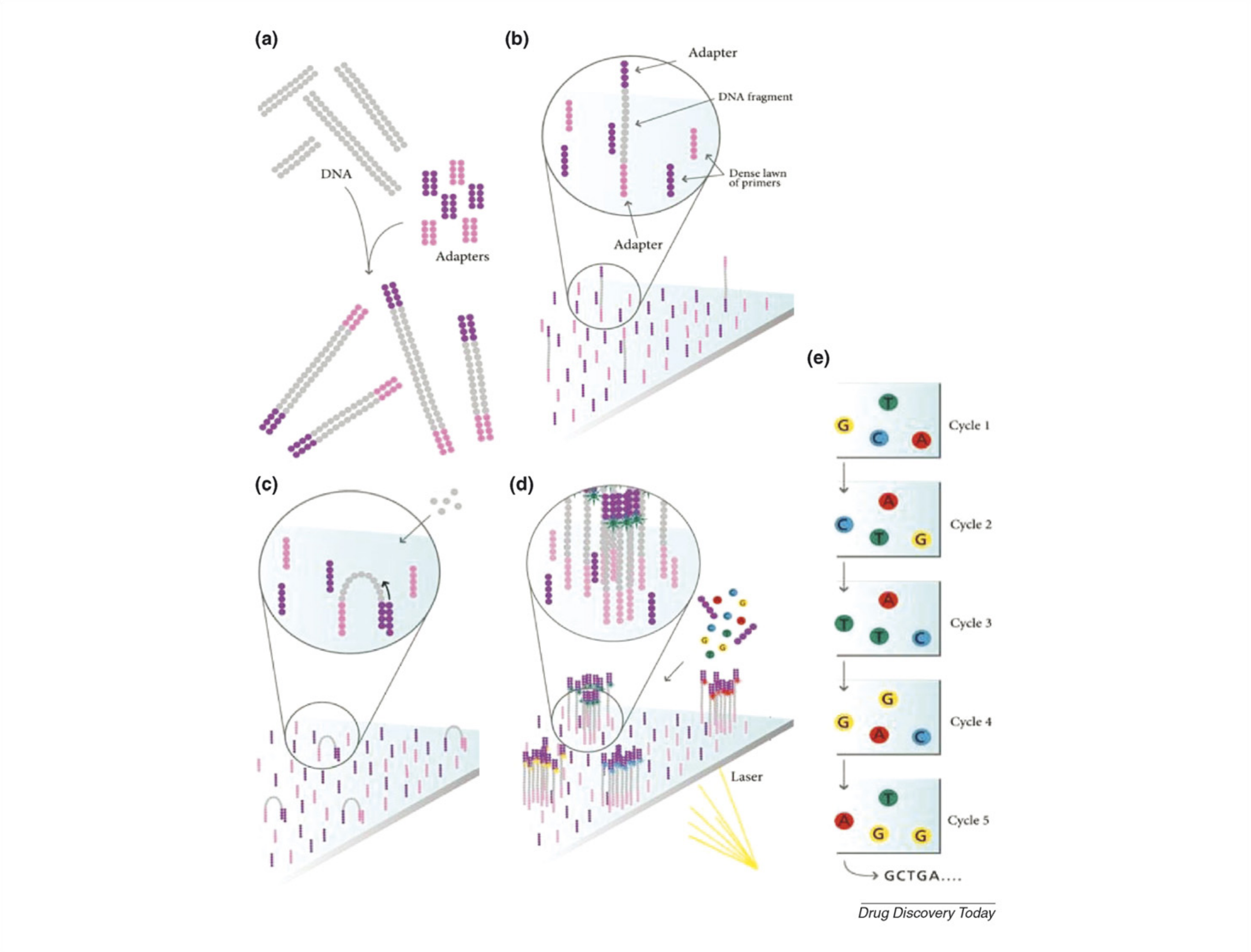


FIGURE 3

Illumina/Solexa process overview. Genomic DNA is randomly fragmented and adapters are added to each fragment end **(a)**. Single-stranded fragments are then attached randomly inside flow cells **(b)**. Bridge amplification is performed to generate double-stranded fragments **(c)**. Following repeated cycles of denaturation and bridge amplification, millions of DNA copies are present in each flow cell **(d)**. DNA sequence is determined with four-labeled reversible terminators primers and polymerase **(e)**. Unincorporated terminators are washed, and the sequence is determined in each flow cell following laser excitation. The blocked 3′ terminus and fluorophore are removed from the incorporated base and the cycle is repeated (figure adapted from http://www.illumina.com).

relationship between CTCF and boundaries of histone methylation domains. Wakaguri and colleagues [55] utilized this technology to generate a very large transcriptional start site (TSS) dataset including more than 19 million 5′ end sequences associated with RefSeq genes. This substantially expanded TSS set is integrated with various informatics tools such as for predicting transcription factor binding sites in the DBTSS.

A substantial opportunity for the Illumina/Solexa approach (and as discussed below for the SOLiD technology) exists based on the very low cost/high-throughput production of DNA sequences. A challenge that exists for each technology is in accomplishing the *de novo* genome assembly employing whole genome shotgun sequences that are relatively short. To meet this challenge, several groups are developing novel assembly algorithms that are specifically tailored to the characteristics of these sequences [59,60]. Even with these new algorithms, limitations for *de novo* assembly will still exist, especially for repetitive regions of genomes. Indeed, such sequences can present significant challenges and result in less than optimal assemblies even for the long 800 bp reads provided by the ABI 3730XL technology.

SOLiD technology

The Supported Oligonucleotide Ligation and Detection Platform (SOLiD) technology developed by Applied Biosystems (ABI) utilizes hybridization-ligation chemistry [22] and is based on the polymerase colony (polony) approach [61] (Fig. 4). The sample preparation aspect of this technology including library construction and clonal amplification of the target DNA by emPCR on beads is very similar in principle to the 454 approach. However, the size of the beads used for emPCR (1 μm for SOLiD versus 26 μm for 454) and the array format (random for SOLiD versus ordered for 454) are different. These differences afford the SOLiD technology the potential of generating a significantly higher density sequencing array (potentially over a few hundred fold higher), as well as

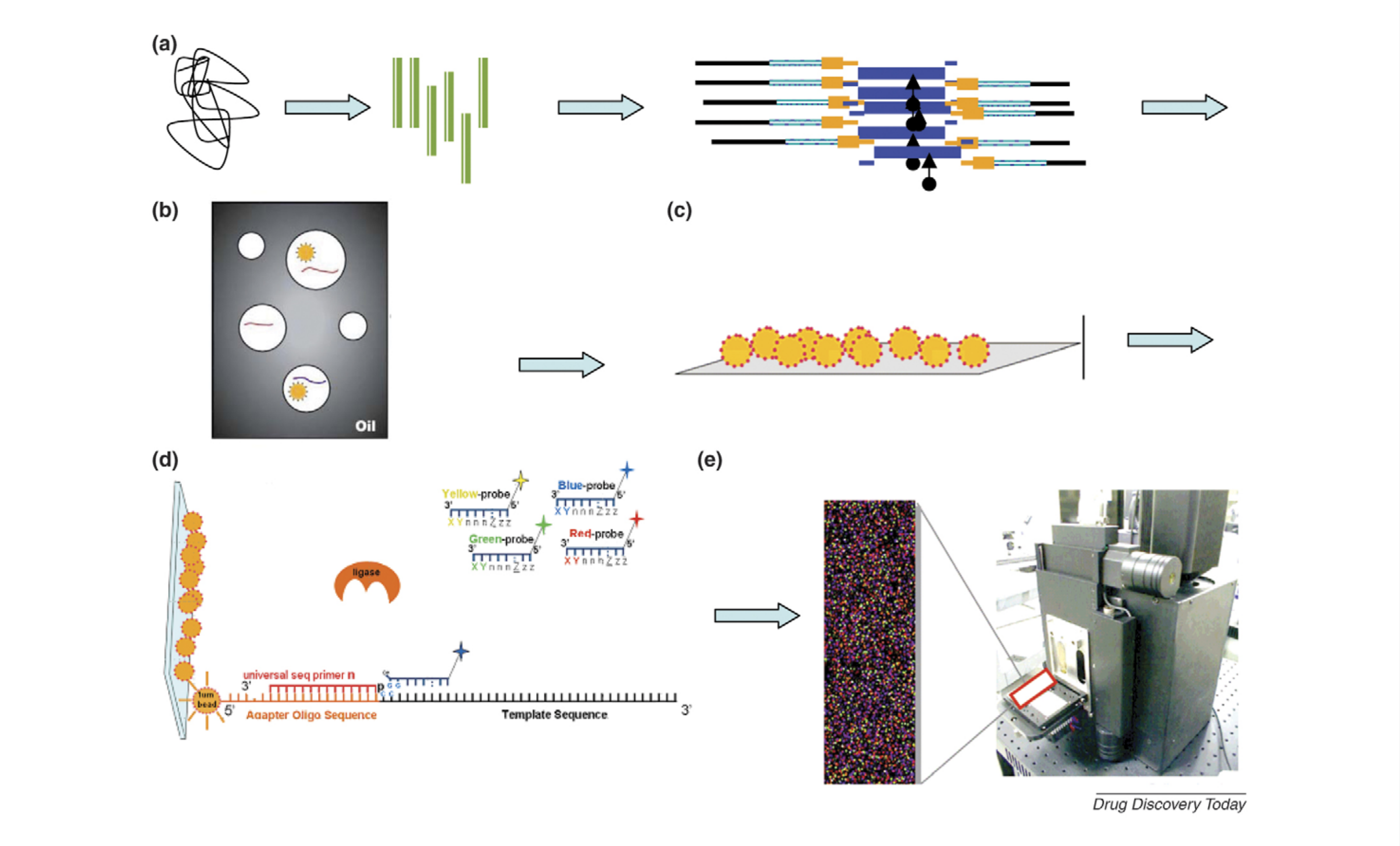


FIGURE 4

ABI Solid process overview. In a manner similar to the 454 technology, DNA libraries are prepared **(a)** and amplified by emulsion PCR **(b)**. The template beads are deposited to form a random array **(c)**, and sequencing is performed by repeated hybridization cycles with sequencing primers **(d)** and fluorescently labeled probes encoding the bases that are being interrogated **(e)** (figure adapted from http://www.appliedbiosystems.com).

more flexibility in terms of sample input format. The sequence interrogation is done through the repeated cycles of hybridization of a mixture of sequencing primers and fluorescently labeled probes. This is then followed by the ligation of the sequencing primers and the probes, subsequently followed by the detection of the fluorescent signals on the probes that encode the bases being interrogated. The initial specifications suggested that the system will be capable of generating 200–300 million bp of sequence data per day or 1–3 G bp per run with a raw base accuracy of 99% and a 25–35 bp read length. As discussed above with respect to the Illumina/Solexa technology, the short read lengths present formidable but potentially solvable challenges for *de novo* whole genome sequence assembly. In addition, the error profiles or the accuracy of this sequencing by ligation approach remains to be fully established.

Although the SOLiD instrument is just becoming accessible to the research community, there are already many demonstrations of the utility based on the underlying polony technology. These include efficient sequencing of human exons captured on programmable oligonucleotide arrays [62], assay of very low abundance cellular transcripts [63], digital polony exon profiling, to study alternative pre-messenger RNA splicing, accurate low-cost resequencing of a bacterial genome [22], multi-locus long-range haplotyping [64] and detection of rare mutations associated with resistance to kinase inhibitors in BCR-ABL oncogene in patients with chronic myelogenous leukemia [65].

Beyond the next generation technologies

The technologies described above are in their infancy and likely to become even higher throughput at reduced cost based on improvements to chemistry, engineering and informatics as has been the case for Sanger technology. However, even with those anticipated improvements there is an ongoing need for revolutionary advances such that individual genomics can indeed become a cost-effective component of standard medical care. Examples of additional sequencing approaches that are currently being designed and developed include a nanofluidic-based sequencing system, several single molecule based approaches [23,66,67] and nanopore sequencing. Aiming to set a new benchmark for the ultimate cost and efficiency limits of Sanger sequencing, the nanofluidic-based sequencing system [68] utilizes microfabrication technologies to integrate thermal cycling, sample purification and capillary electrophoresis, in a nanoliter-scale bioprocessor. The nanopore sequencing technology [69] explores the possibility of analyzing DNA sequences by measuring transient blockades of the monovalent ion current within a pore while the DNA molecules are being transported through. If single-base-pair resolution can be achieved, this approach has the potential to provide an extremely rapid, low-cost method for sequencing and would eliminate the need for cloning and amplification.

The diversity of technological approaches currently being developed will form the basis for a new era of medical practice that will increasingly incorporate the knowledge of human genomics into strategies of health care. The sequencing of complete diploid human genomes is becoming increasingly affordable and these datasets will support the research enterprise toward the development of improved prevention and intervention strategies. There still remains much to be accomplished in order to enable complete genome sequencing as a standard component of a patient's medical care. Current progress and visions for continued evolutionary and revolutionary advances in DNA sequencing technology suggest that individualized genomic medicine will rapidly advance from a vision to reality.

Note added in proof

A recent publication describes the utilization of the 454 Life Sciences technology to generate the genome sequence of James Watson [70].

References

1 Venter, J.C. *et al.* (2001) The sequence of the human genome. *Science* 291, 1304–1351

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

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

4 Levy, S. *et al.* (2007) The diploid genome sequence of an individual human. *PLoS Biol.* 5, e254

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

6 Maxam, A.M. and Gilbert, W. (1977) A new method for sequencing DNA. *Proc. Natl. Acad. Sci. U. S. A.* 74, 560–564

7 Smith, L.M. *et al.* (1986) Fluorescence detection in automated DNA sequence analysis. *Nature* 321, 674–679

8 Bashkin, J. *et al.* (1996) DNA sequencing by capillary electrophoresis with a hydroxyethylcellulose sieving buffer. *Appl. Theor. Electrophor.* 6, 23–28

9 Bashkin, J.S. *et al.* (1996) Implementation of a capillary array electrophoresis instrument. *J. Capillary Electrophor.* 3, 61–68

10 Behr, S. *et al.* (1999) A fully automated multicapillary electrophoresis device for DNA analysis. *Electrophoresis* 20, 1492–1507

11 Prober, J.M. *et al.* (1987) A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. *Science* 238, 336–341

12 McCombie, W.R. *et al.* (1992) Rapid and reliable fluorescent cycle sequencing of double-stranded templates. *DNA Seq.* 2, 289–296

13 Ju, J. *et al.* (1995) Fluorescence energy transfer dye-labeled primers for DNA sequencing and analysis. *Proc. Natl. Acad. Sci. U. S. A.* 92, 4347–4351

14 Lee, L.G. *et al.* (1997) New energy transfer dyes for DNA sequencing. *Nucleic Acids Res.* 25, 2816–2822

15 Hall, N. (2007) Advanced sequencing technologies and their wider impact in microbiology. *J. Exp. Biol.* 210, 1518–1525

16 Myers, E.W. *et al.* (2000) A whole-genome assembly of Drosophila. *Science* 287, 2196–2204

17 Adams, M.D. *et al.* (1991) Complementary-DNA sequencing – expressed sequence tags and human genome project. *Science* 252, 1651–1656

18 Porter, D. *et al.* (2006) SAGE and related approaches for cancer target identification. *Drug Discov. Today* 11, 110–118

19 Kodzius, R. *et al.* (2006) CAGE: cap analysis of gene expression. *Nat. Methods* 3, 211–222

20 Wang, T.L. *et al.* (2002) Digital karyotyping. *Proc. Natl. Acad. Sci. U. S. A.* 99, 16156–16161

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

22 Shendure, J. *et al.* (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* 309, 1728–1732

23 Braslavsky, I. *et al.* (2003) Sequence information can be obtained from single DNA molecules. *Proc. Natl. Acad. Sci. U. S. A.* 100, 3960–3964

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

25 Ronaghi, M. *et al.* (1999) Analyses of secondary structures in DNA by pyrosequencing. *Anal. Biochem.* 267, 65–71

26 Ronaghi, M. *et al.* (1998) PCR-introduced loop structure as primer in DNA sequencing. *Biotechniques* 25, 876–878

27 Nyren, P. *et al.* (1993) Solid phase DNA minisequencing by an enzymatic luminometric inorganic pyrophosphate detection assay. *Anal. Biochem.* 208, 171–175

28 Dressman, D. *et al.* (2003) Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *PNAS* 100, 8817–8822

29 Smith, A.D. *et al.* (2008) Using quality scores and longer reads improves accuracy of Solexa read mapping. *BMC Bioinformatics* 9, 128

30 Huse, S.M. *et al.* (2007) Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol.* 8, R143

31 Brockman, W. *et al.* (2008) Quality scores and SNP detection in sequencing-by-synthesis systems. *Genome Res.*

32 Ewing, B. *et al.* (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* 8, 175–185

33 Bainbridge, M.N. *et al.* (2006) Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach. *BMC Genomics* 7, 246

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

35 Hopkins, K.L. *et al.* (2007) Rapid detection of gyrA and parC mutations in quinolone-resistant *Salmonella enterica* using Pyrosequencing (R) technology. *J. Microbiol. Methods* 68, 163–171

36 Edvinsson, B. *et al.* (2007) Rapid genotyping of *Toxoplasma gondii* by pyrosequencing. *Clin. Microbiol. Infect.* 13, 424–429

37 Hefler, L.A. *et al.* (2007) Vascular endothelial growth factor gene polymorphisms are associated with prognosis in ovarian cancer. *Clin. Cancer Res.* 13, 898–901

38 van der Straaten, R. *et al.* (2007) Exploratory analysis of four polymorphisms in human GGH and FPGS genes and their effect in metho trexate-treated rheumatoid arthritis patients. *Pharmacogenomics* 8, 141–150

39 Braicu, E.I. *et al.* (2007) Polymorphism of IL-1 alpha, IL-1 beta and IL-10 in patients with advanced ovarian cancer: results of a prospective study with 147 patients. *Gynecol. Oncol.* 104, 680–685

40 Zhou, Z.Y. *et al.* (2006) Pyrosequencing, a high-throughput method for detecting single nucleotide polymorphisms in the dihydrofolate reductase and dihydropteroate synthetase genes of *Plasmodium falciparum*. *J. Clin. Microbiol.* 44, 3900–3910

41 Zhou, Z. *et al.* (2006) Pyrosequencing – a high-throughput method for detecting single nucleotide polymorphisms (SNPS) in the dihydrofolate reductase and dihydropteroate synthetase genes of *Plasmodium falciparum*. *Am. J. Trop. Med. Hygiene* 75, 56

42 Swaminathan, K. *et al.* (2007) Global repeat discovery and estimation of genomic copy number in a large, complex genome using a high-throughput 454 sequence survey. *BMC Genomics* 8, 132

43 Korbel, J.O. *et al.* (2007) Paired-end mapping reveals extensive structural variation in the human genome. *Science* 318, 420–426

44 Rodriguez-Canales, J. *et al.* (2007) Identification of a unique epigenetic sub-microenvironment in prostate cancer. *J. Pathol.* 211, 410–419

45 Bollati, V. *et al.* (2007) Changes in DNA methylation patterns in subjects exposed to low-dose benzene. *Cancer Res.* 67, 876–880

46 Liu, T. *et al.* (2007) Regulation of Cdx2 expression by promoter methylation, and effects of Cdx2 transfection on morphology and gene expression of human esophageal epithelial cells. *Carcinogenesis* 28, 488–496

47 Beier, V. *et al.* (2007) Monitoring methylation changes in cancer. pp. 1–11, Springer-Verlag, Berlin

48 Chelbi, S.T. *et al.* (2007) Expressional and epigenetic alterations of placental serine protease inhibitors – SERPINA3 is a potential marker of preeclampsia. *Hypertension* 49, 76–83

49 Goldberg, S.M. *et al.* (2006) A sanger/pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes. *Proc. Natl. Acad. Sci. U. S. A.* 103, 11240–11245

50 Bainbridge, M.N. *et al.* (2006) Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach. *BMC Genomics* 7, 11

51 Cheung, F. *et al.* (2006) Sequencing *Medicago truncatula* expressed sequenced tags using 454 Life Sciences technology. *BMC Genomics* 7, 272

52 Chetverina, H.V. and Chetverin, A.B. (1993) Cloning of RNA molecules in vitro. *Nucleic Acids Res.* 21, 2348–2353

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

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

55 Wakaguri, H. *et al.* (2007) DBTSS: database of transcription start sites, progress report 2008. *Nucleic Acids Res.* 36, D97–D101

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

57 Fields, S. (2007) Molecular biology. Site-seeing by sequencing. *Science* 316, 1441–1442

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

59 Warren, R.L. *et al.* (2007) Assembling millions of short DNA sequences using SSAKE. *Bioinformatics* 23, 500–501

60 Jeck, W.R. *et al.* (2007) Extending assembly of short DNA sequences to handle error. *Bioinformatics* 23, 2942–2944

61 Mitra, R.D. *et al.* (2003) Fluorescent in situ sequencing on polymerase colonies. *Anal. Biochem.* 320, 55–65

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

63 Kim, J.B. *et al.* (2007) Polony multiplex analysis of gene expression (PMAGE) in mouse hypertrophic cardiomyopathy. *Science* 316, 1481–1484

64 Zhang, K. *et al.* (2006) Long-range polony haplotyping of individual human chromosome molecules. *Nat. Genet.* 38, 382–387

65 Nardi, V. *et al.* (2008) Quantitative monitoring by polymerase colony assay of known mutations resistant to ABL kinase inhibitors. *Oncogene* 27, 775–782

66 Levene, M.J. *et al.* (2003) Zero-mode waveguides for single-molecule analysis at high concentrations. *Science* 299, 682–686

67 Rich, A. (1998) The rise of single-molecule DNA biochemistry. *Proc. Natl. Acad. Sci. U. S. A.* 95, 13999–14000

68 Blazej, R.G. *et al.* (2006) Microfabricated bioprocessor for integrated nanoliter-scale Sanger DNA sequencing. *Proc. Natl. Acad. Sci. U. S. A.* 103, 7240–7245

69 Deamer, D.W. and Akeson, M. (2000) Nanopores and nucleic acids: prospects for ultrarapid sequencing. *Trends Biotechnol.* 18, 147–151

70 Wheeler, D.A. *et al.* (2008) The complete genome of an individual by massively parallel DNA sequencing. *Nature* 452, 872–876