Bulletin of Environment, Pharmacology and Life Sciences
Bull. Env. Pharmacol. Life Sci., Vol 7 [11] October 2018 : 200-206

Online ISSN 2277-1808
Journal's URL:http://www.bepls.com
CODEN: BEPLAD
Global Impact Factor 0.876
Universal Impact Factor 0.9804
NAAS Rating 4.95

REVIEW ARTICLE **OPEN ACCESS**

Next-Generation DNA Sequencing: Methodology and Application

Mamata Khandappagol, Savita Shivanna Khandappagol[1], Swapna, Rajashree Biradar, Anupama Patil and Prabhamani Pujar Siddachar.
[1]Central Research Institute for Dryland Agriculture, Hyderabad-59, Telengana, India.
E-mail:agricosavita@gmail.com

ABSTRACT
Next-generation high-throughput DNA sequencing techniques are opening fascinating opportunities in the life sciences. Novel fields and applications in biology and medicine are becoming a reality, beyond the genomic sequencing which was original development goal and application. Serving as examples are: personal genomics with detailed analysis of individual genome stretches; precise analysis of RNA transcripts for gene expression, surpassing and replacing in several respects analysis by various microarray platforms, for instance in reliable and precise quantification of transcripts and as a tool for identification and analysis of DNA regions interacting with regulatory proteins in functional regulation of gene expression. The next-generation sequencing technologies offer novel and rapid ways for genome-wide characterization and profiling of mRNAs, small RNAs, transcription factor regions, structure of chromatin and DNA methylation patterns, microbiology and metagenomics. In this article, development of commercial sequencing devices is reviewed and presently commercially available very high-throughput DNA sequencing platforms, as well as techniques under development, are described and their applications discussed.
Key words: NGS, Illumina, Roche 454, ABSOLiD and Ion Torrent.

Received 12.07.2018 Revised 20.08.2018 Accepted 02.10.2018

INTRODUCTION

The development of DNA sequencing strategies has been a high priority in genetics research since the discovery of the structure of DNA and the basic molecular mechanisms of heredity. However, it was not until the works by Maxam, Gilbert and Sanger in 1977 [10, 15] that the first practical sequencing methods were developed and implemented on a large scale. The first isolation and sequencing of a plant cDNA by Bedbrook and colleagues a few years later initiated the field of Plant Molecular Genetics [1]. Plant biotechnology started shortly thereafter with the successful integration of recombinant DNA and sequencing techniques to generate the first transgenic plants using Agrobacterium [4, 7]. The first genetic map in plants based on restriction fragment length polymorphisms [RFLPs; 2] enabled the capture of genetic variation and started the era of molecular marker-assisted plant breeding. Since then, sequencing methodologies have been essential tools in plant research. They have allowed the characterization and modification of genes and metabolic pathways, as well as the use of genetic variation for studies in species diversity, marker-assisted selection (MAS), germplasm characterization and seed purity. The determination of the reference genomes in Arabidopsis thaliana, rice and maize using Sanger sequencing strategies constituted major milestones that enabled the analysis of genome architecture and gene characterization in plants. More recently, the development and increasing availability of multiple Next-Generation sequencing (NGS) technologies minimized research limitations and bottlenecks based on sequence information. It is difficult to overstate the influence that these massively parallel systems have had in our understanding of plant genomes and in the expansion, acceleration and diversification of breeding and biotechnology projects. At the same time, this influence tends to understate the importance that capillary Sanger sequencing still has in day-by-day research and development work. This review provides a description of major sequencing technologies that are available today, their use as well as future prospects in basic plant genetics research, biotechnology and breeding in crop plants.

NGS SEQUENCING TECHNOLOGIES

Sanger sequence analyzers

For more than 30 years and until recently, sequencing based on the Sanger and Maxam Gilbert chemistries were the only practical methods to routinely determine DNA sequences in plants and other biological systems. During the 80's and 90's Sanger-based platforms increased throughput by orders of magnitude and became the method of choice while the Maxam-Gilbert method remained a low-throughput process. The development of automated Sanger systems was greatly facilitated by technical innovations such as thermal cycle-sequencing and single-tube reactions in combination with fluorescence-tagged terminator chemistry [19]. Additional improvements in parallelization, quality, read length and cost-effectiveness were achieved by the development of automatic base calling and capillary electrophoresis. In the current version of Sanger sequencing a mixture of primer, DNA polymerase, deoxinucleotides (dNTPs) and a proportion of dideoxynucleotide terminators (ddNTP), each labelled with a different fluorescent dye, are combined with the DNA template. During the thermal cycling reaction, DNA molecules are extended from templates and randomly terminated by the occasional incorporation of a labelled ddNTP. DNA is then cleaned up and denatured. Detection is achieved by laser excitation of the fluorescent labels after capillary-based electrophoresis separation of the extension products. The differences in dye excitation generate a "four colour" system that is easily translated by a computer to generate the sequence. Modern Sanger sequencers like the Applied Biosystems ABI3730 have reached a high level of sophistication and can achieve routine read-lengths close to 900 bp and per-base raw accuracies of 99.99% or higher [16].

Roche 454

The 454 platform was the first NGS platform available as a standalone system. DNA templates need to be prepared by emulsion PCR and bound to beads, with 1–2 million beads deposited into wells in a titanium-covered plate. The Roche 454 technology is based on Pyrosequencing and additional beads that have sulphurylase and luciferase attached to them are also loaded into the same wells to generate the light production reaction. DNA polymerase reactions are performed in cyclesbut, unlike Sanger, there are no terminators. Instead, one single dNTP is alternated in every cycle in limiting amounts. Fluorescence after the reaction indicates the incorporation of the specific dNTP used in the cycle (Figure1) [11]. Because the intensity of the light peaks is proportional to the number of bases of the same type together in the template, the fluorescence can be used to determine the length of homopolymers, although accuracy decreases considerably with homopolymer length. The current 454 chemistry is able to produce the longest reads of any NGS system, about 700 bp, approaching those generated by Sanger reads. However, 454 systems can sequence several megabases for less than 100 dollars.

Illumina

The Solexa platform (now owned by Illumina) has become the most widely used NGS system in Plant biotechnology and breeding. Illumina captures template DNA that has been ligated to specific adapters in a flow cell a glass enclosure similar in size to a microscope slide, with a dense lawn of primers. The template is then amplified into clusters of identical molecules, or polonies and sequenced in cycles using DNA polymerase. Terminator dNTPs in the reaction are labelled with different fluorescent labels and detection is by optical fluorescence. As only terminators are used, only one base can be incorporated in one cluster in every cycle. After the reaction is imaged in four different fluorescence levels, the dye and terminator group is cleaved off and another round of dye-labelled terminators is added (Figure 2). A the total number of cycles determine the length of the read and is currently up to 101 or 151, for a total of 101 or 151 bases incorporated, respectively. This technology was able to yield the highest throughput of any system, with one of the highest raw accuracies. One major disadvantage is the short read it produces. However, paired-end protocols virtually double the read per template and facilitate some applications that were originally out of the reach of the technology. The Illumina HiSeq 2000 sequencer is currently able to sequence up to 540-600 Gbp in a single 2-flow cell, 8.5-day run at a cost of about 2 cents per Mbp.

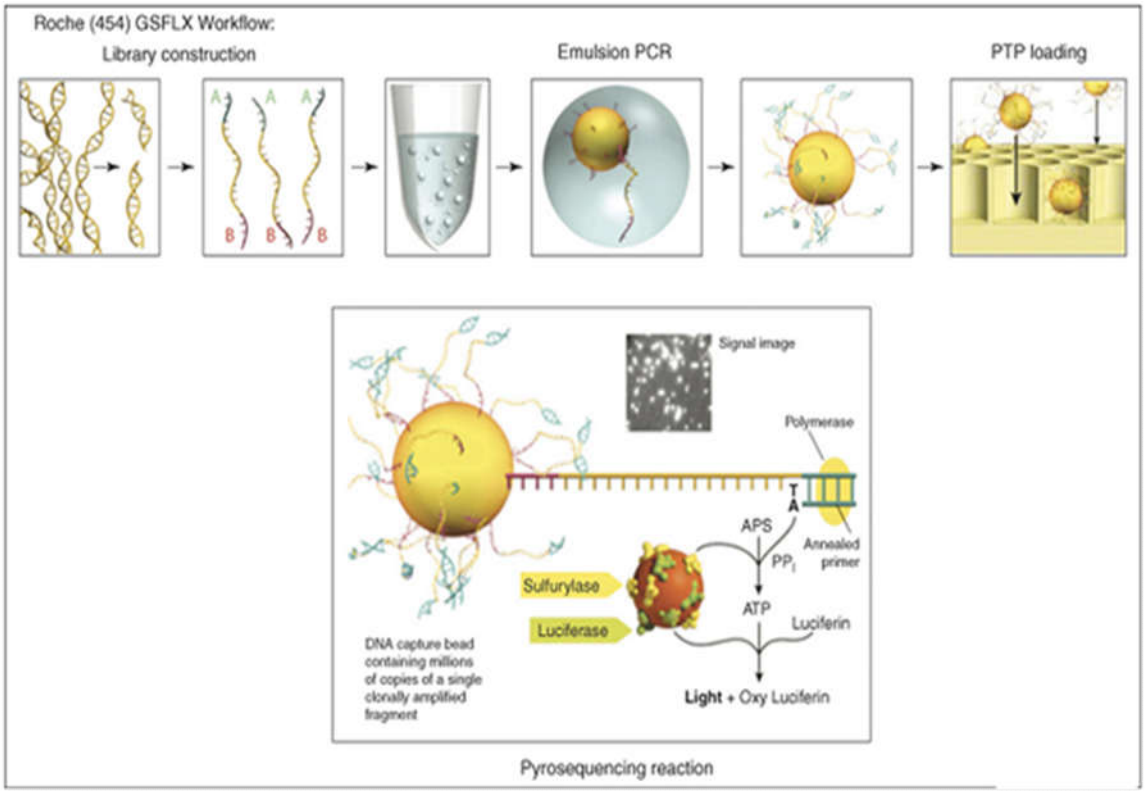

Figure1. 454 Workflow: library construction ligates 454-specificadaptersto DNA fragments and couples amplification beads with DNA in an emulsion PCR to amplify fragments before sequencing. The beads are loaded into the picotiterplate (PTP). The bottom panel illustrates the pyrosequencing reaction that occurs on nucleotide incorporation to report sequencing by synthesis.

Life Technologies SOLiD

ABI has commercialized the SOLiD (Support Oligonucleotide Ligation Detection) platform. This platform is based on Sequencing by Ligation (SbL) chemistry. SbL is a cyclic method but differs fundamentally from other cyclic NGS chemistries in its use of DNA ligase instead of polymerase, and two-base encoded probes instead of individual bases as units. In SbL, a fluorescently labeled 2-base probe hybridizes to its complementary sequence adjacent to the primed template and ligated. Non-ligated probes are then washed away, followed by fluorescent detection. In SOLiD, every cycle (probe hybridization, ligation, detection, and probe cleavage) is repeated ten times to yield ten colour calls spaced in five-base intervals. The extension product is removed and additional ligation rounds are performed with an n–1 primer, which moves the calls by one position. Colour calls from the ligation rounds are then ordered into a linear sequence to decode the DNA sequence (Figure 3) [11]. SOLiD has similar throughput and cost per base to Illumina. It also has the best raw accuracy among commercial NGS systems.

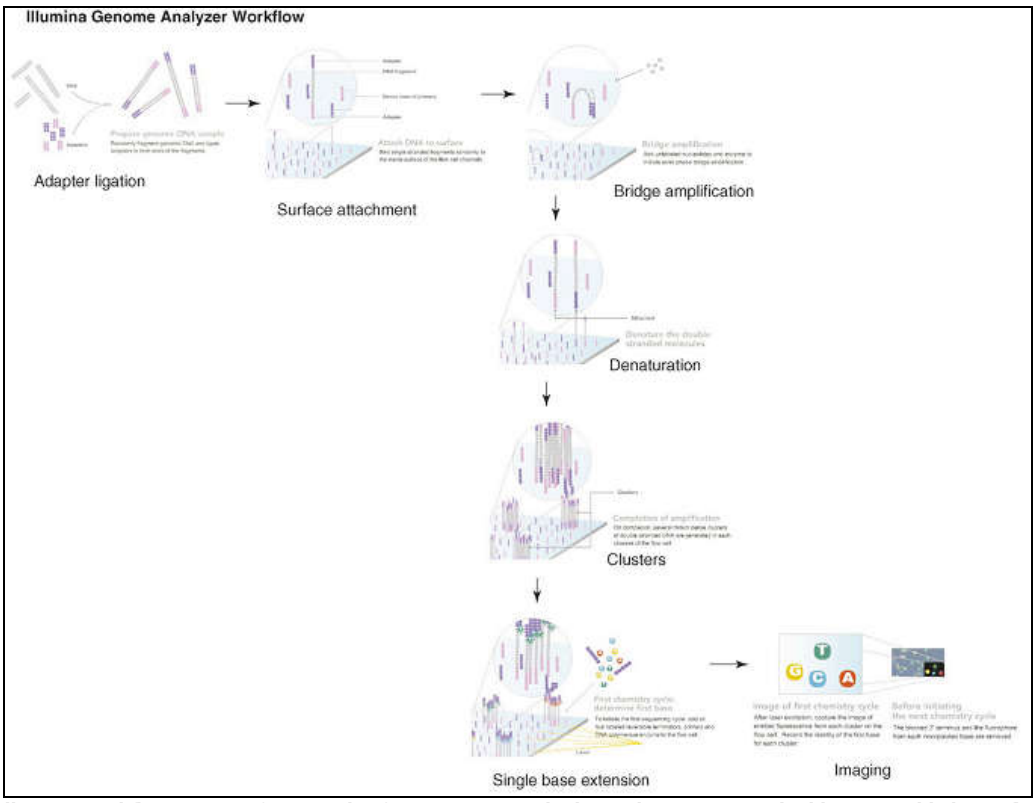

Figure2. Illumina workflow Starting from similar fragmentation and adapter ligation steps, the library is added to a flow cell for bridge amplification (an isothermal process that amplifies each fragment into a cluster).The cluster fragments are denatured, annealed with a sequencing primer and subjected to sequencing by synthesis using 30 blocked labeled nucleotides.

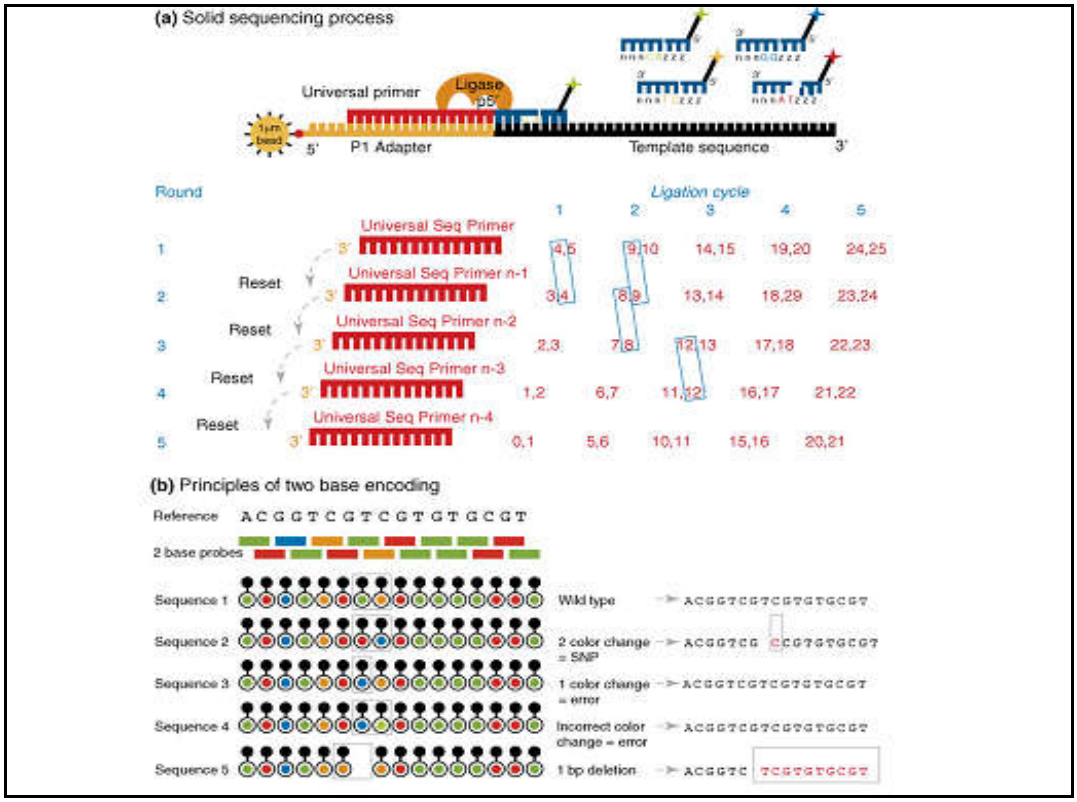

Figure3. ABSOLiD sequencing. (a) ABSOLiD sequencing by ligation first anneals a universal sequencing primer then goes through subsequent ligation of the appropriate labeled 8mer, followed by detection at each cycle. (b)Two bases encoding of the ABSOLiD data greatly facilitates the discrimination of base calling errors from true polymorphisms or indel events. Figures related to the SOLiD(tm) System are reproduced with permission from Applied Biosystems. (c) 2007AppliedBiosystems.Allrights reserved

Life Technologies Ion Torrent

Ion torrent is the commercial name for a new NGS platform now owned by Life Technologies. Currently, its usefulness in being evaluated for a number of applications in plant biotechnology and breeding. Ion Torrent differs from other NGS in that its chemistry does not require fluorescence or chemiluminescence, and for that matter optics (e.g. a CCD camera) to work. Beads, each carrying PCR clones from a single original fragment, are subjected to polymerase synthesis using standard dNTPs on an ion chip. The ion chip is a massively parallel semiconductor-sensing device that contains ion sensitive, field-effect transistor-based sensors (tiny pH meters, essentially), coupled to more than one million wells where the polymerization reaction occurs. Cycles of reactions including one single nucleotide are produced, in a way that is analogous to the Roche 454 system. In each cycle, the electronic detection of changes in pH due to the release of a proton during base incorporation indicates that a base has been incorporated. The IonTorrent has the lowest throughput but also the fastest turnaround times of all commercially available NGS systems. The current Ion Torrent chip can yield several hundred thousand reads with an average length of about 100 bp in less than 2 hours.

Other NGS platforms, Helicos Heliscope, Polonator

There are other NGS systems that have been marketed in the last few years, however, they have had limited use in plant sciences. Helicos developed the first commercial single molecule sequencer, called HeliScope. However, very few units were sold due to the cost of the machine, on-site requirements and other considerations. Currently, Helicos provides sequencing as a service. One additional company, Azco-Biotech is marketing the Max-Seq Genome sequencer (http://www.azcobiotech.com /instruments/maxseq.php). This commercial version of the academic, open-source Polonator can run either sequencing by synthesis or sequencing by ligation protocols,similar to Illumina and SOLiD, respectively, although it generates shorter reads, 35- or 55-bp-long.

Pacific Biosciences and the 3rd generation

Pacific Biosciences has launched the PacBio RS platform, considered the first commercially available 3rd - Generation system. The first early-access instruments were deployed in late 2010. The PacBio system is based on SMRT, a single-molecule sequencing chemistry with real time detection. The sequencing cell has DNA polymerases attached to nano wells and exposed to single molecule templates and labelled NTPs. No terminators are used, although conditions are set to slow polymerization to a level that can be detected by a CCD camera. Each dNTP has a unique fluorescent label that is detected and then cleaved off during synthesis. Polymerization is detected as it happens, several bases per second. Because of this real-time detection and the enzyme processivity, this method has the potential to generate reads in excess of 10 kilobases in a few minutes. The potential of a technology that is able to sequence single molecules and produce long reads is immense. However, the PacBio technology may need to overcome a number of technical challenges before it reaches a widespread use in plant sciences. Average read length in current outputs exceeds 1 Kbp although single-pass error rate has been reported to be 15%, considerably higher than other sequencing platforms [6]. One major source of errors consists of deletions produced during detection. As will be discussed later, improvements in raw quality and further gains in read length will broaden the range of optimal applications for PacBio.

Applications of high-throughput DNA sequencing

Novel fields and applications in biology and medicine are becoming a reality, beyond genomic sequencing as the original development goal and application. Examples include personal genomics with detailed analysis of individual genomic stretches; precise analysis of RNA transcripts for gene expression, surpassing and replacing in several aspects analysis carried out by various microarray platforms, for example in reliable and precise transcript quantification; and as a tool for identification and analysis of DNA regions that interact with regulatory proteins in functional regulation of gene expression. Next-generation sequencing technologies offer novel, rapid ways for genome-wide characterisation and profiling of mRNAs, small RNAs, transcription factor regions, chromatin structure and DNA methylation patterns, in microbiology and metagenomics.

Personal genomics, project human diversity in 1000 genomes

The cost of genome sequencing, an important factor in future studies, is becoming low enough to make personal genomics a close reality. Reduction of cost by two orders of magnitude is needed to be able to realise the potential of personal genomics, for which the goal of $1000 for a human genome sequence has been set. The impressive results obtained so far in various projects with the new technology are very convincing and will lead to lower cost. The analysis of the first two available human genomes [9, 22]. has demonstrated, how difficult it still is to draw medically or biologically relevant conclusions from individual sequences. More genomes need to be sequenced, to learn how genotype correlates with phenotype. A plan for a project to sequence 1000 human genomes has been prepared, which will allow creation of a reference standard for the analysis of human genomic variations that is expected to

contribute to studies of disease (http://www.1000genomes.org/). Illumina, Roche 454 Life Sciences and Applied Biosystems will take part in the project and generate the equivalent of 25 human genomes each per year over a period of three years. This significant sequence contribution will enable the team to analyse the human genome with deeper sequencing and shorten its completion time. The 1000 Genomes Project will identify variants present at a frequency of 1% over most of the genome, and as low as 0.5% within genes

The immediate applications and relevance of next-generation sequencing techniques in the medical field have been demonstrated already, by the ability to detect cancer alleles with deep sequencing of genomic DNA in cancerous tissues (carefully isolated by laser micro dissection and capture techniques), which would have presented a very tedious task for the Sanger technique.

RNA sequencing, analysis of gene expression

The high throughput of next-generation sequencing technology, rapidly producing huge numbers of short sequencing reads, made possible the analysis of a complex sample containing a mixture of a large number of nucleic acids, by sequencing simultaneously the entire sample content. This is now possible without the tedious and time-consuming bacterial cloning, avoiding associated disadvantages. It may also be applied to the characterisation of mRNAs, methylated DNA, DNA or RNA regions bound by certain proteins and other DNA or RNA regions involved in gene expression and regulation. The original SAGE technique [21] demonstrated novelty and powerful analysis, but was limited in applications because of the need for difficult ligation of a huge number of short DNA transcripts, subsequent cloning and Sanger sequencing. Using next-generation technology, the concept of the SAGE method now allows the analysis of RNA transcripts in a biological sample by obtaining short sequence tags, 20–35 bases long, directly from each transcript in the sample. With this technique, transcripts are characterised through their sequence [13] in contrast to the probe hybridisation employed in DNA chip techniques, with their inherent difficulties of cross hybridisation and quantitation. Owing to the huge number of samples analysed simultaneously, sequence-based techniques can detect low abundance RNAs, small RNAs, or the presence of rare cells contained in the sample. Another advantage of this approach is that it does not require prior knowledge of the genome sequence. The technique has been applied recently to transcriptome profiling in stem cells [3] and to RNA-Seq study into alternative splicing in human cells [17].

Chromatin immunoprecipitation, ChIP-Seq technique

Next-generation sequencing technology allowed replacement of microarrays in the mapping step with high-throughput sequencing of DNA binding sites, and their direct mapping to a reference genome in the database [14]. The sequence of the binding site is mapped with high resolution to regions shorter than 40 bases, a resolution not achievable by microarray mapping. Moreover, the ChiP-Seq technique is not biased and allows the identification of unknown protein binding sites, which is not the case with the ChiP-on-chip approach, where the sequence of the DNA fragments on the microarray is pre-determined, e.g. in promoter arrays, exon arrays, etc.

PROSPECTS FOR FUTURE DNA SEQUENCING TECHNOLOGY AND APPLICATION

The availability of ultra-deep sequencing of genomic DNA will transform the biological and medical fields in the near future, in analysis of the causes of disease, development of new drugs and diagnostics. It may become a promising tool in the analysis of mental and developmental disorders such as schizophrenia and autis [5, 12 and 18]. It is anticipated that DNA sequencing of whole genomes for clinical purposes using these new technologies will probably occur in the next couple of decades. Some of the most recent applications can be found in the proceedings of the AGBT conference (Advances in Genome Biology and Technology), February 2009.

The novel sequencing technologies will be also useful in microbial genomics, for example in the metagenomics measuring the genetic diversity encoded by microbial life in organisms inhabiting a common environment [8]. Many microbial sequencing projects have been already completed or are being prepared and several comparative genome analyses are under way to link genotype and phenotype at the genomic level. The proposed Human Microbiome Project (also called The Second Human Genome Project), analysing the collection of microbes in and on the human body, will contribute to understanding human health and disease [20]. Changes in microbial communities in the body have been generally linked to immune system function, obesity and cancer. In future, each individual's microbiome could eventually become a medical biometric.

An important application is planned by the US DOE Joint Genome Institute, JGI (http://www.jgi.doe.gov/), which will focus its sequencing efforts on new plant and microbial targets that may be of use in the development of alternative energies. The JGI plans to sequence the genome of the marine red alga, which may play an important environmental role in removing carbon dioxide from the atmosphere.

Genomics, proteomics and medical research all benefit from recent advances and novel techniques for high-throughput analysis (e.g. DNA and protein microarrays, quantitative PCR, mass spectrometry, novel DNA sequencing techniques and others). Devices with short DNA sequence reads (25–50 bases) have already found many applications, but for genomic sequencing, and for analysis of the ever more important structural genetic variations in genomes, such as copy number variations, chromosomal translocations, inversions, large deletions, insertions and duplications, it would be a great advantage if sequence read length on the original single DNA molecule could be increased to several 1000 bases and more per second. Ideally, the goal would be the sequence determination of a whole chromosome from a single original DNA molecule. Hopes for future in this direction may provide novel developments in several physical techniques (e.g. various advanced AFM methods, electron microscopy, soft X-rays, various spectroscopic techniques, nanopores and nano-edges), with many improvements needed and under intense development.

REFERENCES

1. Bedbrook, J. R., Smith, S. M. and Ellis, R. J. (1980). Molecular cloning and sequencing of cDNA encoding the precursor to the small subunit of chloroplast ribulose 1, 5-bisphosphate carboxylase. *Nature* .287(5784): 692–697.
2. Bernatsky, R. and Tanksley, S. (1986). Toward a saturated linkage map in tomato based on isozymes and random cDNA sequences. *Genetics*. 112(4): 887-898.
3. Cloonan, N. *et al.* (2008). Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat. Methods.*5:613–619.
4. Fraley, R. T., Stephen. G., Rogers., Robert. B., Horsch., Patricia. R., Sanders., Jeffery. S., Flick., Steven. P., Adams. and Michael. L., (1983). Expression of bacterial genes in plant cells. *Proc. Natl. Acad. Sci. USA* .(80)15: 4803–4807.
5. Geschwind, D.H. (2008). Autism – family connections. *Nature*. 454:838–839.
6. Glenn, T. C. (2011). Field guide tonext-generation DNA sequences. *Mol. Ecol. Resour.,* 11(5): 759-769.
7. Herrera-Estrella, L., Depicker, A., Montagu, M. V. And Schell, J., (1983). Expression of chimaeric genes transferred in to plant cells using a Ti plasmid-derived vector. *Nature.* 303: 209-213.
8. Hugenholtz, P. and Tyson, G.W. (2008) Metagenomics. *Nature.*455:481–483.
9. Levy, S., Sutton, G., Pauline, C., Feuk, L., halpern, A. L., Walenz, B. P., *et al*. (2007). The diploid genome sequence of an individual human. *PloS5*, e254
10. Maxam, A. M. and Gilbert, W. (1977). A new method for sequencing DNA*. Proc. Natl. Acad. Sci. USA* .74(2): 560–564
11. Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nature Reviews* .11(1): 31-46.
12. Morrow, E. M., Seung- Yun., Steven, W., Tae- Kyung., Lin, Y., Hill, R. S., Mukaddas, N. M., *et al*. (2008) Identifying autism loci and genes by tracing recent shared ancestry.Science321, 218–223.
13. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. And Wold, Barbara., (2008) Mapping and quantifying mammalian transcriptomes by RNA Seq.Nat. *Methods.*5:621–628.
14. Robertson, G., Wilson, M. D., Spyrou, C., Brown, D. D., Hadfield, J. And Odom, D, T., *(*2007) ChIP-Seq techniques.*Nat. Methods*.4:651–657.
15. Sanger, F., Nicklen, S. and Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. USA*. 74(12): 5463–5467.
16. Shendure, J. and Ji, H. (2008). Next-generation DNA sequencing. *Nat. Biotech.,* 26(10): 1135-1145.
17. Sultan, M., Schulz, M. H., Richard, H., Magen, A., Klingenhoff, A., Scherf, M. and seif, M., *et al*. (2008) A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome.*Science*. 321:956–960.
18. Sutcliffe, J.S. (2008). Insights into the pathogenesis of autism.*Sci.*321:208–209.
19. Trainor, G. L. (1990). DNA sequencing, Automation and the Human Genome. *Analy. Chemi.,* .62(5): 418-426.
20. Turnbaugh, P.J., Ley, R. E., Hamady, M., Fraser-Liggett., Knight, R. and Gordon, J. I., (2007) The human microbiome project.*Nature*. 449:804–810.
21. Velculescu, V. E., Zhang, L., Vogelstein, B., Kinzler, K. W., (1995). Serial analysis of gene expression. *Science*. 270:484– 487.
22. Wheeler, D. A., Srinivasan, M., Egholm, M., Yufeng Shen., Lei Chen., McGuire, A., Wen, H., Yi-Ju, C., 1 *et al*. (2008). The complete genome of an individual by massively parallel DNA sequencing. *Nature*. 452: 872–876

CITATION OF THIS ARTICLE

Mamata Khandappagol, Savita Shivanna Khandappagol, Swapna, Rajashree Biradar, Anupama Patil and Prabhamani Pujar Siddachar. Next-Generation DNA Sequencing: Methodology and Application. Bull. Env. Pharmacol. Life Sci., Vol 7 [11] October 2018: 200-206