CIBTech Journal of Biotechnology ISSN: 2319-3859 (Online)
An Online International Journal Available at http://www.cibtech.org/cjb.htm
2013 Vol. 2 (4) October-December, pp.1-20/Ritesh and Malik
Review Article

NEXT GENERATION SEQUENCING: A REVOLUTION IN GENE SEQUENCING

Ritesh Kaur and *Chander Parkash Malik
School of Life Sciences, Jaipur National University, Jaipur, Rajasthan 302017, India
**Author for Correspondence*

ABSTRACT
The discovery of dideoxynucleotide sequencing by Sanger when combined with innovative physical mapping approaches that helped to set up long-term relationships between cloned stretches of genomic DNA, fluorescent DNA sequencers produced reference genome sequences for model organisms and for the reference human genome. From last several years the revolutionary advances takes place in DNA sequencing technologies, with the introduction of next-generation sequencing (NGS) techniques. With the aid of these NGS methods, now millions of bases to be sequenced in one round, and are very cost effective and we are rapidly moving to the point where every organism is able to sequence. This review entails a brief abridgment of NGS technologies and the development of exemplary applications of such methods in the fields of molecular marker development, metagenomics, transcriptomes investigation, epigenetic analysis, protein-DNA interactions, etc.

Key Words: *NGS, Metagenomic, Transcriptomics, Protein-DNA Interactions*

INTRODUCTION
Genome sequencing is widely appreciated in shaping the future biomedical research. It provides a general outline for assembling fragmentary DNA information into landscape of biological structure and function. The rapid advancement in DNA sequencing technology makes a tremendous shift in high throughput increase, highly reduced per base cost of raw sequence, specialized infrastructure of robotics, bioinformatics, databases and instrumentation and an accompanying requirement for extensive investment in the equipments for proper utilization of the technologies. The landmark achievements of Sanger and Coulson (1975), Maxam and Gilbert (1977) and the development of the dideoxy chain termination method provided the base for the sequence based research for decades known as DNA sequencing technology and commonly referred as Sanger sequencing. The automated Sanger method is considered as a 'first generation' technology and the newer methods are referred to as 'next generation sequencing' (NGS), which constitute various strategies which are based on a combination of one of the many protocols for template preparation, sequencing, imaging, genome alignment and assembly. The commercial NGS technologies have transformed our perspective towards various scientific approaches in basic, applied and clinical research. At present NGS platforms commonly utilize quite different chemistry and base incorporation and detection tools, which comprise two main steps: template library preparation and detection of incorporated nucleotides (Glenn, 2011; Zhang *et al.*, 2011). Recently, six NGS platforms are available which are commercially classified into two groups: the first group consists of PCR based technologies and is comprised of four platforms: which include Roche GS-FLX 454 sequencer (Roche Diagnostics Corp, Branford, CT, USA), Illumina Genome Analyzer (Illumina Inc., San Diego, CA, USA), ABI SOLiD System (Life Technologies Corp., Carlsbad, CA, USA) and Ion Personal Genome Machine (Life Technologies, South San Francisco, CA, USA). The second group includes, the HeliScope (Helicos Bioscience Corp., Cambridge, MA, USA) and PacBio RS SMRT System (Pacific Biosciences Menlo Park, CA, USA) is based on 'single molecule sequencing' technologies hence does not require an amplification step prior to sequencing. Among the six commercially available platforms, the Illumina/Solexa Genome Analyzer, Roche 454 GS-FLX sequencer, Applied Biosystems SOLiD Analyzer and HeliScope (second generation sequencing technologies) leads the market whereas the Pacific Biosciences PacBio RS SMRT system and Ion Personal Genome Machine by Life Technologies (third

generation sequencing technologies) have been introduced recently. Novel approaches of sequencing and also improvement in second and third generation sequencing are in phase of development to bring the cost of human genome sequencing under $1000.These comprises the use of scanning tunneling microscope (SEM), fluorescence resonance energy transfer (FRET), single molecule detector and protein nanopores. With the help of NGS technologies, analysis that was unreachable luxuries just a few years ago is being increasingly enabled at a rapid speed. Large-scale sequencing centers are now swapping to next generation sequencing.

History and Advances in Sequencing Technologies

Sanger sequencing method elaborate the DNA polymerase dependent synthesis of a complimentary DNA strand using natural 2'-deoxynucleotides (dNTPs) and termination of synthesis using 2', 3'-dideoxynucleotides (ddNTPs) that serve as non-reversible synthesis terminators. The introduction of a ddNTP in the growing oligonucleotide chain terminates the DNA synthesis reaction resulting in a set of shortened fragments of varying lengths with an appropriate ddNTP at their 3' terminus. High resolution polyacrylamide gel electrophoresis (PAGE) is used to separate these shortened fragments and analyzed to reveal the DNA sequence.

Further increasing advancements in enzymology, polymer biochemistry and capillary array electrophoresis, fluorescence detection technology and fluorescent dyes helped DNA sequencing reach the current status. Sanger sequenced the first genome, bacteriophage ΦX 174, which is 5375 bases in length. The automated high throughput Sanger sequencer is the most advanced version which has a 96-capillary array format which is capable of sequencing upto 1 kb for 96 individual samples at a time. The major advancements in recombinant protein engineering, fluorescent dye development, capillary electrophoresis, automation, robotics, informatics and process management helps in the success of human genome project. In 2001, the first consensus sequence of human genome and human diploid sequence is obtained from Sanger sequencing (Levy *et al.*, 2007). Since 2004, efforts were directed towards the development of next generation sequencing technologies which left Sanger sequencing with fewer reported advances.

And now, NGS technologies have the ability to potentially produce vast amount of data cheaply. These massively parallel high throughput sequencers are capable of generating sequence reads from fragmented libraries of a particular genome (genome sequencing); from cDNA library fragments generated through reverse transcription and from a pool of PCR amplified products (amplicon sequencing) and have an excess of one billion sequence reads per instrument per run. A major platform in the next-generation sequencing (NGS) is RNA sequencing Costa *et al.,* (2010) provide a broad analysis of the RNA sequencing methodology which accurately define the expression levels of specific genes, differential splicing, and allele-specific expression of transcripts at the transcriptomes level.

The sequencing of micro RNAs in adenovirus type 3 (AD3) infected human laryngeal epithelial (Hep2) cells (Qi *et al.,* 2010). In analysis of micro RNAs profiles, 492 precursor micro RNAs identified in the AD3 infected Hep2 cells and 540 precursor micro RNAs in the control by using the SOLiD sequencing technology. Thus, NGS is effective and influential for micro RNA profiling in the virus-infected cell lines. Cui *et al.,* (2010) also apply SOLiD sequencing to profile micro RNAs involved in the host response to enterovirus 71 (EV71) infections and found 64 micro RNAs with changed expression from more than 2-fold in response to EV71 infection in Hep2 cells.

With the help of another NGS technology, ChIP-Seq, targeted micro RNA genes of a transcription factor, EGR1, in human erythroleukemia cell line K562 (Wang *et al.,* 2010). A comparative genome-wide polymorphism-fixation analysis of human codons (Jiang *et al.,* 2010). As numerous mutation data have been identified by sequencing and many more will be identified by NGS in the near future, such analysis may help us understand mutational process in the recent genome evolution. A time line comparison of different NGS technologies in terms of read length, accuracy and total output enlighten rapid progress in sequencing abilities of NGS platforms are summarized in Table 1.

Table 1: Comparative profile of different Next Generation Sequencing (NGS) technologies

Generation	Company	Platform	Template preparation method	Sequencing method	Detection method	Approx. read length (bases)	Approx. sequencing output/run	Run time (days)
Second	Roche	Roche/454 GS FLX system	Fragment, Mate Pair library/emPCR	Pyrosequencing	Optical	350-450	0.45 Gb	0.35
Second	Illumina	Illumina GA IIx	Fragment, Mate Pair library/Bridge amplification	Reversible terminator/ sequencing by synthesis	Fluorescence/Optical	50-75	≤95 Gb	7-14
Second	ABI	ABI/SOLiD 5500xl system	Fragment, Mate Pair library/emPCR	Cleavage probe/sequencing by ligation	Fluorescence/Optical	35-75	~250 Gb	7-8
Second	Helicos	HeliScope	Fragment, Mate Pair library/single molecule	Single molecule / sequencing by synthesis	Fluorescence/Optical	25-35	~20-28 Gb	≤1
Third	Pacific Biosciences	PacBio RS system	Fragment library/single molecule	Real-time single molecule sequencing	Fluorescence/Optical	1000	~60-75 Mb	0.02
Third	Life Technologies/ Ion Torrent	Personal Genome Machine (with Ion torrent-314, 316 and 318 chip)	Single molecule	Sequencing by synthesis	Change in pH detected by Ion-sensitive Field Effect Transistors (ISFETs)	100-200	≥10 Mb for 314, ≥100 Mb for 316 and ≥1 Gb for 318 chip	0.15 for 314, 0.2 for 316 and 0.23 for 318 chip

Next-generation sequencing methods

Broad steps of all the next generation sequencing technologies are: Template preparation, sequencing, imaging and data analysis and the unique combination of different steps differentiates one technology from the other. Currently, clonally amplified templates and single DNA molecule template are the two approaches used for template preparation. Cyclic reversible termination (CRT), single nucleotide addition (SNA), sequencing by ligation (SBL) and real time sequencing are different methods used for sequencing. Different methods of imaging used by NGS platforms range from measuring bioluminescent signals, pH changes, temperature changes, one color imaging to four color imaging of single molecular events. Moreover, the huge scale of sequencing requires an equally matching scale of computational analysis which comprises of image analysis, signal processing, background subtraction, base calling and quality control to produce the final sequence data from each run. In other words, these analysis place substantial demand on the information technology (IT), data storage, packing and library information management system (LIMS) infrastructures.

Template Preparation

In different NGS technologies, standard approaches used for template preparation are randomly breaking genomic DNA into small fragments which are then used to generate either fragment library or the mate-pair library. The template fragments in libraries are then immobilized on a solid surface or support to allow thousands to billions of sequencing reactions to be undertaken simultaneously.

Clonally Amplified Templates

The two common methods used to prepare clonally amplified templates are emulsion PCR (Dressman *et al.,* 2003) and solid-phase amplification (Fedurco *et al.,* 2006). Each amplified product of a circularized fragment is called a DNA nanoball (DNB). Amplification of templates is required to generate enough signals that can be detected by the imaging systems which are incapable of detecting single fluorescent signals.

Emulsion PCR (emPCR)

Emulsion PCR allows cloned amplification of the templates without the use of bacterial cloning method which usually results in the loss of genomic sequences. In emPCR, fragment or mate pair library is generated, following which adaptor sequences carrying universal priming sites are attached to the library templates to allow their PCR amplification using universal sequencing primers. Adaptor ligated DNA templates are made single stranded and then captured onto the agarose beads whose surfaces are decorated with oligomers complimentary to adaptor sequences under conditions which favor the hybridization of one DNA molecule per bead. Each of the bead carrying a single DNA fragment hybridized to the oligo-decorated surface is then isolated into the individual oil-water micelles containing PCR reagents by mixing bead- DNA complex in water-oil emulsion and vortexing. The micelles are then subjected to emulsion PCR to generate millions of copies of a single template molecule present on the surface of each bead. After the successful amplification and enrichment of emPCR beads, millions of copies of them can either be chemically cross linked to an amino coated glass surface (Life/APG; Polonator), immobilized on polyacrylamide gel on the standard microscope slide (Polonator; Shendure *et al*., 2005) or deposited into individual Pico Titer plate (PTP) wells (Roche/454 Genome analyzer; Leamon, 2003) in which NGS reactions can be performed.

Solid Phase Amplification

In solid-phase, DNA amplification is done by attaching adaptor ligated single stranded library fragments to a solid surface called as single molecule array or flow cell and then conducting solid-phase bridge amplification of these fragments. In bridge amplification, one end of the single stranded DNA fragment is immobilized to a solid surface through an adaptor. The fragments subsequently bend over and hybridize to the complimentary primers (creating the bridge) covalently attached to the solid surface in high density thereby forming the template for the synthesis of their complimentary strands. After amplification, a flow cell with 100-200 million spatially separated clusters are produced, where in each cluster is composed of

millions of copies of a single template molecule which provide free ends to which a universal sequencing primer can be hybridized to initiate the NGS reaction.

Single-molecule Templates

Immobilization of single molecule template on a solid support is done by using any one of three different methods. In the first method, individual primer molecules complimentary to adaptors ligated to library fragments are covalently attached to the solid support (Harris *et al.,* 2008), following which the template molecules prepared by randomly cleaving the starting material into smaller fragments and adding common adaptors to the ends are hybridized to the immobilized primer. In the second method, spatially distributed primer molecules are immobilized on the solid support following which single-stranded adaptor ligated single-molecule templates are hybridized to the immobilized primer and the primer extended resulting in covalently attached single-stranded single-molecule templates. A common primer is then hybridized to the template. In above both methods, DNA polymerase binds to the immobilized primed template configuration to initiate the NGS reaction (Harris *et al.,* 2008). In the third method, spatially distributed single polymerase molecules are attached to the solid support (Eid *et al.,* 2009) to which a single primed template molecule is attached. This method can be used with larger DNA molecules and with real time methods resulting in partially longer read sequences. Clonal amplification generates a population of identical templates, each of which has undergone the sequencing reaction and the signal observed upon imaging represents the consensus of the nucleotides or probes added to the identical templates for a given cycle. Dephasing (both lagging and leading strand) is the major problem with clonally amplified templates and occurs when individual molecules move out of synchronicity and results in increased fluorescence noise, base called errors and shorter reads (Erlich *et al.,* 2008).

Sequencing and Imaging

Different methods of sequencing can be broadly classified as cyclic reversible termination (CRT), sequencing by ligation (SBL), single nucleotide addition (SNA) and real time sequencing are discussed below:

Cyclic Reversible Termination

Cyclic reversible termination (CRT) is a cyclic method which includes combination of modified nucleotide (reversible terminators), imaging of the fluorescence generated by fluorescent dye attached to the nucleotide and cleavage of the fluorescent dye and terminating/inhibiting group. In this method, DNA polymerase binds to primed template and incorporates a single fluorescently modified nucleotide or reversible terminator which is complimentary to the template base and followed by washing of the unincorporated nucleotides. The incorporated nucleotides are then identified by imaging following which the terminating/inhibiting groups and the fluorescent dye are removed by a cleavage step. An additional washing step is also included before switching to the next incorporation step.

Two types of reversible terminators are utilized by CRT method: 3’ blocked terminator and 3’ unblocked terminator. 3’ blocked terminator carries a cleavable group attached to the 3’-oxygen of the 2’-deoxyribose sugar. Blocking groups such as 3’-O-allyl-2’-deoxyribonucleoside triphosphate (dNTPs) (Ju, 2006) and 3’-O-azidomethyl-dNTPs (Guo *et al.,* 2008; Bentley *et al.,* 2008) have been successfully utilized in CRT method. Mutant DNA polymerase is used to facilitate the incorporation of the 3’ blocked terminators. Additionally, while using 3’-blocked terminators, two chemical bonds are cleaved, one to remove the fluorophore from the nucleotide and the other to restore the 3’-OH group.

The need to screen large libraries of mutant DNA polymerases to incorporate 3’ blocked terminators led to the development of 3’-unblocked reversible terminators. The 3’-unblocked terminators show more favorable enzymatic incorporation and can be incorporated using wild type DNA polymerase. A small terminating group attached to the base of a 3’-unblocked nucleotide can act as an effective reversible terminator (Lightning; Laser Gen Inc; Wu *et al.,* 2009), whereas a second nucleoside analogue attached to the base of a 3’-unblocked nucleotide can act as an inhibitor (Virtual terminators; Helicos Biosciences; Bowers *et al.,* 2009). The terminating (Lightning Terminators) or inhibiting (Virtual terminators) groups of the 3’-unblocked terminators need appropriate modifications so that they can terminate DNA synthesis

once a single nucleotide is added. This is essential because the 3'-unblocked terminators contain a free 3'-OH group which is a natural substrate for incorporating the next incoming nucleotide. In case of 3'-unblocked terminators cleavage of a single bond releases both the terminating or inhibiting group and the fluorophore group from the base.

Sequencing by Ligation (SBL)

Instead of DNA polymerase, DNA ligase is used in a cyclic method which comprises probe hybridization, ligation of probe to the primer, fluorescence imaging and cleavage. The method either use oligonucleotide sequence in which one interrogation base is associated with a particular dye or two-base encoded probe. The first step comprises the hybridization of labeled probe to its complimentary sequence adjacent to the primer followed by a DNA ligase mediated joining of the dye labeled probe to the primer. A washing step is then incorporated to wash away the non-ligated probes following which fluorescence is imaged to determine the identity of the ligated probe (Landegren *et al.,* 1988). The next cycle can be started using either cleavable probes to remove the fluorescent dye and regenerate a 5'-PO_4 group for subsequent ligation cycles or by removing and hybridizing a new primer to the template.

Single-nucleotide Addition: Pyrosequencing

In pyrosequencing, pyrophosphate molecule is released when each nucleotide is incorporated by DNA polymerase which initiates a series of downstream enzymatic reactions to produce light by the action of enzyme *luciferase*. The amount of light generated is directly proportional to the number of nucleotides incorporated (Marguiles *et al.,* 2005). In the first step, DNA beads amplified by emulsion PCR are loaded into the individual picotiter plate (PTP) well in such a manner that each PTP well carry a single DNA bead. Smaller magnetic beads containing enzymes, *sulphurylase* and *luciferase* are also loaded into the each well to surround the DNA bead. The slide is then mounted in a flow chamber following which the sequencing reagents along with a single type of 2'-deoxyribonucleoside triphosphate are added to each well. Following the incorporation of complimentary dNTP, DNA polymerase extends the primer and pauses. The addition of next complementary dNTP then reinitiates the DNA synthesis. The order and the intensity of light generated from each PTP well undergoing the pyrosequencing reaction is then recorded as a series of peaks or flowgrams of high resolution using charge-coupled device (CCD) camera planted below the fibre-optic slide and the DNA sequence data revealed.

Real Time Sequencing

Pacific Biosciences introduced the Real time sequencing technology which is the third generation sequencing technology. This technique involves recording the fluorescence emitted during DNA synthesis, as the phosphate chain is cleaved by the continuous incorporation of dye labeled nucleotides by the DNA polymerase (Metzker, 2009). In the Pacific Biosciences platform, sequence information of template DNA is obtained when the single DNA polymerase molecule deposited at the bottom of individual zero-mode waveguide detectors (ZMW) incorporate phosphor linked nucleotides into the growing primer strand.

PCR-based Next Generation Sequencing Platforms

Roche /454 FLX Pyrosequencer

Roche 454 Genome Sequencer was introduced in 2004 and also the first next generation sequencing technology to gain commercial significance and is based on the sequencing by synthesis, pyrosequencing technology. The pyrosequencing approach make use of the pyrophosphate molecule released on each incorporation of a nucleotide by DNA polymerase during DNA synthesis to fuel a set of downstream enzyme cascade that finally produces light from the cleavage of oxyluciferin by *luciferase*. The amount of light emitted is directly proportional to the number of a particular nucleotide incorporated till the level of detector saturation. In this method, the library templates are amplified by the technique of emulsion PCR (Dressman *et al.,* 2003) instead of sequencing in PCR tubes or microtiter plate wells. In emulsion PCR, the library fragments are mixed with agarose beads, surfaces of which carry millions of oligomers attached to them which are complementary to the 454-specific adaptor sequences ligated or PCR generated on both ends of the fragments during library construction. Each of these agarose beads carrying

a single unique DNA fragment then hybridizes to the oligo decorated surface and is separated into the individual oil: aqueous micelles containing PCR reagents. (Fig.1.)The DNA beads are then exposed to emulsion PCR to generate millions of copies of the same fragment covering the surface of each bead. The amplified beads are recovered from the emulsion followed by an enrichment step that retains only the amplified beads discarding the failed ones. The beads (each containing a unique amplified fragment) are then arrayed into the several hundred thousand single wells on the surface of the pico titer plate (PTP) with each well holding a single bead and providing fixed location for each sequencing reaction to be monitored. Subsequently, much smaller magnetic and latex beads of 1 µm diameter, containing active enzymes (*sulphurylase* and *luciferas*e) are added and centrifuged to surround the DNA containing agarose beads in the PTP wells.

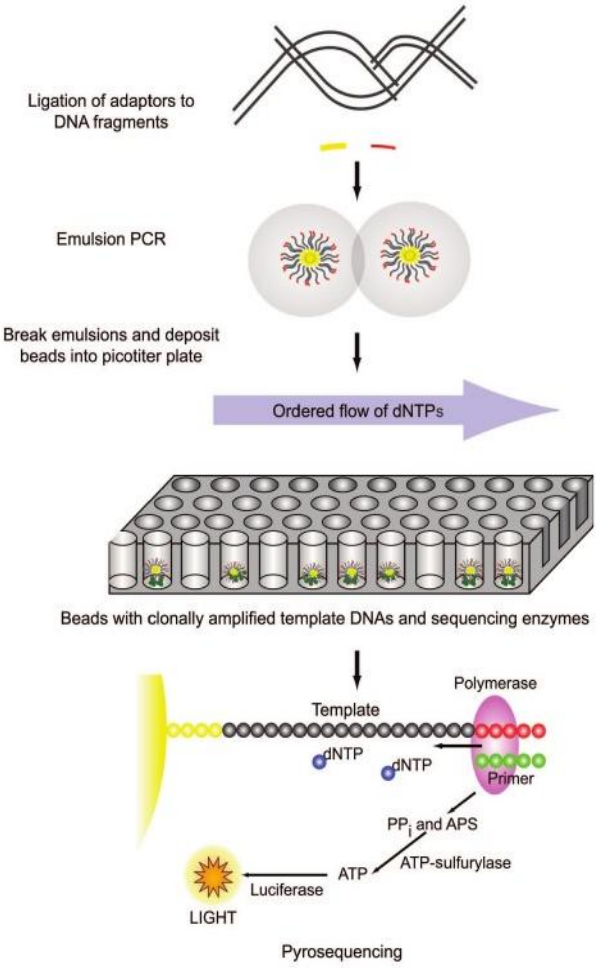

Fig. 1. Roche 454 GS FLX sequencing
(Source- Karl *et al.*, 2009).

PTP is then placed in the sequencer where it acts as a flow cell into which each of the nucleotides and other pyrosequencing reagents are delivered in a sequential fashion. Each incorporation step is then followed by an imaging. In which a CCD camera placed opposite the PTP records the light emitted from each bead due to luciferase activity. A defined single nucleotide pattern in the adaptor sequence adjacent to the universal sequencing primer, which corresponds to the sequence of the first four sequences added, enables the 454-analysis software to calibrate the level of light emitted from single nucleotide incorporation for the downstream base-calling analysis that occurs after the run is completed. For

homopolymeric repeats of up to 6 nucleotides, the number of dNTPs incorporated is directly proportional to the intensity of light. The sequential flow of nucleotides entirely eliminates the occurrence of substitution errors in the Roche/454 sequence reads. The current GS/FLX system provides 200 nucleotides flow cycles giving an average read length of 800 bp during a 7 hr run. These raw signals are processed by 454 pyrosequencing analysis software and then screened by various quality filters to remove poor quality sequences (Mardis, 2008a) resulting in a combined throughput of 100Mb of high quality sequence data. After the processing of the FLX sequences, they are assembled using the assembly algorithm (Neobler). The raw base accuracy reported by Roche is over 99%. Roche 454 Genome Sequencers are currently available in two versions: GS FLX+ system (1 Mb sequence read capacity) and the recently introduced GS Junior system (100 kb sequence read capacity).

Illumina Genome Analyzer

Genome Analyzer introduced by Illumina (formerly known as Solexa) in 2007, currently dominates the NGS market. Illumina platform is based on the concept of 'sequencing by synthesis' (SBS) coupled with bridge amplification on the surface of a flow cell. Single stranded adaptor ligated DNA fragments are attached to the solid surface known as single-molecule array or flow cell using a micro fluidic cluster station (Fig.2.). Each flow cell is an eight channel sealed glass micro fabricated device with their interior having covalently attached oligos complementary to the specific adaptors ligated onto the library fragments. The DNA fragments are hybridized to the oligos using active heating and cooling steps followed by subsequent incubation with the amplification reagents and an isothermal polymerase that result in the generation of discrete areas or clusters of the library fragments.

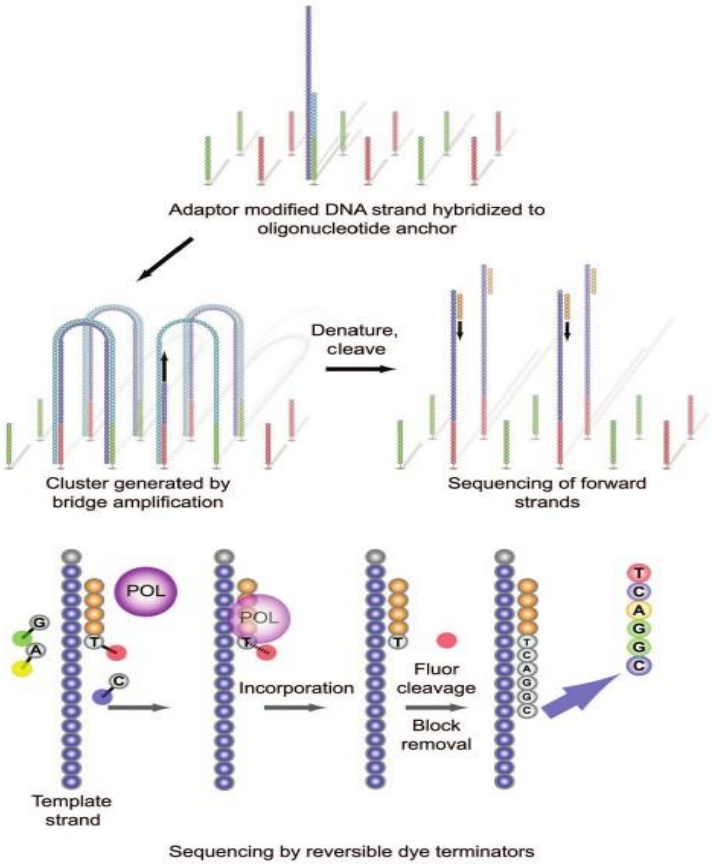

Fig. 2. Illumina Genome Analyzer sequencing
(Source- Karl *et al.*, 2009)

The flow cell is then placed in the fluidics cassette within the sequencer and each cluster is supplied with all four reversible terminators (modified nucleotides) with removable fluorescent moieties and special DNA polymerase that is capable of incorporating the terminators into the growing oligonucleotide chains. Terminators are differentially labeled fluorescent nucleotides with their 3' OH chemically blocked and this modification of 3'OH ensures that only a single base is incorporated per cycle. Each nucleotide incorporation step is then followed by an imaging step to identify the incorporated nucleotides on each cluster following which a chemical treatment cleaves the fluorescent group and de-blocks the 3' end preparing each strand for incorporation of next base in the next flow cycle. This series of steps is continued for a specific number of cycles as determined by user-defined instrument settings, generally is a read length of 25-35 bases. A typical Illumina genome analyzer yields ~35bp reads producing at least 1 GB of sequence per run of 2-3 days with raw base accuracy greater than 99.5%. Illumina approach is incapable of resolving short sequence repeats inspite of being more effective at sequencing homopolymeric stretches than pyrosequencing (Bentley, 2006). Due to the use of modified DNA polymerases and reversible terminators, substitution errors are the most common error types noted in Illumina sequencing data with higher proportion of errors occurring when 'G base' is the previous incorporated nucleotide (Dohm *et al.,* 2008). Additionally, an under representation of AT rich (Dohm *et al.,* 2008; Harismendy *et al.,* 2009) and GC rich regions (Hillier *et al.,* 2008; Harismendy *et al.,* 2009) was revealed by the genome analysis of Illumina data and may probably be due to the amplification bias during the template preparation.

The Illumina Genome Analyzer is the most adaptable and easy to use sequencing platform and because of its superior data quality, proper read length and high capacity made it a system of choice for whole genome sequencing applications, including human and model organisms. At present four versions of illumina sequencers are present in the commercial market: the HiSeq 2000, HiSeq 1000 and Genome Analyzer IIx have sequencing outputs of upto 600, 300 and 95 GB, respectively. Recently introduced MiSeq platform is capable of generating up to 150 bp sequencing reads with a combined throughput of 1.5-2 Gb per run. In 2012, Illumina introduced HiSeq 2500 platform as an upgraded form of HiSeq 2000 and is capable of generating up to 120 Gb of data in 27 h resulting in the sequencing of the entire genome in 24 h (i.e. genome in a day).

SOLiD Analyzer

Applied Biosystems (Life technologies) introduced SOLiD (support oligonucleotide ligation detection) technology in 2007 and commercialized it as their NGS platform. SOLiD uses unique 'sequencing by ligation' (SBL) approach catalyzed by DNA ligase. The procedure used in this platform involves attaching the adaptor-ligated library templates to the 1 µm magnetic beads whose surfaces are covered with the oligos complementary to the SOLiD specific adaptor sequences and then amplifying each of the DNA-bead complexes by emulsion PCR(Fig. 3.). After amplification, beads are covalently attached to the surface of a chemically treated glass slide that is placed into a fluidics cassette within the sequencer. Initiation of ligation based sequencing is marked by the hybridization of a universal sequencing primer complementary to the SOLiD-specific adaptors ligated to the library templates amplified by emPCR following which the semi-degenerate 8-mer fluorescent oligos and the DNA ligase are added in an automated manner within the instrument. DNA ligase seals the phosphate backbone as soon as the matching 8-mer oligo hybridizes to the DNA fragment sequence adjacent to the attached universal sequencing primer at the 3' end. Ligation step is followed by an imaging step, in which a fluorescent readout identifies the ligated 8-mer oligo which corresponds to one of the 4 possible nucleotides. Subsequently, the linkage between the fifth and the sixth base of the ligated 8-mer is cleaved chemically to remove the fluorescent group enabling the subsequent round of ligation. The probe hybridization, ligation, imaging and cleavage cycle is repeated 10 times to yield ten color calls spaced in five base intervals following which the extended primer (synthesized fragment) is stripped from the bound templates by denaturation. The second round of sequencing starts with the hybridization of the n-

positioned universal primer and subsequent cycles of ligation mediated sequencing. This round resets the interrogation bases and the corresponding color calls one position to the left. The same process is repeated with n-2, n-3 and n-4 positioned universal primers. The fluorescence obtained from the five ligation rounds is then decoded with a two-base calling processing software to generate the color calls which are ordered into a linear sequence and aligned to a reference genome to decode the DNA sequence. The use of 'two base-encoded' probes enables extra quality check of reads accuracy in color calling and SNV calling. Additionally, the two base-encoding schemes enable the distinction between a sequencing error and a sequence polymorphism: an error would be detected in only one particular ligation reaction, whereas a polymorphism would be detected in both. In SOLiD system, two slides can be processed per instrument run, one slide receives sequencing reagents as the second is being imaged (Mardis, 2008b) and each slide can be divided to contain different libraries in four or eight quadrants. The read length for SOLiD Analyzer is between 25-35 bp with a combined throughput of 2-4 Gbp per sequencing run. Today, two versions of Applied Biosystems SOLiD sequencers are available, the 5500 systems and 5500 xl system with upto 100 and 250-Gb sequencing capacity respectively and a raw base accuracy of 99.94%.

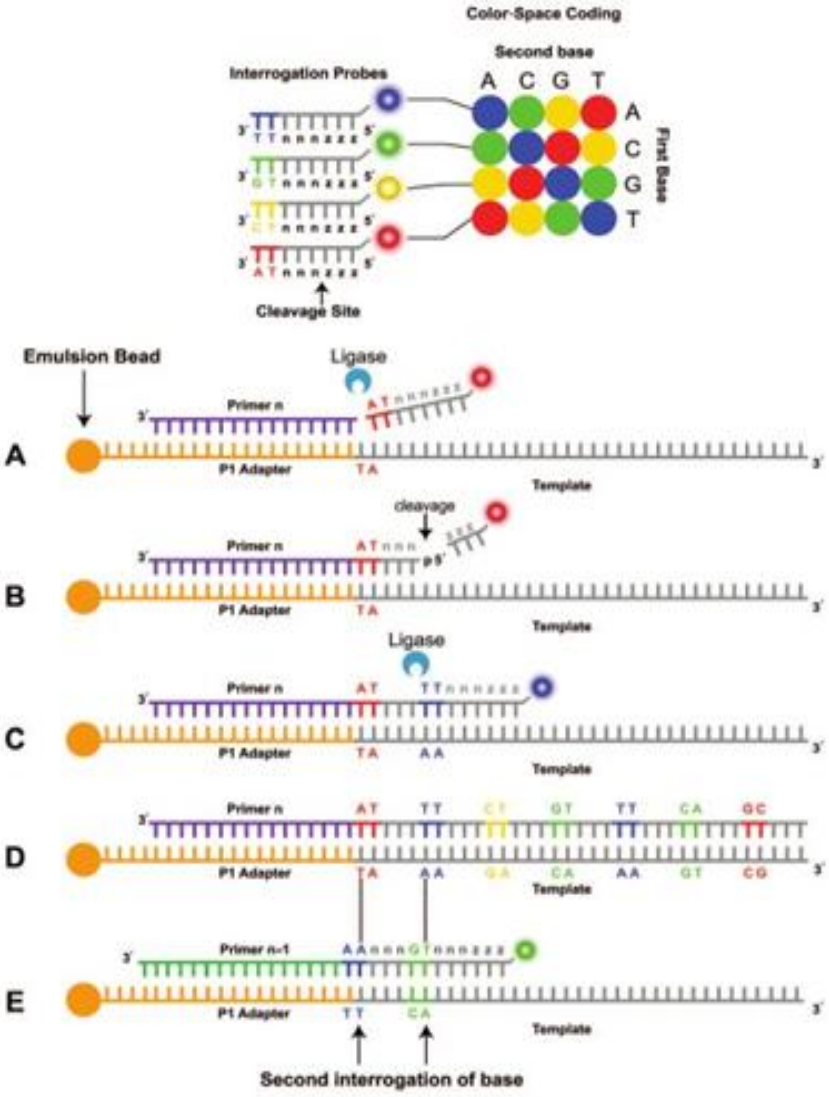

Fig.3. Applied Biosystems SOLiD sequencing by ligation (Source- Karl *et al.*, 2009)

Ion Torrent

This NGS platform can be considered as the world's smallest solid-state pH meter. Ion Personal Genomic Machine (Ion Torrent) which is introduced by Life technologies in 2010, which utilizes pH changes to detect base incorporation event. The system is based on the real time detection of hydrogen ions, by product of nucleotide incorporation into a growing DNA strand by DNA polymerase. Ion Torrent makes use of sequencing chips made up of high density array of microwells where each well acts as an individual DNA polymerization chamber containing a DNA polymerase and the sequencing template. Beneath the layer of microwells is an ion sensitive layer followed by a sublayer of highly dense field effect transistor (acting as ion sensor) (FET) array aligned with the array of microwells and pH change created during nucleotide incorporation is detected by the FET sensors which converts this signal to a recordable voltage change thereby revealing the primary sequence. The change in voltage is directly proportional to the number of nucleotides added at each step.

At present, Ion Torrent offers three different sequencing chips: Ion 314, Ion 316 and Ion 318. Ion 314 chips carry 1.2 million microwells generating roughly 10 Mb of sequence data with an average read length of 100 bases. The Ion 316 carries 6.2 million microwells generating 100 Mbp of sequence information with an average read length of 100 bases whereas the third generation sequencing chip Ion 318 is built with 11.1 million microwells to produce 1 Gb of sequencing data with an average read length of 200 bases. In 2012, life technologies introduced a further new generation of Ion semi conductor sequencers called the Ion Proton bench top sequencers which offer the reasonable price, bench top scale, high throughput sequencing and have the potential of deciphering the human genome or human exome in just few hours which 'democratized' sequencing methods. There will be two versions of Ion Proton chips: Ion Proton I chip built with 165 million wells (about 100 folds more than that Ion 314 chip) and ion Proton II chip having 660 million wells (about 1000 fold more than that of Ion 314 chip) which will be based on CMOS semiconductors technology to record chemistry changes instead of light and translate these changes into digital data. This newly introduced method of sequencing greatly reduces the sequencing cost but have several limitations as far as sequencing complete genome is considered. The first limitation is posed by short read lengths which limits the assembly of de novo sequencing projects as it is unable to read long repetitive regions in the genome. Secondly, due to sequential addition of nucleotides, error accumulation can occur if reaction wells are not properly pinged between reaction steps. Thirdly, in pyrosequencing, sequencing through smaller repetitive regions of the same nucleotide (monopolymeric) regions of 5 to 10 bases can be challenging. Currently, the short read lengths place a large burden on the reassembly process and limit the assembly of de novo sequencing projects due to an inability to read through long repetitive regions in the genome. Also, due to the sequential nature of this sequencing by synthesis method, error accumulation can occur if reaction wells are not properly purged between reaction steps. Finally, as for pyrosequencing in the previous generation, sequencing through smaller repetitive regions of the same nucleotide (homopolymer regions) on the order of 5 to 10 bases can prove challenging. Ion Torrent has reported sequencing accuracy data in which an *E. coli* DH10B sample was sequenced and homopolymer regions were analyzed (Ion Torrent Application Note, Spring, 2011). The sequencing accuracy for a 5-mer homopolymer region was shown to be around 97.5%; however, it was difficult to tell the size of the sample set from which these data were generated. Also, accuracy data for homopolymer lengths greater than 5 bases were not reported.

Single-molecule DNA sequencing platforms

SMRT DNA Sequencer

In 2010, Pacific Biosciences introduced a reliable third generation sequencing platform which is based on the single molecule real time (SMRT) DNA synthesis technology. The technology directly measures the fluorescence emitted by the cleavage of the phosphate chain of fluorescently labeled nucleotides incorporated by DNA polymerase onto a complementary sequencing template. Interestingly, this technology is a dense array of nanostructures called zero-mode waveguide (ZMW) which use electron

beam lithography and ultraviolet photolithography which allow optical interrogation of single fluorescent molecules. ZMW nanostructures are efficiently packed onto a surface. PacBio successfully develop a parallel confocal imaging system that revealed high sensitivity and resolution of fluorescent nucleotides in each of the ZMW nanostructures. The major technical problem with this technology after the development of ZMW array fabrication and detection scheme was to immobilize a single DNA polymerase molecule at the base of each ZMW which can incorporate fluorescently labeled nucleotides efficiently. As a first step is, a set of fluorescently labeled deoxyribonucleoside pentaphosphate (dN5Ps) substrates was synthesized to enable the spectrum differentiation of each base without decreasing the processivity of the DNA polymerase (Korlach *et al.,* 2008). In the second step, the surface of each ZMW nanostructure which was composed of a fused silicon bottom layer and aluminium top layer was chemically treated to allow the selective localization of the DNA polymerase. The derivation of the aluminium surface with polyvinyl phosphonic acid (PVPA) significantly decreased the protein adsorption to the aluminium layer without compromising protein adsorption to the bottom glass layer (Korlach *et al.*, 2008). The SMRT bell templates were generated by PacBio for SMRT sequencing technology which allows consecutive sequencing of both the sense and antisense strand of double stranded DNA fragment by ligating universal hairpin loops to the ends of the fragment. SMRT sequencing technology does not require any amplification step for the template preparation thus reducing the time needed for sample preparation. Additionally DNA fragments over a broad size range can be used to generate SMRT bell templates. The accuracy of sequencing and SNP detection also increases with the use of bell templates. Once the ZMW array fabrication, immobilization of the DNA polymerase and preparation of the SMRT bell templates is completed, by the action of DNA polymerase which is attached at the bottom of each waveguide the complementary DNA strand is synthesized from single stranded template. In this technology, the florescent label is attached to the terminal phosphate group rather than the nucleotide base, leading to the release of the different colored fluorescent moiety with nucleotide incorporation (Pushpendra, 2008; Flusberg *et al.,* 2010). The technology eradicates the need of the washing step between each nucleotide flows accelerating the speed of nucleotide incorporation and improving sequence quality. Additionally, the natural capacity of DNA polymerase to incorporate 10 or more nucleotides per second in several thousand parallel ZMWs (Eid *et al.,* 2009; Zhou *et al.,* 2010) is utilized in this approach. The robustness of the genetic data generated by the PacBio's single nucleotide sequencing arrays was improved by correlating polymerase kinetic data to DNA methylation pattern during DNA sequencing (Flusberg *et al.,* 2010). Additionally, to sequence mRNA strands using this technology, the DNA polymerase attached to the bottom of each ZMW can be replaced with a ribosome and the incorporation of fluorescently labeled tRNAs can be monitored.

Helicos Biosciences HeliScope

In 2008 Helicos Biosciences introduced 'The HeliScope' which was the first commercially available single molecule sequencing (SMS) platform that depend on highly sensitive fluorescence detection system to directly record each nucleotide as it is incorporated. The system utilizes sequencing by synthesis using one-color CRT method on a single DNA molecule template. In this method, template fragments ranges 100-200 bases in size were first attached to a substrate within a microfluidic flow cell. During sequencing, nucleotides bearing fluorescent dye are introduced one species at a time and incorporated by DNA polymerase to the growing complimentary strand. The fluorescent nucleotides have a capability to stop the polymerase extension until the incorporated nucleotides florescence is captured, recorded and analyzed to identify which nucleotide was incorporated into which growing strand with the help of highly sensitive CCD camera connected to the fluorescent microscope. Remaining unincorporated nucleotides and the byproducts of the previous cycles were then washed off following which the fluorescent labels on the extended strands are cleaved by chemical treatment and removed. Another cycle of addition of different species of nucleotide, labeling, cleaving and imaging are then follows (Ewing *et al.,* 1998; Harris *et al.,* 2008; Zhang *et al.,* 2011).

Harris and colleagues (2008) used Cy-5-1255-dNTPs, the earliest versions of their virtual terminators which lacks the inhibiting group and reported that when primer immobilization strategy was used to generate single molecule templates, the deletion errors in homolpolymeric repeat regions were found to be the most common error types. This may be due to incorporation of two or more Cy5-1255-dNTPs in a given cycle. These errors can be reduced to a great extent using two-pass sequencing, which gives ~25 base consensus reads using the template immobilized strategy. The read length obtained is ranges from 30-35 bp with 20-28 Gbp of potential sequence reads per run and a raw base accuracy greater than 9.

Nanopore Sequencing Technologies

Individual base detection is intended to be possible through the measurement of conductivity either across or through a membrane, via a nanoscale pore. These nanopores (4 nm) are slightly larger than the width of a double-stranded DNA molecule, where DNA is threaded through the pore. Nanopores could also be designed to measure tunneling current across the pore as bases, each with a distinct tunneling potential could be read. The nanopore approach remains an interesting potential fourth-generation technology. This "fourth-generation" name is suggested, since optical detection is eliminated along with the requirement for synchronous reagent wash steps. Nanopore technologies may be broadly characterized into two types, biological and solid-state. Protein alpha hemolysin was the first model of biological nanopore, which natively bridges cellular membrane causing lysis. In a biolayer membrane protein was inserted to separate two chambers while sensitive electronics measure the blockade current, which changes as DNA molecules moves through the pore.

The second model is based on the use of nanopores fabricated mechanically in silicon or other derivative. The use of these synthetic nanopores eases the difficulties of membrane stability and protein positioning that accompanies the biological nanopore system Oxford Nanopore has established. In transverse tunneling current scheme, electrodes are positioned at the pore opening and the signal is detected from subnanometer probes (Zwolak *et al.*, 2005).

In capacitance measurements, voltage is detected across a metal oxide-silicon layered structure. The voltage signal is induced in longitudinal direction across the capacitor by the passage of charged nucleotides. The optical recognition of nucleotides is basically performed in two steps. First, each base (A, C, G, or T) in the target sequence is converted into a sequence of oligonucleotides, which are then hybridized to two-color molecular beacons which are fluorophores attached (Soni *et al.*, 2007) Because the four nucleotides (A, C, G, or T) have to be determined, the two fluorescent probes are coupled in pairs to uniquely define each base. For example, for two probes A and B, the four unique permutations will be AA, AB, BA, and BB. As the hybridized DNA strand is threaded through the nanopore, the fluorescent tag is stripped off from its quencher and an optical signal is detected. In this both protein (Sauer *et al.*, 2003) and solid-state nanopores can be used (Mc Nally *et al.*, 2010).

Protein Nanopore Sequencing

Fundamentals of this sequencing are "lab on a chip" technology which integrates multiple electronic cartridges into a rack-like device. In this technology, a single protein nanopore is incorporated in a lipid bilayer across the top of a microwell which is equipped with electrodes. For sample preparation, detection and analysis, multiple microwells are incorporated onto an array chip, and each cartridge holds a single chip with integrated fluidics and electronics. After that a sample is introduced into the cartridge, which is then inserted in an instrument called a GridION node. All nodes communicate with each other and with the user's network and storage system in real time and each node can be used separately or in a cluster. Although the main application of this platform is sequencing of DNA, it can be adapted (by proper modification of the αHL nanopore) for the detection of proteins and small molecules.

Solid-State Nanopore Sequencing

Solid-State Nanopore sequencing also known as man-made nanopores and are considered to be next-generation nanopore technology. Solid-state materials (specifically a metal-dielectric layered structure) such as: silicon nitride, silicon or metal oxides, and more recently graphene are used and artificial pores are fabricated on them. Graphene is a new, single-atom thick material which is known to be the thinnest

possible membrane. This Solid state synthetic nanopore sequencing device comprises of 1–5 nm thick graphene membrane which is suspended in a Si chip coated with 5 µm SiO_2 layer.

Applications of the Current Sequencing Technologies

The continuous advancement of new sequencing technologies over the past few years has initiated a new beginning in the era of structural and functional biology. All of these technologies is based on different principle and has a distinct role and suitability as per the experiment and the organism to be sequenced. However one common thing in all these technologies is they produce data on a scale never imagined before and bypass the tedious cloning procedures. The cost and time required for sequencing has reduced. With the help of these technologies, no organism remains untouched just because of its genome size and complexity and that is a boon to solving complex biological problems whether in fields as varied as agriculture, environment or medicine (Morozova and Marra, 2008; Marguerat *et al.*, 2008; Deschamps and Campbell, 2010; Pareek *et al.*, 2011; Egan *et al.*, 2012). Some of the typical applications have addressed both RNA and DNA and are profiled below.

Whole Genome Sequencing

From past few decades gene sequencing has come a long way from sequencing partial genes to a set of genes, then large chunks of chromosomes and finally whole genome itself.

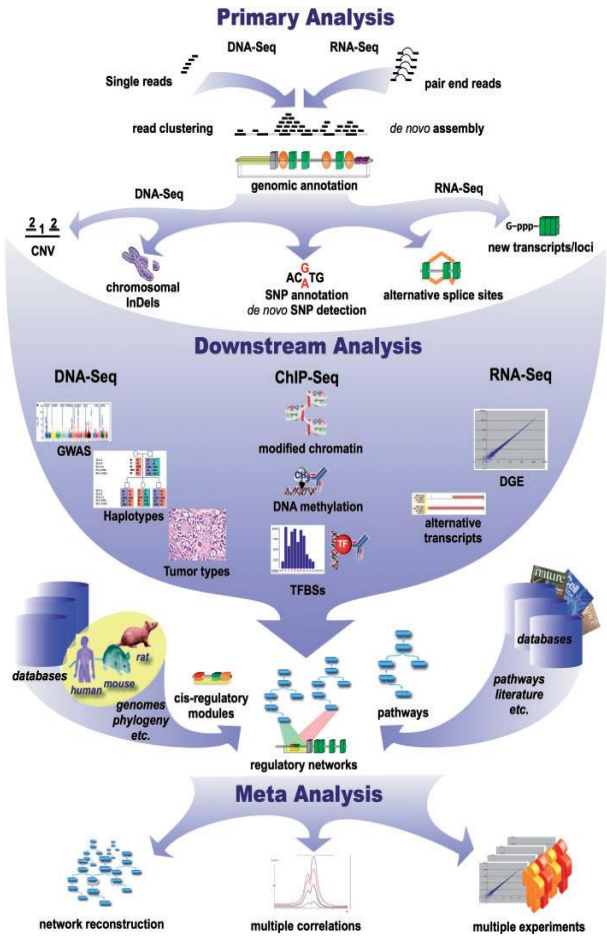

Fig. 4. Diagrammatic representation of the whole genome sequencing events (Source- Thomas Werner, 2010)

This is all because of periodic revolutions in the application of Sanger sequencing methodology. This was observed as the sequencing march initiated from the most simple to the most complex of the organisms. Each organism is complex in its own way. *Arabidopsis*, Human and rice genomes are sequenced and then quick advancements were made in respect to understanding of the gene function, cloning of major QTLs, identification of disease resistant loci etc. in many species. To sequence a polyploid species like wheat or sugarcane with the existing technologies as the genome size inflated the cost and time required for complete sequencing is still a elusive dream. With the introduction of commercial NGS technology in the early part of the last decade it has become clear that all the major organisms are essential to be sequenced in an effort to rapidly understand the organism's adaptation, evolution and thereby accelerate its improvement for the sake of the society. Now in the databases, the data is easily available and there are so many genome projects going on. Truly genome sequencing has now attained global proportions making inroads in sequencing newer organisms and with rapid advancements it is expected that all major species are sequenced before the end of this decade.

Targetted Genomic Resequencing

Sequencing of genomic subregions and gene sets is being used to identify polymorphisms and mutations in genes implicated in regions of the human genome and whole-genome association studies have implicated in disease (Yeager *et al.,* 2008; Ding *et al.,* 2008). Especially in the latter setting, regions of interest can be hundreds of kb's to several Mb in size. Overlapping long-range PCR amplicons (approximately 5–10 kb) can be used for up to several hundred kb's. More recently, enrichment has been achieved by hybridizing fragmented, denatured human genomic DNA to oligonucleotide capture probes complementary to the region of interest and subsequently eluting the enriched DNA (Albert *et al.,* 2007; Hodges *et al.,* 2007; Okou *et al.,* 2007; Porreca *et al.,* 2007). These captured probes can be immobilized on a solid surface. Current Nimble-Gen arrays contain 350 000 oligonucleotides of 60–90 bp in length that are typically spaced 5–20 nucleotides apart, with oligonucleotides complementary to repetitive regions being debarred. In reported studies, up to 5Mb of sequence has been captured on the 350K array, with 60%–75% of sequencing reads mapping to targeted regions; other reads mapping to nontargeted regions reflect nonspecific capture. Nimble Gen in development use an array of 2.1_106 features for capturing larger genomic regions. Agilent's solution-based technology uses oligonucleotides up to 170 bases in length, with each end containing sequences for universal PCR priming and with primer sites containing a restriction endonuclease– recognition sequence. The oligonucleotide library is amplified by the PCR, digested with restriction enzymes, and ligated to adapters containing the T7 polymerase promoter site. With the help of biotinylated UTP, *in vitro* transcription is performed to generate single stranded biotinylated cRNA capture sequences. An another upgrading approach developed by RainDance Technologies uses a novel microfluidics technology in which individual pairs of PCR primers for the genomic regions of interest are segregated in water in emulsion droplets and then assembled to create a "primer library". Distinctly, emulsion droplets are prepared which contain genomic DNA and PCR reagents. Two separate droplet streams are generated, one with genomic DNA/PCR reagents and other with primer-library droplets. The 2 streams are merged and are paired in a 1:1 ratio. Then these paired droplets proceed through the microfluidic channel, they pass an electrical impulse that causes them to physically merge. The merged droplets containing individual primer pairs and genomic DNA/PCR reagents are deposited in a 96-well plate and amplified by the PCR. After amplification, the emulsions are disrupted, and the amplicons are pooled and processed for NGS.

Metagenomics

NGS had an amazing influence on the study of microbial diversity in environmental and clinical samples. Operationally, genomic DNA is extracted from the sample of interest and then converted to an NGS library and then sequenced. The output sequence is aligned to known reference sequences for microorganisms that are expected to be present in the sample. de novo assembly of the data set can yield information to support the presence of known and potentially new species. Quantitative information of individual microbial species can be derived by the analysis of the relative abundance of the sequence

reads. These days, most NGS-based metagenomic analysis have used the Roche 454 technology and for de novo assembly of previously uncharacterized microbial genomes and it is associated with longer read lengths which facilitate alignment to microbial reference genomes. Examples of metagenomic studies comprise the analysis of microbial populations in the ocean (Huber *et al.,* 2007; Sogin *et al.,* 2006) and soil (Urich *et al.,* 2008), the identification of a novel are navirus in transplantation patients (Palacious *et al.,* 2008), and the characterization of microflora present in the human oral cavity (Keijser *et al.,* 2008) and the guts of obese and lean twins (Turnbaugh *et al.,* 2008).

Development of Markers

Human or rice genome sequencing led to a race for identification of important genes required for disease management in humans and both abiotic and biotic stress tolerance in plants. If reference genome is available, it becomes easy to genotype other members of the same species and then do comparative genomics. Molecular markers are basic part of any breeding program and availability of sequenced genomes provides a platform for marker discovery. Commonly used marker are SSRs for mapping purposes and also provide cross transferability within a species. Thus sequencing one member of a species can lead to development of markers for that species with wide usage. One of the most polymorphic and abundant marker for this cause are the SNPs which can be used to arrive at fine linkage maps. Once these maps are available then it can be used for high throughput genotyping, allele mining, association genetics and gene discovery.

Transcriptome Investigations

Before the introduction of NGS technologies, Transcriptome analysis was limited to the powers of the classical cDNA synthesis methods which led to the development of ESTs in diverse species and in different stages of development, environmental interactions, tissue etc. For many genomes, transcript information is available which has been useful in cloning of genes and promoters. EST databases are must for all type of functional genomics. But the limiting factor of Transcriptome analysis is the poor quality of the reads and low representation of the expressed genome. Then, NGS has provided a new powerful approach, termed "RNA-Seq," for mapping and quantifying transcripts in biological samples. A typical protocol of RNA- Seq involve the generation of first strand cDNA via random hexamer–primed reverse transcription and subsequent generation of second strand cDNA with RnaseH and DNA polymerase. The cDNA is then fragmented and ligated to NGS adapters. Reference genome is not required in RNA sequencing and hence in non-model organism it remains a tool of choice. RNA sequencing is slowly taking over other methods of Transcriptome analysis including microarrays, SAGE etc. Some of the advantages of RNA sequencing include identification of alternately spliced transcript and capture of low abundance messages. Thus, provides a snapshot of the expressed messages directly without any bias and truly identify the differentially expressed sequences and expressed SSRs and SNPs for the purpose of allele mining. NGS platforms also result in cloning of full length cDNAs which have a far greater significance in functional genomics. RNA-Seq has been applied to a variety of organisms, including *Saccharomyces cerevisiae, Arabidopsis thaliana*, mice, and human cell (Turnbaugh *et al.,* 2008; Nagalakshmi *et al.,* 2008; Wilhelm *et al.,* 2008; Mortazavi *et al.,* 2008; Lister *et al.,* 2008; Marioni *et al.,* 2008; Morin *et al.,* 2008; Morin *et al.,* 2008; Emrich *et al.,* 2007; Pan *et al.,* 2008; Wang *et al.,* 2008)

Epigenetic Analysis

Complex and dynamic nature of higher eukaryotes which subjected to many forms of regulation within and outside the cell. The cellular form of regulation is well defined and studied in different organisms and provides an vision into its functioning. However, the another form of regulation which is remains elusive and hidden from the naked eye as many times the manifestations are not seen in the phenotype which are the epigenetic variations arising out of many intrinsic and extrinsic factors. These factors can lead to methylation of the genome or modifications in the histone proteins which lead to a combined effect on the expression of the genome in both positive and negative ways. These variations known as epigenetic variations also keep the genome in a state of flux with heterochromatinization of DNA which may or may not be reversible. These variations are of prime importance in cancer biology and the Human Epigenome

Project (HEP) which aims to catalog all the DNA methylation that occurs in the human genome and find out its association with gene regulation during oncogenesis. In plants too, many genes are governed by these factors and the major effect are seen in the hybrids that arise out of contrasting genotypes. Such hybrids are reported to carry lot of epigenetic variations which could be responsible for their hybrid vigor. In contrast to classical ways of finding or mapping these variations, NGS has now the power to map all the methylation pattern in the genome and develop a epigenome map for different species. These maps then provide a tool for researchers to dissect the role of epigenetic variations like CpG methylation in the regulatory pathways that exist in a particular genotype.

Study of Protein-DNA Interactions

In higher eukaryotes chromosomal DNA is generally bound with histone proteins and condensed into chromatin which leads to chromosome packing and then remodeling to allow for expression of the genes. These studies have been carried out using Chip assays in combination with hybridization, PCR and lately microarrays. There are limitations related to presence of restriction sites and thereby narrow down the scope of analysis. The advent of NGS gave rise to a new procedure called Chip-Seq which sequences the immuno-precipitated DNA on a genomic scale. This technique has exposed the detail in histone modifications of the human genome. Other studies that came out of Chip Seq included mapping of nucleosome binding sites, transcription factor binding sites, ribosome binding sites and chromosome conformation capture (3C) method to detect higher orders of chromosome structure.

CONCLUSION

Rapid advances in the sequencing technologies and different chemistries will make next generation sequencing a major force in both structural and functional genomics. In comparison to Sanger method, the next generation techniques will provide a holistic attitude to researchers with the dual advantage of cost and time savings. There are certain issues with small read lengths but the accuracy and volume of data generated somewhat makes up for their limitations. For complex organisms such are plants, of which many are polyploids and having big genomes, resolving the repeat regions is still a challenge and there is need to improve upon the NGS techniques and the computational tools. Next Generation Sequencing is likely to provide highly valuable application(s) in genome sequencing, epigenome mapping, Transcriptome analysis and discovery and profiling of protein-DNA interaction. With the help of latest software with simpler algorithms it will be increasingly easier for researchers to approach the post sequencing data issues for expanding the scope of research and generate novel ideas.

REFERENCES

Albert TJ, Molla MN, Muzny DM, Nazareth L, Wheeler D and Song X (2007). Direct selection of human genomic loci by microarray hybridization. *Nature Methods* **4** 903–5.

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP and Milton J (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456** 53–59.

Bowers J, Mitchell J, Beer E, Buzby PR and Causey M (2009). *Virtual* terminator nucleotides for next-generation DNA sequencing. *Nature Methods* **6** 593–595.

Deschamps S and Campbell M (2010). Utilization of next-generation sequencing platforms in plant genomics and genetic variant discovery. *Molecular Breeding* **25** 553–570.

Ding L, Getz G, Wheeler DA, Mardis ER, McLellan MD and Cibulskis K (2008). Somatic mutations affect key pathways in lung adenocarcinoma. *Nature* **455** 1069 –75.

Dohm JC, Lottaz C, Borodina T and Himmelbauer H (2008). Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Research* **36** 105.

Dressman D, Yan H, Traverso G, Kinzler KW and Vogelstein B (2003). Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proceedings of the National Academy of Sciences of USA* **100** 8817–8822.

Egan AN, Schlueter J and Spooner DM (2012). Applications of next-generation sequencing in plant biology. *American Journal of Botany* **99** 175-85.
Eid J, Fehr A, Gray J, Luong K and Lyle J (2009). Real-time DNA sequencing from single polymerase molecules. *Science* **323** 133–138.
Emrich SJ, Barbazuk WB, Li L and Schnable PS (2007). Gene discovery and annotation using LCM-454 transcriptome sequencing. *Genome Research* **17** 69 –73.
Erlich Y, Mitra PP, DelaBastide M, McCombie WR and Hannon GJ (2008). Alta-Cyclic: a self-optimizing base caller for next-generation sequencing. *Nature Methods* **5** 679–682.
Ewing B, Hillier L, Wendl MC and Green P (1998). Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Research* **8** 175–185.
Fedurco M, Romieu A, Williams S, Lawrence I and Turcatti G (2006). BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. *Nucleic Acids Research* **34** 22.
Flusberg BA, Webster DR, Lee JH, Travers KJ and Olivares EC (2010). Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nature Methods* **7** 461–465.
Glenn TC (2011). Field guide to next-generation DNA sequencers. *Molecular Ecology Resources* **11** 759–769.
Guo J, Xu N, Li Z, Zhang S and Wu J (2008). Four-color DNA sequencing with 3′-*O*-modified nucleotide reversible terminators and chemically cleavable fluorescent dideoxynucleotides. *Proceedings of National Academy of Science of USA* **105** 9145–9150.
Harismendy O, Ng PC, Strausberg RL, Wang X and Stockwell TB (2009). Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biology* **10** R32.
Harris TD, Buzby PR, Babcock H, Beer E and Bowers J (2008). Single-molecule DNA sequencing of a viral genome. *Science* **320** 106–109.
Hillier LW, Marth GT, Quinlan AR, Dooling D and Fewell G (2008). Whole-genome sequencing and variant discovery in *C. elegans: Nature Methods* **5** 183–188.
Hodges E, Xuan Z, Balija V, Kramer M, Molla MN and Smith SW (2007). Genome-wide in situ exon capture for selective resequencing. *Nature Genetics* **39** 1522–7.
Huber JA, Mark Welch DB, Morrison HG, Huse SM, Neal PR, Butterfield DA and Sogin ML (2007). Microbial population structures in the deep marine biosphere. *Science* **318** 97–100.
Ju J, Ruan C, Fuller C, Glazer A and Mathies R (1995). Fluorescence energy transfer dye-labeled primers for DNA sequencing and analysis. *Proceedings of National Academy of Science of USA* **92** 4347–4351.
Karl V, Voelkerding, Shale A, Dames and Jacob D (2009). Next-Generation Sequencing: From Basic Research to Diagnostics. *Durtschi Clinical Chemistry* **55**(4) 641–658.
Keijser BJ, Zaura E, Huse SM, Van der Vossen JM, Schuren FH, Montijn RC (2008). Pyrosequencing analysis of the oral microflora of healthy adults. *Journal of Dental Research* **87** 1016 – 20.
Korlach J, Bibillo A, Wegener J, Peluso P, Pham TT, Park I, Clark S, Otto GA and Turner SW (2008). Selective aluminum passivation for targeted immobilization of single DNA polymerase molecules in zero-mode waveguide nanostructures. *Proceedings of National Academy of Science of USA* **105** 1176–1181.
Landegren U, Kaiser R, Sanders J and Hood L (1988). A ligase-mediated gene detection technique. *Science* **241** 1077–1080.
Lander ES (2011). Initial impact of the sequencing of the human genome. *Nature* **470** 187–197.
Leamon JH (2003). A massively parallel Pico Titer Plate based platform for discrete picoliter-scale polymerase chain reactions. *Electrophoresis* **24** 3769–3777.
Levy S, Sutton G, Ng Feuk and Halpern ALPC (2007). The diploid genome sequence of an individual human. *PLOS Biology* **5** e254.

Lister RO, Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH and Ecker JR (2008). Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell* **133** 523–36.
Mardis ER (2008a). The impact of next-generation sequencing technology on genetics. *Trends in Genetics* **24** 133–141.
Mardis ER (2008b). Next-generation DNA sequencing methods. *Annual Review of Genomics and Human Genetics* **9** 387–402.
Margulies M, Egholm M, Altman WE, Attiya S and Bader JS (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437** 376–380.
Marioni JC, Mason CE, Mane SM, Stephens M and Gilad Y (2008). RNA-Seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research* **18** 1509 – 17.
Maxam AM and Gilbert W (1977). A new method for sequencing DNA. *Proceedings of National of Academy of Science of USA* **74** 560–564.
McNally B (2010). Optical recognition of converted DNA nucleotides for single-molecule DNA sequencing using nanopore arrays. *Nano Letters* **10**(6) 2237–2244.
Metzker ML (2009). Sequencing in real time. *Nature Biotechnology* **27** 150–151.
Morin R, Bainbridge M, Fejes A, Hirst M, Krzywinski M and Pugh T (2008). Profiling the HeLa S3transcriptome using randomly primed cDNA and massively parallel short-read sequencing. *Biotechniques* **45** 81–94.
Morin RD, O'Connor MD, Griffith M, Kuchenbauer F, Delaney A and Prabhu AL (2008). Application of massively parallel sequencing to micro RNA profiling and discovery in human embryonic stem cells. *Genome Research* **18** 610 –21.
Morozova O and Marra MA (2008). Applications of next-generation sequencing technologies in functional genomics. *Genomics* **92** 255-264.
Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M and Snyder M (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320** 1344 –9.
Okou DT, Steinberg KM, Middle C, Cutler DJ, Albert TJ and Zwick ME (2007). Microarray-based genomic selection for high-throughput resequencing. *Nature Methods* **4** 907–9.
Palacios G, Druce J, Du L, Tran T, Birch C and Briese T (2008). A new arenavirus in a cluster of fatal transplant-associated diseases. *New England Journal of Medicine* **358** 991– 8.
Pan Q, Shai O, Lee LJ, Frey BJ and Blencowe BJ (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics* **40** 1413–5.
Pareek CS, Smoczynski R and Tretyn A (2011). Sequencing technologies and genome sequencing. *Journal of Applied Genetics* **52** 413-35.
Porreca GJ, Zhang K, Li JB, Xie B, Austin D and Vassallo SL (2007). Multiplex amplification of large sets of human exons. *Nature Methods* **4** 931–6.
Pushpendra KG (2008). Single-molecule DNA sequencing technologies for future genomics research. *Trends in Biotechnology* **26** 602–611.
Sanger F and Coulson AR (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology* **94** 441–448.
Sauer-Budge AF (2003). *Physical Review Letters* **90**(23) 238101/1–238101/4.
Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, Wang MD, Zhang K, Mitra RD and Church GM (2005). Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309** 1728–1732.
Sogin ML, Morrison HG, Huber JA, Mark Welch D, Huse SM and Neal PR (2006). Microbial diversity in the deep sea and the underexplored "rare biosphere." *Proceedings of National Academy of Science of USA* **103** 12115–20.
Soni GV and Meller A (2007). *Clinical Chemistry* **53**(11) 1996–2001.
Thomas Werner (2010). Next generation sequencing in functional genomics. *Briefings in Bioinformatics* **11**(5) 499- 511.

Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A and Ley RE (2008). A core gut micro biome in obese and lean twins. *Nature* **457** 480–4.

Urich T, Lanzen A, Qi J, Huson DH, Schleper C and Schuster SC (2008). Simultaneous assessment of soil microbial community structure and function through analysis of the meta-transcriptome. *PLOS ONE* **3** e2527.

Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L and Mayr C (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature* **456** 470–6.

Wilhelm BT, Marguerat S, Watt S, Schubert F, Wood V and Goodhead I (2008). Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature* **453** 1239 – 43.

Wu W, Litosh VA, Stupi BP, Metzker ML (2009). Photocleavable labeled nucleotides and nucleosides and methods for their use in DNA sequencing. *US Patent* **11**/567 189.

Yeager M, Xiao N, Hayes RB, Bouffard P, Desany B and Burdett L (2008). Comprehensive resequence analysis of a 136 kb region of human chromosome 8q24 associated with prostate and colon cancers. *Human Genetics* **124** 161–70.

Zhang J, Chiodini R, Badr A and Zhang G (2011). The impact of next-generation sequencing on genomics. *Journal of Genetics and Genomics* **38** 95–109.

Zhou XG, Ren LF, Li YT, Zhang M, Yu YD and Yu J (2010). The next-generation sequencing technology: a technology review and future perspective. *Science China Life Sciences* **53** 44–57.

Zwolak M and Di Ventra M (2005). Electronic signature of DNA nucleotides via transverse transport. *Nano Letters* **5**(3) 421–424.