International Journal of Advanced Biotechnology and Research ISSN 0976-2612,
Vol 1, Issue 2, Dec-2010, pp 120-129 http://www.bipublication.com

ROLE OF HIGH-THROUGHPUT SEQUENCING TECHNOLOGIES IN GENOME SEQUENCING

K.V. Chaitanya, Akbar Alikhan. P, V. Prasanth Reddy, Rishabh Lakhtakia and M. Taraka Ramji
Department of Biotechnology, GITAM Institute of Technology, GITAM University,
Visakhapatnam- 530045

ABSTRACT
The modern biology has undergone a drastic change with the increasingly available genetic information of many organisms from prokaryotes to human beings. This was possible due to the invention of novel DNA sequencing technologies called high through-put sequencing techniques (HTS technologies), which are capable of sequencing the genetic information with increased speed, accuracy and efficiency at a lower cost. Development of these technologies has revolutionalized the traditional Sanger sequencing method and has almost made the sequencing a bench-top instrument. HTS technologies are also used in solving the genome complications, for studying the diversity and genetic variations. These technologies are believed to be useful in novel medical diagnostics and treatment. This review will focus on using different HTS technologies for solving the complexities in the genomes of fossil, prokaryotic and eukaryotic organisms.

Keywords:- Drosophila, Fungi, fossils, genomics, high-throughput sequencing technologies, human, microorganisms, plants.

INTRODUCTION:-

Genome comprises the entire genetic information of an organism encoded either in DNA or RNA consisting of both coding and non-coding regions [1]. Genome constitutes double helical DNA in higher organisms, single circular chains of DNA in bacteria, viruses and in organelle like chloroplast, mitochondria, linear chains of RNA in some viruses and transposable elements. In eukaryotes the entire genome is packed in copies of chromosomes and the copy number varies from two in diploids to four in tetraploids [2]. Study of genome will not only yield the information regarding the total number of genes in an organism, but also about the mechanisms that could have led to the production of great variety of genomes that exists to day by comparing different genomes for their size, codon usage bias, GC content, repeats (STR), duplication of genes etc. The variation in the genetic information especially to the traits of diseases requires comparisons between individuals which makes the genome more complex in the context of biology [3].

The term sequencing refers to determine the order of the nucleotides in the DNA sequencing and amino acids in protein sequencing. Genome sequencing is a technique that determines the complete DNA sequence of an organism's genome at a time which includes the chromosomal DNA, mitochondrial DNA and in the case of plants, chloroplast DNA also [4]. Genome sequencing has created a revolution in the biology research by further opening the doors for studying the molecular processes involved in the complete cellular systems, leading to the concept of systems biology. Genome sequencing has also laid the foundation to the 'omics' technologies such as proteomics and transcriptomics [5]. All this was possible because of the availability of advanced nucleic acid sequencing techniques. The present review focus on the role of different sequencing techniques involved in the genome sequence of different organisms. Much emphasis was given to the contribution of high-throughput

sequencers in decoding the genomes. Attempt was also made to study the role of high-throughput sequencers in understanding the genetic variation and their relationship to biological function.

DNA Sequencing Technologies:-

The first DNA sequences were obtained in early 1970s using laborious methods based on 2-dimensional chromatography. Following the development of dye-based sequencing methods with automated analysis [6], DNA sequencing has become easier and faster. One of the earliest ways of nucleotide sequencing is RNA sequencing. The major landmark of RNA sequencing is the sequence of bacteriophage MS2 complete genome [7]. The development of rapid methods by Sanger, Gilbert and Maxam became the method of choice for the DNA sequencing.

One common challenge faced in all sequences is the poor quality during first 15-40 bases of sequencing and rapid decline in quality after 700 bases, and limitation in size (300-1000 bases) has hampered the quality of sequencing as well as time. Automated DNA sequencers, which can sequence up to 384 DNA samples in a single batch supported by number of software programs, can reduce the low-quality DNA sequencing. These programs score the quality of each peak and remove low-quality base peaks [5]. Invention of large scale sequencers enabled the sequencing of large fragments including whole chromosome, but the assembly of sequence information is complex and difficult, particularly with sequence repeats often causing gaps in genome assembly.

High-throughput sequencing:-

The demand for the invention of high quality sequencers which can sequence the large fragments of the genomes efficiently with low cost has led to the development of high-throughput sequencing technologies that parallelize the sequencing process, producing thousands or millions of sequences at one hit with lower cost beyond what is possible with standard methods [8].

In vitro clonal amplification

This technique uses *In vitro* cloning step for amplification of individual DNA molecules, which can be isolated by Emulsion PCR. Emulsion PCR isolates individual DNA molecules along with primer-coated beads in aqueous droplets within an oil phase. PCR then coats each bead with clonal copies of the DNA molecule followed by immobilization for sequencing. Emulsion PCR is an important part of Polony sequencing, which provides clonal amplifications of a single DNA molecule, grown in a gel matrix [9] and SOLiD (Sequencing by Oligonucleotide Ligation and Detection) capable of generating hundreds of millions to billions of 50 base reads at one time. In *bridge PCR*, fragments are amplified upon primers attached to a solid surface, used in the Illumina Genome Analyzer. Pyrosequencing uses DNA polymerization, adding one nucleotide species at a time, detecting and quantifying the number of nucleotides added to a given location through the light emitted by the release of attached pyrophosphates is capable of generating millions of 200-400 bases reads [10]. Solexa sequencing system can generate hundreds of millions of 50-100 base reads [11]. These methods have reduced the cost from $0.01/base in 2004 to nearly $0.0001/base in 2006 and increased the sequencing capacity from 1,000,000 bases/machine/day in 2004 to more than 5,000,000,000 bases/machine/day in 2009 [12].

Parallelized sequencing

In this process, DNA molecules are bound to a solid surface, used as a template and sequenced parallely by synthesis using DNA polymerase, which determines the base. Reversible

terminator methods uses reversible versions of dye-terminators, by adding one nucleotide at a time, detect fluorescence emission at each position of the base in real time, by repeated removal of the blocking group to allow polymerization of another nucleotide [13]. Sequencing by hybridization uses a single pool of DNA whose sequence is to be determined is fluorescently labeled and hybridized to an array of known sequences. Strong hybridization signals received from a given spot on the array identifies its sequence in the DNA being sequenced [14]). Mass differences between the DNA fragments can also be used to determine the sequence of the fragments using Mass spectrometry [15]). Other sequencing approaches are in the process to label the DNA polymerase and Electron microscopy to identify the position of nucleotides in the chain by labeling them with heavy metals or halogens

Role of HTS technologies in sequencing the microbial genomes :-

Ever since the first microbial genome of *Haemophilus influenza* has been decoded [16] comprising full genetic complement with 1830137·bp of DNA and 1743 predicted genes, numerous microbial genomes were sequenced including *Mycobacterium tuberculosis* [17], one of the most important human bacterial pathogens, *E.coli* [18], malarial parasite *Plasmodium falciparum* [19, 20] and yeast [21]. So far, approximately 300 complete bacterial genomes have been sequenced. In the case of pathogenic microbes, multiple species have been sequenced whose genetic information of each new strain revealed the discovery of new genes. Sequencing of B *Streptococcus* strains eight genomes led to the discovery of 33 new genes from each genome [22], giving rise to a concept of Pan-genome, which focuses on the importance of sequencing the genomes of different strains or species belonging to the same genus whose complete knowledge of genetic complementation will increase the scope of drug design [23].

The major step in sequencing is to amplify the DNA, which was done traditionally by cloning and transformation into *E.coli*, might not give good results in the case of microorganisms that yield toxic compounds. Margulis et.al., method also known as 454 sequencing method is capable of sequencing 25 million bases in 4hours, in which the DNA is amplified using a clonal approach and sequenced using a microfabricated massively parallel platform. Adapters are ligated to the sheared DNA fragments of approximately 300 bases and these tiny DNA fragments are captured on beads of 30 mm in diameter. The reactions are adjusted in such a way that only one DNA fragment will be captured by a bead. Subsequently DNA captured in these beads will be amplified at a rate of 10 million copies of the initial fragment. The beads are then dispensed into open wells of fiber optic slide and pyrosequenced, which detects the luciferase emission of the pyrophosphate release using a realtime monitor. This generates a sequence of 100 bp in length. Sequencing of *Mycoplasma genitalium* genome using this system yielded 96% of the genome coverage with an accuracy of 99.96% was obtained in just 4h.

Another technique for the High throughput sequencing was reported by Shendure et.al., [24]. This approach differs from the Margulie's method with respect to sequencing chemistry and signal detection, which employs an epifluorescence microscope and an array platform. Here, the single DNA molecule is grown on a solid phase to dive rise to polonies by using clonal amplifications. DNA library containing approximately 1.6 million fragments each approximately 135 bases in length are sequenced using 17-18 bp sequence tags derived from the genome. Each fragment was captured on a bead and amplified using emulsion PCR and immobilized on an acrylamide gel. Parallel sequencing is carried out using a four dye

ligation protocol for the identification of each base. For each fragment 26 bp sequence was determined. An *E. coli* strain MG1655 engineered for tryptophan biosynthesis was re-sequenced using this approach with an error rate of one per million bases.

Another approach called single molecule DNA sequencing was used to resequence the M-13 phage genome [25]. The library construction process is simple and fast and does not require PCR, resulting a single stranded, poly (dA)-tailed templates. Poly (dT) oligonucleotides are covalently anchored to glass cover slips at random positions. These oligomers are first used to capture the template strands, and then either as a primer for the template-directed primer extension that forms the basis of the sequence reading or, optionally, for a template replication step before sequencing. Up to 224 sequencing cycles were performed; each cycle consisting of adding the polymerase and labeled nucleotide mixture containing one of the four bases, rinsing, imaging multiple positions, and cleaving the dye labels. This sequencing process was performed simultaneously on more than 280,000 primer-template duplexes. Single-molecule method also enabled to re-sequence each individual template *in situ*, which greatly reduced the ensemble error rate.

Screening for larger mutations and SNPs is one of the best applications of the high throughput sequencing of microbial genomes. Currently, SNPs screening is being done by microarrays. In an approach, polony sequencing was used to screen an auxotroph of *E. coli*. The sequencing result has identified a number of deletions and inversions as well as SNPs [24]. Even though the data obtained per clone was around 26 bases, identification of rearrangements and SNPs in the genome was possible. In a similar study on bacterium *Myxococcus xanthus,* a laboratory-evolved strain that had been selected for a cheating phenotype and re-selected for a cooperative phenotype was shotgun sequenced using 454 sequencing technology. The 454 sequence was able to identify point mutations in the evolved strain compared with the reference strain, which could then be associated with the changes in phenotype as well as identifying errors in the reference [26].

Plant genomes:-

The complexity and size of the plant genomes have limited the ability to obtain their comprehensive genetic information. After the sequence of *Arabidopsis thaliana* genome, (Arabidopsis Genome Initiative, 2000), the complete rice genome (Oryza sativa; International Rice Genome Sequencing Project, 2005), poplar genome (*Populus trichocarpa*) have also been completed and, a draft sequence of the 2,300-Mb maize genome was released [27]). In addition, the National Center for Biotechnology Information Entrez Genome Projects website reports that sequencing of several more plant genomes is in progress. The first wave of plant genome sequencing has passed, in which all these genomes were sequenced using the traditional approaches, constructing the sequence libraries from individual segments of the genome and are sequenced by Sanger's method. A whole-genome shotgun (WGS) strategy, with improved assembly algorithms, has been used for several recent plant genomes, in which the sequencing libraries are made directly from genomic DNA [28-30].

We are now entering a new era in plant genomics research. Large and complex plant genomes such as sugarcane, wheat and barley are difficult to decode because of historic duplication events, the amplification of families of transposable elements, and polyploidisation conditions. The development of next-generation DNA sequencing technologies has revolutionized plant genome research and stimulated the analysis and sequencing of large plant genomes. These novel DNA sequencing

technologies provide ultra-high throughput at a substantially lower cost with huge increases in sequencing throughput and, perhaps more importantly, the ability to avoid the handling of individual clones from shotgun libraries. There are currently four commercially available high throughput technologies, which are used in sequencing the plant genomes. 454 Life Sciences (acquired by Roche), Solexa (acquired by Illumina), ABI SOLID (acquired from Agencourt Biosciences), and Helicos Biosciences. These technologies can be grouped in to two classes based on the lengths of the sequence reads produced. Solexa, ABI SOLID, and Helicos all produce very short reads in very large quantities, while the 454 platform can produce a more moderate amount of sequence, but with much longer read lengths [31].

The data produced through Solexa, ABI SOLID, and Helicos have very short read-length sequences, making de novo assembly extremely challenging. Development of novel methods for sequencing assemblies of large genomes from short-read sequences have solved this problem and successfully assembled the cucumber genome of size 367Mb [32] and the genome sequences of Chinese cabbage (500Mb) potato (830Mb) are under way. The development of these assembly methods creates new opportunities for building reference sequences and carrying out accurate analyses of unexplored genomes in a cost effective way. Approaches are being made for the identification of genes and promoters by genome analysis of key species and wild crops, which reduces the complexity of genome assembly to establish reference genomes for the major crops. Genes and markers with important agricultural implications can be discovered and further applied in breeding programs for crop improvement. The list of plant genomes sequenced was displayed in Table 1.

High throughput DNA sequencing technology is capable of identifying new micro RNAs (miRNAs) in plant species. Several miRNA genes in the plant kingdom are ancient, with conservation extending between angiosperms and the mosses, whereas many others are more recently evolved. Solexa sequencing approach was used to identify eight new micro RNAs in one of the model legume species, *Medicago truncatula* [33]. Using deep sequencing and computational methods, 48 non-conserved miRNA families, nearly all of which were represented by single genes, were identified in *Arabidopsis thaliana* [34].

With increasing information on plant genomes, comparative analyses of genomes have become an important aspect in identification of genes and their function. Genome comparisons are more useful in the understanding of plant biology and evolution specifically at the intra kingdom levels, but the plant species for which genome sequences are available span only 200 million years of land plant evolution. Sequencing of the pteridophytes genome like *Selaginella moellendorffii* and moss genome *Physcometriella patens* which is in the wedge of completion [35], will add another 200 million years of evolutionary history to comparative plant genomics.

Fungal genomes:-

Blumeria graminis is one of the most significant pathogens of cereal crops, which causes the powdery mildew and can reduce crop yields by as much as 40%. In order to support the developments regarding the crop protection and disease resistance, the genome of this pathogen has been fully sequenced and is annotated [36]. One of the basic experimental challenges posed by *Blumeria graminis* is that it can only be grown on its host; thus, the supply of biological material is very limited and may be contaminated by host tissues. The advent of high-throughput DNA sequencing platforms has revolutionized the depth at which transcriptomes can be analyzed, and the development of robust and efficient protocols for generating cDNA that

can be introduced directly in the sequencing pipeline is of huge importance [37]. One of the requirements of genome annotation is a collection of full-length cDNA sequences from as many diverse stages of the organism as possible to use in *ab initio* gene discovery by programs such as EuGene and FGENESH.

Full length cDNA was synthesized and was sequenced using 454 pyrosequencing (one run on GS-FLX), which yielded 247,306 reads, comprising a total of 50.8 M bases corresponding to an average read length of 205 bases. The data were assembled (using MIRA;www.chevreux.org/projects_mira.html), clustered, and combined with ESTs available in public repositories. The number of unique *B. graminis* genes identified by cDNA sequencing was increased from 4,584 to 7,727. When the cDNA sequences were compared to genomic DNA, it became evident that there was a marked heterogeneity in the RNA populations of *B. graminis*. Some bases were added in the transcript at a considerable distance from the beginning of the mature transcript. The majority of the additions include adenosines, but thymine, cytosine, and guanosine were also found. Similar observations were reported in the sequencing of other related fungi *Magnaporthe oryzae* (syn. *grisea*) and other eukaryotic organisms [38].

Several high-throughput techniques are available for sequencing of genes in a genome, but their function can be inferred only on the basis of sequence motifs or sequence similarity. The immediate challenge after the collection of genome sequences is to investigate how a genome and interactome (interactions of protein-protein or protein-DNA) determines the phenome of an organism [39]. To study these interactions, technologies for genome-wide analysis of gene expression such as microarray hybridization are used. However, the complexity of higher eukaryotic genomes makes interactome analysis more complicated and difficult. One approach to counteract these difficulties is synthetic genetic array (SGA) analysis [40]. Out of the 6,000 genes of yeast genome, 5,000 genes have shown to be non-essential in a genome-wide single-gene-knockout project, but the double mutants of these non-essential genes produced lethal phenotypes. SGA analysis allows the identification of genetic interactions, because if a double mutant has a synthetic lethal phenotype the two corresponding wild-type genes often have a functional relationship. Among the 5,000 non-essential genes, the function of 132 genes were tested by making double mutants with each other, it was determined that each gene has an average of 30 synthetic genetic interactions and was hypothesized that there can be 100,000 such interactions in the yeast genetic network [41]. Using cluster analysis of SGA results, the function of an unknown gene can also be predicted on the basis of the genes with which it is connected in the SGA network. SGA analysis gives a much more complex network of the yeast interactome than previously reported.

Structural variations in Drosophila genome:-

The complete genome sequences of 10 species of Drosophila have been published in 2007, to compliment with already published *Drosophila melanogaster* [43] and *D. pseudoobscura* [43]. One common thing in comparative genomes of these 12 drosophila species is structural variation. Copy number variation (CNV), is a type of structural variation includes deletions, duplications, insertions, and genomic rearrangements which will affect the number of occurrences of a specific DNA sequence present in the genome [44]. CNV is known to occur extensively in the Drosophila genome with functionally significant consequences [45]. Until recently, comparative genomic hybridization with whole-genome tiling arrays (array-CGH) was the primary method for characterizing CNVs [46]. However, several limitations for this

platform reduce its efficacy and efficiency and make this a costly affair. High throughput sequencing provides an ideal and cost-effective platform for CNV characterization by overcoming the inherent limitations of cross-hybridization and provides a digital count of sequence representation without prior knowledge or design work. Using HTS, deletions in three deficiency fly stocks were successfully characterized and the associated breakpoints were accurately determined using the Illumina genome analyzer [47]. Sequence reads obtained were mapped to the Drosophila reference genome release 5.1 using the vendor provided Eland pipeline.

Human Genome:-

High-throughput sequencing (HTS) technologies such as Illumina/Solexa and AB SOLiD, have enabled the sequence a full human genome considerably at a lower cost of at least 200-fold less than the regular methods [48]. Several other sequencing technologies are currently under development and are in the early stages of commercialization [49].

Selective genome sequencing

Re-sequencing of genomes is one of the regularly used approaches for obtaining the contrasting results during biomedical research and clinical practice, for which the underlying cost remains prohibitive [50]. To reduce the cost, in most of the genome experiments specific portions of the genome will be sequenced. Multiplex PCR is one of the regularly used approach for sequencing the specific portions [51]. However, the specificity of multiplex PCR is relatively poor when >10-20 amplicons are simultaneously amplified. Another approach is long range PCR with which a continuous genomic region up to ~40kb can be specifically amplified [52]). However, broad application of this technology is restricted largely due to its low amplification efficiency and inability to multiplex [53].

Several innovative technologies like Selector [54], Gene-Collector [55] MegaPlex PCR [56], Multiplex Exon Capture technique [57] and Sequence-capture Array [58] have recently been developed for selective genomic sequencing. The first three (Selector, Gene-Collector and MegaPlex PCR) are suitable for selecting genomic loci on the order of a few hundreds. In contrast, the Multiplex Exon Capture technology and Sequence-capture Array strategy are more suitable for applications at the genome scale because tens of thousands of target regions can be simultaneously captured. As far as the performance is concerned, Sequence-Capture Array (65-77%) tend to be less specific than medium-throughput strategies (in general >90%) because of increase in the assay complexity with an exception of Gene-Collector strategy which has a specificity of only 58%. Strikingly, the Sequence-capture Array strategy showed the best, enrichment distribution up to 90% bases within a 5-fold range due to the simplicity of the overall procedure with minimum amplification steps involved. On the other hand, low uniformity was observed for the Multiplex Exon Capture approach, which is reflected in its high dropout rate of 72% ([59].

Sequencing of degraded and ancient fossil DNA

The ancient DNA isolated from the fossils is always less and is heavily fragmented, chemically modified and contaminated with environmental DNA making the sequencing difficult [60]. Development of HTS technologies have bought solutions to some of these problems, making a significant contribution to the emerging field of paleogenomics [61], enabling the sequencing of the genomes from extinct organisms [62]. Recently, high-throughput sequencing techniques have been applied to sequence the fossil mitochondrial DNA (mtDNA) genome [63]. Using 454 sequencing, mtDNA genome has been recovered

from bone of Neandertal man excavated from Vindija Cave, Croatia [64].

Conclusions:-

In this review, we have described the tremendous progress made in sequencing the genomes of different organisms using HTS technology. Still, a long way to go in the genome sequencing to identify the genetic variations in the genomes of closely related species. A number of new technologies are under development that can reinvigorate the genomics field by increasing the efficiency of the sequence and massively decreasing the cost, which takes the genome sequencing to the next level of mutation screening, evolutionary studies and environmental profiling.

REFERENCES:-

1. Parfrey LW, Lahr DJG, Katz LA (2008). The Dynamic Nature of Eukaryotic Genomes. Mol. Biol. Evol. 25: 787-794.
2. Glass JI, Assad-Garcia N, Alperovich N, Yooseph S, Lewis MR, Maruf M, Hutchison CA 3rd, Smith HO, Venter JC (2006). Essential genes of a minimal bacterium. Proc. Natl. Acad. Sci. USA. 103: 425–30.
3. Evan EE, Jonathan F, Greg G, Augustine K, Suzanne ML, Jason HM, Joseph HN (2010). Missing heritability and strategies for finding the underlying causes of complex disease. Nature 11: 446-450
4. Mailman MD, Feolo M, Jin Y, Kimura M, Tryka K, Bagoutdinov R, et al. (2007). The NCBI dbGaP database of genotypes and phenotypes. Nat. Genet. 39: 1181–1186.
5. Hall N (2007). Advanced sequencing technologies and their wider impact in microbiology. J. Exp. Biol. 210: 1518–1525.
6. Olsvik O, Wahlberg J, Petterson B, et al. (1993). Use of automated sequencing of polymerase chain reaction-generated amplicons to identify three types of cholera toxin subunit B in Vibrio cholerae O1 strains. J. Clin. Microbiol. 31: 22–25.
7. Fiers W, Contreras R, Duerinck F, Haegeman G, Iserentant D, et al. (1976). Complete nucleotide-sequence of bacteriophage MS2-RNA - primary and secondary structure of replicase gene, Nature 260: 500-507.
8. Schuster SC (2008). Next-generation sequencing transforms today's biology. Nature Meth.. 5: 16–18.
9. Adessi C, Matton G, Ayala G, Turcatti G, Mermod J-J, Mayer P, Kawashima E (2000). Solid phase DNA amplification: characterisation of primer attachment and amplification mechanisms. Nucleic Acids Res. 28:e87/1–e87/8.
10. Ronaghi M, Karamohamed S, Pettersson B, Uhlen M, Nyren P (1996). Real-time DNA sequencing using detection of pyrophosphate release. Anal. Biochem 242: 84–89.
11. Valouev A, Ichikawa J, Tonthat T, et al. (2008). A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. Genome Res. 18: 1051–1063.
12. Tang F, Barbacioru C, Wang Y, et al. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. Nature Meth. 6: 377–382.
13. Margulies M, Egholm M, Altman WE et al. (2005). Genome sequencing in microfabricated high-density picolitre reactors. Nature 437: 376–380.
14. Hanna GJ, Johnson VA, Kuritzkes DR, et al. (2000). Comparison of sequencing by hybridization and cycle sequencing for genotyping of human immunodeficiency virus type 1 reverse transcriptase. J. Clin. Microbiol. 38: 2715–2721.
15. Edwards JR, Ruparel H, Ju J (2005). Mass-spectrometry DNA sequencing. Mut. Res. 573: 3–12.
16. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM. et.al.(1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269: 496-512.
17. Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, Gordon SV, Eiglmeier K, Gas S, Barry CE, III et.al. (1998). Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. Nature 393: 537-544.
18. Blattner FR, Plunkett G, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF, et.al. (1997). The complete genome sequence of *Escherichia coli* K-12. Science 277: 1453-1474.
19. Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, Carlton JM, Pain A, Nelson KE, Bowman S et.al. (2002). Genome sequence of

the human malaria parasite *Plasmodium falciparum*. Nature 419: 498-511.
20. Hall N, Pain A, Berriman M, Churcher C, Harris B, Harris D, Mungall K, Bowman S, Atkin R, Baker S. et.al. (2002). Sequence of *Plasmodium falciparum* chromosomes 1, 3-9 and 13. Nature 419: 527-531.
21. Goffeau A, Aert R, Agostini-Carbone ML, Ahmed A, Aigle M, Alberghina L, Albermann K, Albers M, Aldea M, Alexandraki D et.al. (1997). The yeast genome directory. Nature 387: 100-105.
22. Tettelin H, Masignani V, Cieslewicz MJ, Donati C, et al. (2005). Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". Proc. Natl. Acad. Sci. USA. 102: 13950-13955.
23. Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R (2005). The microbial pan-genome. Curr. Opin. Genet. Dev. 15: 589-594.
24. Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, Wang M. D, Zhang K, Mitra RD, Church GM (2005). Accurate multiplex polony sequencing of an evolved bacterial genome. Science 309: 1728-1732.
25. Harris TD, Phillip RB, Hazen B, Eric B, Jayson B, Ido B, et al. (2008). Single molecule DNA sequencing of a viral genome. Science 320:106-109.
26. Velicer G J, Raddatz G, Keller H, Deiss S, Lanz C, Dinkelacker I, Schuster SC (2006). Comprehensive mutation identification in an evolved bacterial cooperator and its cheating ancestor. Proc. Natl. Acad. Sci. USA. 103: 8107-8112.
27. Pennisi E (2008). Plant sciences. Corn genomics pops wide open. Science 319:1333.
28. Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, et al. (2006). The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science. 313:1596–1604.
29. Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, et.al. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 449: 463–467.
30. Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, et.al. (2009). The *Sorghum bicolor* genome and the diversification of grasses. Nature 457: 551–556.
31. Rounsley S, Pradeep Reddy M, Yeisoo Y, Ruifeng H, Nick S, et al. (2009). De novo next generation sequencing of plant genomes. Rice 2:35-43.
32. Huang S, Li R, Zhang1 Z, etal. (2009). The genome of the cucumber, *Cucumis sativus* L. Nature Gen. 41: 1275-1281.
33. Szittya G, Simon M, Dulce MS, Runchun J, Manuel PSF, Vincent M, Tamas D (2008). High-throughput sequencing of *Medicago truncatula* short RNAs identifies eight new miRNA families. BMC Genomics 9:593-601.
34. Fahlgren N, Howell MD, Kasschau KD, Chapman EJ, Sullivan CM, et al. (2007). High-throughput sequencing of Arabidopsis microRNAs: evidence for frequent birth and death of MIRNA genes. PLoS ONE 2: e219.
35. Ralf R, Frank W (2005). Moss (*Physcomitrella patens*) functional genomics. Gene discovery and tool development, with implications for crop plants and human health. Briefings in functional genomics and proteomics. Vol 4, 48–57.
36. Both M, Michael C, Michael PH, Stumpf[c] Pietro DS (2005a). Gene Expression Profiles of Blumeria graminis Indicate Dynamic Changes to Primary Metabolism during Development of an Obligate Biotrophic Pathogen. *The* Plant Cell 17:2107-2122.
37. Both M, Sabine EE, Michael C, Elisabeth M, George D, Pietro DS (2005b). Transcript profiles of *Blumeria graminis* development during infection reveal a cluster of genes that are potential virulence determinants. Mol. Plant-Microbe Interact. 18:125-133.
38. Gowda M, Haumeng L, Joe A, Feng C, Richard P, Guo-Liang W (2006). Robust analysis of 5'-transcript ends (5'-RATE): a novel technique for transcriptome analysis and genome annotation. Nucleic Acids Res. 34: e126.
39. Rabinowicz PD, Rensink W (2005). Ways to get from plant genomes to phenomes: via yeast. Genome Biol. 6:310-311.
40. Scarcelli JJ, Susan V, Christine AH, Catherine VH, David CA, Charles NC (2008). Synthetic genetic array analysis in *Saccharomyces cerevisiae* provides evidence for an interaction between *RAT8/DBP5* and genes encoding P body components. Genetics 179: 1945-1955.
41. Dixon C, Neal M, Richard M Z, Donnie AD, David SG (2005). α-Synuclein Targets the Plasma Membrane via the Secretory Pathway and Induces Toxicity in Yeast. Genetics 170: 47-59.
42. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD et.al. (2000). The genome sequence of *Drosophila melanogaster*. Science 287: 2185–2195.
43. Richards S, Liu Y, Bettencourt BR, Hradecky P, Letovsky S, et al. (2005). Comparative genome sequencing of *Drosophila pseudoobscura*:

chromosomal, gene, and cis-element evolution. Genome Res. 15: 1–18.
44. Redon RS, Ishikawa KR, Fitch LF, Perry GH, et. al. (2006). Global variation in copy number in the human genome. Nature 444: 444–454.
45. Zhou Q, Zhang G, Zhang Y, Xu S, Zhao R, et al. (2008). On the origin of new genes in Drosophila. Genome Res. 18: 1446–1455.
46. Carter NP (2007). Methods and strategies for analyzing copy number variation using DNA microarrays. Nat. Genet. 39: S16–S21.
47. Daines B, Hui W, Yumei L, Yi H, Richard G, Rui C (2009). High-throughput multiplex sequencing to discover copy number variants in Drosophila. Genetics 182: 935–941.
48. Dalca A, Brudno M (2010). Genome Variation Discovery with High Throughput Sequencing Data. Briefings in Bionformatics, 11:3-14..
49. Ashkenasy N, Sanchez-Quesada J, Bayley H, Ghadiri MR (2005). Recognizing a single base in an individual DNA strand: A step toward DNA sequencing in nanopores. Angew. Chem. Int. Ed. Engl. 44: 1401-1404.
50. Holt RA, Jones SJM (2008). The new paradigm of flow cell sequencing. Genome Res. 18: 839-846.
51. Chen TL, Siu LK, Wu RC, et al. (2007). Comparison of one-tube multiplex PCR, automated ribotyping and intergenic spacer (ITS) sequencing for rapid identification of Acinetobacter baumannii. Clin. Microbiol. Infect. 13: 801-806.
52. Yu CE, Devlin B, Galloway N, et al. (2004). ADLAPH: A molecular haplotyping method based on allele-discriminating long-range PCR. Genomics 84: 600-612.
53. Conway BR, Savage DA, Brady HR, Maxwell AP (2007). Association between haptoglobin gene variants and diabetic nephropathy: Haptoglobin polymorphism in nephropathy susceptibility. Nephron Exp. Nephrol. 105: e75-79.
54. Dahl F, Stenberg J, Fredriksson S, et.al. (2007). Multigene amplification and massively parallel sequencing for cancer mutation discovery. Proc. Natl. Acad. Sci. USA. 104: 9387-9392.
55. Fredriksson S, Baner J, Dahl F etal. (2007). Multiplex amplification of all coding sequences within 10 cancer genes by Gene-Collector. Nucleic Acids Res. 35: e47.
56. Meuzelaar LS, Lancaster O, Pasche JP, et al. (2007). MegaPlex PCR: A strategy for multiplex amplification. Nature Meth. 4: 835- 837.
57. Porreca GJ, Zhang K, Li JB, et al. (2007). Multiplex amplification of large sets of human exons. Nature Meth. 4: 931-936.
58. Okou DT, Steinberg KM, Middle C, et al. (2007). Microarray-based genomic selection for high-throughput resequencing. Nature Meth. 4: 907-909.
59. Ting N, Han W, Shen S, Mark J, Jun Z (2009). Selective gene amplification for high-throughput sequencing. Recent patents on DNA & gene sequences 3: 29-38.
60. Stiller M, Michael K, Udo S, Michael H, Matthias M (2009).Direct multiplex sequencing (DMPS): A novel method for targetedhigh-throughput sequencing of ancient and highly degraded DNA. 19: 1843-1848.
61. Poinar HN, Schwarz C, Qi J, Shapiro B, Macphee RD, et al. (2005). Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA Science 311: 392-394
62. Miller W, Drautz DI, Ratan A, Pusey B, Qi J, Lesk AM, Tomsho LP, et al. (2008). Sequencing the nuclear genome of the extinct woolly mammoth. Nature 456: 387-390.
63. Green RE, Krause J, Ptak SE, Briggs AW, Ronan MT, Simons JF, Du L, Egholm M, Rothberg JM, Paunovic M et al. 2006. Analysis of one million base pairs of Neanderthal DNA. *Nature* 444: 330-336.
64. Malez, M, Ullrich, H (1982). Neuere plaa¨ anthropologische untersuchungen am material aus der ho¨ hle vindija. Palaeontologia Jugoslavia 29: 1–44.

Table 1:- List of plant genomes sequenced

Organism	**Genome size**	**Year of completion**
Arabidopsis thaliana	125 Mb	2000
Oryza sativa ssp indica	420 Mb	2002
Oryza sativa ssp japonica	466 Mb	2002
Populus trichocarpa	550 Mb	2006
Vitis vinifera	490 Mb	2007
Elaeis guineensis	1800 Mb	2007
Physcomitrella patens	500 Mb	2008
Carica Papaya	372 Mb	2008
Cucumis sativus	367 Mb	2009
Zea mays	2800 Mb	2009
Brassica napus	1100 Mb	2009
Glycine max	1100 Mb	2010
Brachypodium distachyon	272 Mb	2010