Journal of Advances in Science and Technology

Vol. VIII, Issue No. XVI, February-2015, ISSN 2230-9659

AN ANALYSIS UPON VARIOUS TECHNIQUES AND APPLICATIONS OF HIGH-THROUGHPUT DNA SEQUENCING

AN INTERNATIONALLY INDEXED PEER REVIEWED & REFEREED JOURNAL

www.ignited.in

Journal of Advances in Science and Technology
Vol. VIII, Issue No. XVI, February-2015, ISSN 2230-9659

An Analysis upon Various Techniques and Applications of High-Throughput DNA Sequencing

Shaioqua Imam[1] Dr. Surendra Sarsaiya[2] Dr. Akhilesh Kumar[3]

[1]Research Scholar, Sri Satya Sai University of Technology & Medical Sciences, MP

[2]Supervisor

[3]Co-Supervisor

Abstract – Recent advances in DNA sequencing have revolutionized the field of genomics, making it possible for even single research groups to generate large amounts of sequence data very rapidly and at a substantially lower cost. These highthroughput sequencing technologies make deep transcriptome sequencing and transcript quantification, whole genome sequencing and resequencing available to many more researchers and projects.

The recent progresses of high-throughput sequencing (HTS) technologies enable easy and cost-reduced access to whole genome sequencing (WGS) or re-sequencing. HTS associated with adapted, automatic and fast bioinformatics solutions for sequencing applications promises an accurate and timely identification and characterization of pathogenic agents.

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 revolutionized 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.

Standardization of DNA extraction is a fundamental issue of fidelity and comparability in investigations of environmental microbial communities. Commercial kits for soil or feces are often adopted for studies of activated sludge because of a lack of specific kits, but they have never been evaluated regarding their effectiveness and potential biases based on high throughput sequencing.

----------♦----------

INTRODUCTION

DNA sequencing is the act of determining the nucleotide sequence of given DNA molecules — from a short segment of a single molecule, such as a regulatory region or a gene, up to collections of entire genomes. In the early 1970s, the first DNA sequences were obtained through extremely laborious techniques. An example is the sequencing of the two dozen base pairs of the lac operator. The first revolution in the DNA sequencing field took place in the second half of the 1970s with methods published by Allan Maxam and Walter Gilbert (Maxam and Gilbert 1977) and Frederick Sanger and colleagues (Sanger et al. 1977).

Both techniques greatly increased the throughput of sequencing DNA. The method from Gilbert and Maxam, however, was more complex and involved the use of hazardous chemicals. The Sanger method, on the other hand, offered overall higher efficiency after a series of optimisations, in particular switching from radioactive to dye labelling of nucleotides and using capillary electrophoresis instead of slab gels.

This technique dominated DNA sequencing for the following decades and lead to the determination of a complete1 human reference genome sequence at the turn of the millennium (The International Human Genome Sequencing Consortium 2001).[1,2]

In 1977 the first genome, that of the 5,386 nucleotide (nt), single-stranded bacteriophage wX174, was completely sequenced using a technology invented just a few years earlier. Since then the sequencing of whole genomes as well as of individual regions and genes has become a major focus of modern biology and completely transformed the field of genetics.

Over the last decade, alternative sequencing strategies have become available which force us to completely redefine ''high-throughput sequencing.'' These technologies outperform the older Sanger-sequencing technologies by a factor of 100–1,000 in daily throughput, and at the same time reduce the cost of sequencing one million nucleotides (1 Mb) to 4–0.1% of that associated with Sanger sequencing. To reflect these huge changes, several companies, researchers, and recent reviews use the term ''next-generation sequencing'' instead of high-throughput sequencing, yet this term itself may soon be outdated considering the speed of ongoing developments.[3-6]

Twenty years ago, microbiological research had been radically transformed by the advent of whole-genome sequencing (WGS), which gave rise to the pathogenomics era. More recently, high-throughput sequencing (HTS) has revolutionized sequencing approaches through sequencing platforms and technologies based on multiparallelized shotgun sequences, allowing one to obtain a whole microbial genome in one run or even a fraction of a run. Having access to WGS has significantly increased the knowledge of microorganisms and multiplied the clinical investigations in microbiology. The current standard procedure of clinical microbiology follows a sequential approach, generally consisting of pathogen isolation and in identification before, in some cases, performing drug susceptibility testing and/or epidemiological typing. HTS can be applied in the two major areas of clinical microbiology: diagnostic microbiology (management of infected patients) and the epidemiology of infectious diseases. Obtaining and combining results from time-consuming and heterogeneous methods into a single one is the promise of HTS, which can be foreseen as the future standard tool for a genome-based diagnosis.[7]

Along with the development of the low cost, next generation high throughput sequencing techniques, the EarthMicrobiome Project has been launched in 2011, aiming to reveal the gigantic, unexplored microbial genetic resource in soil, seawater, freshwater, the atmosphere, and other environments on our planet. At least 200,000 samples will be analyzed according to this ambitious plan. To maximize the comparability among the different studies, it needs standardized protocols for every operation step, including DNA extraction, PCR, sequencing, and data processing. Extraction of DNA of high quality is the first key step to profile microbial community with high fidelity (Martin-Laurent et al. 2001). However, the diversity of environmental sample types makes it impossible to simply apply one protocol or kit for DNA extraction.

Unlike soil or other environmental samples, activated sludge (AS) is almost composed of bacterial cells or their products, mostly extracellular polymeric substances (EPS) (Liu and Fang 2003). Generally, 1 g of dry mass of AS contains over 1–~10×1012 bacterial cells.This value is over 100-fold higher than the microbial density in soil samples. Its abundance guarantees that biomass is not a concern, and only hundreds of microliters to several milliliters of sludge are enough for DNA extraction. However, the complex biopolymers that constitute a very large portion of AS and other organic or inorganic matters adsorbed on AS are difficult to separate thoroughly from DNA during extraction. Moreover, the EPS are innate protectors of bacterial cells (Flemming et al. 2007). Breaking apart of the cell should be efficient for such samples, with the precondition that it should not result in over-fragmentation of DNA.

To the best of our knowledge, there is currently no well accepted commercial DNA extraction kit designed for AS samples, which is distinct from all other environmental samples. Thus, the cross-use of commercial kits made for other sample types (such as soil and stool) should be evaluated for their applicability to AS samples, although they had been randomly selected in previous AS studies. On the other hand, for AS samples containing bulking water, ethanol fixation is usually adopted during transportation and storage. As far as we know, the effect of this processing on the bacterial community profiling has not yet been evaluated.[8-10]

HIGH-THROUGHPUT SEQUENCING TECHNOLOGIES IN GENOME SEQUENCING

Genome comprises the entire genetic information of an organism encoded either in DNA or RNA consisting of both coding and non-coding regions. 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. 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 today 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.

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.

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. 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 highthroughput sequencers in understanding the genetic variation and their relationship to biological function.[11,12]

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, 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. 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. 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.[13]

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 highthroughput 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.[14]

SOIL DNA ANALYSIS

Ail alternative to morphological identification of soil biota is DNA analysis. To date microbes, particularly bacteria, have been the target for soil DNA analyses as they can provide discriminative DNA profiles; however, the taxonomic resolution of bacterial DNA fingerprint methods is limited. DNA fingerprint techniques rely on differences in fragment lengths between species in a sample to generate a profile and so individual species present are not identified. In contrast, the recent development of DNA metabarcoding (PGR amplification of DNA mixtures using universal primers) and high- throughput sequencing (HTS) enables rapid species identification and offers the potential to improve soil discrimination by targeting non-culturable microorganisms (i.e. those that cannot be cultured on routine microbiological media) and alternative soil taxa such as eukaryotes.

The majority of forensic soil DXA studies to date have examined microbial diversity using DNA fingerprinting methods including denaturing gradient gel electrophoresis. amplicon length heterogeneity PGR. and terminal restriction fragment length polymorphism. T-RFLP is extensively used in forensic soil science and is done by amplifying a region of the 16S ribosomal RNA encoding gene (rRNA) and digesting it with restriction endonucleases. The 16S rRNA fragments of varying length are separated by gel electrophoresis and analysed to provide a distinct profile (fingerprint) dependent upon the species composition within the sample. This method was introduced by Liu *et al.* in 1997 to characterise bacterial communities in activated sludge, bioreactor sludge, aquifer sand and termite guts.

Shortly after Horswell. J. demonstrated that T-RFLP could generate discriminative microbial DNA profiles from soil. The benefits of T-RPLP include small soil sample sizes (< 1 g). use of conun011 forensic equipment, ease of automation, and low cost, allowing rapid and reproducible soil community fingerprinting. Furthermore, T-RPLP analysis has been shown to provide higher discriminatory power between sites than elemental analysis and allows a more powerful analysis than culture-based techniques that account for only *2%* of the total bacteria present in a sample. Although T-RFLP is a

useful forensic tool. resolution and taxonomic identification are limited by co-migration of multiple species appearing as a single species during electrophoresis. In contrast. DXA metabarcoding can provide a standardised species identification method from complex soil communities, whilst increasing the discriminatory power between locations for forensic application.

DNA analysis from soils is far more complex than analysis from a single animal or plant, as a soil contains DNA from multiple specimens. The application of DNA barcoding to assess a DNA mixture is termed DNA metabarcoding. DNA metabarcoding involves PGR amplification of a target gene using universal primers to extract genetic information from a DNA mixture that ideally represents the diversity of a particular group of taxa within a sample.

DNA metabarcoding can be used to profile soil communities by targeting specific taxonomic groups. Currently, the most commonly utilised markers for environmental samples are 16S ribosomal RNA gene region (bacteria). 18S ribosomal RNA gene region (eukaryotes). and the internal transcribed spacer 1 (ITS1) (fungi). For each of these markers, curated reference databases are regularly updated to enable robust taxonomic identification of the sequences present in a sample: Greengenes. 1SS and ITS, respectively. Traditional barcoding of single taxa using a single specimen means that high quality DNA extracts can be used, and therefore long barcode regions (>500 bp) can be amplified with high discrimination capacity. In contrast, metabarcoding generally utilises shorter regions since DNA is more degraded in environmental samples. For example, the recommended *mark* region for plants is not used as large fragment lengths (760 bp) are not easily amplified from degraded DNA in soil. Instead, the identification of plant material from soils most commonly utilises the *rbcL* and *trnL* gene regions: however, unfortunately no curated database has yet been developed. Curated databases are advantageous for assigning accurate taxonomic identifications as the sequences are regularly monitored to ensure that all entries are reliable. However, for environmental samples such as soils, many sequences may return as 'unknown' as many will not previously have been sequenced and so show no match to the curated database entries.

DNA metabarcoding coupled with high-throughput sequencing (HTS) offers the potential to drastically increase taxonomic resolution within soil samples compared to DNA fingerprinting and thus increase the discriminatory power between different geographic locations. Multivariate analysis methods based on a distance matrix are commonly used to visualise the similarity between different metagenomic samples. Multidimensional Scaling (MDS) used throughout this study is an example of unconstrained multivariate analysis and illustrates the similarity between all samples. As a result, sample variation within a single site can be observed and Analysis of Similarity (ANOSIM) statistics can be applied to determine significant differences between samples. Recent advances in HTS have revolutionized the field of genomics, making it possible to rapidly generate large amounts of sequence data at a substantially lower cost. Many sequencing platforms are available, utilising different combinations of template preparation, sequencing and data analysis. In all of these different approaches, platform-specific adapters and unique tags (termed indexes) are incorporated into the sequences during PGR amplification; the unique tag enables multiple samples to be sequenced simultaneously. This technology has been successfully used to identify communities from soils and has shown to detect higher diversity than traditional DNA barcoding. For example. HTS of nematode diversity in tropical rainforest identified 7700 individuals, whereas traditional barcoding from individual specimens only identified 360 individuals. This indicates the potential of HTS to generate a more detailed composite picture of a source area. However, it is important to ensure that bioinfomiatic analysis does not increase diversity estimates due to sequencing errors. Despite the clear potential to explore alternative soil communities and increase taxonomic resolution. DNA metabarcoding and HTS cannot be utilised in casework without prior validation and consideration of potential limiting factors.[15-18]

DNA SEQUENCING USING BIOINFORMATICS ANALYSIS

Bioinformatics analysis of sequencing data can be divided into several stages. The first step is technology dependent, and deals with processing the data provided by the sequencing instrument. Downstream analysis is then done ad hoc to the type of experiment. When sequencing new genomes, de novo assemblies are required, which are possibly followed up with genome annotations. Re-sequencing projects use the short reads for aligning (or mapping assembly) against a reference sequence of the source organism; these alignments are then analyzed to detect events relevant to the experiment being conducted (e.g. mutation discovery, detection of structural variants, copy number analysis). The first step of bioinformatics analysis starts during sequencing, and involves signal analysis to transform the sequencing instruments fluorescent measurements into a sequence of characters representing the nucleotide bases. As sequencers image surfaces densely packed with the DNA sequencing templates and sequencing products, image processing techniques are required for detection of the nascent sequences and conversion of this detected signal into nucleotide bases.

Most technologies assign a base quality to each of the nucleotides, which is usually a value representing the confidence of the called bases. Although each vendor has methods specific to their technology to evaluate base quality, most provide the user with a

Phred-like Score value: a quality measurement based on a logarithmic scale encoding the probability of error in the corresponding base call.

To achieve contiguous stretches of overlapping sequence (contigs) in de novo sequencing projects, software that can detect sequence overlaps among large numbers of relatively short sequence reads is required. The process of correctly ordering the sequence reads, called assembly,is complicated by the short read length; the presence of sequencing errors; repeat structures that may reside within the genome; and the sheer volume of data that must be manipulated to detect the sequence overlaps. To address such complications, hybrid methods involving complimentary technologies have been successful.

For example, by mixing 200 bp 454 sequences reads with Sanger sequences, Goldberg et al. successfully sequenced the genomes of several marine organisms. A different approach eliminated the need for Sanger sequencing by mixing two distinct next generation sequencing technologies.

These studies provided practical examples of how the strengths of different technologies can be used to alleviate their respective short comings. Homology with previously sequenced organisms can help when sequencing new genomes. The use of this strategy was demonstrated during sequencing of the mouse genome; by taking advantage of the conserved regions between mouse and human, establishing a framework for further sequencing. A similar approach can be used to produce better assemblies with next generation sequencing. For example, to sequence the genome of the fungus *Sordaria macrospora*, short reads from 454 and Illumina instruments were first assembled using Velvet, and the resulting contigs were then compared to draft sequences of related fungi (*Neurospora crassa*, *N. discreta* and *N. tetrasperma*).

This process helped produce a better assembly by reducing the number of contigs from 5,097 to 4,629, while increasing the N50 (the contig length N, for which 50% of the genome is contained in contigs of length N or larger), from 117 kb to 498 kb.

More recently, new algorithms have been developed, which can assemble genomes using only short reads. Most of these methods are based on de Bruijn graphs. Briefly, the logic involves decomposing short reads into shorter fragments of length k (k-mers). The graph is built by creating a node for each k-mer and drawing a link, or “edge,” between two nodes when they overlap by k-1 bp. These edges specify a graph in which overlapping sequences are linked. Sequence features can increase the resulting graph’s complexity. The graph can, for example, contain loops due to highly similar sequences (e.g. gene family members or repetitive regions), and so-called bubbles can be created when single base differences (e.g. due to polymorphisms or sequencing errors) result in the creation of non-unique edges in the graph, which yield not one, but two possible paths around the sites of the sequence differences.

Graph complexity and size increase for large genomes, and given that the graph needs to be available in memory for efficient analysis, not all implementations can handle human size genomes. Some publicly available implementations, such as Velvet and Euler-SR, have been successfully used to assemble bacterial genomes.

Another implementation, ABySS, makes use of parallel computing through the Message Passing Interface (MPI), to distribute the graph between many nodes in a computing cluster. In this way, ABySS can efficiently scale up for the assembly of human size genomes, using a collection of inexpensive computers.

For re-sequencing experiments, high-throughput aligners are required to map reads to the reference genome. Many applications have long been available for sequence alignments; however, the amount and size of the short reads created by next generation sequencing technologies required the development of more efficient algorithms.[]19-21]

CONCLUSION

Current high-throughput sequencing technologies provide a huge variety of sequencing applications to many researchers and projects. Given the immense diversity, we have not discussed these applications in depth here; other reviews with a stronger focus on specific applications and data analysis are available.

New technologies on the horizon, SMRT by Pacific Biosciences, BASE by Oxford Nanopore, and other technologies such as that suggested by IBM, demonstrate the major future directions in the field of DNA sequencing: the ability to use individual molecules without any library preparation or amplification, the identification of specific nucleotide modifications, and the ability to generate longer sequence reads.

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] W. Gilbert and A. Maxam. The nucleotide sequence of the lac operator. Proc Natl Acad Sci U S A, 70(12):3581–3584, Dec 1977.

[2] The International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature, 409(6822):860–921, Feb 2001. doi: 10.1038/35057062.

[3] Bentley DR, Balasubramanian S, Swerdlow HP, et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456: 53–9.

[4] Clarke J, Wu HC, Jayasinghe L, et al. 2009. Continuous base identification for single-molecule nanopore DNA sequencing. Nat Nanotechnol 4: 265–70.

[5] Ansorge WJ. 2009. Next-generation, DNA sequencing techniques. Nat Biotechnol 25: 195–203.

[6] Shendure JA, Porreca GJ, Church GM. 2008. Overview of DNA sequencing strategies. Curr Protoc Mol Biol Chapter 7: Unit 7.1.

[7] Loman, N.J.; Constantinidou, C.; Chan, J.Z.; Halachev, M.; Sergeant, M.; Penn, C.W.; Robinson, E.R.; Pallen, M.J. High throughput bacterial genome sequencing: An embarrassment of choice, a world of opportunity. Nat. Rev. Microbiol. 2012, 10, 599 606.

[8] Martin-Laurent F, Philippot L, Hallet S, Chaussod R, Germon JC, Soulas G, Catroux G (2001) DNA extraction from soils: old bias for new microbial diversity analysis methods. Appl Environ Microbiol 67:2354–2359

[9] Liu Y, Fang HHP (2003) Influences of extracellular polymeric substances (EPS) on flocculation, settling, and dewatering of activated sludge. Crit Rev Env Sci Technol 33:237–273

[10] Flemming HC, Neu TR, Wozniak DJ (2007) The EPS matrix: the “House of Biofilm cells”. J Bacteriol 189:7945–7947

[11] Parfrey LW, Lahr DJG, Katz LA (2008). The Dynamic Nature of Eukaryotic Genomes. Mol. Biol. Evol. 25: 787-794.

[12] Hall N (2007). Advanced sequencing technologies and their wider impact in microbiology. J. Exp. Biol. 210: 1518–1525.

[13] Fiers W, Contreras R, Duerinck F, Haegeman G, Iserentant D, et al. (1976). Complete nucleotidesequence of bacteriophage MS2-RNA – primary and secondary structure of replicase gene, Nature 260: 500-507.

[14] Schuster SC (2008). Next-generation sequencing transforms today's biology. Nature Meth.. 5: 16– 18.

[15] Gangneux, C., M. Akpa-Vinceslas, H. Sauvage, S. Desaire, S. Houot, and K. Laval. 2011. Fungal, bacterial and plant dsDNA contributions to soil total DNA extracted from silty soils under different fanning practices: Relationships with chloroform-labile carbon. Soil Biology & Biochemistry 43:431-437.

[16] Horswell, J. 2002. Forensic comparison of soils by bacterial community DNA profiling. Journal of Forensic Sciences 47:350-353.

[17] Macdoiiald, C. A., R. Aug, S. J. Cordiner, and J. Horswell. 2011. Discrimination of Soils at Regional and Local Levels Using Bacterial and Fungal T-RFLP Profiling*. Journal of Forensic Sciences 56:61-69.

[18] Logares, R.. T. H. Haverkamp, S. Kumar, A. Lanzen, A. J. Kederbragt, C. Quince, and H. Kauserud. 2012. Environmental microbiology through the lens of liigh-tliroughput DXA sequencing: synopsis of current platforms and bioinfomiatics approaches. J. Microbiol. Methods 91:106-113.

[19] Zerbino DR, Birney E (2008) Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18: 821-829.

[20] Chaisson MJ, Pevzner PA (2008) Short read fragment assembly of bacterial genomes. Genome Res 18: 324-330.

[21] Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, et al. (2009) ABySS: A parallel assembler for short read sequence data. Genome Res 19: 1117- 1123.