Journal of Advances in Science and Technology

Vol. 10, Issue No. 21, February-2016, ISSN 2230-9659

A RESEARCH ABOUT THE ROLE OF HIGH THROUGHPUT DNA SEQUENCING TECHNIQUES IN GENOMIC SEQUENCING: METHODS AND APPLICATIONS

Journal of Advances in Science and Technology
Vol. 10, Issue No. 21, February-2016, ISSN 2230-9659

A Research about the Role of High Throughput DNA Sequencing Techniques in Genomic Sequencing: Methods and Applications

Shaiqua Imam[1]* Dr. Surendra Sarsiya[2]

[1]Research Scholar, SSSUTMS, Sehore

[2]UTD, SSSUTMS, Sehore

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 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.

This enormous growth of available DNA data has been a tremendous boon to large-scale genomics studies and has rapidly advanced fields such as environmental genomics, ancient DNA research, population genomics and disease association. On the other hand, however, researchers and sequence archives are now facing an enormous data deluge. Critically, the rate of sequencing data accumulation is now outstripping advances in hard drive capacity, network bandwidth and processing power.

---------------------------♦-----------------------------

INTRODUCTION

In the year 2000, researchers announced the first whole genome sequence of a plant species. Sequencing of Arabidopsis thaliana was a cutting-edge achievement in the field of plant genomics. The impact of that study was so great that it boosted the demand for genomic information. However, using the conventional Sanger method (first generation technology), sequencing a whole genome is time-consuming, laborious, and expensive work. In 2005, sequencing-by-synthesis technology developed by 454 Life Sciences revolutionized sequencing technology and started the second-generation sequencing era. Both required previous amplification in vivo (molecular cloning) or in vitro (e.g., polymerase chain reaction (PCR)). This was followed by the third-generation sequencing platforms, capable of sequencing single molecules without previous amplification. The sequencing generations following Sanger's approach are also known as next-generation sequencing (NGS), although this is rather ambiguous terminology for obvious reasons. The new sequencing strategies greatly reduced the necessary effort, time, and cost, also allowing for unprecedented throughput.

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 and Frederick Sanger and colleagues.

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 optimizations, in particular switching from radioactive to dye labeling 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.

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 tetrapod's . 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 high throughput sequencers in understanding the genetic variation and their relationship to biological function.

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.

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.

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 .

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 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. Solexa sequencing system can generate hundreds of millions of 50-100 base reads. 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.

GENOME-SEQUENCE ANALYSIS TOOLS

While developments in sequencing technology make it possible to obtain large-scale sequence data in a short time, the assembly and analysis of sequences remains a challenging task. Thus, much of the effort in recent years has been dedicated to developing and improving bioinformatics tools.

Different scenarios may cause erroneous base-calling in the sequencing platforms. For instance, most of the errors that come from indels in 454 reads are caused by incorrect photopolymer length calls. On the other hand, the sequencing chemistry of Illumina ensures that only one nucleotide is incorporated in each cycle, avoiding such photopolymer issues. However, this technology may suffer from wrong identification of the incorporated nucleotide.

Finally, areas in the genome with a high single-nucleotide polymorphism (SNP) density may get lower coverage with the ABI/SOLiD system. Thus, the sequencing data are managed and analyzed with advanced bioinformatics tools. Currently, a number of bioinformatics software packages are available, which are essentially used for different purposes, including alignment, assembly, annotation, and sequence-variation detection (e.g., identification of SNPs) (Paszkiewicz and Studholme, 2010; Bao et al., 2011).

The first step of assembly is to control the quality of the raw sequences. Since most of the machines produce the data in FASTA or FASTQ formats, the FASTX-Toolkit and FastQC emerge as useful tools for the preprocessing steps. After quality check and trimming (such as removing adapter sequences and short reads), the next step of sequencing data analysis is assembly of the sequences. The genome-assembly process can be divided into 2 steps: draft assembly and assembly improvement (finishing). In the majority of the cases, 98% of the genome is covered by draft assembly with an error rate of 1/2000 b, while this ratio is 5-fold lower in finished assemblies.

Usually, before assembly, repetitive elements are identified and filtered out from the dataset. Repetitive elements are one of the challenging issues for assembly procedures. In fact, the majority of the gaps in an assembly are caused by repeated sequences.

Sequencing with longer reads emerges as a good way out. Paired-end sequencing is also commonly used for this purpose. Depending on availability, repetitive elements are computationally detected by homology searches to known repeat sequences. REPuter, Tandem Repeat Finder, and Repeat Masker are among the most common programs for detecting such repetitive elements. When there is a lack of a reference genome, repetitive elements are identified de novo. The basic workflow pipeline is composed of masking the known repeats, de novo repeat finding on the masked genome, and classification of the newly identified repeats. Detailed de novo repeat discovery tools are mentioned elsewhere. RECON (Bao and Eddy, 2002), Repeat Modeler, Repeat Scout (Price et al., 2005), and REPET (Flutre et al., 2011) are examples of the best-known software packages for this purpose.

Presently, a number of assembly approaches are applied for short-read assemblies. The first assemblers are based on a simple strategy known as the greedy algorithm, which is an implementation of finding the shortest common super sequence (Narzisi and Mishra, 2011). The algorithm proceeds as follows: 1) pairwise comparison of all sequences is done to identify overlapping sequences and merge the best overlapped sequences; and 2) these steps are repeated until no more sequences are left to be merged. The greedy algorithm has been used mainly for assembling small genomes. On the other hand, since the algorithm needs local information at each step, the presence of complex repeats may lead to misassembles. The most accepted packages based on this method are TIGR, PHRAP, CAP3, PCAP, Phusion, SSAKE, and VCAKE.

With the advent of sequencing technologies, new assemblers have been developed, particularly for more complex genomes. The overlap-layout-consensus (OLC) approach analyzes the overlap graph of the sequencing reads and searches for a consensus genome. When applied to short reads, the main drawback of this approach is that it shows low performance, as too many overlaps have to be calculated. Examples of genome-assembly software packages applying the OLC approach are ARACHNE and Atlas.

CURRENT LIMITATIONS AND CHALLENGES

The utility of HTS for genome-based diagnosis in clinical microbiology no longer needs to be proven; the challenge is more, nowadays, in the transfer from pilot studies to routine use in a clinical context. Although HTS is the more powerful method, PCR-based methods currently dominate the market in routine clinical laboratories, because they are more cost-effective, and their workflow and produced data are easier to manage. For the HTS technologies, steps upstream of the sequencing run are not trivial; to prepare a library is an expert user task, and that is why a majority of sequencing platform providers tries to simplify and automate this time-consuming step. Moreover, the newly-arrived HTS methods are facing regulatory hurdles, which may limit their routine use in genome-based diagnosis.

Currently, HTS is at a technological crossroads: bench top sequencing has been commercially launched in 2011, between the second and third generation of sequencers. These bench top sequencers open the way for a wider dissemination; in other words, their compact and easy access format facilitates the generalization of these technologies. Today, technology is still improving capacity, and reliability is constantly evolving, which factually explains the difficulty in adopting and establishing this technology for genome-based diagnosis. The format and the power of sequencing platforms to be used for routine diagnosis have not been totally agreed upon in the community. These specification issues depend on the modality (culture-dependent or culture-independent) of future genome-based diagnosis, on the commercial strategy of sequencer suppliers and on the future certification of these sequencing platforms (FDA and CE-IVD approval).

According to a survey among physicians, which has tried to enlighten existing barriers for the integration of personalized medicine into clinical practice, the access to specific training and clear guidelines were two major items that could influence this adoption of a new clinical practice. A consequence of a young technology, such as HTS, is the lack of academic training, which can result in less adhesion to the new technology.

High throughput sequencers produce huge amounts of data. Long-term archiving of this data is not a trivial task, and it is evident that the HTS community is facing a storage problem. A reflection on how to store proprietary data concerning patients or research centers has to be made.

A significant part of HTS analysis is based on the comparison with diverse databases. An improvement of databases used to analyze HTS data in order to obtain well-annotated reference databases is urgently needed. Even if some centers organized and distributed some data, a lot of databases exist, and there is a wealth of data, which is left mostly inaccessible and unexplored. Better centralization and data collection should be developed.

Another limitation with HTS data is the lack of format standardization. Even if some formats are widely used in the HTS field, such as FASTQ, SAM or BAM formats, standardized formats and procedures are still lacking. The establishment of standards would help data sharing and connecting tools. In clinical applications, it is also important to integrate and standardize meta-data and include them in the analysis. The integration of results with other types of data, such as the sample collecting place or phenotypic data, is necessary for implementing personalized medicine. Since the technology has been in constant evolution and the algorithms are evolving with it, there is currently no stable pipeline for the analysis of HTS data. Use of a pipeline often implies downloading and installing a lot of software, which requires a minimum of computational skills and computational resources, which are limited in a hospital. The lack of an intuitive graphical interface is one of the main limitations in HTS bioinformatics. Indeed, the use of analysis tools is often too complex for most researchers and clinical staff, which choose to use more straightforward approaches, potentially sacrificing the quality of their results.

ROLE OF HTS TECHNOLOGIES IN SEQUENCING THE MICROBIAL GENOMES

Ever since the first microbial genome of Haemophilus influenza has been decoded comprising full genetic complement with 1830137·bp of DNA and 1743 predicted genes, numerous microbial genomes were sequenced including Mycobacterium tuberculosis , one of the most important human bacterial pathogens, E.coli , malarial parasite Plasmodium falciparum and yeast . 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, 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.

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 micro fabricated 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 pyro sequenced, 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., . 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 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. 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.

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.

Many new de novo and resequenced plant genomes are expected in the near future for plants in general and crop species in particular, using second- and mostly third generation sequencing platforms. Further work is needed to complete the biggest and most complex genome drafts while achieving high-quality reference sequences for most plant genomes. This genome knowledge will be coupled with deep gene-expression analyses (RNA-Seq and true RNA sequencing), uncovering alternative splicing, copy number variations, etc. ChIP-Seq and microRNA-Seq availability for an increasing number of crops will further expand the emerging field of epigenetics. These are all necessary tools for food production and security in a climate-change scenario.

REFERENCES

Ahn S.M., Kim T.H., Lee S., Kim D., Ghang H., Kim D.S., et al. (2009). The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. Genome Res 19: pp. 1622–1629

Bansal V. (2010). A statistical method for the detection of variants from next-generation resequencing of DNA pools. Bioinformatics 26: i318–i324

Bao S., Jiang R., Kwan W., Wang B., Ma X., Song Y.Q., (2011). Evaluation of next-generation sequencing software in mapping and assembly. J Hum Genet 56: pp. 406–414.

Bao Z., Eddy S.R., (2002). Automated de novo identification of repeat sequence families in sequenced genomes. Genome Res 12: pp. 1269–1276.

Boers, S.A.; van der Reijden, W.A.; Jansen, R. (2012). High-throughput multilocus sequence typing: Bringing molecular typing to the next level. PLoS One, 7, e39630.

Dalloul R.A., Long J.A., Zimin A.V., Aslam L., Beal K. et al. (2010). Multiplatform next-generation sequencing of the domestic turkey (Meleagris Gallopavo): genome assembly and analysis. PLoS Biol 8:e1000475

Flutre T., Duprat E., Feuillet C., Quesneville H. (2011). Considering transposable element diversification in de novo annotation approaches. PLoS One 6: e16526.

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

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

Loman, N.J.; Misra, R.V.; Dallman, T.J.; Constantinidou, C.; Gharbia, S.E.; Wain, J.; Pallen, M.J. (2012). Performance comparison of benchtop high throughput sequencing platforms. Nat. Biotechnol. 30, pp. 434–439.

Margulies M., Egholm M., Altman W.E., et al. (2005). Genome sequencing in microfabricated high-density picolitre reactors. Nature 437: pp. 376–380.

Narzisi G, Mishra B (2011). Comparing de novo genome assembly: the long and short of it. PLoS One 6: p. 19175.

Paszkiewicz K., Studholme D.J., (2010). De novo assembly of short sequence reads. Brief Bio informs 11: pp. 457–472.

Price A.L., Jones N.C., Pevzner P.A., (2005). De novo identification of repeat families in large genomes. Bioinformatics 21 (Suppl. 1): pp. 351–358.

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

Corresponding Author

Shaiqua Imam*

Research Scholar, SSSUTMS, Sehore

E-Mail – chairman.iab@gmail.com