A Study of Next Generation Sequencing Data, Workflow, Application and Platform Comparison

Aswathy Thankachan and Mr Bino Thomas
1 P G Scholar, Department of Computer Science and Engineering, St. Joseph’s College of Engineering and Technology, Palai, Kerala, India.
2 Assistant Professor, Department of Computer Science and Engineering, St. Joseph’s College of Engineering and Technology, Palai, Kerala, India.
E-mail: aswathy911@gmail.com, binothomas@sjcetpalai.ac.in

Abstract.
DNA sequencing determines the precise order of nucleotides within a DNA molecule. Next generation Sequencing (NGS) is an efficient parallel high throughput DNA Sequencing Technology which revolutionizing the genomic research. Earlier methods that are quite expensive give rise to different sequence comparison techniques. NGS used for faster detection of variants in human genome and give rise to accelerated response to disease detection like cancer, hepatitis etc. In this paper work flow of data analysis part of NGS are clearly discussed. The different NGS platforms, its applications and different sequencing comparison techniques are also mentioned here.

Keywords: Sanger sequencing, Maxam-Gilbert Sequencing, Needleman-Wunsch algorithm

1. Introduction
DNA Sequencing is a fundamental terminology in Bioinformatics. Genes are the fundamental element in a cell that makes up the organism. The process of determining the precise order of nucleotides Adenine(A), Guanine(G), Cytosine(C), Thymine(T) within a DNA molecule is referred to as DNA sequencing. The order of the bases along DNA contains the complete set of instructions that make up the genetic inheritance. The key principle of DNA sequence generation method is to evaluate the sequencing with very high accuracy and reliability.

Next generation sequencing (NGS) Technology which is a high through put parallel sequencing technology which accelerated the genomic research. NGS deals with enormous amount of data as the different combination of nucleotides which is then converted into a string problem in computer science.[12] Applications of NGS includes personalized medicine, Genetic Disorder, clinical diagnosis of different viral disease like cancer, hepatitis etc. It provide greater opportunities for fast and low cost sequencing in the modern times. Viral transmission like hepatitis, cancer, HIV can be efficiently detected by using NGS Technologies. International research project like The Human Genome Project (HGP) is a remarkable achievement of NGS. It is a program that aims to provide complete mapping of all set of genomes and will act as a world resource of detailed information about the function, structure and organization of the complete set of human genomes. After HGP the time, cost of comparison started decreasing...
while amount of scientific knowledge has continued growing exponentially. So NGS has its own significance to understand the code of life.

2. Old Sequencing Methods

Human genome consist of 3 billion base which encapsulated 23 pairs of chromosomes inside the nucleus of a cell. DNA is having a double helix structure get attached with each other based on the complimentary pairs (A-T) and (C-G). Sequencing of the Single DNA strand can be done using different methods.

2.1. Sanger sequencing

In 1977 Federic Sanger [1] did First DNA Sequencing called Sanger sequencing or chain termination method. It uses enzymatic synthesis to fragments DNA strands. DNA template, primer, polymerise and florescent nucleotides are the commonly used enzymatic synthesis. Strands are subjected to electrophoresis and visualized by UV light. The output obtained is termed as reads ie the Sequence of DNA strand is obtained as the combination of A C T G. The method is expensive time consuming more over it provides high quality reads. The use of dideoxynucleotide triphosphates (ddNTPs) as DNA chain terminators is the key point. Non-specific binding of the primer to the DNA results in deformalities in reads-out of the sequence.

2.2. Maxam-Gilbert

Maxam-Gilbert Chemical Sequence Method is more effective yet still limiting method used chemical sequencing. Strands are subjected to gel electrophoresis and precious order is determined. The method is best suited for heterogeneous DNA sequences. But it employs harmful radioactive chemicals and was difficult to scale up for high-throughput, robust sequencing. Sangers Sequencing is most preferred Compared with Maxam-Gilberth method because of its higher efficiency and through put.

2.3. Dye-Terminator Sequencing

Other methods include Dye-terminator sequencing utilizes labelling of the chain terminator ddNTPs, which permits sequencing in a single reaction. The variations in dye effects results in the unequal peak heights and shapes in the electronic DNA sequence trace chromatogram due to the incorporation of the dye-labelled chain terminators into the DNA fragment. The common challenges of DNA sequencing include poor quality in the first 15-40 bases of the sequence and can sequence upto 700- 900 bases.

3. Next Generation Sequencing Data

Next generation Sequencing (NGS) is an efficient parallel high throughput DNA Sequencing Technology. High Throuput indicates that large scale replication of genomes is possible for further evaluation. NGS replaces enzymes synthesis by machines and makes the processing much faster. Fig 1 shows NGS pipeline

3.1. NGS Library Preparation

Sample DNA fragments are taken to library preparation. It is the process of preparation of targeted DNA into a form compatible with sequencing data. Library preparation consists of steps like DNA fragmentation, adaptor sequencing, size selection and final library quantification and QC. After this fragments are taken to DNA sequencing machines.
3.2. **NGS Platforms**

NGS systems are typically represented by SOLiD/Ion Torrent PGM from Life Sciences, Genome Analyzer/HiSeq 2000/MiSeq from Illumina, and GS FLX Titanium/GS Junior from Roche. Depending upon the application, quality and number of reads and run time each sequencer is selected. Table 1.1 shows some platform compared based on read length, its prons and cons.

| Platforms                  | Read Length       | Advantage          | Disadvantage                           |
|----------------------------|-------------------|--------------------|----------------------------------------|
| 454 GS Roche               | 700-1000bp        | Long read          | Expensive                              |
| HiSeq Illumina             | 2000- 2*150 bp    | Low cost           | Short read, long structural variants   |
| Life Technology Solid 4    | 35-50 bp          | High accuracy      | Expensive, long run time               |

3.2.1. **Advanced NGS platforms** Pacific Biosciences (Pac Bio) Single Molecule Real Time (SMRT) sequencing, the Illumina Tru-seq Synthetic Long-Read technology, and the Oxford Nanopore Technologies sequencing platform are the advanced version of NGS. This provide direct sequencing of single DNA molecule having read lengths and run time which simplifies resequencing of genomes, to achieve a higher level of accuracy. Nanopore released a device called MinION which is a palm sized sequencing device that helps the real time analysis of single molecules.

3.3. **Application**

There are different applications for NGS that includes personalized medicine, Genetic Disorder, clinical diagnosis of different viral disease like cancer, hepatitis etc. Applications of NGS such as whole genome sequencing, target sequencing, gene expression profiling, chromatin immune precipitation sequencing, and small RNA sequencing that are used to accelerate biological and biomedical research [2]. The Human Genome Project (HGP) is an international research program that aims is to provide complete mapping of all set of genomes and will act as a world resource of detailed information about the function, structure and organization of the complete set of human genomes.
3.4. Work flow and data Analysis
The Graphical Pipeline for Computational Genomics (GPCG) [4] represents the workflow for basic sequence alignment and data quality control, analysis of single nucleotide polymorphism. Analytical steps starts with the quality assessment of raw data, read alignment, variant identification calling, annotation and visualization of result in a phylogenetic tree.

The first analysis step after completing the sequencing run is to evaluate the quality of raw reads and to remove, trim or correct reads that do not meet the defined standards. The NGSQC Toolkit and PRINSEQ are able to handle FASTQ files that are capable of filtering and trimming reads. Those who meet certain quality standard are aligned to an existing reference genome. The University Of Santa Cruz (UCSC), and the Genome Reference Consortium (GRC) are two main sources for the human reference genome assembly which provide alignment programs to refer the human genome for MAQ, BWA, SOAP, Bowtie/Bowtie2 are examples of alignment programs.

If the sequence doesn’t have a reference genome then it is called Denovo assembly. It creates a transcriptome without the aid of a reference genome and sequenced separately then added to database for further observations. It identifies structural variants and complex rearrangements, such as deletions, inversions, or translocations and generally produces fragmented assemblies Ex: DNASTAR. In the variant identification, SAM File is obtained and compressed into BAM file, which is a universal file format for mapped sequenced reads that contains the sequence score of each reads

Predicting the functional impact of Data is done by using Computer-aided annotation that enables research groups to filter and prioritize potential disease-causing mutations for further analysis. The annotation tools and strategies that have been developed to retrieve information and test hypotheses about the functional role of variants present in the human genome. NGS visualization tools should support users by displaying aligned reads, mapping quality and identified mutations combined with annotations from various public resources.

3.5. Database
Tools have been developed for the effective and efficient mining of sequence information. These include Locus-Link, which provides a gene-centric view of sequence-based information. The National Centre for Biotechnology Information Map Viewer, the University of California Santa Cruz Genome Browser, and the European Bioinformatics Institutes Ensemble system are some of major genome browsers which consists of DNA,RNA,Protein,Disease,Expression databases[5][6].

3.6. Comparison Techniques
Sequence comparisons lie at the heart of all bioinformatics. Edit distance is a way of quantifying how similar two strings were. It is the minimum number of operations required to transform one string into the other. Commonly used sequence comparison techniques were:

Multiple Sequence Alignment (MSA)[8]: used for homologous sequence comparisons of three or more biological sequences, generally protein,DNA,or RNA. MSA programs use heuristic methods because identifying the optimal alignment between more than a few sequences of moderate length is computationally expensive. Several high speed multiple sequence alignment program includes MAFFT (Multiple Alignment Using Fast Fourier Transform) can align up to 500 sequences [9].

Needleman-Wunsch algorithm [10] is a smart way to reduce the massive number of possibilities of sequence that are needed to be considered. It is developed by IBM for dynamic programming that runs in O(nm) time. The basic idea is to build up the best alignment by using optimal alignments of smaller sub sequences. It makes use of the concept of longest common substring.

The P-Bigram method [8] is method used to find the similarity problem in DNA sequence by integrating the dynamic programming concept. Editing operations include deletion, insertion,
substitution two characters. This method provided an efficient algorithm that locates all minimum operation in a string. The proposed method was tested on 5 DNA matching datasets with representative two characters based string measures which runs in $O(n^2)$ time.

De Bruijn graph also [11] called Kmer graph, is another efficient way to represent a DNA sequence in terms of its k-mer components. It is based on assembly approach and was proposed to handle the assembly of repetitive regions better. Sequence is constructed from an NGS library, and then the genome is derived from the de Bruijn graph. The reads are k-mers and each one corresponds to one edge in the graph. The vertices are (k-1) mers that appear in some read, and edges defined by overlap of k-2 nucleotides. Small values of k produce small graphs and thus eliminate repetitive sequences and thus remove the computational complexity.

4. Conclusion
The use of Next generation sequencing Data has its own relevance in the field of Bioinformatics to understand the code of life. The high through-put sequence comparison helps to analysis the sequence similarities of human genome. Different platforms are selected based on the application with which we are dealing with. The work flow and data analysis of sequences are gone through different stages in order to determine the precise sequence. Sequencing similarity can be compared by techniques speed up the faster detection of abnormal behavior in human genome. The use of NGS include efficient detection and faster response rate of viral transmission like Hepatitis, tumor etc. A web system called the Global Hepatitis Outbreak and Surveillance technology (GHOST)[3] developed for advanced molecular detection (AMC) for Hepatitis C virus transmission is a remarkable application of NGS. GHOST aims to improve the accuracy and response time of transmission detection by integrating epidemiological evidence. The tool is available to public health laboratories to identify outbreaks by simply uploading viral sequences.

References
[1] F. Sanger, S. Nicklen, and A. R. Coulson 1977 DNA sequencing with chain-termination inhibitors. J. Mol. Biol. 74(1):546-57.
[2] Chien-Yueh Lee, Yu-Chiao Chiu, Liang-Bo Wang, Yu-LunKuo 2013 Common applications of next-generation sequencing technologies in genomic research. Translational Cancer Research Vol 2, No 1
[3] Inna Rytsareva, David S. Campo, Yuei Zheng, Sharma V. Thangachan, 2015 Efficient detection of viral transmissions with Next-Generation Sequencing data. Fifth IEEE International Conference on Computational Advances in Bio and Medical Sciences (ICCABS 2015))
[4] Federica Torri Ivo D. Dinov Alen Zamanyan Sam Hobel Alex Genco Petros Petrosyan, 2012 Next Generation Sequence Analysis and Computational Genomics Using Graphical Pipeline Workflows. Genes (Basel) 2012 Sep doi: Genes (Basel) 3(3): 543575.
[5] Andreas D Baxevanis, 2003, Using Genomic Databases for Sequence-Based Biological Discovery. BMC informatics, Mol Med 2003 Sep
[6] Dong Zou, Lina Ma, Jun Yu, ZhangZhang, 2015 Feb “Biological Databases for Human Research”, Proteomics Bioinformatics 2015 Feb; 13(1): 5563.
[7] Veronique G. LeBlanc, and Marco A. Marra ). 2015 Sep. Next-Generation Sequencing Approaches in Cancer. BMC Bioinformatics (Basel). 2015 Sep: 7(3): 19251958.
[8] Patsaraporn Somboonsak, Mud-ArmeenMunlin , 2011, AA New Edit Distance Method for Finding Similarity in Dna Sequence, World Academy of Science, Engineering and Technology International Journal of Biological, Biomolecular, Agricultural, Food and Biotechnological Engineering Vol:5, No:10
[9] JurateDaugelaitė Aisling O, Driscoll and Roy D. Sleator, 2013 An Overview of Multiple Sequence Alignments and Cloud Computing in Bioinformatics. BMC Genomics Volume 2013 , Article ID 615630, 14 pages
[10] http://vlab.amrita.edu/?sub=3brch=274sim=1431cnt=1
[11] http://www.homolog.us/Tutorials/index.php?p=1.4s=1
[12] https://en.wikipedia.org/wiki/DNA sequencing