Special Proceedings of 20th Annual Conference of SSCA
held at Pondicherry University, Puducherry during January 29-31, 2018; pp 27-34

Statistical Methods for Next Generation Sequence Data – An Introduction

P. Venkatesan
Department of Statistics, ICMR National Institute for Research in Tuberculosis, Chennai

Final Version Received on August 28, 2018

Abstract:

Next Generation Sequencing (NGS) is the latest high throughput technology which revolutionized genomic research. NGS methods are highly parallelized enabling to sequence thousands to millions of molecules simultaneously. This technology results into huge amount of data, which need to be analyzed to conclude valuable information. To extract signals from high-dimensional NGS data and make valid statistical inferences and predictions, novel data analytic and statistical techniques are needed. Analysis of NGS data unravels important clues in quest for the treatment of various life-threatening diseases and other scientific problems related to human welfare. This paper presents a brief review of approaches to statistical methods and models for working with NGS data. The topics range from basic preprocessing and analysis with NGS data to more complex genomic applications such as copy number variation,expected to deal with genomic data in basic biomedical research, genomic clinical trials and personalized medicine.

Key words: Denova assembly, alignment, base calling algorithm, high throughput data, gene expression,variant detection, HMM, Poisson, inflated Poisson, negative binomial.

1. Introduction

Deoxyribonucleic acid (DNA) is the carrier of genetic materials for all living organisms and many viruses. It is the most essential component of chromosomes and plays an important role in developing and functioning organisms. It consists of four kinds of nucleotides: Adenine (A), Cytosine (C), Guanine (G), and Thymine (T). The important role of DNA leads people to explore and research on DNA. This gives rise to the rapid development of DNA sequencing technology. DNA sequencing is a process for determining the exact type and order of nucleotides for a fragment of genome or the whole genome. Evolving from the traditional sequencing technology, Sanger method (Sanger &Coulson, 1975) to the currently widely used next-generation technologies (NGS) (Metzker, 2010; Mardis, 2013) and the next next-generation sequencing technologies (next-NGS), DNA sequencing technologies are rapidly developing and moving towards to the direction with low-cost, high-speed and high-accuracy.

With the help of DNA sequence, researchers can illuminate genetic information from a biological system. Deciphering DNA sequences is necessary for almost all branches of life sciences and its understanding has grown exponentially in the past decades. The first-generation sequencing method, called Sanger sequencing, was developed by Edward Sanger in 1975. Sanger sequencing was considered as the gold standard for DNA sequencing for

Corresponding author: P. Venkatesan
E-mail: venkaticmr@gmail.com

around three decades (Sanger etal, 1977). The first major breakthrough of first-generation sequencing was the Human Genome Project (HGP), 13-years long, $3 million project, completed in 2003. Due to inherent limitations in throughput, speed, scalability and resolution of first-generation Sanger sequencing approach, second-generation of sequencing method, or Next-Generation Sequencing (NGS) has been developed to cater high demand for cheaper as well as faster sequencing technology. The NGS, also known as Massive Parallel Sequencing, allows the complete genome (of human) to be sequenced in less than a day. More recently, third generation sequencing (TGS) has evolved(Raza& Ahmed 2017).

2. NGS Capabilities

The NGS is a powerful, flexibly, indispensable and universal biological tool that infused several areas of biological study. Some of the promises and capabilities of NGS are(Grada&Weinbrecht 2013;Lee et al 2011, Raza& Ahmed 2017):

- High-throughput sequencing (HTS) of the human genome lets us discover genes and regulatory pathways associated with disease.
- Targeted sequencing of specific genes or genomic region helps in the identification of disease-causing mutations. It helps in the faster diagnosis, and outcome of disease-targeted sequencing may help in better therapeutic decision-making for several genetic diseases, including many cancers.
- RNA-Seq (NGS of RNA) provides entire transcriptomic information of a samplewithout any need of previous knowledge related to genetic sequence of an organism.
- RNA-Seq provides a strong alternative approach to Microarrays for gene expression studies, and let the researchers visualize RNA expression in the form of sequence.
- Variant study is common in medical genetic, where DNA sequence and data are compared with a reference sequence to catalogue the differences. These differencesmay range from single nucleotide polymorphisms (SNPs) to complex chromosomal rearrangement.

2.1. NGS Platforms

High throughput next-generation sequencing (NGS) technologies arecapable of generating massive amounts of data in the form of paired-end orsingle-end reads with either fixed or variable lengths. The size of data files isoften in the magnitude of mega- or giga-bytes (up to 1000 giga base pairs or Gbin a single sequencing run) and is likely to further increase in the coming years.While sequencing costs have dropped precipitously and sequencing speed and efficiency have raised exponentially, the development of computational tools for preliminary analysis of these gigantic datasets have lagged compared to the data generation. Hence, there is an increasing demand for efficient and user-friendly tools.

The generations of NGS is classified as follows:

1) First generation sequencing technology:
    a) Sanger Sequencing
2) Second generation sequencing technology
    a) Roche - 454
    b) Illumina – GA II
    c) SOLiD
3) Third generation sequencing technology
    a) Helicose
    b) PacBio

c) Illumina = HiSeQ, MiSeq
d) Ion Torrent
e) Oxford Nanopor

There are four predominant commercially available next-generation sequencing technology: Roche/454, Illumina/Solexa, AB/SOLiD and ABI/Ion Torent. Roche 454 makes the longest read and also the fastest method with high accuracy; Illumina GA also has a wide range of read length from 50bp to 250bp and high throughput with median running speed, while AB SOLiD utilizes a shortest read. A comparison of these methods is discussed by Liu et al(2012) and Mardis (2013).These massively parallel DNA sequencing technologies have been applied to transcriptome sequencing (RNA-Seq), *de novo* genome sequencing, and genome re-sequencing. RNA-Seq is a widely used approach totranscriptomic profiling (Martin and Wang, 2011). The Next-NGS of third generation sequencing is the recent development. Helicos single molecule sequencing technologies, is SMRT technolog from Pacific Bioscience and single-molecule nanopore sequencing technology from Oxford Nanopore Technologies Company. This sequencing technology is in the direction of high-throughput, low cost, and long read length. However, the most commonly available databases is from Illumina and this paper concentrates only on illumina platform.

2.2. NGS Read Types

In DNA sequencing technology, there are three types of reads: single-end reads, paired-end reads, and mate pair reads. Single-end reads are the result of sequencing one end of the fragments, while paired-end reads and mate pair reads obtain both ends of the DNA fragments while sequencing. The difference between paired-end and mate pair refers to how they make the sequencing library and how the DNA fragment is sequenced.

FASTA File Format*:* FASTA is a standard text-based format for sequencing. Each sequence contains two lines. The first line starts with a '>' character and is followed by the sequence identifier and/or description. The second line is the sequence containing A, C, G,T, or N (unknown base) (Fig.1).

```
@sequence_id
GATTCCTGTAAGCTTAAAGCTCCATTGTACCCG
ATATACGCCTTT
```

Figure 1: FASTA File Format

FASTQ File Format*:* Although the nucleotide is determined by collecting the fluorescence signal, the final sequence output is in another widely used file format called FASTQ (Cock et al., 2010). FASTQ format is a text-based file format. It contains all of the nucleotide sequences and its corresponding quality scores. Next followsan example of the FASTQ format. FASTQ format adopts four lines to represent a sequence. The first line starts with "@" character and is followed by the sequence identifier. The second line is nucleotide sequences letters. The third line begins with a "+" character and optional description. The fourth line is the quality score(Fig.2). Each score represents the quality of its corresponding base in the first line. Therefore, the number of qualities should be the same as the number of letters in the sequence. The quality score for the Illumina GA platform can be calculated by the following formula:

$$Q_{solexa(prior\ to\ v.1.3)} = -10\log_{10}(p/(1-p))$$

where p is the probability that the corresponding base is incorrect.

```
@sequence_id
GATTCCTGTAAGCTTAAAGCTCCATTGTACCCG
ATATACGCCTTT
+
&??#55CCFF%%>>>>>6615%%+++***09@??=><
<=++@@AB
```

Figure 2: FASTQ File Format

2.3. NGS Assembly

Sequence assembly is to merge some short DNA sequence reads with certain overlapping bases into a longer DNA sequence in order to reconstruct the original structure of DNA. This process is vital because current sequencing technologies are unable to sequence the whole genome at one time. The whole genome needs to be cut into small fragments and then sequenced. There are two different types of assembly: de-novo assembly and mapping assembly. De-novo assembly is assembling short reads to create longer sequences, while mapping assembly is assembling reads to an existing backbone sequence template and then building a similar sequence as the backbone. The process of de-novo assembly is explained in the following Fig 3. S1, S2, S3, and S4 are four short sequence reads with overlapping. The longer sequence, namely "contig," can be obtained by assembling these four reads.

```
S1          ACCTGTTA
S2              TGTTACCA
S3                  ACCAGATA
S4                      ATACGCGG

Contig    ACCTGTTACCAGATACGCGG
```

Figure 3: De-novo Assembly

There are number of assembly tools that are freely available. Some of the important free assembly softwares are MIRA, SOAPdenovo, ABySS, EULER, RAY and commercial softwares are CLC, Newbler, etc.

2.4. NGS Sequence Alignment

Sequence alignment is to compare the similarity of two sequences. The theoretical basis of sequence alignment is Darwin's theory of evolution. If two sequences share high similarity, they are speculated to evolve from the same ancestor through the process of nucleotide replacement, sequence fragments, and missing and genetic variations. In sequence alignment,two or more sequences are put together in a way that the same nucleotide bases are aligned in the same column. Occasionally, gaps are inserted into the sequence in order to obtain the best alignment result. There are a number of alignment tools, most of which utilize one of the alignment algorithms: Needleman-Wunsch algorithm and Smith-Waterman algorithm. These two algorithms are both based on dynamic programming with the difference that Needleman-Wunsch algorithm is a global alignment technique, whereas Smith-Waterman algorithm is a general local alignment method.Widely used alignment software includes BLAST, BWA, MOSAIK, BFAST, Bowtie, SOAP and SSAHA.

Prior to the in-depth analysis of NGS deep sequencing data (differentialgene expression and alternative splicing analysis for RNA-Seq studies, structural variants

identification for genome re-sequencing studies, and genome assemblyfor *de nove*genome sequencing studies), the major concern is about the following issues: (1) basic statistics of a sequencing run such as total numbers of raw, cleaned, and unique reads as well as the degree of reads redundancy; (2) sequencing library quality, i.e., whether the library truly represents the genome of the re-sequencing organism, and (3) the number of sequencing runs required, i.e., how many runs are necessary to attain a full representation of the sequencing library or to suffice a *de-novo* genomeassembly. Further, the primary data is usually discarded soon after run. The secondary and tertiary data maintained on fast access disk during analysis, then moved to slower access disk afterwards.

3. NGS Data Analysis

The NGS genomic data analysis classification is given in Fig.4. The analysis is based on data from genomics, transcriptomics and epigenomics studies. At the genomic level NGS data, we look for point mutations, smallindels, copy number variation and structural variation. At transcriptomics level NGS data, we look for differential gene expression, gene fusion, alternatie splicing and RNA editing. The epigenomics NGS data is used for methylation, histone modification and transcription factor binding. The general flow of analysis is given in Fig.5. To answer a specific biological question, the huge volume of raw data is to be preprocessed, assembly/alignment algorithms should be used to get the sequence and the comparisions should be made to answewr the question.

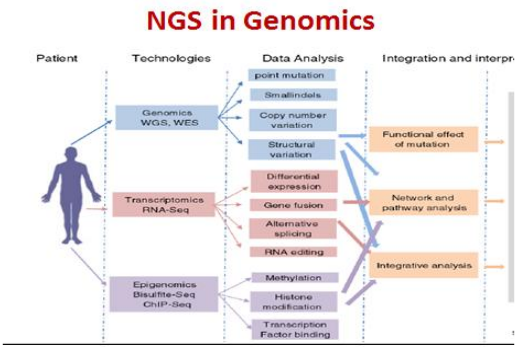

Figure 4: NGS Data Analysis(Wu et al.2012)

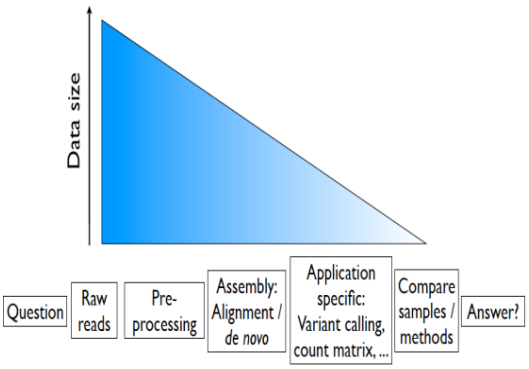

Figure 5: Data Size

The NGS data analysis is carried out in three stagesas given in Fig.6. The first stage is the primary analysis. The Illumina Genome Analyzer sequencing system produces image files as primary output from a sequence run. The conversion of these images into sequence files require the following steps:

Image analysis: where clusters are located on the image and cluster intensities, positions and noise estimates are calculated.

Base calling: where the sequenceof bases is read from each cluster, a confidence level for each base is calculated and read filtering is performed.

Sequnce analysis: where files in FASTQ format are generated. The second stage analysis is application specific about the alignment/assembly. The workflow for alignment and assembly are given in Fig.7(a) and 7(b) and resulting sequence in Fig.8(a) and 8(b). The third stage focues on tertiary analysis to answer the scientific question.

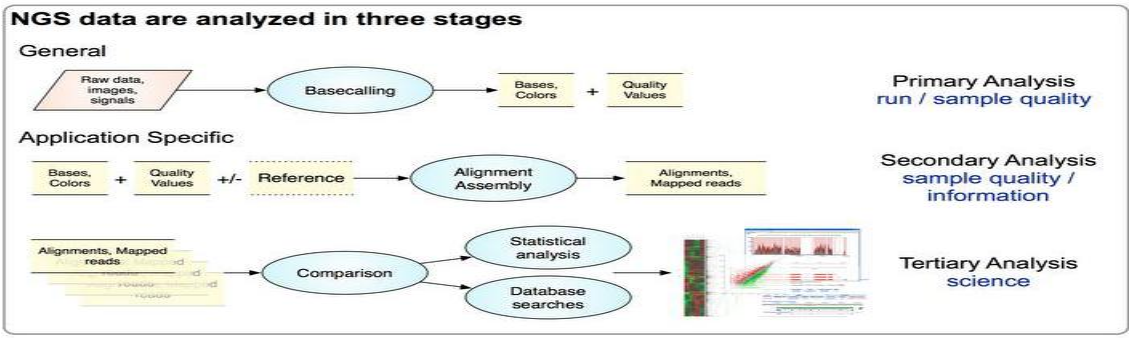

Figure 6: Stages of NGS Data Anmalysis (Sanchez- 2009)

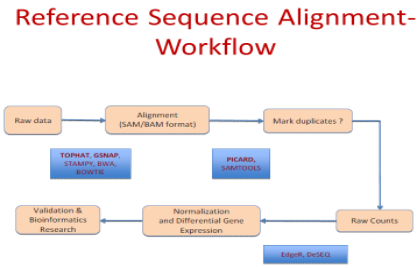

Figure 7(a): Alignment workflow

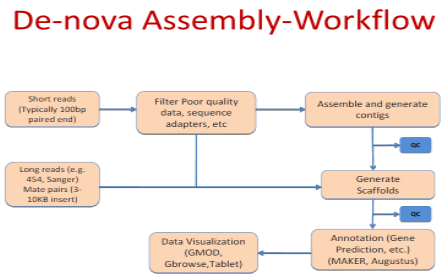

Figure 7(b): Assembly workflow

Figure 8(a): Matching (Wu et al. 2012)

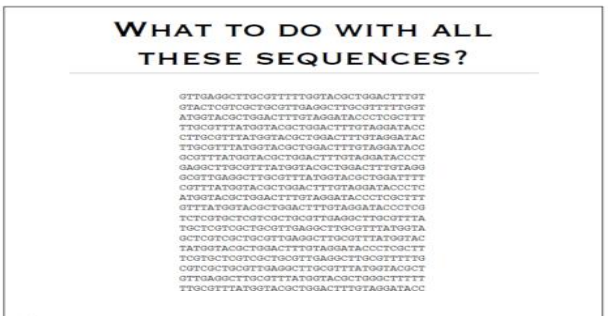

Figure 8(b): Aligned sequence

4. Biological Questions

Many biological questions can be asswered more confidently using NGS. In this paper we restrict our attention only to four questions one from each important areas. Different matching softwares have varied speed. For example, for matching 10,000,000 32BPS reads the BLAST takes more than six moths, BLAT takes two months, MAQ takesone and half days where as Bowtie takes about 17 minutes even though the underlying algorithm Borrows-Wheeler is almost similar. Bowtie is the most commonly used matching algorithm.

***Mapping*:** The first question is to determinethe read point of origin with respect to the reference. Given a read, we can find many matches in the reference sequence. We formulate as a testing problem as illustreated in Fig.9(a). The hypothesis with minimum mismatches and gaps is accepted.

SNPs: The next important application is the identification of SNPs using NGS data. The reads given in Fig.8(a) are arranged as shown in Fig. 9(b) to identify the SNP.

Differential Gene ***Expression*****:**RNA-seq is a methodology for RNA profiling based on NGS that enables to measure and compare gene expression patterns at an unprecedented way much efficient than microarray based methods. Fig.10(a) gives the methodology to find differentially expressed genes.

Variant detection**:** NGS technology makes it possible to comprehensively catalogue genetic variation in population samples, creating a foundation for understanding human disease, ancestry and evolution. A simplified method is given in Fig.10(b)

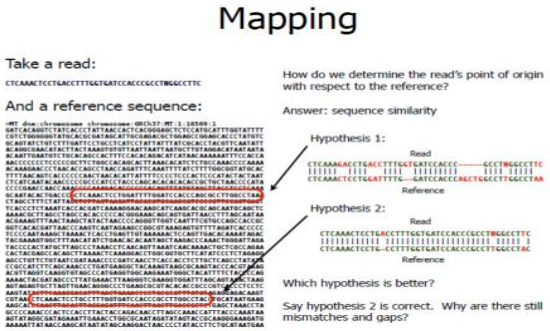

Figure 9(a): Mapping(Wu et al.2012)

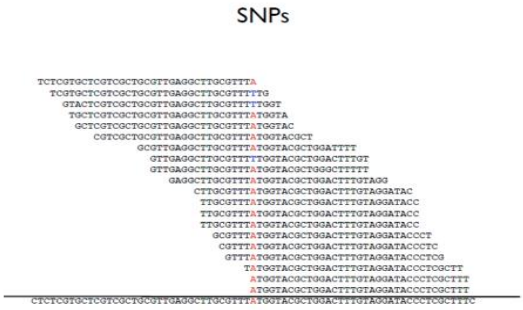

Figure 9(b): SNP Analysis

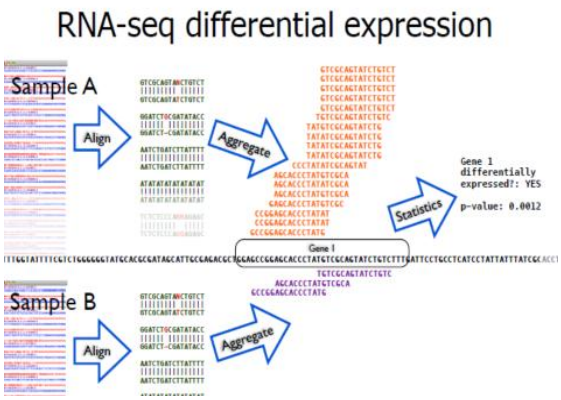

Figure 10(a): Gene expression (Wu et al 2012)

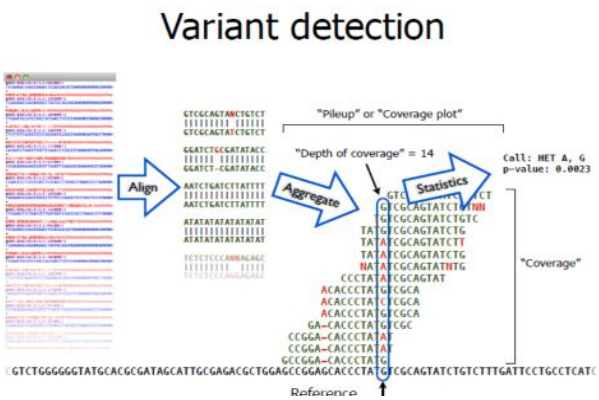

Figure 10(b): Variant detection

5. Statistical models

Since NGS data are basically count data, only the discregte models are used for analysis. The most important models are the Poisson, Inflated Poisson, Generalized Poisson, Negative binomial. The hidden Markov models (HMM) and hierarchical HMM the models most commonly used. However, the NGS data presents a significant challenge for data storage, analyses, and solutions. Average NGS experiment generates gigabytes to terabytes of raw data, making data analysis and management of data problematic. NGS sequencing benefits will not be fully appreciated until extremely high-performance computing and intensive bioinformatics support is available. Efficiently alignjng short reads to a reference genome, especially when developing new algorithms to handle ambiguities is a challenge. Compared to reference-based assembly, de novo assembly is even more challenging. Software development with the ability to distinguish low-frequency alleles descendent from

ancient ancestors from de novo or extremely rare mutations is a challenging problem. Many programs are open source and additional programming may be needed to modify the program to the needs of a specific NGS project

6. Discussion

The availability of high quality data NGS datastill throughput data, does not preclude biological replicates particularly in an swering genetic questios. The high dimensionality of data makes direct use of classical statistical techniques difficult. The success comes mostly from machine learning approached and Bayesian approaches. In this article, an introduction to statistical methods for processing, variation discovery and other aspects are introduced and discussed. The impending arrival of yet more NGS technologies makes even more important modular, extensible frameworks that produce high-quality variant and genotype calls despite distinct error modes of multiple technologies for many experimental designs.

Acknowledgements
All sources quoted in this article are duly acknowledged.

References:

Cock, P.J.A., Fields, C.J., Goto, N., Heuer, M.L. and Rice, P.M. (2010).The Sanger FASTQ file format for sequencequality score and the Solexa/Illumina FASTQ varients. *Nucleic Acid Research*, **38**, 1767-71.

Grada, A. and Weinbrecht, K. (2013). Next-generation sequencing: methodology and application. *Journal of Investigative Dermatology*, 133:e11; doi:10.10*38/jid.2013.248.*

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L. and Law, M.(2012). Comparison of next generation sequencing systems. *Journal of Biomedicine and Biotechnology*, ID 251364, 1-11.

Lee, H.C., Lai, K., Lorenc, M.T., Imelfort, M., Duran, C. and Edwards, D. (2011). Bioinformatics tools and databases for analysis of NGS data. *Briefing in Functional Genomis*, **11**, 12-24.

Martin, J.A. and Wang, Z. (2011). Next-generation transcriptome assembly. *Nature Reviews Genetics*, **12,** 671-682.

Mardis, E.R. (2013). Next-generation sequencing platforms. *Annual Review of Analytical Chemistry*, **6**, 287-303.

Metzker, M.L. (2010). Sequencing technologies-the next generation. *Nature Reviews Genetics*, **11**, 31-46.

Raza, K. and Ahmed, H. (2017). Recent advancement in next generation sequencing techniques and its computational analysis. *International Journal of Bioinformatics Research and Applications,***13**,1-14.

Sanger, F. and Coulson, A.R. (1975). A rapid method for determing sequences inDNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology*, **94**, 441-48.

Sanger, F., Nicklen, S. and Coulson, A.R. (1977). DNA sequencing with chain terminating inhibitors. *Proceedings of the National Academy of Sciencesof the United States of America*, **74**, 5463-5467

Wu, Z., Hansen, K. and Irizarry, R.A. (2012). Statistical methods for next generation sequence. 1-48 (http://www.biostatjhsph.edu/~khansen/enar2012html)