Informatics in Medicine Unlocked 11 (2018) 75–82

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: www.elsevier.com/locate/imu

Computational analysis of next generation sequencing data and its applications in clinical oncology

Rucha M. Wadapurkar, Renu Vyas*

MIT School of Bioengineering Sciences and Research, MIT ADT University, Raj Baugh Campus, Loni Kalbhor, Pune 412201, Maharashtra, India

ARTICLE INFO

Keywords:
Next generation sequencing
Mutations
Cancer
Sanger sequencing
Variant identification and annotation
Data analysis

ABSTRACT

Next generation sequencing (NGS) has made great strides in sequencing technology as it enables sequencing of genes in a high throughput manner with low cost. Various NGS platforms such as Illumina, Roche, ABI/SOLiD are used for wet-lab analysis of NGS data and computational tools such as BWA, Bowtie, Galaxy, SanGeniX are used for dry-lab analysis of NGS data. One of the important aspects of NGS data is its usage in early disease diagnosis especially in cancer which was earlier not possible with conventional sequencing technologies such as Sanger sequencing, NGS can identify all those mutations which cannot be identified using conventional sequencing technologies as researchers can now sequence the whole genome, exome or transcriptome. Exome sequencing is preferred, as a higher number of mutations are found to exist in the exome part of genes. The present comprehensive review encompasses the complete NGS data analysis workflow that includes alignment of NGS reads, identification and annotation of mutations and visualization, discussion of software tools for variant identification and annotation, evaluation of structural variation in NGS data, and study of different DNA sequencing technologies. In the field of clinical oncology, NGS has already proven its usefulness, and the mortality rate has been reduced, as now doctors can suggest a proper treatment to their patients by checking the complete genomic profile. However, data storage and the complexity in interpreting enormous amounts of data obtained with NGS still remain a computational challenge to researchers, as for each sample, the number of different and very large analysis files are generated directly from the raw sequence read file to the final result file. NGS resultant data is very complex, and its interpretation requires expert bioinformatics assistance, as a large number of mutations are identified from samples, but to differentiate clinically significant mutations among them with appropriate use of validation methods is a challenging task. This review is intended to provide researchers with a complete overview of NGS along with knowledge of how the tools will be employed, and insight into identification and interpretation of cancer mutations for clinical diagnostics.

1. Introduction

Next generation sequencing (NGS) has created a noteworthy paradigm shift in the clinical diagnostic field. It refers to an aggregate collection of methods in which various sequencing reactions occur at the same time, bringing about vast amounts of sequencing data for a little division of the cost of Sanger sequencing. With the help of NGS methods, base-pair level sequencing of whole genome or exome can be performed with minimum errors and at a lower cost. The Human Genome Project had determined the sequence of ~3 billion base pairs and identified around ~25,000 human genes based on the principle of Sanger sequencing, which had given rise to the release of the human reference genome [1–3]. A couple of years after the fact, as sequencing methods turned out to be more exact and affordable; the 1000 Human Genome Project [4] was launched for sequencing of 1092 human genomes that were published. The identification of genetic variations among individuals sampled from 14 populations has been carried out by the International HapMap Project which has utilized data from six distinct nations in Europe, East Asia, sub-Saharan Africa and the Americas [5]. Visualization of different plant species around the globe was the core part of 1000 Plant Genome Project [6]. Approximately ~125,000 species out of 370,000 green plants were recorded as gene entries in GenBank by 1000 Plant Genome Project [7]. With the 10,000 Genome Project [8], the Genome 10K Community of Scientists (G10KCOS) intend to make an accumulation of DNA specimens and tissues for 10,000 vertebrate species particularly assigned for WGS i.e. whole genome sequencing. Using next generation sequencing (NGS), Cheng JH and co-workers have identified 118 target genes for miRNAs of interest in articular cartilage and 214 target genes in subchondral bone [9]. Apart from this, a number of biomarkers in practice today are

* Corresponding author.
E-mail address: renu.vyas@mituniversity.edu.in (R. Vyas).

https://doi.org/10.1016/j.imu.2018.05.003
Received 26 March 2018; Received in revised form 15 May 2018; Accepted 15 May 2018
Available online 17 May 2018

clinically validated and detected by NGS testing e.g., BRAF mutations in melanoma, EGFR mutations and ALK fusions in NSCLC. Through NGS we can screen a broad range of genes in a single test utilizing scarce biopsy tissue of patients [10]. Several cancer related mutations have been identified using NGS e.g. mutations in tumour suppressor gene p53 or in one of the RAS proto-oncogenes. HRAS [Harvey rat sarcoma viral oncogene homolog], KRAS [Kristen rat sarcoma viral oncogene homolog] and NRAS [neuroblastoma RAS viral (v-ras) oncogene homolog] are used as biomarkers for lung cancer, ovarian cancer and breast cancer [11].

The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC) were launched in 2005 and 2008 respectively to comprehensively understand cancer genetics. TCGA is a comprehensive atlas of cancer genomic profiles and catalogue major cancer-causing genome alterations in around 30 human tumours generated through genome sequencing. Genomic abnormalities in 50 different cancer types are made available through the ICGC data portal [12]. Recently, Oxford Nanopore Technologies have successfully sequenced the human reference genome for the GM12878 Utah/Ceph cell line using the MinION nanopore sequencer. 91.2 GB of sequence data has been produced, the alignment of which has detected large structural variants and epigenetic modifications [13]. In the future, NGS will take into account the accurate pan-genomes detection to depict the complement of a considerable number of genes in strains of species, commonly connected to archaea and bacteria [14].

The present review describes the past five years of research in next generation sequencing. The reviews published previously [15,16] were focused on particular aspects of NGS, the present review attempts to provide readers with a holistic approach to understand NGS. According to different technological and usage aspects, the review is divided into seven sections viz. sequencing technologies, sequencing and assembly of DNA, NGS data analysis workflow, the role of NGS in variant identification, cancer genomics, transcriptomics and proteomics, handling of NGS data and the current scenario in NGS.

1.1. A brief overview of sequencing technologies: first, second and third generation

There are three generations of sequencing technologies that have evolved so far, (Fig. 1).

1.1.1. First generation

Sanger sequencing belongs to the first generation technologies which was invented in 1977. It is based on the principle of DNA chain termination, which labels dideoxynucleotide triphosphates (ddNTPs) with four fluorochromes. The labeled fragments are then separated by gel electrophoresis and the base identification is carried out using fluorescence detection. This method generates long contiguous sequence reads (> 500 nucleotides) and currently it is being used for validating results of NGS studies. It is a time consuming process and can sequence only a few thousands of nucleotides in a week.

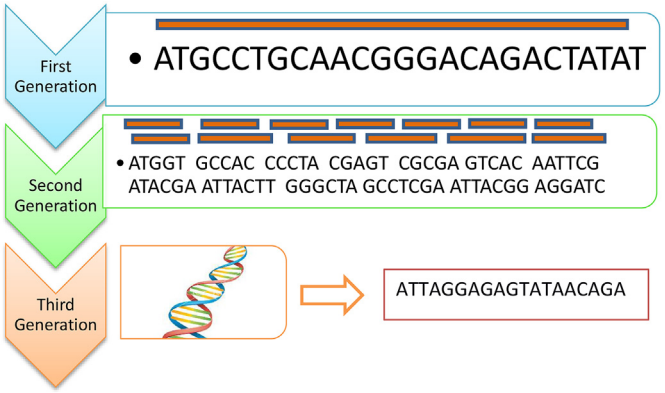

Fig. 1. A schematic depicting evolution of NGS technologies.

1.1.2. Second generation

Using Second generation sequencing, which is also known as a high-throughput sequencing or Next generation sequencing technology, thousands or millions of short sequence reads can be created at a very high speed, with more exactness and within a few hours [17]. These have been created by Illumina [18,19], Roche 454 [20] and Biotechnologies/SOLiD [21]. Researchers can sequence more than five human genomes at ~30x coverage all the while or ~100 exome samples in a single run with the most widely used Illumina platform, as it produces a large number of sequence reads with high precision [22].

1.1.3. Third generation

Third generation sequencing can sequence a human genome requiring little to no effort within a matter of hours and is under development currently. It sequences genes at the molecular level i.e. instead of performing DNA fragmentation and sequencing by amplification and synthesis used by second generation sequencing methods, it can sequence a single DNA molecule. Pacific Biosciences (http://www.pacificbiosciences.com/), Helicos BioSciences (http://www.helicosbio.com/), Complete Genomics (http://www.completegenomics.com/), and Oxford Nanopore (http://www.nanoporetech.com/) are involved in the third generation of sequencing technologies [23,24].

Today, original sequencing is not utilized because of its prohibitive cost and time utilization, though second generation sequencing technologies are utilized more due to their low cost and time proficiency. These advanced technologies have made a substantial reduction in time and cost of sequencing since last few years. The graph shown in Fig. 2 depicts DNA sequencing data of last two decades which we extracted from the site of National Human Genome Research Institute [25].

1.2. Applications of NGS

NGS technologies have many applications such as DNA-sequencing and assembly to determine an unknown genome without any preparation or search for variations among genome samples, RNA-sequencing [26,27], to analyze gene expression [28] and to predominantly identify DNA regions of DNA binding proteins, for example, transcription factors etc. The most important application of NGS is in identifying mutations. Commonly, short i.e. 50–250 bp NGS reads are initially mapped to a reference genome and after that from the mapped data, variations are detected. While most of the NGS applications concentrate on identification of single nucleotide variations (SNVs) or small insertions/deletions (indels), structural variation including translocations, bigger indels, and copy number variation (CNV) can also be recognized from similar data. Structural variation discovery can be performed from whole genome NGS data or “targeted” data including exomes or gene panels. While targeted sequencing incredibly increments sequencing coverage or depth of specific genes, it might present predispositions in the data that require particular computational analysis. Since the past few years, there have been extensive advances in methods used to identify structural variations and a full coverage of variations from SNVs; balanced translocations to CNV can now be identified with reasonable sensitivity from either whole genome or targeted NGS data. Such methods are connected to clinical testing where they can supplement fluorescence *in situ* hybridization or array-based testing. The identification of structural DNA variation has since quite a while ago assumed a part in the diagnosis of cancer and Mendelian disorders, originating before the approach of current DNA sequencing [29,30]. Structural DNA variation is found in a DNA region larger than 1 kb and incorporates a few classes, for example, translocations, inversions, insertions/deletions (indels) and copy number variations (CNVs) [31].

NGS-based diagnostics implement some portion of the clinical genomic testing in which a limited set of genes are targeted and not the