ISSN No: 1008-0562

Identification of Cancer Mutation Using Whole Genome Sequencing

[1]Dr. M. Sheerin Banu, [2]P.R.Kavitha, [3]R.Keerthana, [4]Sabrish Surender

[1]Professor and Head, Department of Information Technology, R.M.K Engineering College.

[2]Final Year B.Tech Information Technology, R.M.K Engineering College.

[3]Final Year B.Tech Information Technology, R.M.K Engineering College

[4]Head, Center of Excellence Google Cloud platform & AI,Virtusa Software Services Pvt.Ltd.

Abstract

Genome Sequencing is the most promising area of research that helps to improve understanding of disease causes and treatments. Whole genome sequencing involves raw genome data as input that contains several genome information.The raw data is initially analyzed to remove unwanted data reads in order to obtain the mutation sequence .The analysis involves several quality checks to make sure that the mutation sequence thus obtained does not contain duplicate data reads.This mutation sequence is further used for predicting the disease type and appropriate treatment to be provided to the patients.

Keywords: *Genome sequencing,Mutation sequence,Disease prediction*

I. Introduction:

Whole genome sequencing has been considered as a research tool, and it is presently being introduced to clinical sequencing purposes. The Whole Genome sequencing is generally connected to **"decoding,"** but a genome sequence is still considered in code. In a clear-cut manner, a genome sequence is generally a long string of letters in an inscrutable language. **Genome Sequencing** allows researchers to extract the genetic information that is obtained in the DNA of humans, animals and even plants.

When we read a sentence, it contains words that give different meaning for each and every word in the sentence. Similarly, the human genome has more information in it. Whole-genome sequencing is an extensive technique for analyzing. Human genome is a larger sequence and it contains both wanted and unwanted information for analysis. For this we use genomic information for identifying the inherited disorders, finding the mutations and identifying the disease. The sequencing cost is rapidly dropping and with the use of those large volumes of data with today’s sequencers makes whole genome sequencing a powerful tool for genomic research.

II. Probabilities and Challenges in Whole-Genome Sequencing:

Determining the future using the genome sequence is one of the aspects. It is useful for individuals or to the physician or to the scientist. The normal genome or reference genome considers the value related to the point and it is improved by the study of whole genome sequencing. The understating of the genome information and it associate with the value and the complexity factor and the biological information and other details. In human Cell, genetic functions, networks and pathways are superior and improving. It is very large enough. The individual genome value is very limited during translation, if the individual is

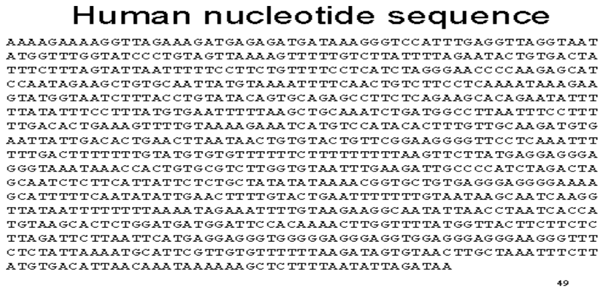
ISSN No: 1008-0562

healthy. On the other hand, the combined variant analysis of the affected individual genome identically appealed in finding the appropriate variant.

The NGS technology is the most authoritative application for whole genome sequencing in cancer genomics for studying the mutation and normal genome. At first, the method will identify genome wide detection and changes in the characterization in hundreds of cancer genomes. The use of DNA sequencing in the clinical laboratory to identify the mutations in the cancer genome that gives rise to one or more particular therapies. By the genome sequencing technology, identifying the mutation status for the particular diagnosis and this method helps for the oncologist to make decisions and they will give the appropriate treatment for the particular individual.

In the initial stage of the whole genome sequencing NGS platform is not used for clinical DNA sequencing. This is due to the library framework complication and operational instruments and it does not tell the data analysis uncertainty and the difficulty. Before communicating results to the physician and treating them, the whole genome sequencing and analysis must be performed. The production and analysis of data must be accomplished by the clinical laboratory with important experts and specialists will verify and approve the laboratory for using different methods. If the instruments of genome sequencing are not appropriate to the clinical sequencing, it leads to finding any other “third- generation sequencing” platform suited. The sequencing platform is either used at the initial stage of commencement or under development. Sequencing single molecules of DNA in a single plex or in multiple fashion and identifying the real time sequence data. These instruments have the run time measured in minutes or hours and not in days and these instruments require small sample preparation before sequencing. Both the features can be applied to clinical laboratory instrumentation.

The genome sequencing generally involves deciding the order of bases, nucleotide sub-nets adenine(A), guanine(G), cytosine(C) and thymine(T), found in DNA nucleotides. Finding the mutation sequence in the original DNA sequence is difficult. It takes hours to complete the processing.



Human nucleotide sequence

```
AAAAGAAAAGGTTAGAAAGATGAGAGATGATAAAGGGTCCATTTGAGGTTAGGTAAT
ATGGTTTGGTATCCCTGTAGTTAAAAGTTTTTGTCTTATTTTAGAATACTGTGACTA
TTTCTTTAGTATTAATTTTTCCTTCTGTTTTCCTCATCTAGGGAACCCCAAGAGCAT
CCAATAGAAGCTGTGCAATTATGTAAAATTTTCAACTGTCTTCCTCAAAATAAAGAA
GTATGGTAATCTTTACCTGTATACAGTGCAGAGCCTTCTCAGAAGCACAGAATATTT
TTATATTTCCTTTATGTGAATTTTTAAGCTGCAAATCTGATGGCCTTAATTTCCTTT
TTGACACTGAAAGTTTTGTAAAAGAAATCATGTCCATACACTTTGTTGCAAGATGTG
AATTATTGACACTGAACTTAATAACTGTGTACTGTTCGGAAGGGGTTCCTCAAATTT
TTTGACTTTTTTTGTATGTGTGTTTTTTCTTTTTTTTTAAGTTCTTATGAGGAGGGA
GGGTAAATAAACCACTGTGCGTCTTGGTGTAATTTGAAGATTGCCCCATCTAGACTA
GCAATCTCTTCATTATTCTCTGCTATATATAAAACGGTGCTGTGAGGGAGGGGAAAA
GCATTTTTCAATATATTGAACTTTTGTACTGAATTTTTTTGTAATAAGCAATCAAGG
TTATAATTTTTTTTAAAATAGAAATTTTGTAAGAAGGCAATATTAACCTAATCACCA
TGTAAGCACTCTGGATGATGGATTCCACAAAACTTGGTTTTATGGTTACTTCTTCTC
TTAGATTCTTAATTCATGAGGAGGGTGGGGGAGGGAGGTGGAGGGAGGGAAGGGTTT
CTCTATTAAAATGCATTCGTTGTGTTTTTTAAGATAGTGTAACTTGCTAAATTTCTT
ATGTGACATTAACAAATAAAAAAGCTCTTTTAATATTAGATAA
                                                     49
```

FIGURE 1. Sample Human Genome

III. Whole Genome Sequencing Analysis on Cancer Genomics:

Next generation sequencers are rapidly increased and arithmetic analyses have the study of the protein changed mutations and it contains the details of the cancer types. Although, there is a bounded knowledge in somatic mutations in the regions that are not coded. The somatic variation contains only limited information of the cancer types. But the whole genome sequence approach contains all the information and genomic alterations of cancer. It helps us to understand mutations and its signatures and implications etc… This review describes the technologies for cancer identification and mutation interpretation and

ISSN No: 1008-0562

precision medicine. The analysis of large-scale WGS is required for non-coding and structure variants and it is integrated with RNA-sequence, immune-genomic, and some clinic-pathological information. The WGS sequence is a large sequence and it contains all information for analyses. The sequence contains all the information and it is used for cancer identification and precision medicine. The sequence contains some unrelated data and other information. It must be removed during analysis.

IV. Problem Solution:

A. Index Reference Genome:

Before the start of the analysis, we first need to download the human reference genome and index it with the tools that will be aligning the short-read sequences to it. The genome reference used here is human (hg19) chromosome 20 from UCSC. This indexing step only needs to be done once. There are a lot of different DNA aligners out there.

B. Quality control metrics of short reads:

One of the first things you need to do before doing any sort of alignments is to check the quality of the short reads you received from your sequencing facility. A great tool that runs a variety of quality metric checks is **Fast QC.**

C. Trim short reads:

Trimming poor quality bases from your short reads is beneficial. Poor quality bases can cause misalignments or cause the short read to not align to the reference genome. The advantages of Trim Galore! Is that you can remove bases from either side of the short read, that fall below a certain quality threshold, remove primer adapter contamination, as well as remove biased positions, and much more. Trim Galore! Creates new fast qc files that are the final product of the trimming options specified above. A trimming report file is created that gives a description of which primer adapter was found and removed. We set Trim Galore to run FastQC on the trimmed files and so it will have html and zip files for each of the samples. It is wise to check the QC metrics on the trimmed Fast QC files to make sure everything looks better than the raw data.

D. Assembly of sequencing reads:

The genomes are sequenced in different lengths of DNA fragments and the other remaining sequence must be put back together. This is called “assembly” or “reassembly” of the sequence. There are two common approaches: De-novo assembly is used to identify the overlapping regions in the DNA sequences, sequence aligning and joining together to form the genome. This is done without any comparison. Reference genome is used for mapping. In which the aligned new sequence is compared with the reference genome. This de-novo assembly is more challenging. This is the only one method for sequencing new organisms. De-novo assembly introduces less bias than mapping to a reference genome. The genome reference can be charted easily and requires less contiguous reads, during this it may arise or unexpected sequences can be lost. The result of the method is good, only the reference genome is good. It provides better identification of Single nucleotide polymorphisms. Many institutions and organizations have invested more time and effort to create the good reference genomes. The multiple reference genomes have been created for various purposes/testing.

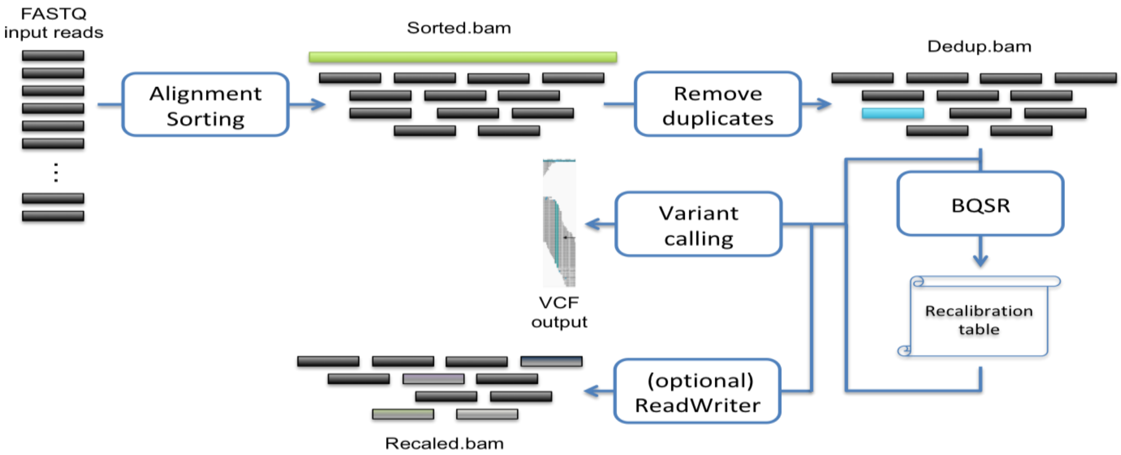
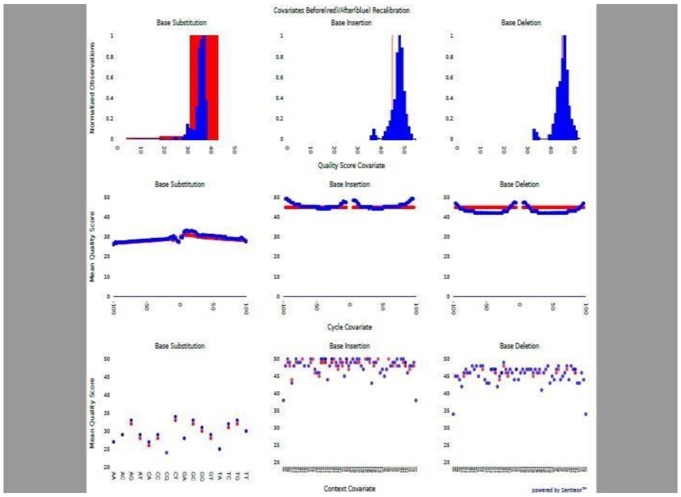


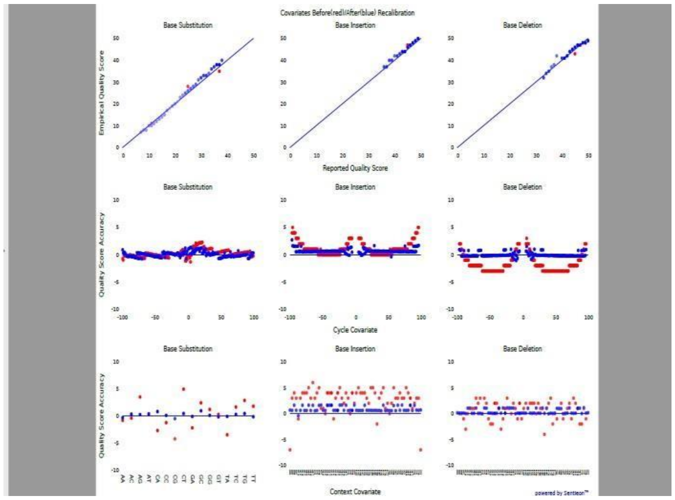
FIGURE 2. Flow Diagram For Whole Genome Sequencing

V. Experimental Results:



A. Base score recalibration

FIGURE 3. ***Base Score Recalibration***



B. Base Recalibration

FIGURE 4. ***Base Recalibration***

VI. Haplotype Caller:

Haplotype Caller is generally used to call the indels and SNPs parallelly through a local group of haplotypes in a region that is active. If a program finds a region exhibiting symptoms of variation, it banishes the mapping information in existence and entirely simulates the data reads in that area. Due to this, the Haplotype Caller is very much precise during the call of regions that are generally arduous to call, for instance when it includes various kinds of variants in a closer manner. Haplotype Caller has the capability of handling non-diploid organisms along with data to be experimented. Although the Haplotype Caller algorithms utilized to evaluate variant resemblance is not suited for extreme gene frequencies. Haplotype Caller also has the capability of handling junctions that form RNA sequence, which is generally an exception for many other callers

VII. Result:

The result of (FIGURE 2) gives mutation sequence. Based on the mutation sequence the cancer type can be predicted. There are various types of mutation that cause various types of cancer. In order to predict the type of cancer and the appropriate treatment, we require three types of file formats. The original data sequence contains many unwanted gene information. Removing this unwanted data and finding the mutation sequence is a challenging task.

VIII. Development:

The explication and analysis of genomic data are still in the beginning stages. It is expected that the thorough knowledge of the human genome will for sure provide a new path for proceedings in medicine and biotechnology. For instance, organizations, such as Myriad Genetics, started providing simple ways to manage genetic tests that can illustrate inclination to a variety of diseases, including breast cancer, liver diseases and many other diseases. Even the causes for cancers, Alzheimer's disease and rest of the areas of clinical interest are thought as the benefits from genomic information and it can also result in the long-term significant advances in their maintenance.

IX. Requirements:

The Whole Genome Project requires original human data for finding the cancer type. It is the most challenging task. To run and store the data, it requires a large amount of space. It takes more time to run and execute the results. Once the processing is finished, the sequence must be compared with the reference sequence in order to find the mutation. The mutation may be known or unknown or it may produce new diseases. Getting the reference genome for different kinds of mutation is very difficult.

X. Summary:

The whole genome sequencing application is created with all features available for both patients and clinical specialists. This application also recommends what type of treatment and prescribes medicine. The users or patients can easily use this for finding the diseases.

ISSN No: 1008-0562

References:

[1]. R.D. Fleischmann, "Whole-Genome Random Sequencing and Assembly of H. Influenzae", *Science*, vol. 269, no. 5223, pp. 496-512, 1995.

[2]. J. Weber, W. Myers, "Human Whole Genome Shotgun Sequencing", *Genome Research*, vol. 7, no. 5, pp. 401-409, 1997.

[3]. A.M. Maxam, W. Gilbert, "A New Method for Sequencing DNA", *Proc. Nat'l Academy of Science*, no. 2, pp. 560-564, 1997.

[4]. J.C. Venter, H.O. Smith, L. Hood, "A New Strategy for Genome Sequencing", *Nature*, vol. 381, no. 6581, pp. 364-366, 1996.

[5]. E. Anson, E. Myers, "ReAligner: A Program for Refining DNA Sequence Multi alignments", *J. Computational Biology*, vol. 4, no. 3, pp. 369-383, 1997.

[6]. "Privacy and Progress in Whole Genome Sequencing", 2012,[online]Available:www.bioethics.gov/cms/sites/default/files/PrivacyProgress508.pdf.

[7]. A. Weston, L. Hood, "Systems Biology Proteomics and the Future of Healthcare: Toward Predictive Preventive and Personalized Medicine", *J. Proteome Research*, vol. 3, no. 2, pp. 179-196, 2004.

[8]. K. Wetterstrand, *DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP)*.

[9]. A. Tarasov et al., "Sambamba: fast processing of NGS alignment formats", *Bioinformatics 2015*.

[10]. S. Andrews, *FastQC: a quality control tool for high throughput sequence data*, 2010