International Journal of Technical Innovation in Modern Engineering & Science (IJTIMES)
Impact Factor: 5.22 (SJIF-2017), e-ISSN: 2455-2585
Volume 4, Issue 6, June-2018

EXOME SEQUENCE ANALYSIS USING COMPUTATIONAL TOOLS FOR CLINICAL DIAGNOSIS

Krupa Mehta[1], Dr. Devarshi Mehta[2], Dr. Vishal Dahiya[3]

[1]*Faculty of Computer Technology, GLS University, Ahmedabad. krupa.mehta@glsuniversity.ac.in*
[2]*Faculty of Computer Technology, GLS University, Ahmedabad. devarshi.mehta@glsuniversity.ac.in*
[3]*Institute of Information and Communication Technology, Indus University, Ahmedabad. vishaldahiya.mca@indusuni.ac.in*

Abstract - All the pieces of an individual's DNA that provide instructions for making proteins are known as exons. A collection of all the exons of a genome is known as exome and the process to sequence these exomes are known as whole exome sequencing. To predict any disease, the exome sequencing analysis focuses on some common steps that are discussed in this paper. Many tools are available to accomplish the task of each step.

Keywords: exome sequencing, Mendelian, phenotypes, SNPs, indels, FASTQ, FASTA, BAM, SAM, VCF.

I. INTRODUCTION

The current trend in bioinformatics focuses on the study of the exome sequences of the human. Since 2005, next generation DNA sequencing platforms have become widely available due to major reduction in the cost of DNA sequencing. The development methods for coupling targeted capture and massively parallel DNA sequencing has made it possible to determine cost efficiency of almost all of the coding variant present in an individual human genome. This process of DNA sequencing is known as exome sequencing [1].

With next-generation sequencing, it is now feasible to sequence large amounts of DNA. The exome (the protein-coding region of the human genome) represents less than 2% of the genome, but contains ~85% of known disease-related variants, making whole-exome sequencing a cost-effective alternative to whole-genome sequencing [1]. This method allows variations in the protein-coding region of any gene to be identified, rather than selecting only a few genes. Most known mutations that cause disease occur in exons, thus whole exome sequencing is thought to be an efficient method to identify possible disease-causing mutations.

The study of exome sequences allows early prediction of disease in human. Exome sequencing is the powerful and accurate method of predicting diseases. Perhaps the most widely used targeted sequencing method is exome sequencing. Sequencing the cancer exome provides useful information about the coding mutations that contribute to tumor progression [4]. The exome sequencing is rapidly proving to be powerful new strategy for finding the cause of known or suspected Mendelian disorders for which the genetic basis has yet to be discovered [1].

Exome sequencing has proven to be not only a cost-effective method to detect disease-causing variants and discover gene targets but also an attractive option for the traditional targeted gene panel sequencing for clinical diagnosis.

II. IMPORTANCE OF EXOME SEQUENCING

The exome is a source of rare disease related variants. Exome sequencing is used to identify genes and mutations that influence risk for human diseases [2]. One of the immediate applications of exome sequencing will be facilitating the accurate diagnosis of individuals with Mendelian disorders that are difficult to confirm using clinical or laboratory criteria alone. It is well justified strategy for discovering rare alleles underlying Mendelian phenotypes and perhaps complex traits like [1]:

- **Positional cloning:** Positional cloning studies that are focused on protein coding sequences have proved to be highly successful at identifying variants for monogenic diseases.
- **Protein coding disorder:** This process also identifies most alleles that are known to underline Mendelian disorders disrupt protein coding sequences.
- **Functional disorder:** The exome sequencing predicts the functional consequences that can cause the damage by focusing on a large fraction of rare, protein-altering variants.

Exome sequencing is rapidly proving to be a powerful new strategy for finding the cause of known or suspected Mendelian disorders for which the genetic basis has yet to be discovered. The drop in per-base sequencing price is expected to drive the generation of immense amount of exome sequence data, creating a big data challenge in bioinformatics. Exome sequencing experiments produce millions to billions of short sequence reads at a high speed [3]. This generated exome sequences needs to be analysed to identify the disease. Computer science comes into the picture to mange and analyse such bulky and complex sequences.

III. EXOME SEQUENCE ANALYSIS USING COMPUTATIONAL TOOLS

A sequence needs to be passed from many stages before generating the final result. There are many tools available which can perform various types of analysis on exome sequence. Each tool focuses on single aspect of processing. Thus, to generate final analysis combination of tools are required. The general process to analyse any exome sequence is shown in Fig. 1.

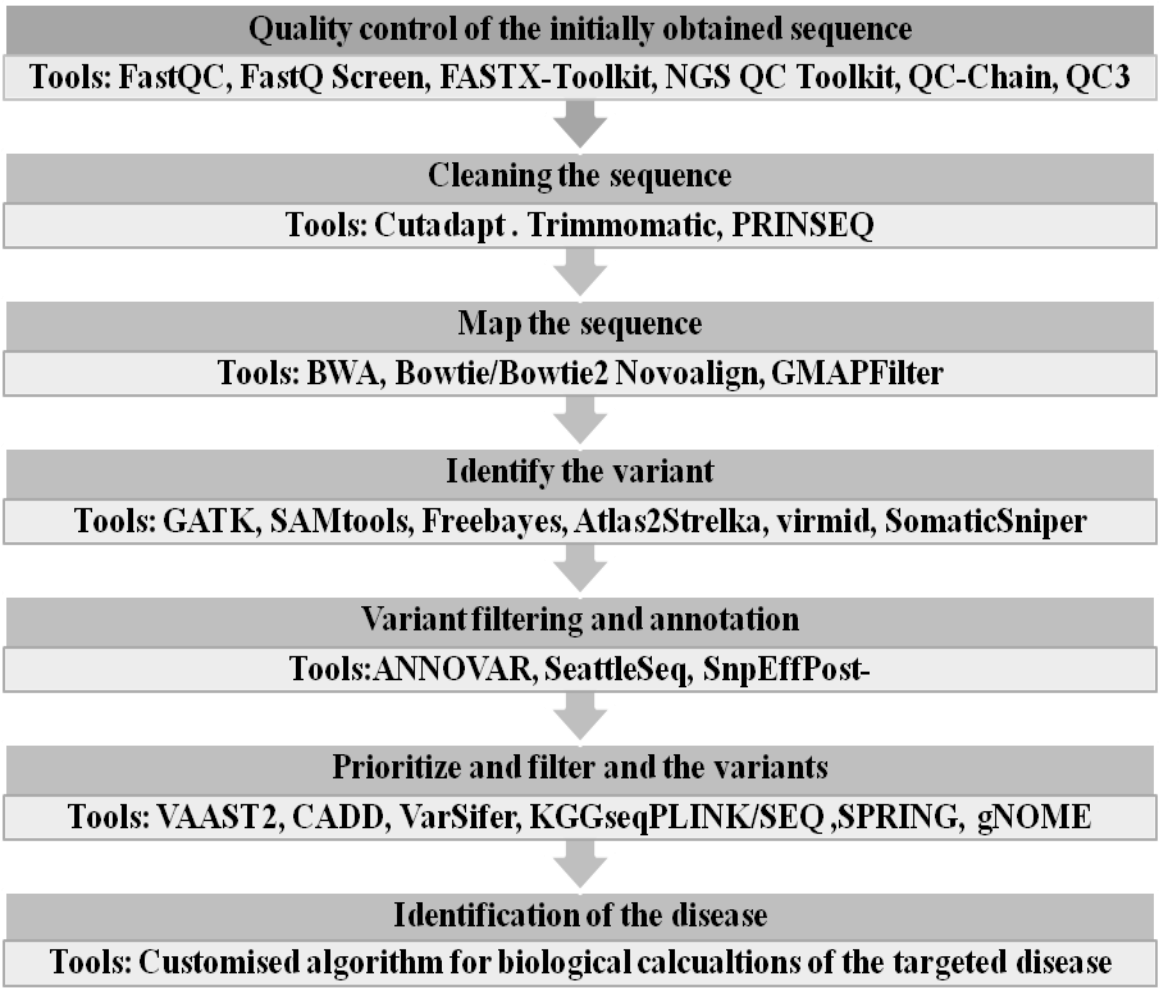

Fig. 1 General steps to analyse exome sequence along with related tools

A. Quality control of the initially obtained sequence

The first and very important step is to quality control the raw data. The raw sequence may have some low quality or contaminating reads which can lead to false analysis. There are many tools like FastQC, FastQ Screen, FASTX-Toolkit, NGS QC Toolkit, PRINSEQ, QC-Chain, and QC3 are available which ensures the quality of the raw sequences for further analysis. These tools works either on FASTQ or FASTA file. The obtained exome sequence may follow either FASTQ or FASTA format. These two formats are widely accepted as a raw sequence data [3][4]. Completion of this step ensures that there is no low quality or contaminating reads in the sequence.

B. Cleaning the sequence

After eliminating the defective reads from the sequence, the next step is to clean the unwanted portions of the sequence. This step includes removal, trimming or correcting the reads that do not meet the defined standards. Raw data generated by the sequencing platforms are compromised by the sequence artifacts such as base calling errors, indels, poor quality reads and adaptor contamination [5]. The aim of cleaning the sequence is to identify and quantify the source of contaminations, filter the contaminating reads, and obtain the processed reads as clean as possible [6]. Standard pre processing procedure includes adapter removal and trimming of low quality bases at the ends of the reads. Depending on the study design and use of the data, redundant reads and undesired sequences such as contamination from primers, adaptors, or other species may be removed at this point. Several tools are available to perform those tasks, such as Cutadapt and Trimmomatic. PRINSEQ and QC on the other hand, provide both QC and cleaning functions as a suite [3].

C. Map the sequence

After performing quality control and cleaning process, the next step is to map short nucleotide reads to the reference genome and with high efficiency and accuracy which is known as sequence alignment and the result of the process is stored in a SAM (Sequence Alignment/Map) / BAM (Binary Alignment/Map) file. As each of the millions of short reads must be compared to the 3 billion possible positions within the human genome, this computational step is not only trivial but also a critical step. Any errors in alignment to the reference genome will be carried through to the rest of the analysis. This step is thus computationally intense and time consuming [7]. Finding the optimal alignment for a sequence read requires an alignment algorithm that is tolerant to imperfect matches, where genomic variations may occur. Moreover, the algorithm needs to be able to align millions of reads at a reasonable speed [3]. The sequencing technologies are constantly pushing the lengths of generated reads-requiring new and improved algorithms [5]. There are various s/w programs that can be used to perform sequence reads alignment for e.g. Bowtie/Bowtie2, BWA, MAQ, Novoalign[2].

D. Identify the variant

After mapping of the short reads to the reference genome, the next step is comparing the aligned sequences with known sequences to determine which positions deviate from the reference position [7]. This process is used to identify the variant sites where the aligned sequences deviate from the known sequences at the reference position. A variant call is a conclusion that there is a nucleotide difference of some reference at a given position in an individual genome [2]. Since the short reads are already aligned, the sample genome can be compared to the reference genome and variants can then be identified. These variants may be responsible for disease, or they may simply be genomic noise without any functional effect. Variant call format (VCF) is the standardized generic format for storing sequence variation including SNPs, indels, larger structural variants and annotations [7].

E. Variant filtering and annotation

After the identification of variant, the next step is to determine which of these variants are likely to contribute to the pathological process under study. This step combines two process, first filtering and second annotation. The process of filtering removes variants that are fit for specific genetic models or are not present in normal tissue. The process of annotation is used to seek out the information about variants and to identify the variants that are fit for the biological process [7]. This process of filtering and annotating is used to reduce variant sites to a smaller set of genes with possible function and activity [2]. With the large amount of data produced by exome sequencing, the possibility of predicting the functional impact of variants in an automated fashion is becoming increasingly important. Computer aided annotation enables to filter and prioritize potential disease – causing mutations for further analysis [5]. Many tools exist to examine relevant variants by referencing previously known information about their biological functions and inferring potential effects based on their genomic context [7]. The available tools implement different methods for variant annotation. Most of them focus on the annotation of SNPs, since the can be easily identified and analysed. indels are also covered by some tools, whereas annotation of structural variants is limited to CNVs and only performed by recently developed applications. The most common annotation is to provide database links to various public variant databases [5].

F. Prioritize and filter the variants

After reducing the variant site by performing filtering and annotating, the next step is to prioritize and filter the variants. This step is used to prioritize variants relative to the disease [3]. Different types of genomic variants including SNVs, indels, CNVs, and large SVs can be detected from the sample by comparing the aligned reads to the reference genome. In cancer studies, it is important to distinguish somatic from germline variants as the two classes of variants often play distinct roles in tumor development. Germline variants are inherited mutations present in the germ cells, which are related to patient family history. Somatic variants are mutations that are present only in somatic cells and can be tissue-specific [3].

G. Identification of the disease

The final step is to predict the disease according to the comparison of sequences available in existing databases. This step also requires implementing some algorithm that focuses on the biological calculations for the targeted disease. Exome sequencing enables the identification of more novel genetic variants than previously possible, but it still requires computational and experimental approaches to predict whether a variant is deleterious [7].

IV. IMPLEMENTATION OF EXOME SEQUENCING FOR CLINICAL DIAGNOSIS

There are many cases where exome sequences played a pivotal role as a diagnostic tool for rare or novel diseases. The above discussed steps(Fig: 1) are successfully implemented by Elizabeth A. Worthey et.al.[8] for the clinical diagnosis of a single child with a life threatening but previously undefined form of inflammatory bowel disease.

The team tried various treatment methods but was not able to get proper results. Ultimately, exome capture was performed using 5 μg of the patient's genomic DNA. The captured fragments were amplified and sequenced. The sequence reads obtained were aligned to the human genome reference sequence, and variants (nucleotides from the patient that differed from the reference sequence including insertions, deletions, and substitutions) were identified. All identified variants were subsequently annotated with information for identification of candidate mutations by applying an algorithm which derives and displays a variety of data for each variant. Data including the depth of coverage, conservation across species, percentage of reads with the variant, novelty, potential splice site alteration, known disease association, and likelihood that a variant is deleterious to the protein were extracted from reference data sets or computed in bulk for all variants. These data were stored in a database to query the data. Filtering of the variants using a number of queries based on likely modes of inheritance was performed to identify candidate mutations.

On the basis of analysis of the exome sequence, this child has been diagnosed as having an XIAP mutation and resulting immunodeficiency.

V. CONCLUSION

Exome sequencing is proving to be the successful and accurate method for the diagnosis of rare and novel diseases. Study reveals that exome sequence analysis comes to rescue where traditional diagnosis methods fail. To analyse the sequence some general steps like quality control, sequence cleaning, mapping the sequence, variant identification, variant prioritization and annotation are needed. According to the annotated variants the mutation is identified and biological computations are applied to identify the targeted disease.

REFERENCES

1. Michael J. Bamshad, Sarah B. Ng, Abigail W. Bigham, Holly K. Tabor, Mary J. Emond, Deborah A. Nickerson and Jay Shendure. Exome sequencing as a tool for Mendelian disease gene discovery. Nature Reviews: Genetics, Translational Genetics. Volume 12, November 2011.
2. Zuoheng Wang, Xiangtao Liu, Bao-Zhu Yang and Joel Gelernter. The role and challenges of exome sequencing in studies of human diseases. Frontier in Genetics. Volume 4, Article 160, August 2013.
3. Bao et al. Review of Current Methods, Applications, and Data Management for the Bioinformatics Analysis of Whole Exome Sequencing. Cancer Informatics 2014:13(S2) 67–82 doi: 10.4137/CIN.S13779.
4. Zhou Q, Su X, Wang A, Xu J, Ning K (2013) QC-Chain: Fast and Holistic Quality Control Method for Next-Generation Sequencing Data. PLoS ONE 8(4): e60234. doi:10.1371/journal.pone.0060234
5. Stephan Pabinger, Andreas Dander, Maria Fischer, Rene Snajder, Michael Sperk, Mirjana Efremova, Birgit Krabichler, Michael R. Speicher, Johannes Zschocke and Zlatko Trajanoski. A survey tool for variant analysis of next generation genome sequencing data. Briefings in Bioinformatics. Volume 15. No 2 256-278. January 2013.
6. Qian Zhou, Xiaoquan Su, Anhui Wang, Jian Xu, Kang Ning. QC-Chain: Fast and Holistic Quality Control Method for Next-Generation Sequencing Data. PLOS ONE. April 2013, Volume 8, Issue 4, e60234.
7. Marisa P. Dolled-Filhart, Michael Lee Jr., Chih-wen Ou-yang, Rajini Rani Haraksingh, and Jimmy Cheng-Ho Lin. Computational and Bioinformatics Frameworks for Next-Generation Whole Exome and Genome Sequencing. Hindawi Publishing Corporation The Scientific World Journal Volume 2013, Article ID 730210.
8. Elizabeth A. Worthey, Alan N. Mayer, Grant D. Syverson,Daniel Helbling, Benedetta B. Bonacci, Brennan Decker, Jaime M. Serpe, Trivikram Dasu, Michael R. Tschannen, Regan L. Veith, Monica J. Basehore, Ulrich Broeckel, Aoy Tomita-Mitchell, Marjorie J. Arca, James T. Casper, David A. Margolis, David P. Bick, Martin J. Hessner, John M. Routes, James W. Verbsky, Howard J. Jacob, and David P. Dimmock. Making a definitive diagnosis: Successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. Genetics IN Medicine. Volume 13, Number 3, March 2011.
9. Tony Shen, Stefan Hans Pajaro-Van de Stadt, Nai Chien Yeat, Jimmy C.-H.Lin. Clinical applications of next generation sequencing in cancer: From panels, to exomes, to genomes. Frontiers in Generics.6,215(2015).
10. Jennifer D. Hintzsche, William A. Robinson and Aik Choon Tan. A Survey of Computational Tools to Analyze and Interpret Whole Exome Sequencing Data. Hindawi Publishing Corporation International Journal of Genomics Volume 2016, Article ID 7983236.
11. Ravi K. Patel, Mukesh Jain. NGS QC Toolkit: A Toolkit for Quality Control of Next Generation Sequencing Data. PLoS ONE. February 2012, Volume 7, Issue 2, e30619.
12. https://www.illumina.com[15 May, 2018]
13. Ayman Grada1 and Kate Weinbrecht2. Next-Generation Sequencing: Methodology and Application. Journal of Investigative Dermatology (2013) 133, e11; doi:10.1038/jid.2013.248.