Annovar Pipeline in the detection of genetic variants in patients with Retinal Dystrophies in the Saudi Population

H. A. Aljafar[2], A. Bin Owayn[2], M. Al Rashed[3], R. Almohammad[1], E. Alsolme[1], F. Alqubaishi[2], G. M. Alanezi[1], S. Al Shahrani[1], S. Zameer[1], F. Al Adel[4], M. Abed Althagafi5; L. Abu Safieh1

1 King Fahad Medical City, Riyadh, Saudi Arabia, 2 King Abdulaziz City for Science and Technology, Riyadh, Saudi Arabia, 3 King Saud University, Riyadh, Saudi Arabia, 4 princes Noura University hospital, Riyadh, Saudi Arabia, 5 King Fahad Medical City, King Abdulaziz City for Science and Technology, Riyadh, Saudi Arabia.

Background

Inherited retinal dystrophies are a common cause of blindness, characterized by loss of photoreceptor function. They can be inherited in different patterns but due to the nature of the Saudi population, autosomal recessive is the major form of inheritance. We aim to identify rare homozygous variants and to characterize the genetic cause of syndromic and non-syndromic retinal dystrophy in an attempt to understand the disease pathogenesis to develop screening, counseling, prevention protocols and new potential therapeutic targets. Annovar pipeline [1] was selected for analysis. It is a powerful tool for the study of high-throughput sequencing data or (NGS) technologies.

Methods

Patients DNA was used for whole-exome sequencing (WES). Genomic variants were annotated using ion-reporter exome analysis pipeline. Quality-assessment and data analysis was done using FastQC0.11.8. Data-Alignment was carried out using Burrows-Wheeler Aligner0.7.17-r1188. Different tools were used to check Duplicated reads, and create in-depth mapping statistics alignment review, sequencing quality and Genome-assembly. Quality was checked by optimization contigs/scaffolds, total assembly size, maximum scaffold size, N50value, median contig-length, and GC-content. Generated Variant Calling Format(VCF) file was assembled for genome purity and SNPs. SNP-analysis results were extracted using multiple parameters, such as: SNPs, MNPs, Indels, Transitions/Transversions, Total Het/Hom. Functional annotation of SNPs was done by using Annovar pipeline.

Results

Whole exome sequencing (WES) of 4 cases of retinal dystrophy were annotated. The cases were screened for 283 genes related to autosomal recessive retinal dystrophies. 170 related variants were identified while 113 were undetected when GeneCard (online source) were used for annotation. However, when Annovar (offline) tools were used, 24 genes that were previously undetected were then identified, increasing the total of genes identified to 194. Nevertheless, there were genes detected online that were not detected offline.

References

1 .Wang, K., etal, **ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data.** *Nucleic Acids Res* 2010, **38**:e164.
2. Simon A. **FastQC:a quality control tool for high throughput sequence data.** http://bioinformatics.babraham.ac.uk/projects/fastqc, . 2010.
3. Li H, Durbin R. **Fast and accurate short read alignment with Burrows-Wheeler transform**. *Bioinformatics (Oxford, England)* 2009, **25**(14):1754-1760.
4. Okonechnikov K, et al. **Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data**. *Bioinformatics (Oxford, England)* 2016, **32**(2):292-294.
5. rtg-tools . **RTG tools: utilities for accurate VCF comparison & manipulation.** Java. Real time genomics. 2017. available under BSD license at:https://github.com/RealTimeGenomics/rtg-tools
6. Cleary JG, et al.: **Comparing Variant Call Files for Performance Benchmarking of Next-Generation Sequencing Variant Calling Pipelines.** *BioRxiv* 2015, 023754. **10**.1101/023754.
7. Digital image, accessed 21 August 2020, <https://student.unsw.edu.au/clinq-images-and-tables-found-online>.
8. J. Clin. Med. 2020, 9 (1), 132; https://doi.org/10.3390/jcm9010132

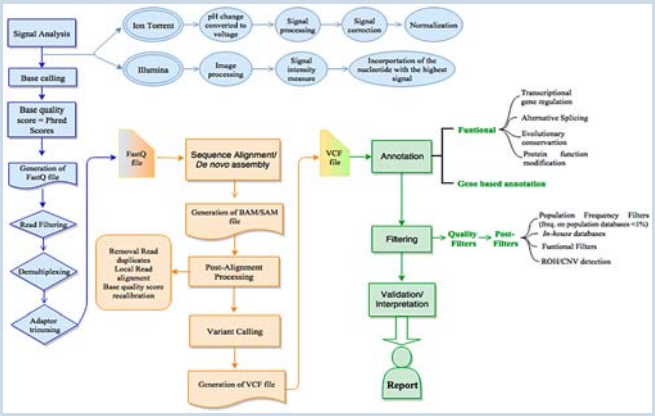

Fig 1: Schematic overview of Next-Generation Sequencing (NGS) of an open-sourced bioinformatics pipeline for the processing data that is whole genome sequence (WGS) or whole exome sequence (WES) analysis in Clinical Genetics [8].

Data analysis and statistics

The obtained reads were analyzed and a quality assessment was done using FastQC version 0.11.8 [2]. Alignment of the data was carried out by using Burrows-Wheeler Aligner version 0.7.17-r1188 [3]. SAMtools version 1.9 was used to remove duplicate reads, and create in-depth mapping statistics. Following alignment, QualiMap version 2.2.1 was used to review the alignment and sequencing quality-control [4]. The quality of the genome assembly was checked by optimization of the number of contigs/scaffolds, total assembly size, maximum scaffold size, N50value, median contig length, and GC content. To analyze the purity of the assembled genome, a Variant Calling Format (VCF) file was generated. SNP variants were compared to a reference genome. The results of SNPs analysis were extracted with the optimization of the total number of SNPs, MNPs, Insertions, Deletions, Indels, SNP Transitions/Transversions, Total Het/Hom ratio, SNP Het/Hom ratio, MNP Het/Hom ratio, Insertion Het/Hom ratio, Deletion Het/Hom ratio, Indel Het/Hom ratio, Insertion/Deletion ratio and Indel/SNP+MNP ratio a quick stats with Real Time Genomics (RTG) Tools software package [5, 6]. Functional annotation of the SNPs were done by using Annovar pipeline [1].

Conclusions

Four cases were analyzed in reference to 11 Annovar databases. These variants are associated with retinal dystrophies. Further validation and functional assessment are needed to establish exact variants role in pathogenesis. In the cases evaluated comparative WES have shown to be more reliable with the Annovar pipeline.