**A standard tool to analyze, interpret and report pathological sequence variants related to recessive diseases in NGS data**

Background

Although individually uncommon in general populations, Mendelian diseases are collectively reported to account for ~20% of infants mortality and ~10% of pediatric hospitalizations (Bell et.al,2013). More than 2000(Autosomal and X-linked) recessive disorders have been identified. Prevalence of an offspring affected with a recessive disease is higher among consanguineous couples. Detection of carrier status enables identification of couples with 25% risk of affected offspring (Sallevelt et.al,2017). Therefore, preconceptual detection of carrier status and genetic counseling enables prevention of disease and providing them with informed reproductive choices.

However, using the best method to conduct preconception career screening, interpretation and responsible reporting of the results are vital to prevent misinformation and ethical dilemma. Traditional carrier screening methods focus on certain ethnic/geographical populations with a higher prevalence of certain recessive diseases using targeted screening. More recently, lower costs and higher accuracy of NGS based technologies have enabled to test for higher number of conditions using carrier screening gene panels. Targeted analysis with gene panels have been used on untargeted whole exome sequencing data but only very few studies have been done to identify systemic assessment of whole exome sequencing data to for preconception carrier screening. Sallevelt. et al.(2017) describes a comprehensive filter method that could be used on NGS data of consanguineous as well as non-consanguineous couples to detect important pathogenic variants causing recessive diseases which is mainly the basis for this tool.

This tool can be developed further not only to be used in preconception carrier screening but also in general identification of pathogenic variants in NGS output of patients presenting with mendelian diseases. The advantage in this kind of application is the ability to compare the suggested pathogenic variant related disease symptoms with the actual symptoms of the patients to confirm the diagnosis. Some of these pathogenic findings could be actionable.

Functionality

The Input is GATK output vcf file annotated with RefSeq, dbSNP, Clinvar and/or HGMD, OMIM and converted to a tsv file. The tool will use a comprehensive filter strategy mainly described in Sallevelt. et al.(2017) to identify pathogenic variants responsible for mendelian diseases. The resulting output will show the pathogenic variants and the relevant fields that were collectively assessed for the filtration.

Technologies used,

1. SQL relational database
2. Python based CGI, regex
3. HTML/CSS/java script/JSON

Sequence of the tool functionality

1. HTML/CSS form is presented to the user for the user input tsv file
2. tsv file is imported to the SQL relational database.
3. Filter strategy is used as sql queries and python code in python-based CGI script which performs the queries on the newly imported table on the relational database, parse out necessary information using regex and produce the pathogenic variants with the relevant fields.
4. The output is displayed to the user on the web interface.

Filter strategy

A picture containing graphical user interface

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Source – Figure 2. Sallevelt et.al(2017)

Design and Development

An SQL database is created in the Mysql server(final project database). When the user input is provided the content is imported to a table in the final project database.The NGS data in the form of fastq file is separately run through a pipeline of tools and annotated with refSeq, dbSNP,Clinvar,OMIM prior to using the file as an input of the tool.The resulting vcf file is also converted to a tab separated values file separate from the tool.

HTML/CSS/Javascript has the interface of a form to take the input file from the user and also the interface to visualize the output (pathogenic variants and the relevant fields) from the input -whole exome sequence.

Python based CGI script will be first used to import the content of the input tsv file in to the SQL database in the Mysql server.The filtration method in the form of sql queries in the CGI script will be performed on the newly imported sql table.The resulting output stored in a data structure might be parsed out further using regex by python code and will be either tied to a JSON or transferred as a data structure such as a list of dictionaries to the HTML/CSS for user visualization.

The fields in the tsv input that are filtered in the tool

1.Consequence – type of variant

Diagram

Description automatically generated

Source - <https://useast.ensembl.org/info/genome/variation/prediction/predicted_data.html>

2. Impact – High/ Moderate/Low/Modifier

3. Strand

4. Canonical

5. genomeAD allele frequency

6. Clinical Significance

Table

Description automatically generated

Source - https://www.ncbi.nlm.nih.gov/clinvar/docs/maintenance\_use/

7. Depth

8. Heterozygocity/Homozygocity

Sallevelt, Susan, C.E.H et.al.(2017).A comprehensive strategy for exome-based preconception carrier screening.Genetics in Medicine,volume19,number 5.DOI: [10.1038/gim.2016.153](https://doi.org/10.1038/gim.2016.153)

Bell, J.C.(2013). Carrier Testing for Severe Childhood Recessive Diseases by Next-Generation Sequencing.Sci Transl Med, 3(65). DOI:10.1126/scitranslmed.3001756

Dorchner M.O.(2013). Actionable, Pathogenic Incidental Findings in 1,000 Participants’

Exomes.Am J Hum Genet,93(4).DOI: 10.1016/j.ajhg.2013.08.006

Mahjoubeh Jalali Se5d Dashti & Junaid Gamieldien.(2018). A practical guide to /ltering and prioritizing genetic variants.BioTechniques ,vol 62,no 1