European Review for Medical and Pharmacological Sciences 2021; 25 (1 Suppl): 1-6

Integration of VarSome API in an existing bioinformatic pipeline for automated ACMG interpretation of clinical variants

E. SORRENTINO[1], F. CRISTOFOLI[1], C. MODENA[1], S. PAOLACCI[2], M. BERTELLI[1,2], G. MARCEDDU[1]

[1]MAGI EUREGIO, Bolzano, Italy
[2]MAGI'S LAB, Rovereto (TN), Italy

Abstract. – OBJECTIVE: While the bioinformatic workflow, from quality control to annotation, is quite standardized, the interpretation of variants is still a challenge. The decreasing cost of massively parallel NGS has produced hundreds of variants per patient to analyze and interpret. The ACMG "Standards and guidelines for the interpretation of sequence variants", widely adopted in clinical settings, assume that the clinician has a comprehensive knowledge of the literature and the disease.

MATERIALS AND METHODS: To semi-automatize the application of the guidelines, we decided to develop an algorithm that exploits VarSome, a widely used platform that interprets variants on the basis of information from more than 70 genome databases.

RESULTS: Here we explain how we integrated VarSome API into our existing clinical diagnostic pipeline for NGS data to obtain validated reproducible results as indicated by accuracy, sensitivity and specificity.

CONCLUSIONS: We validated the automated pipeline to be sure that it was doing what we expected. We obtained 100% sensitivity, specificity and accuracy, confirming that it was suitable for use in a diagnostic setting.

Key Words:
VarSome, ACMG, Bioinformatic pipeline.

Introduction

The massively parallel sequencing technology known as next-generation sequencing (NGS) has revolutionized genomic research. The fast and affordable simultaneous interrogation of thousands of target regions for genetic variants is allowing many gene-disease associations to be discovered, increasing our understanding in various fields of medicine, ranging from genetics to oncology and microbiology. Unlike previous diagnostic sequencing technologies, NGS can deliver an analysis of DNA sequences of a sample in a single test, thus providing a better idea of the diagnosis, but also an enormous quantity of data to analyze. Interpreting the clinical significance of hundreds and thousands of variants produced per individual by NGS-based tests is a big challenge.

The ACMG "Standards and Guidelines for the Interpretation of Sequence Variants"[1], widely adopted in clinical settings, assumes that clinicians have a comprehensive knowledge of the literature and the disease. We decided to develop an algorithm that exploits VarSome[2], a widely used platform for variant interpretation that uses information from more than 70 genome databases[3].

Many difficulties need to be addressed when doing variant interpretation: combination of different types of information[4], frequency threshold estimation[5], appropriate interpretation of clinical context[6], continuous incorporation of updated knowledge and lack of inter-laboratory consistency in interpreting variants[7]. To tackle these difficulties and to semi-automate application of the guidelines, we selected VarSome, as it integrates data from different databases, keeping it updated, and defines well-tested thresholds for the different criteria of the ACMG guidelines.

When analyzing genetic data for diagnostic purposes, the workflow, from sequencing to bioinformatic analysis, must be precise, reproducible and validated. To reduce human error, our lab developed an automated pipeline for bioinformatic analysis[8]. By introducing semi-automated variant interpretation, by means of VarSome, we made variant interpretation part of our precise reproducible workflow and enabled addition of expert knowledge from the scientific community.

Corresponding Author: Stefano Paolacci, Ph.D; e-mail: stefano.paolacci@assomagi.org

In this paper we describe how we integrated VarSome into our existing validated, standardized, reproducible analysis pipeline for NGS data to interpret variants and how we assessed the results produced.

Materials and Methods

After the sequencing phase, the pipeline (Figure 1) analyzes the sequences to recognize variants and annotate them. After annotation, variants must be interpreted to distinguish common and benign ones from those with pathogenic potential. In this section we describe how we perform sequencing, bioinformatic analysis and variant interpretation with the inclusion of VarSome.

Sequencing and Bioinformatic Analysis

The NGS data analysis workflow includes DNA extraction, library preparation, sequencing and bioinformatic analysis of the data. The DNA samples were processed using a MiSeq personal sequencer (Illumina, San Diego, CA, USA) with paired-end long reads of 150 bp, according to the manufacturer's instructions. The probe set to capture the target regions, comprising coding exons and 15 bp flanking regions of each gene in the panel, was designed using software from Twist Bioscience[9] and was based on the hg38 genome version.

Bioinformatic analysis includes the steps quality control, mapping against the genome and final annotation of variants, as previously described[8]. Briefly, raw reads undergo a series of quality controls by Fastx-toolkit[10] to reduce the number of error-prone reads and improve the quality of subsequent steps. The sequences generated are aligned against the reference sequence (hg38[11]) by BWA software[12,13], to identify variants in the sample by GATK[14,15] variant calling.

Then, the variants are annotated using VEP (Variant Effect Predictor)[16]: the gene and transcript related to the macro-areas, along with their location, are assigned to each variant. Other important information, like minor allele frequency (MAF) from GnomAD Exome[17], is also retrieved for sub-populations and the whole population, for use in subsequent steps. The APPRIS database[18] is used to select and store only annotations associated with transcripts indicated by the known Ensembl dataset, excluding other transcripts.

VarSome

VarSome (varsome.com) is a search engine, aggregator and impact analysis tool for human genetic variations designed to share global expertise in human variants using data from over 70 genome databases[2,3].

VarSome API[19] is the high-performance variant annotation Application Programming Interface (API) of VarSome, designed to provide a tool that can be queried to extract information. All the information from VarSome used in our software was collected using the Stable API environment[20], which is updated four times per year. We decided to use this environment because it allows greater interpretative stability in selecting variants to report. Stable API documentation can be found at this link: https://stable-api.varsome.com/.

VarSome implements most of the rules of the guidelines on the basis of thresholds that have been carefully adjusted by statistical regression

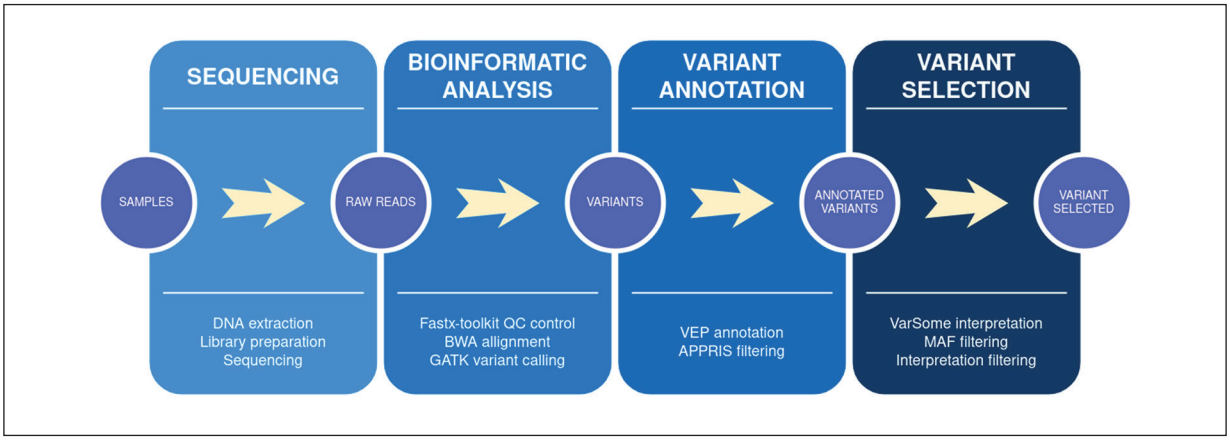

Figure 1. Workflow from samples to variant selection.

against a large population of reliably curated variants[21]. When querying VarSome API, the data is returned using JSON, which is easily accessed from Python, from which our script extracts the requested information. On how we installed and used VarSome API, see https://github.com/saphetor/varsome-api-client-python__documentation.

Users can also submit their own contributions to VarSome, linking variants to phenotypes, diseases or articles, and can make their own pathogenicity assessments. VarSome considers these submissions during the interpretation phase, combining knowledge of experts with data from genome databases.

Algorithm for VarSome Integration

"Standards and guidelines for the interpretation of sequence variants" was published in 2015 by Richards et al[1]. It describes a set of rules for classifying '*Benign*', '*Likely Benign*', '*VUS*' (Variant of Uncertain Significance), '*Likely Pathogenic*' or '*Pathogenic*' variants. The interpretation phase (Figure 2) needs to distinguish common and benign from potentially pathogenic variants. This distinction is done by filtering to exclude variants found in a large proportion of the healthy population.

To do this filtering, we first calculate MAF (Minor Allele Frequency, called decisionMAF), namely the frequency with which the second most common allele occurs in a given population. This frequency is calculated by integrating frequencies from dbNSFP[22], VEP and gnomAD[17] and then inverting it, by doing 100%-decisionMAF, when decisionMAF is over 90%. This step is done because variants having a high decisionMAF are probably wildtype alleles in the reference genome.

The MAF filter excludes variants having a *decisionMAF* over 3%, as they probably have a pathogenic role. This threshold was chosen as a good compromise between 1% of the definition of polymorphism, that excludes some variants that cause diseases, and the computational cost of analyzing all the variants.

Then, variant interpretation and MAF are extracted from VarSome. The newly extracted MAF (called *definitiveMAF*) allows us to integrate the existing *decisionMAF* on splicing variants with data from the scSNV database[23]. Then, we filter '*Benign*' or '*Likely Benign*' variants except those in our exception files. This file contains a list of variants, from Ghosh et al 2018[24], that are known to be pathogenic even if they have a high allele frequency. For the code for this project, see https://gitlab.com/magieuregio/automate_variant_interpretation.

Results

To check whether our pipeline was working as expected, we tested the first 10 samples analyzed (Table I). We first ascertained that our pipeline discarded variants with a *decisionMAF* and a *definitiveMAF* greater that 3%, except for exception variants. Then, we checked that all '*VUS*', '*Likely Pathogenic'* and '*Pathogenic'* variants were selected and that '*Benign'* and '*Likely benign'* were not (except the exceptions). Then we checked that variants in exceptions were correctly marked.

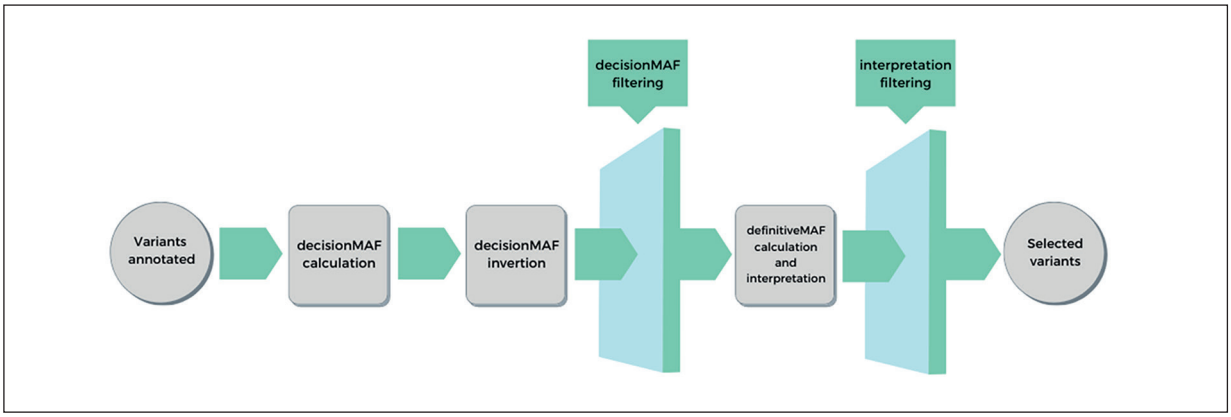

Figure 2. Already annotated variants go through different steps in order to be interpreted: calculation, inversion and filtering on decisionMAF, calculation of definitive MAF, calculation of variant interpretation and interpretation filtering to finally obtain selected variants.

Table I. Results of evaluation on 10 samples.

Sample	FP	FN	TP	TN	Sensitivity	Specificity	Accuracy
RE1756.2021	0	0	1	7	100	100	100
RE1757.2021	0	0	1	10	100	100	100
RE1758.2021	0	0	2	8	100	100	100
RE1759.2021	0	0	1	6	100	100	100
RE1760.2021	0	0	2	4	100	100	100
RE1755.2021	0	0	2	30	100	100	100
RE1752.2021	0	0	4	134	100	100	100
RE1753.2021	0	0	4	118	100	100	100
RE1754.2021	0	0	1	11	100	100	100
RE1761.2021	0	0	3	150	100	100	100
Total	0	0	21	478	100	100	100

FN = False negative; FP = False positive; TN = True negative; TP = True positive.

We checked the number of *True Positive* (TP), *True Negative* (TN), *False Positive* (FP) and *False Negative* (FN) variants to calculate sensitivity , specificity and accuracy We obtained 100% sensitivity, specificity and accuracy, allowing the script to be integrated into the pipeline.

We also tested 70 retinal dystrophy, macular dystrophy and cone dystrophy samples (see **Supplementary Table I** for a list of the genes analyzed) to determine the impact of our algorithm. No variants were selected in 15.7% of the 70 samples (Figure 3), meaning that for these patients a negative report can be issued. For 50% of samples, we found at least one 'Pathogenic' or 'Likely pathogenic' variant, which after endorsement by an expert geneticist, can lead to a conclusive positive test (see **Supplementary Table II** for the list of variants selected by the algorithm).

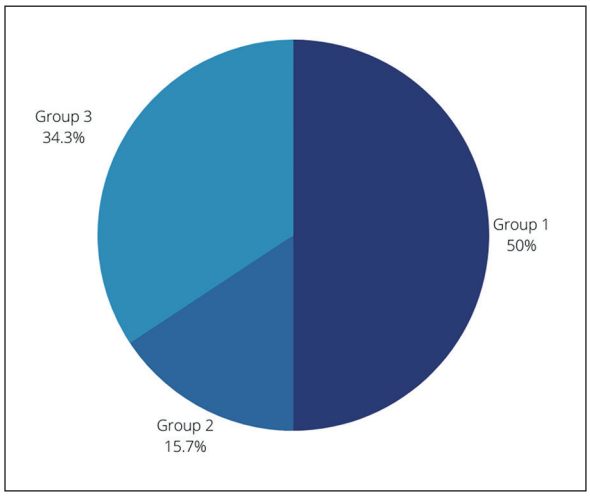

Figure 3. Group 1 includes samples in which at least one 'Likely Pathogenic' or 'Pathogenic', variant was found; Group 2 includes samples in which no variant was selected; Group 3 includes samples in which only 'Variants of Uncertain Significance' were selected.

Discussion

Variant interpretation is a challenging task because it must integrate information from different sources, determine frequency thresholds and understand clinical context. The most widely used clinical guidelines are those of the ACMG: "Standards and guidelines for the interpretation of sequence variants", which assume that clinicians have comprehensive knowledge of the literature and the disease.

To overcome these difficulties, we integrated VarSome, a tool that interprets variants on data from different sources and on thresholds adjusted by statistical regression against a large population, into our diagnostic pipeline. In particular, we chose VarSome API which allows a higher level of customization of the analysis and cost reduction with respect to other commercial solutions.

The tool was integrated in our existing PipeMAGI, a standardized, validated, replicable bioinformatic pipeline for the analysis of NGS panel data developed by MAGI. Though developed for internal use, the pipeline integrated with VarSome introduced inputs from the international scientific community.

Conclusions

We validated the automated pipeline to be sure that it was doing what we expected. We obtained 100% sensitivity, specificity and accuracy, con-

firming that it was suitable for use in a diagnostic setting. We also tested it on 70 retinal dystrophy, macular dystrophy and cone dystrophy samples, obtaining at least one potentially pathogenic variant that led to a positive result in 50% of samples, while for 15.7% of samples we already had a conclusive negative report at this step.

How we adapted VarSome automatic variant interpretation to our clinical workflow, integrating the most recent guidelines (such as ACGS Best Practice Guidelines for Variant Classification 2019-2020[25]) and broadening some criteria, will be discussed in a future paper.

Conflict of Interest

The Authors E. Sorrentino, F. Cristofoli, C. Modena, M. Bertelli and G. Marceddu are employed at MAGI EUREGIO.

References

1) Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, Hegde M, Lyon E, Spector E, Voelkerding K, Rehm HL; ACMG Laboratory Quality Assurance Committee. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med 2015; 17: 405-424.
2) Kopanos C, Tsiolkas V, Kouris A, Chapple CE, Albarca Aguilera M, Meyer R, Massouras A. VarSome: the human genomic variant search engine. Bioinformatics 2019; 35: 1978-1980.
3) VarSome.com. Available online: https://varsome.com/varsome-editions/varsome/ (accessed on 1 March 2021).
4) Bodian DL, Kothiyal P, Hauser NS. Pitfalls of clinical exome and gene panel testing: alternative transcripts. Genet Med 2019; 21: 1240-1245.
5) Whiffin N, Minikel E, Walsh R, O'Donnell-Luria AH, Karczewski K, Ing AY, Barton PJR, Funke B, Cook SA, MacArthur D, Ware JS. Using high-resolution variant frequencies to empower clinical genome interpretation. Genet Med 2017; 19: 1151-1158.
6) Basel-Salmon L, Orenstein N, Markus-Bustani K, Ruhrman-Shahar N, Kilim Y, Magal N, Hubshman MW, Bazak L. Improved diagnostics by exome sequencing following raw data reevaluation by clinical geneticists involved in the medical care of the individuals tested. Genet Med 2019; 21: 1443-1451.
7) Amendola LM, Jarvik GP, Leo MC, McLaughlin HM, Akkari Y, Amaral MD, Berg JS, Biswas S, Bowling KM, Conlin LK, Cooper GM, Dorschner MO, Dulik MC, Ghazani AA, Ghosh R, Green RC, Hart R, Horton C, Johnston JJ, Lebo MS, Milosavljevic A, Ou J, Pak CM, Patel RY, Punj S, Richards CS, Salama J, Strande NT, Yang Y, Plon SE, Biesecker LG, Rehm HL. Performance of ACMG-AMP Variant-interpretation guidelines among nine laboratories in the Clinical Sequencing Exploratory Research Consortium. Am J Hum Genet 2016; 98: 1067-1076.
8) Marceddu G, Dallavilla T, Guerri G, Manara E, Chiurazzi P, Bertelli M. PipeMAGI: An integrated and validated workflow for analysis of NGS data for clinical diagnostics. Eur Rev Med Pharmacol Sci 2019; 23: 6753-6765.
9) Twist Bioscience. Available online: https://www.twistbioscience.com/products/ngs (accessed on 8 April 2021).
10) Fastx-Toolkit. FASTQ/A Short-Reads Pre-Processing Tools. Available online: http://hannonlab.cshl.edu/fastx_toolkit/index.html (accessed online 8 April 2021).
11) UCSC Genome Browser. Available online: http://genome.ucsc.edu/ (accessed on 8 April 2021).
12) Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009; 25: 1754-1760.
13) Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 2010; 26: 589-595.
14) Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, Banks E, Garimella KV, Altshuler D, Gabriel S, DePristo MA. From FastQ Data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. Curr Protoc Bioinformatics 2013; 43: 11.10.1-11.10.33.
15) McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The Genome Analysis Toolkit: A MapReduce Framework for analyzing next-generation DNA sequencing data. Genome Res 2010; 20: 1297-1303.
16) McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, Flicek, P, Cunningham F. The ensembl variant effect predictor. Genome Biol 2016; 17: 122.
17) Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birnbaum DP, Gauthier LD, Brand H, Solomonson M, Watts NA, Rhodes D, Singer-Berk M, England EM, Seaby EG, Kosmicki JA, Walters RK, Tashman K, Farjoun Y, Banks E, Poterba T, Wang A, Seed C, Whiffin N, Chong JX, Samocha KE, Pierce-Hoffman E, Zappala Z, O'Donnell-Luria AH, Minikel EV, Weisburd B, Lek M, Ware JS, Vittal C, Armean IM, Bergelson L, Cibulskis K, Connolly KM, Covarrubias M, Donnelly S, Ferriera S, Gabriel S, Gentry J, Gupta N, Jeandet T, Kaplan D, Llanwarne C, Munshi R, Novod S, Petrillo N, Roazen D, Ruano-Rubio V, Saltzman A, Schleicher M, Soto J, Tibbetts K,

Tolonen C, Wade G, Talkowski ME; Genome Aggregation Database Consortium, Neale BM, Daly MJ, MacArthur DG. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature 2020; 581: 434-443.

18) Rodriguez JM, Rodriguez-Rivas J, Di Domenico T, Vázquez J, Valencia A, Tress ML. APPRIS 2017: Principal isoforms for multiple gene sets. Nucleic Acids Res 2018; 46: D213-D217.
19) The VarSome API. Available online: https://varsome.com/information/varsome-api-info/ (accessed on 8 April 2021).
20) VarSome API Environments. Available online: https://docs.varsome.com/en/varsome-api-environments (accessed on 8 April 2021).
21) VarSome ACMG Implementation. Available online: https://varsome.com/about/resources/acmg-implementation/ (accessed on 8 April 2021).
22) Liu X, Li C, Mou C, Dong Y, Tu Y. dbNSFP v4: A comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. Genome Med 2020; 12: 103.
23) Jian X, Boerwinkle E, Liu X. In silico prediction of splice-altering single nucleotide variants in the human genome. Nucleic Acids Res 2014; 42: 13534-13544.
24) Ghosh R, Harrison SM, Rehm HL, Plon SE, Biesecker LG; ClinGen Sequence Variant Interpretation Working Group. Updated recommendation for the benign stand-alone ACMG/AMP criterion. Hum Mutat 2018; 39: 1525-1530.
25) Association for Clinical Genomic Science. Available online: https://www.acgs.uk.com/quality/best-practice-guidelines/ (accessed on 8 April 2021).