iVar, an interpretation-oriented tool to manage the update and revision of variant annotation and classification

Sara Castellano[1], Federica Cestari[2], Giovanni Faglioni[2], Elena Tenedini[3], Marco Marino[3], Lucia Artuso[3], Rossella Manfredini[1], Mario Luppi[1], Tommaso Trenti[3], and Enrico Tagliafico[1]

[1]University of Modena and Reggio Emilia
[2]Nabla2 s.r.l
[3]University Hospital Modena

November 23, 2020

Abstract

The rapid evolution of Next Generation Sequencing in clinical settings and the resulting challenge of variants interpretation in the light of constantly updated information, requires robust data management systems and organized approaches to variant reinterpretation. In this paper, we present iVar: a freely available and highly customizable tool provided with a user-friendly web interface. It represents a platform for the unified management of variants identified by different sequencing technologies. iVar accepts, as input, VCF files and text annotation files and elaborates them, optimizing data organization and avoiding redundancies. Updated annotations can be periodically re-uploaded and associated to variants as historicize attributes. Data can be visualized through variant-centered and sample-centered interfaces. A customizable search functionality can be exploited to periodically check if pathogenicity related data of a variant are changed over time. Patient recontacting ensuing from variant reinterpretation is made easier by iVar through the effective identification of all patients present in the database and carrying a specific variant. We tested iVar by uploading 4171 VCF files and 1463 annotation files, obtaining a database of 4166 samples and 22569 unique variants. iVar has proven to be a useful tool with good performances for collecting and managing data from medium-throughput

Introduction

The use of Next Generation Sequencing in clinical contexts has rapidly evolved: automated wet lab procedures, as well as sequencing platforms and data analysis pipelines, have become increasingly reliable, producing more and more economic amount of genomic data. How the ever-growing number of genetic variants relates to disease, i.e. variants' interpretation, has now become the major challenge for clinical genomics laboratories. Hence, as well as the evolution of knowledge, the databases for variant interpretation are constantly growing. Consequently, information regarding the pathogenicity of genomic variants, coming from biological and clinical data, raise dynamically as new evidences are acquired. It means that the interpretation of pathogenicity may change over the time and obliges clinical diagnostic laboratories to manage the reinterpretation of variants and to concern its ethical issues. This fact is particularly relevant for variants classified as of uncertain -or unknown- significance (VUS). These variants, for which the current scientific knowledge does not allow classification as either pathogenic/likely pathogenic or benign/likely benign, are the most challenging for both patients' clinical and psychological management. Changing the classification of a variants previously stated as VUS may represent a major issue, especially when genomics is used for the diagnosis of hereditary diseases and it may impact the management of the affected carriers and the choice to extend the testing to all the potential carriers within a family.

Multigene panels for hereditary cancer risk assessment have shown an overall VUS range of 34-41% (Frey et al., 2015; Lincoln et al., 2015). In particular, it has been reported that the 7.3 % of patients who underwent NGS testing with an hereditary cancer multi-gene panel, turned out to harbor variants that have been reclassified and that 94% of them caused a change in their clinical management (Turner, Rao, Morgan, Vnencak-Jones, & Wiesner, 2019).

More in general, the issue of reinterpretation of variants in molecular genetics laboratories is leading to the need for guidelines and tools (Bombard et al., 2019; Chisholm et al., 2018). The potential benefits that reclassification of variants can bring to patient care, sharpen the need for robust data management systems and organized approaches to variant reinterpretation (Appelbaum, Parens, Berger, Chung, & Burke, 2020).

Currently, there is no consensus on how and how frequently a clinical laboratory should revise the classification of variants in patients tested in the past. However, despite the absence of a clear legal duty to re-contact patients after revision of genomic test results, it is a shared opinion that the laboratory has the ethical responsibility to inform the clinicians about events of variants reclassification. Pursuing this goal, involves an optimization of the limited resources that are currently available (Appelbaum et al., 2020; Carrieri et al., 2019; David et al., 2019). Appelbaum and collaborators recently proposed to conceptualize the ethical duty to reinterpret genetic variants by identifying four elements: data storage, initiation of reinterpretation, data reinterpretation and re-contact of patients (Appelbaum et al., 2020). Accessibility to data is obviously a basic prerequisite to reinterpretation. Initiation of reinterpretation regards the triggering of the process, which can be periodical, when a certain number of changes in interpretation are accumulated, or with a stakeholder decision. Data reinterpretation involves both a data analytic pipeline and human judgment. Recontacting of patients regards the responsibility of reaching out to clinicians that have in charge the affected or healthy carriers to communicate them the reinterpreted results.

Laboratories should therefore be able to identify previously analysed patients harbouring certain variants, in order to start the process of recontacting. In this context, suitable informatics tools can simplify the process, mak\souting recontacting as efficient as possible.

Here we present iVar, a freely available tool with a user-friendly web interface. This tool can help to fulfil the aforementioned duties by providing a platform for the unified management of variants identified by different sequencing technologies. Most importantly, iVar represents a useful tool for easily setting up an automated process of periodic re-annotation of variants that allows users to check if pathogenicity of a variant may have changed. Furthermore, patients' recontacting is made easier through the effective identification of all the patients, present in the database, who carry a specific re-annotated variant. In addition, a high level of customization is provided, giving the possibility to upload potentially all the VCF files generated by different tools, including free (e.g. GATK or samtools) and commercial softwares (e.g. Torrent Suite Software or MiSeq Reporter), and the annotation files arising from any bioinformatic pipeline. Since the database can easily be queried by users even without bioinformatics expertise, it can work as a valuable tool to assist geneticist and clinicians in retrieving data for statistical analysis.

Implementation and Overview

iVar workflow

iVar package is publicly available at Github repository (https://github.com/CGR-UNIMORE/iVar).

iVar takes VCF files and annotation files as input (figure 1). The Variant Call Format (VCF) is a text file format containing meta-information lines, a header line, and then data lines each containing information about a position in the genome. The file also has the ability to contain genotype information on samples for each position. Since different variant caller software can generate slightly different VCF files, users must define in advance, through a simple web interface, the specific format of the VCF files to be imported. Moreover, a predefined "gene panel" has to be associated with each imported VCF file indicating the genes included. Additionally, a "virtual panel" filter can be set up to import from VCF files only data lines containing variants included in a predefined gene list. This ensures to import only gene variant data complying to

informed consent, if extended gene panels are utilized for sequencing The second type of file that can be imported is the annotation file: a text file containing data obtained from custom annotation pipelines or commercial tools for variant annotation and classification. As for VCF, the import format for the annotation file can be customized.

Imported and structured data can be viewed through different interfaces. In particular, users can examine a list of variants to evaluate the pathogenicity class label by accessing the annotation information, which relies on the annotation file previously defined, and associate variants to tested samples. Additionally, it is possible to visualize a list of samples and check their variants for pathogenicity classification, allele frequency, and genotype.

Variants annotations can be kept up to date through a process that we term "reannotation". This consists of three main steps: (i) export of variants from iVar in VCF file format; (ii) annotation of the exported variants outside iVar, through custom annotation pipelines or other existing annotation tools; (iii) import of the new annotation file back into iVar. Annotation values are historicized, i.e., modifications are recorded, whenever an updated value is imported, thus keeping track of all changes over time.

To assess if something relevant changes upon reannotation, or at a particular point in time, a customizable search functionality is provided. With this tool, users can specify search conditions for attributes of interest and assess changes to the values of these attributes (e.g changes in the ClinVar attribute from "benign" or "uncertain significance" to "pathogenic").

Software and hardware implementation

The iVar database was developed under Ubuntu 18.04 LTS Linux operating system 64 bit, (\souta 32 bit Linux has a max DB table limit of 2GiB, which is too small). The software was implemented using Python (version 2.7), Web2py Framework (version 2.18.5), Bootstrap4 toolkit, and MariaDB (version 10.3.18) SQL Database backend. Apache (version 2.4.29) and phpmyadmin (version 4.6.6-5) were used as development tools. iVar was built as a platform for collaborative developing with responsive interface for both PC and Mobile.

For database development and testing, a workstation with the following hardware specifications was used: Intel(R) Xeon(R) CPU E3-1231 v3 @ 3.40GHz; 8 GiB RAM; 1TiB Disk, Regarding security: All HTTPS with Let's Encrypt Authority X3 certificate; MariaDB Data-at-Rest Encryption for backup and hdd disposal safety; FS Encryption for VCF File.

Patient consent

According to international and local guidelines, written informed consent for clinical targeted sequencing was obtained from all patients.

Results

In order to test all iVar features, 4,171 VCF files from different analysis platforms 1,463 annotation files produced both by our custom annotation pipeline and by SOPHiA DDM annotation pipeline, were uploaded in order to populate the attributes of variants. During the uploading step, VCF files were filtered according to the patient's obtained informed consent and, where appropriate, a virtual gene panel filter was added. A total of 14 gene panels and 44 virtual gene panels, comprising 301 genes were setup to filter variants associated with different clinical suspects: breast and ovarian hereditary cancer, dyslipidemic disorders, epidermolysis bullosa, hemochromatosis, nephropathies, and retinitis pigmentosa.

VCF files upload

In particular, for testing purposes VCF files generated both by the Torrent Suite Software (TSS) (Saxena et al.) (VCF version 4.1) and by SOPHiA DDM (DDM) (VCF version 4.2) were used.

The TSS VCF were generated using a custom hotspot file containing 5425 variants, including pathogenic variants from ClinVar, Enigma, and LOVD. Therefore, each TSS VCF file includes a large number of variants,

resulting from the hotspot file, where allele frequency is 0 and genotype is 0/0. Hence, a row filter excluding all the lines where the sub-field GT is 0/0 was set when uploading TSS VCF files. Also, we defined the sub-fields allele frequency (AF) and genotype (GT), included in the VCF file "FORMAT", as the attributes linked to the sample, when importing a VCF file. This shows the allelic frequency and the genotype of the variants found within each sample.

For DDM generated VCF files, no filters were applied because no AF sub-field are present. Therefore, to make it congruent with the sample attributes defined in the TSS VCF file type, we set up the SOPHiA DDM VCF file type by exploiting the sub-fields allelic depth (AD) and read depth (DP). The resulting sample attribute is defined as AD/DP*100. Furthermore, considering that DDM VCF files contain additional information in the INFO field, we took advantage of some sub-fields to link these attributes to the variant when importing the VCF file. Specifically, we obtained the gene symbol of the variant from the "SGVEP" sub-field, ClinVar and other database\souts information from the "DBXREF" sub-field, and mutation type from the "TYPE" sub-field.

After defining the VCF file types, 2798 TSS VCF files and 1373 DDM VCF files were uploaded. Next, we checked the number of variants for each imported sample and performed random checks on variant attributes and sample attributes, paying particular attention to variants present in both VCF file types. All VCF files were correctly elaborated by iVar and, importantly, we ascertained that when the same variant is imported from two different VCF file types, the common attributes are overwritten if identical, or historicised if different, preventing redundancies.

Annotation files upload (o text files upload)

Annotation files were uploaded after the VCF files. In particular, using the iVar annotation files definition tool, 9 nine different variants annotation types were defined:

1. AT1: 4 annotation types for files generated by our customized annotation pipelines, which annotate IonTorrent VCF files interrogating databases related to 4 different pathologies: breast and ovarian cancer, dyslipidaemias, epidermolysis bullosa, and hemochromatosis;
2. AT2: 3 annotation types for files generated by our customized annotation pipelines, which annotate SOPHiA DDM VCF files interrogating databases related to 3 different pathologies: breast and ovarian cancer, dyslipidaemias, and nephropathies.
3. AT3: 1 annotation type for files generated by the SOPHiA DDM software;
4. AT4: 1 annotation type for files coming from periodical reannotation with our custom pipeline of VCF files exported from iVar and containing unique variants.

Different annotation types may have common attributes; therefore, we defined the annotation types setting the same names for common attributes. This allows to prevent redundancies and to historicise values when they are different. Additionally, row filters and break conditions can be set if the annotation files contain lines to be skipped, or if only a part of the file is to be imported.

We imported the following annotation files: 54 of (i), of these, 45 were breast and ovarian cancer, 3 dyslipidaemias, 5 epidermolysis bullosa, and 1 hemochromatosis; .61 of type (ii), of these, 48 were breast and ovarian cancer, 7 dyslipidaemias, and 6 nephropathies; 1344 of type (iii); and 4 of type (iv).

Random checks performed on variants attributes showed that all the annotation files were correctly imported and elaborated in iVar.

VCF and annotation uploading performance testing

We tested VCF file uploading and elaboration performance for both types of VCF files considered. A typical TSS VCF file (5000 rows, 3.2 Mb) is imported in 1 and is elaborated in 5 sec. Due to the row filter described above, only 15 variants out of 5000 were uploaded on average. A typical SOPHiA DDM VCF file (500 rows, 150 Kb) also takes 1 second to be imported, but the elaboration time depends largely on the applied virtual panel type. Specifically, it takes took on average 4 seconds to elaborate a VCF file with a 2 genes virtual

panel (50 uploaded variants on average); 6 seconds for a VCF file with a 5 genes virtual panel (70 uploaded variants on average); and 21 seconds for a VCF file with a 22 genes virtual panel (400 uploaded variants on average).)

These results indicate that for larger files containing over 5000 variants, but associated with a very strict row filter, processing times are short. While elaboration times are slightly longer for a smaller VCF file when more variants are uploaded. With our default innodb_page_size of 16KiB, our maximum tablespace size is 64 TiB, which is greater than our hard-disk size. In a particular case, in order to import a 24MiB VCF file, we had to increased max_allowed_packet to 1G and key buffer to 64M. iVar's limitations are inherited from MariaDB and includes limitations on schema, on size, on tables, and on transactions and locks (https://mariadb.com/kb/en/innodb-limitations/).)

For all the annotation files types, uploading took 1 second per file. Elaboration times vary, depending mainly on the number of attributes in the annotation file type, the number of variants, and the number of non-empty attributes in the annotation file. We observed 25 seconds on average to import an AT1 annotation file (with varying numbers of annotations for 90 variants); 2 minutes for an AT2 annotation file (with varying numbers of annotations for 500 variants) and 1 minute for an AT3 annotation file (including, on average, a with varying numbers of annotations for 500 variants).) Differences between elaboration times for AT2 and AT3 annotation files, which include, the same average number of variants, is due to the AT3 annotation type containing about half as many attributes as the AT2.

As a result, we obtained a database containing 4,166 samples and 22,569 unique variants with about 283,659 annotation attributes and, among them, 22,501 variants associated with at least 1 attribute (Table 1).) The total size of all iVar's database tables is about 1.2 GiB.

Using the five tier International Agency for Research on Cancer (IARC) classification system (Plon et al., 2008), and following the guidelines of the American College of Medical Genetics and Genomics (ACMG) for the interpretation of sequence variants (Richards et al., 2015) we assigned a pathogenicity classification, to 1,016 variants (Table 2).) For these, classification was batch uploaded using a tab-delimited text file, for which an appropriate text file type was previously set. Variants classification, however, can be either imported along with the other variant's attributes within an annotation file, or set directly in iVar via the web interface.

Queries functionality testing

A set of common queries were executed for performance testing. In particular, a simple variant search on 22,569 variants was completed in 1 sec, while searching for all variants classified as C5 (330 resulting variants) required 2.3 sec. Searches for variants and their attributes took, on average, 2 sec.

iVar variants export

The "export vcf for reannotation" functionality allows users to export all the unique variants present in the database as a VCF file, readily parsable with any annotation pipeline. We timed this feature by exporting the variants after data upload. It took 1 sec to export a zip file containing 3 VCF files, including up to 10,000 variants each.

Reannotation

To test the reannotation pipeline for all variants in the database, we exported the 22,569 unique annotated variants of iVar database, annotated them using our pipeline, and re-imported the resulting file back into iVar. It took 3 sec to import the file (AT4 annotation file type) into iVar and 30 minutes to elaborate it.

Functionality for annotation changes check

The "variants and attributes search" functionality can assess whether some attribute's values changed after reannotation. This is a customizable search functionality that allows users to set up different types or searches to answer clinical questions. The functionality allows search conditions for variants, attributes,

and previous attributes. To time it, we defined search criteria to identify variants in which ClinVar values changed from "benign" or "uncertain significance" to "pathogenic". This search took about 3 sec and produced 11 results. This way identified, among the 11 selected variants, for example, a BRCA2 variant (chr13|32936829|A|G, NM_000059.3:c.7975A>G) that was consequently reclassified from C3 (uncertain significance) to C5 (pathogenic) allowing us to enrol a new patient in the surveillance program.

Discussion

iVar has proven to be a useful platform for collecting and managing data from our diagnostic laboratory, where genomic data related to hereditary diseases are regularly produced using different sequencing technologies. This first version of the software was designed, as a proof of concept, neglecting possible optimizations of the database structure and showed good performances in the management of medium throughput laboratory like ours. In particular, we uploaded 4171 VCF files produced by our laboratory since 2014, using 44 different virtual panels. We also tested our tool handling a volume of data about 10 times larger, with acceptable performances. The management of massively larger amounts of data, is very likely to require further database optimizations, such as the sizes of caches, and buffers or other parameters. In some instances, it could require more powerful hardware.

The iVar structure was designed to be highly customizable and it is particularly suitable to handle information from different types of input files, to optimize data organization, and to prevent redundancies. We uploaded VCF files, produced by two different technologies and a total of 1463 annotation files produced by different annotation pipelines, comprising both custom and commercially provided software.

The functionality for evaluating annotation changes is a powerful tool to manage reinterpretation of variants in molecular genetics laboratories. Exploiting this functionality, we were able to identify a BRCA2 variant, which was previously classified as "uncertain significance" and to reclassify it into a "pathogenic" variant on the basis of the updated ClinVar annotation.

In the current version, this process is performed manually after each reannotation or whenever changes occurring after a certain date are to be evaluated. In the future, we plan to implement an automatic reannotation process, to be run in the background periodically. This would enable notification to the users whenever significant changes in variants annotations occur. Users will then verify the extent of these changes to assess possible variant reclassifications, identify patients, and start the recontacting process.

iVar could be further improved by linking the database to an existing database containing clinical data, using a unique identifier for each patient. This would allow the integration of clinical and genomic data and consequently make iVar a useful tool to investigate family relationships, and to study genotype-phenotype correlations.

Acknowledgments

This work was supported by Progetto POR-FESR "Hologene 7 2.0: L'Epidermolisi Bollosa (EB) a Modena dalla diagnosi alla terapia genica" (PG/2018/631674), a valere sul Bando Regione Emilia Romagna 986/2018 "Bando per progetti di ricerca industriale strategica rivolti agli ambiti prioritari della Strategia di Specializzazione Intelligente" CUP E51F18000380009".

Disclosure statement: The authors declare no conflict of interest

References

Appelbaum, P. S., Parens, E., Berger, S. M., Chung, W. K., & Burke, W. (2020). Is there a duty to reinterpret genetic data? The ethical dimensions.*Genet Med, 22* (3), 633-639. doi:10.1038/s41436-019-0679-7

Bombard, Y., Brothers, K. B., Fitzgerald-Butt, S., Garrison, N. A., Jamal, L., James, C. A., . . . Levy, H. P. (2019). The Responsibility to Recontact Research Participants after Reinterpretation of Genetic and Genomic Research Results. *Am J Hum Genet, 104* (4), 578-595. doi:10.1016/j.ajhg.2019.02.025

Carrieri, D., Howard, H. C., Benjamin, C., Clarke, A. J., Dheensa, S., Doheny, S., . . . European Society of Human, G. (2019). Recontacting patients in clinical genetics services: recommendations of the European Society of Human Genetics. *Eur J Hum Genet, 27* (2), 169-182. doi:10.1038/s41431-018-0285-1

Chisholm, C., Daoud, H., Ghani, M., Mettler, G., McGowan-Jordan, J., Sinclair-Bourque, L., . . . Jarinova, O. (2018). Reinterpretation of sequence variants: one diagnostic laboratory's experience, and the need for standard guidelines. *Genet Med, 20* (3), 365-368. doi:10.1038/gim.2017.191

David, K. L., Best, R. G., Brenman, L. M., Bush, L., Deignan, J. L., Flannery, D., . . . Committee, A. S. E. L. I. (2019). Patient re-contact after revision of genomic test results: points to consider-a statement of the American College of Medical Genetics and Genomics (ACMG).*Genet Med, 21* (4), 769-771. doi:10.1038/s41436-018-0391-z

Frey, M. K., Kim, S. H., Bassett, R. Y., Martineau, J., Dalton, E., Chern, J. Y., & Blank, S. V. (2015). Rescreening for genetic mutations using multi-gene panel testing in patients who previously underwent non-informative genetic screening. *Gynecol Oncol, 139* (2), 211-215. doi:10.1016/j.ygyno.2015.08.006

Lincoln, S. E., Kobayashi, Y., Anderson, M. J., Yang, S., Desmond, A. J., Mills, M. A., . . . Ellisen, L. W. (2015). A Systematic Comparison of Traditional and Multigene Panel Testing for Hereditary Breast and Ovarian Cancer Genes in More Than 1000 Patients. *J Mol Diagn, 17* (5), 533-544. doi:10.1016/j.jmoldx.2015.04.009

Plon, S. E., Eccles, D. M., Easton, D., Foulkes, W. D., Genuardi, M., Greenblatt, M. S., . . . Group, I. U. G. V. W. (2008). Sequence variant classification and reporting: recommendations for improving the interpretation of cancer susceptibility genetic test results. *Hum Mutat, 29* (11), 1282-1291. doi:10.1002/humu.20880

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., . . . Committee, A. L. Q. A. (2015). 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, 17* (5), 405-424. doi:10.1038/gim.2015.30

Turner, S. A., Rao, S. K., Morgan, R. H., Vnencak-Jones, C. L., & Wiesner, G. L. (2019). The impact of variant classification on the clinical management of hereditary cancer syndromes. *Genet Med, 21* (2), 426-430. doi:10.1038/s41436-018-0063-z

Tables (each table complete with title and footnotes);

Table 1

	Number of variants with at least 1 attribute
Exonic regions	6084
Exonic splicing junctions	3
Exonic regions of noncoding transcripts	52
Intronic regions	12801
Intronic regions of noncoding transcripts	524
Splicing junctions	90
3'-UTR	1648
5'-UTR	415
Missense	5026
Nonsense	63
Synonymous SNV	3169
Variants affecting splicing	720
frameshift	711

Table 2

Pathogenicity Classification of 1,100 annotated variants	Total Number (%)
C5 (Pathogenic)	330 (32.5%)
C4 (Likely Pathogenic)	38 (3.7%)
C3 (Uncertain Significance)	333 (32.8%)
C2 (Likely Benign)	206 (20.3%)
C1 (Benign)	109 (10.7%)

Figure and table legends;

Figure 1. Overview of iVar, a platform for the unified management of variants identified by different sequencing technologies

Table 1. Description of regions and types of variants uploaded in iVar

Table 2. Summary of classified variants included in iVar