Informatics and Technology in Clinical Care and Public Health
J. Mantas et al. (Eds.)

doi:10.3233/SHTI210855

Development of an Expert Knowledge-Based Genomic Variant Prioritisation Platform

Aurora SUCRE[a,b], Gregory MACLAIR[a], Iride MARTINEZ[c], Concha VIDALES[c], Gorka EPELDE[a,b] and Alba GARIN-MUGA [a,b]
[a] *Vicomtech Foundation, Basque Research and Technology Alliance (BRTA), Mikeletegi 57, Donostia-San Sebastián, Spain*
[b] *Biodonostia Health Research Institute, eHealth Group, Donostia-San Sebastián, Spain*
[c] *I+D+I department, DNAData, Donostia-San Sebastián, Spain*

Abstract. Considering the growing interest towards next generation sequencing (NGS) and data analysis, and the substantial challenges associated to fully exploiting these technologies and data without the proper experience, an expert knowledge-based user-friendly analytical tool was developed to allow non-bioinformatics experts to process NGS genomic data, automatically prioritise genomic variants and make their own annotations. This tool was developed using a user-centred methodology, where an iterative process was followed until a useful product was developed. This tool allows the users to set-up the pre-processing pipeline, filter the obtained data, annotate it using external and local databases (DBs) and help on deciding which variants are more relevant for each study, taking advantage of its customised expert-based scoring system. The end users involved in the project concluded that CRIBOMICS was easy to learn, use and interact with, reducing the analysis time and possible errors of variant prioritisation for genetic diagnosis.

Keywords. Genomics, genetic variants prioritisation, user-centred design, standalone tool.

1. Introduction

Genetic variants refer to the variations found on a DNA sequence that may alter the individual's anatomy, physiology, psychology, disease predisposition and drugs susceptibility. The detection and screening of these variants is gaining a good reputation, as it is becoming a very powerful tool in personalised medicine, specifically for diagnosis. Different tools that help researchers doing semiautomatic variant prioritisation exist in the market, but most of them simply do not fulfil all the users' requirements that were detected during this project. Therefore, many researchers and biologists are relying on time-consuming manual mechanisms for variant prioritisation, so their approaches are far from being optimised, having a huge impact on the geneticists' productivity.

In this paper, we present CRIBOMICS, a novel variant prioritisation tool based on expert knowledge. This tool was designed following a user-centred-design (UCD) methodology. UCD methodologies have been applied extensively to develop clinical applications, and have been found to be successful in this field (1). CRIBOMICS allows to process raw NGS data, call genomic variants (SNPs and CNVs), filter them by various features, annotate them using well-known DBs and automatically prioritise the kept variants according to their clinical relevance. Additionally, a customizable DB is embedded in the tool to allow experts to record their own findings in CRIBOMICS and have them available in future analysis. The automatic prioritisation process is based on the formalisation of experts' knowledge into rules. Expert end-users confirmed that the tool facilitates variant prioritisation for genetic disease diagnosis.

2. Methods

The main objective of CRIBOMICS was to create an automatic tool for variant prioritisation that could facilitate the genetic expert workflow. To do so, an iterative UCD methodology was followed, where a multidisciplinary expert committee was involved in the entire development process, in order to conduct an evaluation after every iteration, to provide feedback regarding the tool usability and effectiveness in real conditions. The panel was established at the beginning of the project following Nielsen recommendations (2).

During the initial stages, the main requirements of the end-users were collected and contrasted with the variant analysis guidelines defined by the ACMG (3), to make sure the tool would be compliant with existing standards. Afterwards, the engineering experts were able to define the user task scenarios, the different needed functionalities, and the overall system architecture. Regarding the algorithm for automatic prioritisation, its logic was determined by having the geneticists explain their decision-making process and the developers formalise it into rules that can be interpreted by the software. All the relevant parameters, thresholds, decision-making conditions were collected and translated to be embedded in the knowledge core of the system.

On every development iteration, the experts analysed the novel features and the modifications in the GUI, and provide their feedback following a preestablished protocol, so developers can exploit the data and implement a list of enhancements on every iteration, until the tool fulfilled all the user requirements. The initial versions of the software focused on the secondary analysis, which comprises the filtering, annotation, and prioritization of SNPs. Then, after several iterations the committee detected the need to include new modules for primary analysis (alignment, quality control, variant calling), which was achieved by implementing the GATK pipelines for data pre-processing and germline variant calling (4). Finally, the same happened for CNV detection and analysis, and the necessary modules were developed following GATK best practices.

3. Results

The UCD methodology allowed us to tailor the solution to the experts' needs and enabled the automatization of the process to ease their workflow by formalising their reasoning into a rule-based system. The iterative methodology encouraged having constant interactions with end-users and facilitated the development of a functional user-friendly standalone application that automates the prioritisation of genomic variants and can be exploited by specialists with no bioinformatics knowledge.

The obtained solution combines a GUI developed using QT and Python, different analytical modules based on Python and other third-party tools for genome data processing and prioritisation (i.e. GATK, Samtools, Picard and BWA).

The tool allows users to analyse the data associated to a single patient's genetic sample and also includes modules for duo (5) and trio analyses. In single analyses, it is possible to analyse both SNPs and CNVs and the process varies based on the maturity of the input data and given file format (BCL, FQ, BAM, VCF). CRIBOMICS includes pre-processing modules to perform all the steps from base calls deconvolution to variant calling, in order to obtain the patients' SNPs and CNVs in VCF format. This pipeline is based on the GATK best practices for germline variant discovery (4). Users can tune the analytical pipeline by modifying several parameters; however, default parameters are set to ease the analysis.

Once the variants are stored in the VCF format, the tool extracts all the relevant data regarding the detected variants and performs an initial annotation step to obtain basic information that may be missing (RS IDs, genes, genomic location and the effect they may have in the proteins). Regarding CNV analysis, also the genes that were completely or partially affected by each variant are collected.

Then, the data is shown in summarised tables, and users are able to filter and annotate the detected variants according to various categories.

The variants can be annotated from various sources:

- NCBI: Clinical significance according to ClinVar, diseases associated with the variant or with the gene and publications found in PubMed.
- Ensembl: SIFT prediction, Polyphen prediction and HGVSc IDs.
- dbNSFP: MT2 prediction.
- Internal DB: Users can store their findings in a customisable DB and their annotations will be available for future analyses.

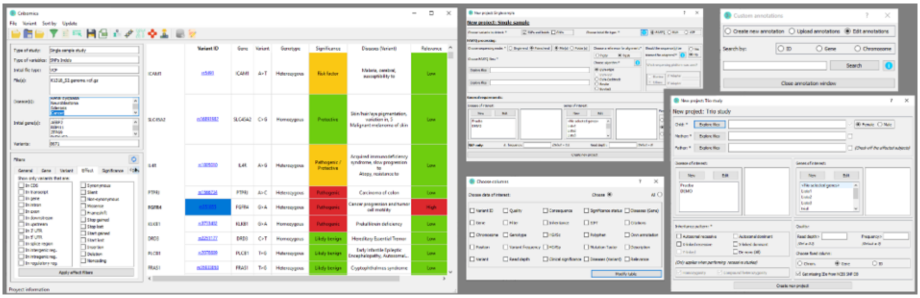
Initially, the SNPs can be filtered according to their 1) genotype, 2) chromosome, 3) affected genes, 4) different quality parameters, 5) genomic location and 6) protein consequence. Once they are annotated, they can also be filtered by their 7) clinical significance and 8) predicted pathogenicity.

CRIBOMICS also includes a proprietary scoring system that helps prioritize the variants during each analysis. This system is based on the experts' decision-making process, that was successfully formalised into logical rules that can be interpreted by the software.

The scoring system set of rules considers 1) the diseases of interest, 2) the clinical significance of each variant, 3) the pathogenicity predictions, and 4) the diseases that have been found associated to the variants or the genes, to objectively classify all detected variants and give them a high, medium, or low relevance tag. Depending on the available information for each analysis, the relevance tags are assigned following different rules. For instance, if no diseases of interest are defined during the analysis, the algorithm only focuses on the clinical significance and prediction results, but if the user provides any disease of interest, the related diseases feature gains relevance and constrains the output of the algorithm.

CRIBOMICS final output includes all the initial relevant data, the annotated information and the relevance score assigned to all the kept variants. This can be stored as an MS Excel file or in a CRIBOMICS-specific format, to be further analysed in the future.

The current analytical workflow was designed for the existing standalone application (Figure 1), but as it was conceived as a modular, customizable solution, it is currently being migrated to a cloud environment without requiring extensive resources.



Type of study: Single sample study
Type of variables: SNPs Indels
Initial file type: VCF

	Variant ID	Gene	Variant	Genotype	Significance	Diseases (Variant)	Relevance
ICAM1	rs5491	ICAM1	A>T	Heterozygous	Risk factor	Malaria, cerebral, susceptibility to	Low
SLC45A2	rs16891982	SLC45A2	C>G	Homozygous	Protective	Skin/hair/eye pigmentation, variation in, 5 Malignant melanoma of skin	Low
IL4R	rs1805010	IL4R	A>G	Heterozygous	Pathogenic / Protective	Acquired immunodeficiency syndrome, slow progression to Atopy, resistance to	Low
PTPRJ	rs1566734	PTPRJ	A>C	Heterozygous	Pathogenic	Carcinoma of colon	Low
FGFR4	rs351855	FGFR4	G>A	Heterozygous	Pathogenic	Cancer progression and tumor cell motility	High
KLKB1	rs3733402	KLKB1	G>A	Homozygous	Pathogenic	Prekallikrein deficiency	Low
DRD3	[illegible]	DRD3	C>T	Homozygous	Likely benign	Hereditary Essential Tremor	Low
PLCB1	[illegible]	PLCB1	T>G	Heterozygous	Likely benign	Early infantile Epileptic Encephalopathy, Autosomal...	Low
FRAS1	[illegible]	FRAS1	T>G	Heterozygous	Likely benign	Cryptophthalmos syndrome	Low

Figure 1. CRIBOMICS graphical user interface

4. Discussion

The UCD approach that was followed proved to be effective when developing this kind of solution, where the final-user contributions are so relevant in order to meet their needs.

During development it was mandatory to exploit standard or at least the most commonly-used tools in the field, therefore the GATK best practices (4) were followed during primary analysis, and the NCBI and Ensembl DBs were used for data annotation.

According to the end-user's evaluation, CRIBOMICS is indeed an intuitive tool that ease the SNP and CNV detection and prioritisation processes, and the proprietary scoring system seems to be the most useful feature, that simplifies their interpretation of the data and allow them to reach conclusions faster, even without any bioinformatic skill. Also the usability of the DB was emphasised, as it helps on keeping track of new findings.

Regarding future work, an authoring tool is currently being developed to allow experts to easily formalise their own prioritisation rules, and make the system customisable.

It is important to highlight that this tool does not make decisions regarding diagnosis but provides experts with recommendations based on the available information and needs.

Even if the tool has been tested by the expert committee and they have provided promising feedback, the tool it is still not publicly available, as it first need to overcome validation. This process will be performed by a larger group of experts that will evaluate: 1) software usability, 2) results accuracy and 3) software efficiency.

References

1. Luna D, Quispe M, Gonzalez Z, Alemrares A, Risk M, Otero C, et al. User-centered design to develop clinical applications. Literature review. Stud Health Technol Inform. 2015;216:967–967.
2. Nielsen J, Landauer TK. A Mathematical Model of the Finding of Usability Problems. In: Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems [Internet]. New York, NY, USA: Association for Computing Machinery; 1993. p. 206–13. (CHI '93). Available from: https://doi.org/10.1145/169059.169166
3. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. 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(5):405–23.
4. Van der Auwera GA, O'Connor BD. Genomics in the Cloud: Using Docker, GATK, and WDL in Terra. O'Reilly Media; 2020.
5. Sucre A, Garcia-Longarte S, Garin-Muga A. DONOR-FINDER: A Web Tool for the Automatic Comparison of Genomic Profiles in ART. Stud Health Technol Inform. 2020 Jun 26;272:139–42.