An open-source tool for managing time-evolving variant annotation

Ilio Catallo(1), Eleonora Ciceri(1), Stefania Stenirri(1), Stefania Merella(2), Alberto Sanna(1), Maurizio Ferrari(2,3,4), Paola Carrera(2,3), Sauro Vicini(1)

(1) e-Services for Life and Health, San Raffaele Scientific Institute, Milan, Italy
(2) Clinical Molecular Biology Laboratory, San Raffaele Scientific Institute, Milan, Italy
(3) Unit of Genomics for Human Disease Diagnosis, San Raffaele Scientific Institute, Milan, Italy
(4) Chair of Clinical Pathology Vita-Salute San Raffaele University, Milan, Italy
Email: last.first@hsr.it

Keywords: variant annotation, reclassification, genomics, genetics, next-generation sequencing.

Abstract. Genomics is drawing more and more attention in these last years, thanks to the introduction of fast and accurate sequencing strategies. Accumulation of data is fast and the amount of information to be managed and integrated is snowballing. While new variants (i.e., variation in an organism's genome sequence) are discovered every day, we still do not know enough about the human genome to have a final understanding of all the implications that they could have from a clinical point of view. When inherited diseases are considered, variants clinical classification may change over time, in relation to new discoveries. In this scenario, software solutions that help operators in the analysis and maintenance of constantly changing genomic data are relevant in the field of modern molecular medicine. In this paper we present GLIMS, an open-source laboratory information management system for genomic data that allows to deal with time-evolving variant annotations. This solution answers to the need of genomic laboratories to keep up with their knowledge about variants and annotations, so as to provide patients with up-to-date reports. We illustrate the architecture of GLIMS modules that are in charge of keeping the database of variants updated and reclassifying patients variants. Then, we demonstrate (via the use of GLIMS) that variant clinical classifications are changing rapidly even in ClinVar, one of the most known and cited genomic databases, thus underlining the need for a tool that tracks changes over time.

1 Scientific Background

A *genome* is the genetic material of an individual. The genetic instructions it contains are used in the growth, development, functioning and reproduction of individuals, and define one's phenotype (that is, one's observable characteristics or traits). Genomes contain *genes*, i.e., regions of DNA that encode specific functions. Genes can acquire mutations in their sequence of nucleotides, leading to different *variants* in the population. Every variant comes with a set of *genomic annotations*, which state its semantics (e.g., specifying whether it is associated with an increased probability of developing a pathology) and its biological structure. Over the years, scientists have published several archives of genomic variants and annotations. A well-known example is ClinVar[1], a freely available archive of clinically significant relationships among human variations and phenotypes. Such archives are not limited to specific pathologies and constantly updated to reflect the knowledge that researchers acquire over time, providing benefits on two main aspects. First, the constant update permits the gathering of new information about Variants of Unknown Significance (VUS), which are known to change very

[1] https://www.ncbi.nlm.nih.gov/clinvar/

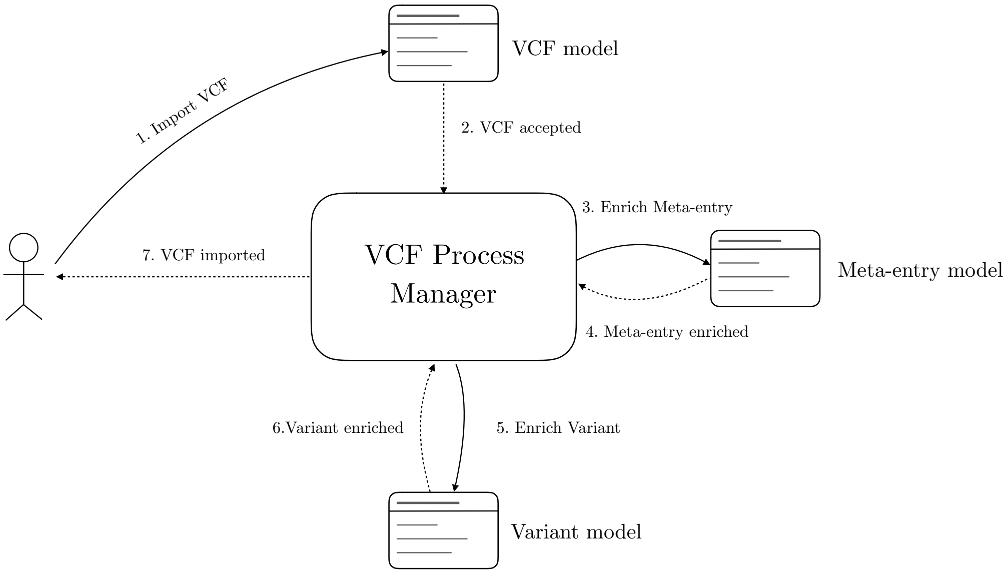

Figure 1: The sequence of operations needed for enriching the internal database

frequently towards a clear pathological significance [1]. As a matter of fact, an uncertain finding can be frustrating for clinicians and patients alike, who may decide for drastic treatment measures (e.g., surgical decisions) only in the name of the persistent fear of contracting a disease [2]. Second, since multiple sources of genomic information may differently interpret the clinical significance of a variant [3], archives of genomic variants often include errors and misinterpretations [4] and present classification inconsistencies (even for well-studied genomic panels [5]). In this regard, a continuous update guarantees that exams will be carried out against the most coherent annotations.

Thus, it is vital for laboratories to: *i)* keep their knowledge about variant classifications updated with respect to the current literature; *ii)* track changes in variant classifications in order to identify which patients' past genetic results are in need of a review. Several published works already outlined the necessity of notifying patients when their genetic test clinical significance changes as a consequence of a variant reclassification, arguing about the clinical impact and ethical duties of such actions [6, 7]. From a technical point of view, the introduction of an automatic solution seems to be the best option at scale, as the exponential growth of interest in the genomic field has brought to the production of an unprecedented mass of genomic data and tests, and laboratories would incur in high costs if they had to manually reclassify patients' exams. Some works in the state of the art already presented automatic solutions that allow laboratories to update database variants and genetic test reports upon reclassification of variants [8, 9, 10]. Still, these solutions are mainly proprietary, and work with non-standard (if not unknown) data format.

In this paper, we present GLIMS, an open-source laboratory information management system for genomic data, that, in addition to automatizing alignment and annotation tasks, also provides reclassification capabilities for time-evolving variant annotations. GLIMS promotes interoperability between systems by adopting the well-known Variant Call Format (VCF) standard as the exporting and importing format for genomic variants.

2 Materials and Methods

GLIMS takes care of two important steps in the analysis of Next-Generation Sequencing (NGS) data, i.e., sequence alignment and variant annotation, in a privacy-compliant cloud environment (in order to reduce the burden of such expensive computations). On top of that, GLIMS also provides two more distinctive functionalities, namely, that of *i)* growing an internal database of variants encountered in patients, which

Table 1: Sequence of events stored into GLIMS in a given moment in time

Event	Timestamp
VariantEnriched(id=4c0e94, info=CLNDBN, ...)	2017-04-18T17:14:39.097Z
VariantEnriched(id=4c0e94, info=CLNSIG, ...)	2017-01-10T11:10:28.023Z
VariantCreated(id=4c0e94, chrom=chr17, ...)	2016-03-27T16:37:02.862Z

can be exported in VCF format; and *ii)* supporting periodic reclassification of variants so as to update patients' record whenever there is a change in their variant annotations.

The management of the internal database of variants and the annotation process in GLIMS are supported by the usage of the well-known VCF file format. A VCF file is a text file used for storing gene sequence variations, and it has been widely used in the last years, with the support of large-scale DNA sequencing projects such as ExAC[2] and 1000 Genomes[3]. The diffusion of the VCF format among the most known genomic databases allows GLIMS to import any information they provide, and keep it updated whenever a new version of such databases is released. A VCF file encodes variant characteristics by subdividing the information in two sections: meta-entries and variants. A *meta-entry* provides metadata describing the file content (e.g., the semantics of a variant annotation). A *variant* provides the description of a genomic variant, defining its characteristics (e.g., chromosome, position on the chromosome, expected nucleotide according to the reference genome, found alternative) and decorating it with its annotations.

The process of enriching the internal database with new variants (and their related annotation) is depicted in Figure 1. As shown, the act of importing a VCF is organized as a sequence of *commands* and *events*. When a user asks for a VCF file to be imported, GLIMS stores her request as a new *VCF* model, which in turn causes a *VCF accepted* event to be fired[4]. Such an event is captured by the *VCF Process Manager* component, which initiates and thereupon orchestrates the enriching process. At this point, depending on their position and content, lines in the input VCF are converted to *Enrich Meta-entry* and *Enrich Variant* commands, which are then directed to the proper *Meta-entry* and *Variant* models.

GLIMS stores every event happening in the system. An example of possible events is reported in Table 1. As shown, each event describes how the system has been altered as a consequence of its occurrence. This means that the state of any model can be reconstructed by simply re-applying what happened in the past. In this respect, the presence of a *timestamp* allows every model to be restored to a specific moment in time.

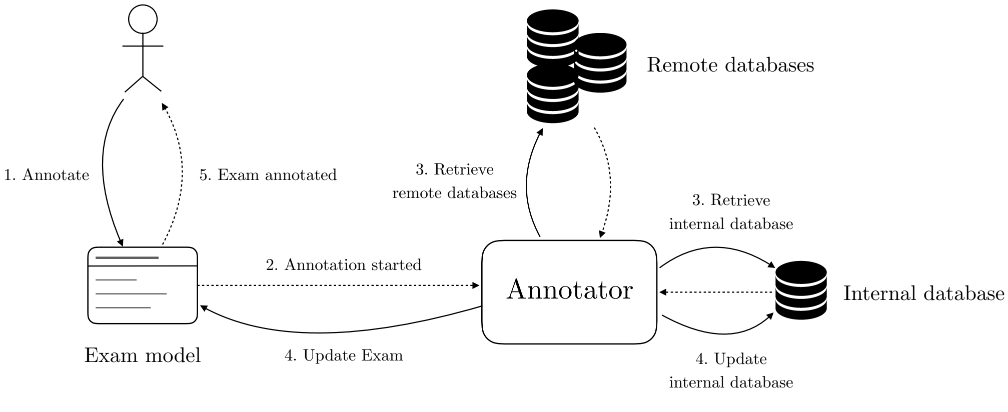

Figure 2: The sequence of operations needed for the annotation of variants in a patient's exam

[2] http://exac.broadinstitute.org/

[3] http://www.internationalgenome.org/

[4] It is important to note that *VCF accepted* events are always fired asynchronously, as the process of importing even small-sized VCF files has to be considered a long-running operation.

Table 2: Dataset: ClinVar releases

2014	**2016(1)**	**2016(2)**	**2017(1)**	**2017(2)**
February 2014	January 2016	November 2016	February 2017	April 2017

A pictorial representation of the annotation process is presented in Figure 2. This process can be used for creating a new exam, as well as for its periodic reclassification, in that reclassification can be seen as a sequence of two annotations on the same exam. Like the enriching process, the interaction between components is driven by commands and events. A user triggers the annotation of a patient's exam by issuing an *Annotate* command. The *Annotator* component reacts to such a request by first making a local copy of the internal database, fetching an up-to-date version of each requested remote database, and then proceeding with the actual annotation. Upon completion, the output VCF may contain novel information with respect to both the internal database and the current state of the exam. Therefore, the *Annotator* takes care of submitting the VCF to the internal database, as well as updating the *Exam* model of interest. Finally, the user is notified with an *Exam annotated* event, which informs her (e.g., via a notification through the user interface) about relevant changes in the annotation of patients' variants.

As anticipated, the crucial aspect for the *Annotator* component is to annotate the patient's variants against the appropriate version of the internal database. To this end, the exam model maintains a list of references to every database that has ever been used for its annotation. When performing a reclassification, the *Annotator* component can therefore instruct the internal database to provide only those variants that have been subject to change since the last reclassification.

3 Results

In this section we demonstrate the need of an automatic solution, such as the one proposed by GLIMS, by evaluating the variability of genomic variants and annotations contained in ClinVar.

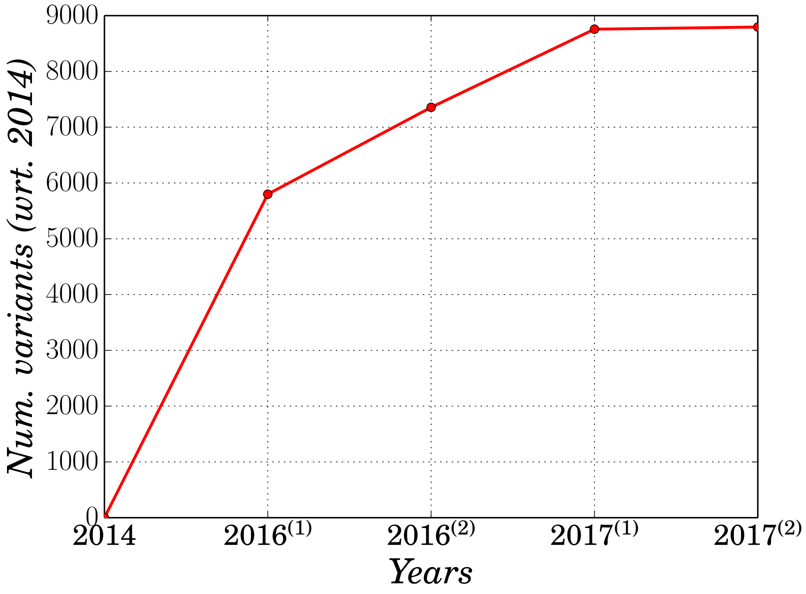

Figure 3: New variants added to ClinVar since release 2014

We retrieved five ClinVar GRCh37 (human `hg19`) releases (Table 2), retained only the variants belonging to the `BRCA1`/`BRCA2` genes and finally imported the resulting VCF files in GLIMS. For each considered release, we then computed: *i)* the number of new variants introduced with respect to release 2014; *ii)* the number of variants whose clinical significance (CLNSIG) annotation or variant disease name (CLNDBN) annotation changed with respect to previous releases.

Figure 3 reports the number of new variants added to ClinVar since release 2014. It is shown that the volume of data in ClinVar continues to grow, gaining up to 360% of the initial volume in two years. This underlines the necessity of maintaining the knowledge on variants constantly updated, so that when a patient requires a new genomic test, her variants can be correctly interpreted at the best of researchers' knowledge.

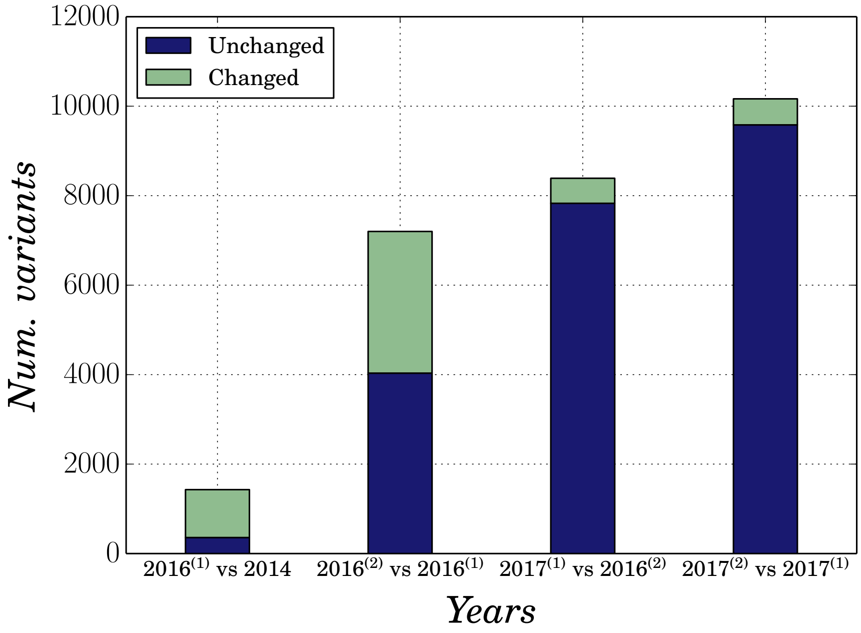

Figure 4: Variants with changed and unchanged CLNSIG or CLNDBN annotations

Figure 4 reports the number of release-to-release changes in CLNSIG or CLNDBN annotations. It is shown that changes happen frequently also for variants that are already known to the scientific community, updating up to 75% of variant classifications with respect to what was reported in the previous release. This highlights the need of updating periodically the outcomes for already performed genomic tests, so that if an important change on the clinical significance of one's variants is found, then she can be timely notified with an updated report.

These results are further confirmed if the updated variants are analyzed in detail, as in the case of the four variants presented in Table 3. In all these cases, the provided update would impact on the final interpretation on one's variants, and thus this links to the necessity of providing patients with an updated report upon reclassification. Reclassification of variants may have an impact on patients' follow-ups (in terms of surveillance, preventive or pharmacological treatments) as well as on the evaluation of risk for their relatives.

Table 3: Variants evolving over time (human reference genome release: GRCh37/hg19). The reported HGVS nomenclature is the standard for the description of sequence variations.

Variants (HGVS nomenclature)	**Gene**	**Position**	CLNSIG	
			2014	**$2017^{(2)}$**
NM_007294.3(BRCA1) : c.190T > G	BRCA1	41258495	VUS	Pathological
NM_007294.3(BRCA1) : c.3119G > A	BRCA1	41244429	VUS	Benign
NM_000059.3(BRCA2) : c.1889C > T	BRCA2	32907504	VUS	Benign
NM_000059.3(BRCA2) : c.5428G > A	BRCA2	32913920	Pathological	VUS

4 Conclusion

In this paper we presented GLIMS, an open-source laboratory information management system for genomic data on germline mutations, and focused our attention on two distinctive functionalities: the capability of maintaining an evolving database of variants over time, and its support to the periodic reclassification of patients' variants. These two functionalities, when combined, offer the possibility of providing patients with fresh and up-to-date information about their genomic variants, which are known to be always changing and in need of refinements (as also proven by our pilot study on the `BRCA` genes in the ClinVar database). The pilot will be expanded including an extensive analysis of the tool with larger panels, in cooperation with the Clinical Molecular Biology Laboratory at San Raffaele Hospital. This project may include other genomic variations such as Copy Number Variations (CNV), somatic mutations and data from other OMICS studies (e.g., epigenetics, proteomics, transcriptomics). Machine learning techniques may be developed for the automatic classification of VUS clinical significance and the computation of genotype-phenotype correlation. These applications would be extremely important and helpful not only for the integration of data, correlation to the clinical phenotype and formulation of hypotheses, but also in the process of harmonization and standardization of protocols. Heterogeneity of the conclusions drawn by different operators is in fact a variable with an important impact on the classification of genetic variants as well as on the clinical management.

Acknowledgments

This work is partially funded by the EU H2020 Framework Programme under project WITDOM (project no. 644371).

References

[1] Narravula, A. et al. (2016). Variants of uncertain significance in newborn screening disorders: implications for large-scale genomic sequencing. *Genetics in Medicine*.

[2] Murray, M. L. et al. (2011). Follow-up of carriers of brca1 and brca2 variants of unknown significance: variant reclassification and surgical decisions. *Genetics in Medicine*, 13(12):998–1005.

[3] Garber, K. B. et al. (2016). Reassessment of genomic sequence variation to harmonize interpretation for personalized medicine. *The American Journal of Human Genetics*, 99(5):1140–1149.

[4] Salgado, D. et al. (2016). How to identify pathogenic mutations among all those variations: Variant annotation and filtration in the genome sequencing era. *Human Mutation*.

[5] Lincoln, S. E. et al. (2017). Consistency of brca1 and brca2 variant classifications among clinical diagnostic laboratories. *JCO Precision Oncology*, 1:1–10.

[6] Otten, E. et al. (2014). Is there a duty to recontact in light of new genetic technologies? a systematic review of the literature. *Genetics in Medicine*, 17(8):668–678.

[7] Dheensa, S. et al. (2017). A'joint venture'model of recontacting in clinical genomics: Challenges for responsible implementation. *European Journal of Medical Genetics*.

[8] Bean, L. J. et al. (2013). Free the data: one laboratory's approach to knowledge-based genomic variant classification and preparation for emr integration of genomic data. *Human mutation*, 34(9):1183–1188.

[9] Aronson, S. J. et al. (2012). Communicating new knowledge on previously reported genetic variants. *Genetics in medicine*, 14(8):713–719.

[10] Wilcox, A. R. et al. (2014). A novel clinician interface to improve clinician access to up-to-date genetic results. *Journal of the American Medical Informatics Association*, 21(e1):e117–e121.