VariOtator, A Software Tool for Variation Annotation with the Variation Ontology.

Schaafsma, Gerard; Vihinen, Mauno

Published in:
Human Mutation

DOI:
10.1002/humu.22954

2016

VariOtator, A SOFTWARE TOOL FOR VARIATION ANNOTATION WITH THE VARIATION ONTOLOGY

Gerard C. P. Schaafsma and Mauno Vihinen*

Protein Structure and Bioinformatics, Department of Experimental Medical Science, Lund University, BMC B13, Lund SE-221 84, Sweden

*To whom correspondence should be addressed.

Tel: +46 72 526 0022 Email: mauno.vihinen@med.lu.se

ABSTRACT

The Variation Ontology (VariO) is used for describing and annotating types, effects, consequences, and mechanisms of variations. To facilitate easy and consistent annotations, the online application VariOtator was developed. For variation type annotations, VariOtator is fully automated, accepting variant descriptions in Human Genome Variation Society (HGVS) format, and generating VariO terms, either with or without full lineage, that is, all parent terms. When a coding DNA variant description with a reference sequence is provided, VariOtator checks the description first with Mutalyzer and then generates the predicted RNA and protein descriptions with their respective VariO annotations. For the other sublevels, function, structure, and property, annotations cannot be automated, and VariOtator generates annotation based on provided details. For VariO terms relating to structure and property, one can use attribute terms as modifiers and evidence code terms for annotating experimental evidence. There is an online batch version, and stand-alone batch versions to be used with a Leiden Open Variation Database (LOVD) download file. A SOAP Web service allows client programs to access VariOtator programmatically. Thus, systematic variation effect and type annotations can be efficiently generated to allow easy use and integration of variations and their consequences.

KEY WORDS

annotation; ontology; Variation Ontology; bioinformatics; software; database; LSDB; variation; LOVD; mutation

Introduction

Information on (genetic) variation is being collected in a widerange of databases. Central databases comprise, for example, UniProtKB [UniProt Consortium, 2015], ClinVar [Landrum et al., 2014], and Ensembl Variation [Cunningham et al., 2015]. Other types of variation databases include locus-specific databases (LSDBs), of which those using the Leiden Open Variation Database (LOVD) management software [Fokkema et al., 2011] form the majority. LSDBs are generally considered as the most reliable source of variation information as these resources are typically curated by experts in the genes and diseases.

A systematic representation of information facilitates data integration, comparison of data, automated searching within and across databases, and the development of dedicated software tools. Published recommendations for LSDBs [Cotton et al., 2008] include the use of a standardized nomenclature. Systematic gene names and symbols are implemented and approved by the HUGO Gene Nomenclature Committee (HGNC) [Gray et al., 2013]. Standardized reference sequences in the Locus Reference Genomic (LRG) sequence format [Dalgleish et al., 2010] are being created and curated at the European Bioinformatics Institute (EBI) and the National Center for Biotechnology Information (NCBI). The LRG records contain stable fixed reference DNA sequences along with all relevant transcript and protein sequences essential to the description of gene variants, and an exon numbering system [MacArthur et al., 2014]. Efforts are being made to standardize variant descriptions, such as the use of the Human Genome Variation Society (HGVS) nomenclature [den Dunnen and Antonarakis, 2000]. Guidelines for establishing [Vihinen et al., 2012] and curating [Celli et al., 2012] LSDBs highlight the importance of systematics in gene variant databases.

Gene variant databases would benefit from a standardized annotation of variant descriptions, so that automated searches and analyses within and across databases would become possible and/or much easier. One way to create a standardized annotation is to use an ontology, a controlled vocabulary conceptualizing a knowledge domain by defining the central terms and their relationships. The use of consistent terminology is essential to guarantee that the information, the message, is correctly understood [Vihinen, 2015a]. The Gene Ontology (GO) [Ashburner et al., 2000] and the Sequence Ontology (SO) [Eilbeck et al., 2005] are widely used for describing gene products in terms of their associated biological processes, cellular components, and molecular functions (GO) or for describing features pertinent to sequence annotation (SO). These ontologies have a very broad scope. The Variation Ontology (VariO) [Vihinen, 2014a] was developed as a specific ontology for describing and annotating types, effects, and mechanisms of variations. Note that VariO does not contain clinical terms apart from pathogenicity association.

VariO can be used for describing variations and their consequences only, that is, it cannot be utilized for annotation of normal or wild-type situations. It should be possible to describe any type

of variation and effect with VariO. The ontology is work in progress and new terms will be added whenever necessary, for example, to facilitate annotations based on novel technologies. VariO annotations are made by combining terms. VariO is organized in three main levels: DNA, RNA, and protein, each of which has four sublevels: variation type, function, structure, and property. Each of these sublevels has then more detailed terms. Variation type terms describe the origin and classification of variations, including such terms as "VariO:0136 DNA substitution" and "VariO:0147 epigenetic DNA variation." General functions affected by variation can be described with function terms. Structure terms are for describing affected DNA, RNA, and protein structural features, and vary substantially between the three levels. Property terms are used for diverse characteristics, such as conservation of DNA variation site or effect on protein abundance. In addition to these four sublevels, attribute terms can be used as modifiers of the structure and property terms, for instance, to describe effects on quantity or affected interactions due to the variation. Since the function terms are general, modifying these with attribute terms does not addmuch information and thus attribute terms are not used with these. Specific descriptions can be made with property terms.

Guidelines for how the annotation is made and how to use VariO in different situations have been published [Vihinen, 2014a, 2015b). The flowchart for steps in VariO annotation has been described [Vihinen, 2014b]. Briefly, the database curator collects all relevant information about the variant and its effects, mechanisms, and consequences. This may include data from laboratories, databases, and literature. The precise position of the variant in the three reference sequences for DNA, RNA, and protein (where relevant) will be obtained and annotated with variation type annotations. For function annotations, the user has to choose the appropriate terms at the relevant levels DNA, RNA, and/or protein. Structural changes due to variation can be explained in detail. Depending on the annotation level, the property annotation items can vary. Both the structure and property annotations can be modulated by using attributes to make the annotations more detailed. Further, Evidence Ontology (ECO) terms can be added to provide users with the possibility to evaluate the strength of evidence behind annotations. ECO describes the methods used to obtain the annotated results [Chibucos et al., 2014].

To facilitate easy annotation and to enhance the consistency in the use of the VariO terms for annotating variants, a user-friendly online application called VariOtator was developed. For variation type annotations the tool is fully automated, it accepts variant descriptions according to the HGVS nomenclature and generates VariO annotation, either with or without the full ontology lineage. When the user provides a full codingDNA variant description, that is,with the reference sequence, the description is first checked with the Mutalyzer Name Checker (https://mutalyzer.nl)

tool [Wildeman et al., 2008]. For the other sublevels, function, structure, and property, annotations cannot be automated. VariOtator generates annotation at these levels based on provided details.

Implementation

Several implementations of the VariOtator tool are available at http://variationontology.org/ (Fig. 1). There is a Graphical User Interface (GUI) at http://variationontology.org/VariOtator.php for interactive submissions. For variation type annotations, there are three batch versions at http://variationontology.org/VariOtatorBatch.php, one with aWeb interface that requires as input a tab-delimited file and two stand-alone versions that use as input an LOVD download file. These stand-alone versions (Linux and Windows) can be downloaded and installed locally. TheWeb service is available at http://variationontology.org/VariOService/?wsdl.

The core of the VariOtator is a Python (2.7) script with the RDFLib package (4.2.0), in combination with vario.owl and eco.owl, the Web Ontology Language (OWL 2) versions of VariO and ECO, respectively (see Fig. 1). The most recent version of vario.owl can be downloaded from http://variationontology.org/download.shtml, (at the moment version 1.04), the latest eco.owl file (release 2015- 01-12) was downloaded from the Evidence Ontology Website (http://evidenceontology.googlecode.com/svn/trunk/eco.owl).

The VariOtator Web interface (http://variationontology.org/VariOtator.php) was developed using PHP5 and JavaScript (Fig. 1). Checking variant descriptions withMutalyzer is done through their SOAP Web service (https://mutalyzer.nl/webservices). The VariOtator SOAP Web service makes use of the python-soaplib package (0.8.1), the client example uses the suds.client package (0.4).

The stand-alone batch versions (LOVD VariOtator) for use with LOVD download files were developed with the cx-Freeze package (4.3.1) (http://cx-freeze.sourceforge.net). The Linux version is a ready-to-use package, the Windows version is available as a Windows Installer package (msi file). All the software is freely available and released and distributed under the terms of the GNU Affero General Public License version 3 (GPLv3).

Features

Web Interface

The user interface is organized according to the levels in VariO. First, the user chooses between annotation for variation type, function, structure, and property, and whether full lineage is desired or not (Fig. 2). If full lineage is chosen, all VariO terms down to the root term (“VariO:0001 variation”) in the ontology are provided. The resulting VariO annotations are shown on screen, and

can be downloaded as a text file. After each annotation step, the user can choose either to restart or to continue, in which case the new annotations are added to the previous one(s). The annotation can also be downloaded in a text file.

Variation Type Annotation

When choosing for variation type, the user can enter a variant description in HGVS format (Fig. 2). If a reference sequence is provided with a coding DNA description, the variant description will be checked withMutalyzer and predicted RNA and protein descriptions and their annotations are provided, as well.VariOtator accepts the variation details in several formats including codingDNA, RNA, and protein descriptions, with or without the reference sequence. Both one- and three-letter amino acid codes are accepted. Examples of input are LRG_1:c.72G>A, NG_008680.1(PAX2):c.412A>G, NM_003990.3:c.412A>G, chr15:g.40702997G>A, r.(21a>u), and p.Trp26Cys. An example of VariOtator output with full lineage for variation type annotation can be found in Figure 3.

Since the terms "VariO:0313 transition," "VariO:0314 pyrimidine transition," "VariO:0315 purine transition," and "VariO:0316 transversion" can be used for annotation of both DNA and RNA substitutions, ancestor terms are included in the annotation also when full lineage is not chosen. So, if the resulting VariO terms are either "VariO:0315 purine transition" or "VariO:0314 pyrimidine transition," the ancestor terms "VariO:0136 DNA substitution" and "VariO:0313 transition" are added in the case of DNA or "VariO:0312 RNA substitution" and "VariO:0313 transition" are added in the case of RNA. Similarly, when the final VariO term is "VariO:0316 transversion," either "VariO:0136 DNA substitution" or "VariO:0312 RNA substitution" is added.

Annotation of Variation Affecting Function

The first step for annotation of variations affecting function is to choose the molecular level (DNA, RNA, or protein) (Fig. 2). Then, an overview of VariO terms on the specific level is displayed. The relevant term is chosen by clicking it after which the annotation is generated. If necessary, more than one term can be chosen. An example of function annotation can be found in Figure 3.

Annotation of Variation Affecting Structure

As with functional annotation, the user first chooses the molecular levels (Fig. 2). If the selected term has sublevels, they are shown. This way the user can make very detailed annotations. Once the structure terms are chosen, it becomes possible to pick an attribute term, to modify and specify the annotation. For example, the quantity of the structure terms can be specified by using quantity change attributes including those for increased, decreased,missing, and not changed. In the next

phase, an ECO term can be added to annotate the (experimental) evidence and method based on which the annotation is made. An example of structure annotation can be found in Figure 3.

Annotation of Variation Affecting Property

Variation properties are annotated similar to structural variations, including the use of attribute terms and ECO annotations. Both structure and property terms are specific for the molecular levels. The protein level allows for the largest number of choices due to the very wide spectrum of effects. An example of property annotation with attribute and ECO terms can be found in Figure 3.

Batch Versions

As variation type annotations can be automated and there are numerous databases containing very large numbers of variants, effective tools are needed for their annotation. For this purpose, we developed batch versions (http://variationontology.org/VariOtatorBatch.php) (see Fig. 1). The Web-based batch version requires a tab-delimited file with variant descriptions (coding DNA and optionally protein variant descriptions) as input. The results are provided in a tab-delimited file containing the original variant descriptions and predicted RNA variant descriptions, and the VariO terms for the variation type annotations at the DNA, RNA, and optionally at protein levels. The VariO terms are given including the full lineage up to but not including the root term "VariO:0001 variation." In contrast to the Web interface, the terms are not checked with Mutalyzer. There is a batch version for that purpose.

For annotating databases using the LOVD management system, stand-alone batch versions are available, both for Linux and Windows. These can be downloaded from http://variationontology.org/VariOtatorBatch.php. These versions are for variation type annotation only, and the variant descriptions are not checked with Mutalyzer. The user has to add columns for VariO annotations at the three molecular levels, DNA, RNA, and protein, to the LOVD database prior to using the VariOtator tool, as the tool uses an LOVD download file as input and adds the VariO terms to the relevant columns in that file. As how to add columns to the database, we refer to the LOVD documentation and VariOtator help files. Annotations are automatically added to the database when uploading the LOVD download file. If for some reason the annotation cannot be made an error log is provided, so that the users can solve the problematic cases.

Web Service

For programmatic access to the VariO annotation type tool, a SOAP Web service was developed, with which VariO annotation generation can be fully integrated into other software. An example client script for how to use this Web service can be found at http://variationontology.org/VariO-client-suds.py. It is a Python script that takes one variant description at a time, and generates the

VariO annotations for that description, including the full lineage up to the root term (not included). A Web Service Definition Language (WSDL) description is available at http://variationontology.org/VariOService/?wsdl for easy generation of client programs in many languages. Again, the variant descriptions are not checked with Mutalyzer, this can be done with their Web services, if desired. Because of this, predicted RNA and protein descriptions are not provided when entering a coding DNA variant description, as is the case with the web interface.

Discussion

The VariOtator tool was developed to guarantee the consistency of annotations with VariO terms and to help with the task of variation annotation. By using the fully automated scripts, variation type annotation systematics in databases can be significantly improved. The specific batch versions for use with LOVD databases allows addition of variation type annotations to be included to the resources in the most widely used LSDB environment. For the three other annotation levels, function, structure, and property, the tool provides an easy-to-use and user-friendly interface. A guide with detailed instructions for annotators is available [Vihinen, 2014b] and examples of annotation with VariO can be found in [Vihinen, 2015b] or at http://www.variationontology.org/Examples.shtml.

An example of the use of VariO in an LSDB can be found at http://databases.lovd.nl/shared/genes/BTK. BTKbase [Väliaho et al., 2006; Schaafsma and Vihinen, 2015] is a database for Bruton agammaglobulinemia tyrosine kinase (BTK) variants causing X-linked agammaglobulinemia, a rare primary immunodeficiency [Tsukada et al., 1993; Vetrie et al., 1993]. BTKbase has been previously maintained with MUTbase software [Riikonen and Vihinen, 1999; Väliaho et al., 2006]. As part of the conversion to the LOVD database management system, some novel features such as VariO annotations were included. In the LOVD installation, the VariO annotations can be found in the columns "VariO Annotation DNA level," "VariO Annotation RNA level," and "VariO Annotation protein level." With the generated batch tools, variation type annotations can be automatically added to all the other LOVD-based LSDBs. It is up to the database curators to make these annotations as VariOtator is not integrated with LOVD. Once the annotations are added, it will become possible to make new kinds of analyses over known variation space. The other types of annotations are so variable and complex that they cannot be automated. VariO annotations are currently being added to the remaining 130 IDbases [Piirila et al., 2006] and are already available in NDDVD, NeuroDegenerative Diseases Variation Database (http://bioinf.suda.edu.cn/NDDvarbase/LOVDv.3.0/genes) containing information for 126 genes in 49 diseases [Yang et al., in preparation].

Acknowledgment

Disclosure statement: The authors declare no conflict of interest.

References

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, et al. 2000. Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25: 25–29.

Celli J, Dalgleish R, Vihinen M, Taschner PE, den Dunnen JT. 2012. Curating gene variant databases (LSDBs): toward a universal standard. Hum Mutat 33:291–297.

Chibucos MC, Mungall CJ, Balakrishnan R, Christie KR, Huntley RP, White O, Blake JA, Lewis SE, Giglio M. 2014. Standardized description of scientific evidence using the Evidence Ontology (ECO). Database (Oxford) 2014: bau075.

Cotton RG, Auerbach AD, Beckmann JS, Blumenfeld OO, Brookes AJ, Brown AF, Carrera P, Cox DW, Gottlieb B, Greenblatt MS, Hilbert P, Lehväslaiho H, et al. 2008. Recommendations for locus-specific databases and their curation. Hum Mutat 29:2–5.

Cunningham F, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S, Gil L, Giron CG, et al. 2015. Ensembl 2015. Nucleic Acids Res 43:D662–D669.

Dalgleish R, Flicek P, Cunningham F, Astashyn A,Tully RE, Proctor G, Chen Y, McLaren WM, Larsson P, Vaughan BW, Beroud C, Dobson G, et al. 2010. Locus Reference Genomic sequences: an improved basis for describing human DNA variants. Genome Med 2:24.

den Dunnen JT, Antonarakis SE. 2000. Mutation nomenclature extensions and suggestions to describe complex mutations: a discussion. Hum Mutat 15:7–12.

Eilbeck K, Lewis SE, Mungall CJ, Yandell M, Stein L, Durbin R, Ashburner M. 2005. The Sequence Ontology: a tool for the unification of genome annotations. Genome Biol 6:R44.

Fokkema IFAC, Taschner PEM, Schaafsma GCP, Celli J, Laros JFJ, den Dunnen JT. 2011. LOVD v.2.0: the next generation in gene variant databases. Hum Mutat 32:557–563.

Gray KA, Daugherty LC, Gordon SM, Seal RL, Wright MW, Bruford EA. 2013. Genenames.org: the HGNC resources in 2013. Nucleic Acids Res 41:D545–D552.

Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, Maglott DR. 2014. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res 42:D980–D985.

MacArthur JA, Morales J, Tully RE, Astashyn A, Gil L, Bruford EA, Larsson P, Flicek P, Dalgleish R, Maglott DR, Cunningham F. 2014. Locus Reference Genomic: reference sequences for the reporting of clinically relevant sequence variants. Nucleic Acids Res 42:D873–D878.

Piirilä H, Väliaho J, Vihinen M. 2006. Immunodeficiencymutation databases (IDbases). Hum Mutat 27:1200–1208.

Riikonen P, Vihinen M. 1999. MUTbase: maintenance and analysis of distributed mutation databases. Bioinformatics 15:852–859.

Schaafsma GCP, Vihinen M. 2015. Genetic variation in Bruton tyrosine kinase. In: Plebani A, Lougaris V, editors. Agammaglobulinemia. Switzerland: Springer International Publishing. p 75–85.

Tsukada S, Saffran DC, Rawlings DJ, Parolini O, Allen RC, Klisak I, Sparkes RS, Kubagawa H, Mohandas T, Quan S, Belmont JW, Cooper MD, et al. 1993. Deficient expression of a B cell cytoplasmic tyrosine kinase in human X-linked agammaglobulinemia. Cell 72:279–290.

UniProt Consortium. 2015. UniProt: a hub for protein information. Nucleic Acids Res 43:D204–D212.

Väliaho J, Smith CIE, Vihinen M. 2006. BTKbase: the mutation database for X-linked agammaglobulinemia. Hum Mutat 27:1209–1217.

Vetrie D, Vořechovský I, Sideras P, Holland J, Davies A, Flinter F, Hammarström L, Kinnon C, Levinsky R, Bobrow M, Smith CIE, Bentley DR. 1993. The gene involved in X-linked agammaglobulinemia is a member of the src family of protein-tyrosine kinases. Nature 361:226–233.

Vihinen M. 2014a. Variation Ontology for annotation of variation effects and mechanisms. Genome Res 24:356–364.

Vihinen M. 2014b. Variation Ontology: annotator guide. J Biomed Semantics 5:9.

Vihinen M. 2015a. Muddled genetic terms miss and mess the message. Trends Genet 31:423–425.

Vihinen M. 2015b. Types and effects of protein variations. Hum Genet 134:405–421.

Vihinen M, den Dunnen JT, Dalgleish R, Cotton RG. 2012. Guidelines for establishing locus specific databases. Hum Mutat 33:298–305.

Wildeman M, van Ophuizen E, den Dunnen JT, Taschner PE. 2008. Improving sequence variant descriptions in mutation databases and literature using the Mutalyzer sequence variation nomenclature checker. Hum Mutat 29:6–13.

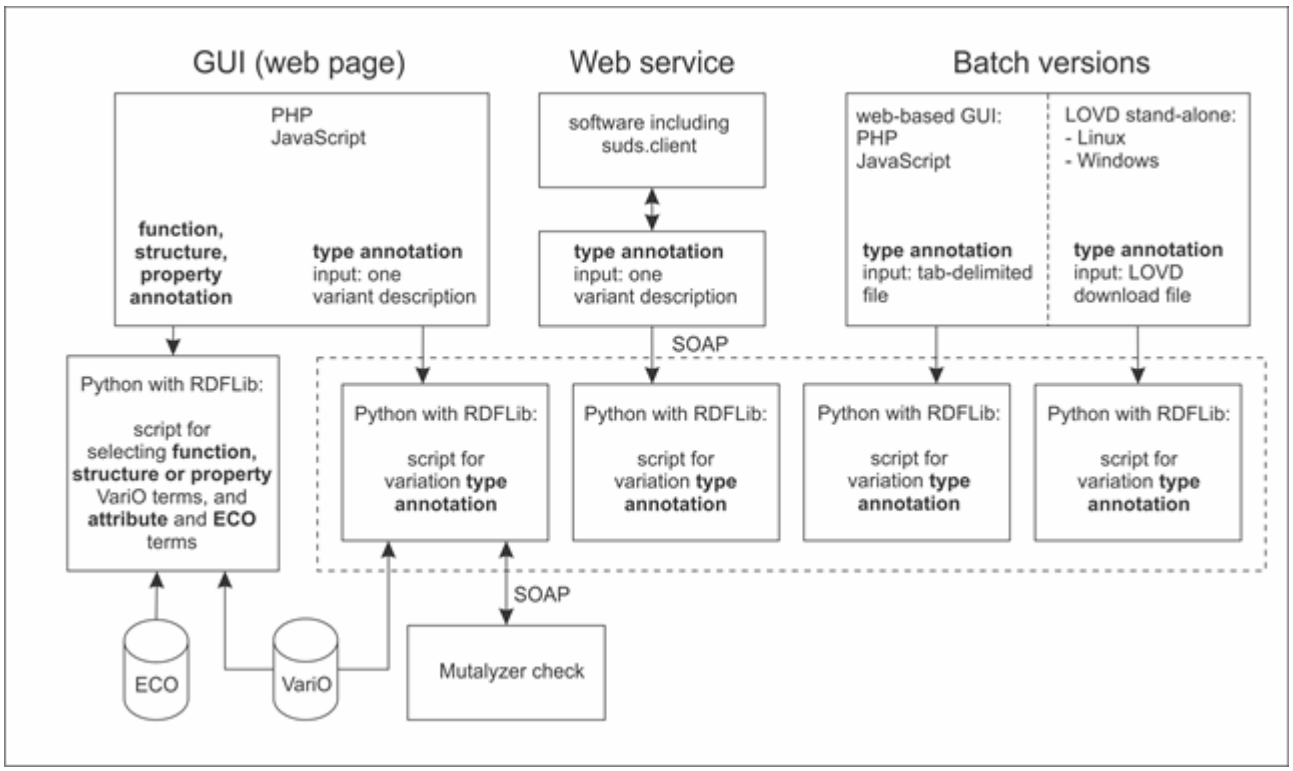

Figure 1: Overview of the VariOtator implementations. The scripts in the dotted box all have the same functionality, but are optimized for different purposes (mainly regarding to I/O). ECO: Evidence Ontology, format eco.owl; VariO: Variation Ontology, format vario.owl; GUI: Graphical User Interface; LOVD: Leiden Open Variation Database; SOAP: computer network messaging protocol; RDFLib: Python library for working with the Resource Description Framework (RDF).

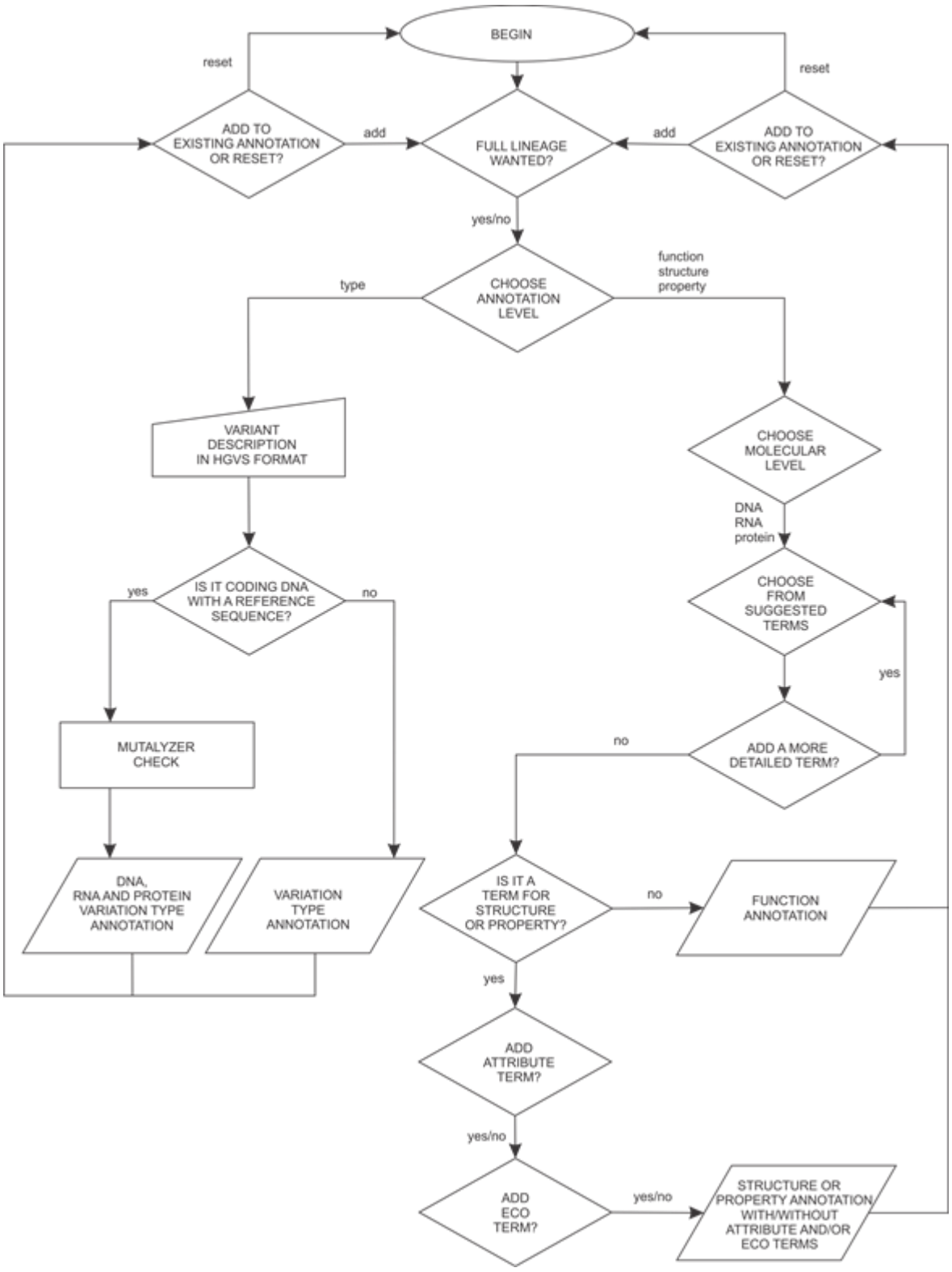
BEGIN
reset
reset
ADD TO EXISTING ANNOTATION OR RESET?
add
FULL LINEAGE WANTED?
add
ADD TO EXISTING ANNOTATION OR RESET?
yes/no
CHOOSE ANNOTATION LEVEL
type
function
structure
property
VARIANT DESCRIPTION IN HGVS FORMAT
CHOOSE MOLECULAR LEVEL
DNA
RNA
protein
yes
IS IT CODING DNA WITH A REFERENCE SEQUENCE?
no
CHOOSE FROM SUGGESTED TERMS
yes
MUTALYZER CHECK
no
ADD A MORE DETAILED TERM?
DNA, RNA AND PROTEIN VARIATION TYPE ANNOTATION
VARIATION TYPE ANNOTATION
IS IT A TERM FOR STRUCTURE OR PROPERTY?
no
FUNCTION ANNOTATION
yes
ADD ATTRIBUTE TERM?
yes/no
ADD ECO TERM?
yes/no
STRUCTURE OR PROPERTY ANNOTATION WITH/WITHOUT ATTRIBUTE AND/OR ECO TERMS

Figure 2: Flowchart of the annotation process with VariOtator

Your variant description was checked with Mutalyzer which provided the following information:

Coding DNA description: NM_000061.2(BTK_v001):c.1574G>A
Description relative to transcription start: NM_000061.2:n.1767G>A
Affected protein(s): NM_000061.2(BTK_i001):p.(Arg525Gln)

VariO terms for **variation type**:

c.1574G>A
VariO:0128 variation affecting DNA
VariO:0129 DNA variation type
VariO:0322 DNA variation classification
VariO:0135 DNA chain variation
VariO:0136 DNA substitution
VariO:0313 transition
VariO:0315 purine transition
r.(1574g>a)
VariO:0297 variation affecting RNA
VariO:0306 RNA variation type
VariO:0328 RNA variation classification
VariO:0312 RNA substitution
VariO:0313 transition
VariO:0315 purine transition
VariO:0308 missense variation
p.(Arg525Gln)
VariO:0002 variation affecting protein
VariO:0012 protein variation type
VariO:0325 protein variation classification
VariO:0021 amino acid substitution

VariO terms for **variation function**:

VariO:0002 variation affecting protein
VariO:0003 variation affecting protein function
VariO:0008 effect on catalytic protein function

VariO terms for **variation structure**:

VariO:0002 variation affecting protein
VariO:0060 variation affecting protein structure
VariO:0064 effect on protein 3D structure
VariO:0070 effect on protein tertiary structure
VariO:0118 effect on protein interaction site
VariO:0120 effect on protein catalytic site

VariO terms for **variation affecting property**:

VariO:0002 variation affecting protein
VariO:0032 variation affecting protein property
VariO:0053 effect on protein activity; has_quality VariO:0292 missing; has_evidence ECO:0000005 enzyme assay evidence

Figure 3: VariOtator web interface output. Variation type annotation with full lineage using NM_000061.2:c.1574G>A as input, and variation function, structure and property annotation on protein level, with full lineage and attribute and ECO terms.