Cunningham et al. Journal of Biomedical Semantics (2015) 6:32
DOI 10.1186/s13326-015-0030-4

JOURNAL OF BIOMEDICAL SEMANTICS

SHORT REPORT Open Access

Improving the Sequence Ontology terminology for genomic variant annotation

Fiona Cunningham[1], Barry Moore[2], Nicole Ruiz-Schultz[3], Graham RS Ritchie[1,4] and Karen Eilbeck[3*]

Abstract

Background: The Genome Variant Format (GVF) uses the Sequence Ontology (SO) to enable detailed annotation of sequence variation. The annotation includes SO terms for the type of sequence alteration, the genomic features that are changed and the effect of the alteration. The SO maintains and updates the specification and provides the underlying ontologicial structure.

Methods: A requirements analysis was undertaken to gather terms missing in the SO release at the time, but needed to adequately describe the effects of sequence alteration on a set of variant genomic annotations. We have extended and remodeled the SO to include and define all terms that describe the effect of variation upon reference genomic features in the Ensembl variation databases.

Results: The new terminology was used to annotate the human reference genome with a set of variants from both COSMIC and dbSNP. A GVF file containing 170,853 sequence alterations was generated using the SO terminology to annotate the kinds of alteration, the effect of the alteration and the reference feature changed. There are four kinds of alteration and 24 kinds of effect seen in this dataset. (Ensembl Variation annotates 34 different SO consequence terms: http://www.ensembl.org/info/docs/variation/predicted_data.html).

Conclusions: We explain the updates to the Sequence Ontology to describe the effect of variation on existing reference features. We have provided a set of annotations using this terminology, and the well defined GVF specification. We have also provided a provisional exploration of this large annotation dataset.

Findings

Background

The Sequence Ontology (SO) [1] provides terminology to define sequence features. These features are the building blocks of sequence annotation, and allow biologically meaningful regions to be assigned between coordinates of sequences such as genome assemblies and transcripts. The relationships between the terms in SO provide for the annotation of multi-part features such as gene models, composed of multiple transcripts, exons, introns and UTR features. Reference genome annotations are often shared using a flat file format GFF3, developed by the GMOD community [2], which stipulates that SO terms describe each annotated feature, thus many genome annotation tools use SO to describe reference genome features. While terms to describe variants have long been part of the Sequence Ontology, increased need for new variation terms to describe the predicted effect of sequence alterations on existing genomic features lead to the development of new terms. This has been driven by the proliferation of software tools that predict the effect of sequence alterations such as Ensembl's Variant Effect Predictor (VEP) [3] and the VAAST suite tool: Variant Annotation Tool (VAT) [4]. In this mansucript, SO terms are italicized and written without underscores.

Next generation sequencing (NGS) technologies have provided an enormous expansion in our understanding of the landscape of genetic variation [5, 6] as well as the impact of that variation on human health [7–9]. These datasets create a significant burden in computational analysis and data storage, but established work-flows for analysis are emerging [3] and well established data formats exist for each stage of the process. The original base calls from the sequencer are converted to FASTQ files [10] that contain the sequence data; the SAM format [11] captures the alignment of the sequence to a

* Correspondence: keilbeck@genetics.utah.edu
[3]Department of Biomedical Informatics, University of Utah, Salt Lake City, UT, USA
Full list of author information is available at the end of the article

reference genome and the Variant Call Format [12] has become widely adopted by variant calling tools to report variants and the information needed to call them. However, knowing the type and genomic location of a sequence change is just the first step in understanding its clinical or biological consequences. Variant annotation then begins the process of adding additional knowledge about the structural and functional consequences of those variants through the impact on reference sequence features and ultimately on phenotype.

The Genome Variation Format (GVF) [13] is a variant file format for the detailed annotation of genetic variation. GVF is a community supported format that uses established ontologies such as the Sequence Ontology [1] to describe the variant data. GVF does not replace existing variant nomenclature systems such as HGVS [14] and ISCN [15] that provide effective ways to unambiguously describe individual variants in the literature. GVF provides the infrastructure to support inclusion of these nomenclatures along with other detailed variant annotations in a format capable of supporting genome scale variant data. GVF is used in the community for exchange of variant annotations between Ensembl [16], DGVa and dbVar [17] and is compatible with existing GFF3 software [2, 18] as well as emerging domain specific tools [4, 19].

User requirements and ontology development

Upon the release of the specification for variant genome annotation, GVF used terms from the Sequence Ontology release 2.4.3. While this resource provided 101 terms to describe the effects of a sequence alteration on genomic features, it was still missing sufficiently specialized terms to fully capture the kinds of variation annotated by the Ensembl variation pipeline [20]. A requirements analysis was undertaken to establish the terminology and relationships between terms to accomplish annotation and facilitate queries of annotated datasets. Ensembl uses 34 terms [21] to describe the effect of variation, 21 of which were new to SO, and 2 required an ammendement to the name. Figure 1 shows a subset of the terms in SO that describe sequence variants, with the Ensembl terms highlighted.

In the SO, the sequence alteration and the effects of the alteration are separated. A *sequence alteration* defines the nucleotide change observed in an individual sequence, in relation to a reference sequence. Examples of alterations are *insertion*, *deletion*, *substitution* and *SNV*. The effect of a *sequence alteration* is the observed or predicted change to annotated reference seqeunce features. These effects of sequence alterations are defined as *sequence variants* in SO and are outlined in Fig. 1. Examples of these terms are *missense variant*, whereby

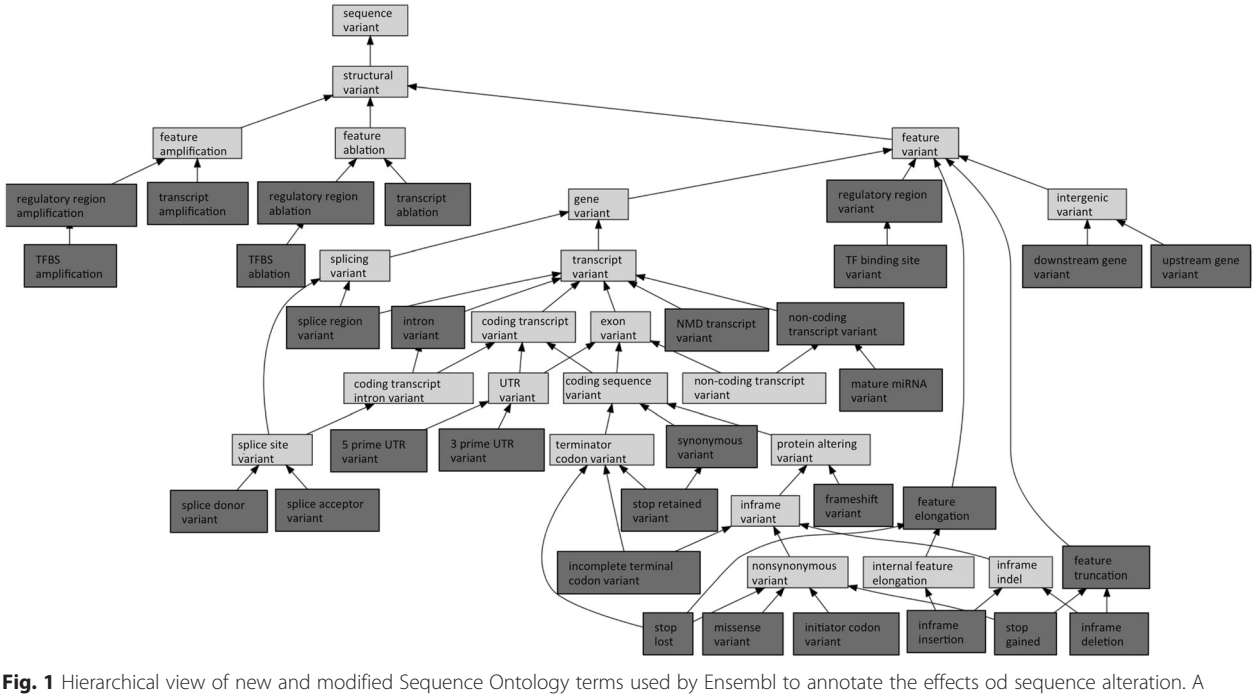

Fig. 1 Hierarchical view of new and modified Sequence Ontology terms used by Ensembl to annotate the effects od sequence alteration. A portion of the SO *sequence variant* subsumption hierarchy is shown, with terms used by Ensembl in dark grey. *Feature variant* terms define cases where the sequence alteration occurs within or overlaps an annotated reference feature such as a transcript or exon, whereas the kinds of *feature ablation*, *feature amplification*, define cases where an entire feature is altered. Definitions for these terms are available from the miSO browser: http://sequenceontology.org/browser/obob.cgi and http://ensembl.org/info/genome/variation/predicted_data.html

codon bases are modified in such a way as the resuling amino acid would change, and *splice donor variant* where by the alteration changes the two-base pair region at the 5′ end of an *intron*.

One of the advantages of using an ontology for the annotation of data, is that given the related nature of the terms, there are options to annotate data to the level of detail afforded by the evidence. Under the *sequence variant* node, SO provides two high level nodes in the ontology: *structural variant* and *functional variant*. Structural variants pertain to changes with regard to annotated sequence features, and are the output of automated variant effect predition tools such as VEP [3]. Functional variants however describe the cellular effect of a sequence alteration and are generally manually curated. These functional terms have largely been absorbed into the Variation ontology [22] and are not automatically assigned by variant effect prediction tools. With regards to *structural variants*, the alteration can either internally modify a sequence feature, when the alteration falls within the extent of a reference sequence feature such as an exon *(feature variant*), or the alteration can be greater than the extent of the sequence feature, causing the ablation or amplification of an entire genomic feature such as a transcript.

The *feature variant* node in the ontology subsumes the terms that describe changes internal to genomic features such as those affecting genes, transcripts and introns. The majority of the sequence alterations currently annotated by Ensembl cause *feature variants*. These feature variant terms are shown in Fig. 1, where the terms used in Ensembl annotations are highlighted in dark grey. There are five subtypes: *intergenic variant*, *gene variant, feature truncation, feature elongation* and *regulatory region variant*. Of these terms, *gene variant* has 77 direct and indirect subtypes and includes most of the terms that describe structural sequence variants caused by substitutions and small insertions and deletions. This portion of the SO contains terms with multiple parents, to allow for effective querying of the annotations. For example, the term *stop retained variant* is both a *synonymous variant* and a *terminator codon variant*. Users are thus able to query the Ensembl data for all terminator codon variants or all synonymous variants.

Annotated variants

GVF formated variant genome annotations for 19 organisms, typed using SO are available within the Ensembl databases [23] and for download (ftp://ftp.ensembl.org/pub/release-69/variation/gvf/). Included in this set is a GVF file of 170,853 human variant annotations, with data from dbSNP [24] and COSMIC [25] using the described terminology. There are four kinds of sequence alterations reported, corresponding to 158205 *SNVs*, 7575 *deletions*, 3097 *insertions* and 1876 *substitutions*. There are 24 kinds of variant_effect reported in the file, and five kinds of genomic feature affected (*mRNA*, *miRNA*, *transcript*, *primary_transcript* and *ncRNA*). There are 1,485,317 reported variant effects with corresponding genomic features, as a single alteration may perturb many annotated genomic features. For example an SNV may intersect two alternate transcripts, one in an exon, the other in an intron. Figure 2 shows a tree map of the proportion of variant effects annotated to each kind of sequence alteration in this dataset. As can be seen, each

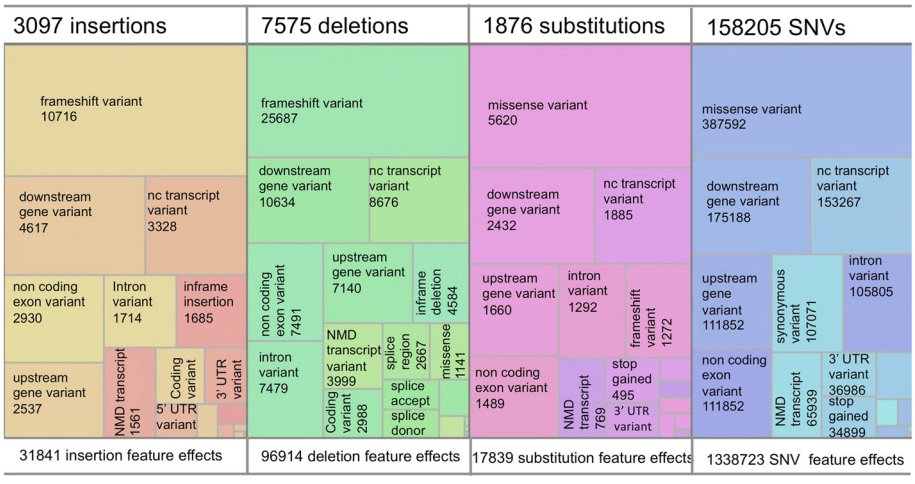

Fig. 2 Treemap of the proportion of variant affect atributed to each kind of sequence alteration in Ensembl human GVF dataset (release 69). A treemap displays hierarchcal data as nested rectangles. In this dataset there are four kinds of sequence alteration annotated: *insertion*, *deletion*, *substitution* and *SNV*, each with a different color. For each sequence alteration, the annotated variant effects are shown with the size of the rectangle proportional to the number of occrurences of that annotation, and the count is provided where space permits. The treemap was generated using the IBM Manyeyes tool (http://www-958.ibm.com/)

kind of alteration causes proportionally different effects upon the genome features; insertions and deletions cause more frameshift variants, where as the SNVs and other substitutions cause more missense variants.

Discussion and conclusions

Detailed annotation of sequence variation is complicated because reference genome annotations are complex. Genes may produce multiple transcripts, may overlap each other on opposite strands, or even be nested within introns of other genes, therefore a variant may influence multiple genomic features. Capturing the effect of a sequence alteration on the genomic features with which it intersects is an important step towards understanding the implication of the variant sequence. The terminology described here provides a basis with whch to categorize and define sequence variation and the flexibility to annotate the effect with respect to the feature intersected. This ontology provides very specific leaf terms, with which to automatically annotate genomic sequence but also useful mid level terms for querying.

Future developments to the ontology will include developing relationships between the sequence variant terms and the sequence features that are affected. There has been significant uptake of these variant effect terms by the genomic variant annotation community. The UCSC genomic browser uses this termnology in variant annotation [26] as does the NCBI's ClinVar data dictionary and dbVar database [17]. New terms will be added as required. New terms and updates to the ontology may be requested using the term tracker (https://sourceforge.net/p/song/term-tracker/). Development of the SO is collaborative, incorporating community discussion via our mailing list and the term tracker as well as the results of focused working groups.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

FC and GR performed requirements analysis. KE, FC, NRS, GR and BM contributed to ontology development and term definition. All authors contributed to manuscript.

Acknowledgements

This work was supported by the National Human Genome Research Institute [R01HG004341 to KE] and National Libarary of Medicine training grant [T15 LM007124-18, NRS]. Ensembl receives majority funding from the Wellcome Trust (grant numbers WT095908 and WT098051) with additional funding for specific project components from the National Human Genome Research Institute (U41HG007234, 1R01HD074078, and U41HG007823), the Biotechnology and Biological Sciences Research Council (BB/K009524/1, BB/L024225/1, BB/M018458/1 and BB/M020398/1), the Centre for Therapeutic Target Validation (CTTV) and the European Molecular Biology Laboratory. The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 282510 (BLUEPRINT). The research leading to these results has received funding from the European Union's Seventh Framework Capacities Specific Programme under grant agreement n° 284209 (BioMedBridges). This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement n° 634143 (MedBioinformatics)

Author details

[1]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK. [2]Department of Human Genetics, University of Utah, Salt Lake City, UT, USA. [3]Department of Biomedical Informatics, University of Utah, Salt Lake City, UT, USA. [4]Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK.

Received: 19 February 2013 Accepted: 22 July 2015
Published online: 31 July 2015

References

1. Eilbeck K, Lewis SE, Mungall CJ, Yandell M, Stein L, Durbin R, et al. The Sequence Ontology: a tool for the unification of genome annotations. Genome Biol. 2005;6:R44.
2. Generic Model Organism Database (GMOD). [http://gmod.org].
3. McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics. 2010;26:2069–70.
4. Yandell M, Huff C, Hu H, Singleton M, Moore B, Xing J, et al. A probabilistic disease-gene finder for personal genomes. Genome Res. 2011;21:1529–42.
5. 1000 Genomes Progect Structural Variant group specification. [http://www.1000genomes.org/wiki/Analysis/Variant%20Call%20Format/VCF%20(Variant%20Call%20Format)%20version%204.0/encoding-structural-variants].
6. MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, Walter K, et al. A systematic survey of loss-of-function variants in human protein-coding genes. Science. 2012;335:823–8.
7. Rope AF, Wang K, Evjenth R, Xing J, Johnston JJ, Swensen JJ, et al. Using VAAST to identify an X-linked disorder resulting in lethality in male infants due to N-terminal acetyltransferase deficiency. Am J Hum Genet. 2011;89:28–43.
8. Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA, et al. Exome sequencing as a tool for Mendelian disease gene discovery. Nat Rev Genet. 2011;12:745–55.
9. Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, et al. Exome sequencing identifies the cause of a mendelian disorder. Nat Genet. 2010;42:30–5.
10. Cock PJ, Fields CJ, Goto N, Heuer ML, Rice PM. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res. 2010;38:1767–71.
11. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25:2078–9.
12. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27:2156–8.
13. Reese MG, Moore B, Batchelor C, Salas F, Cunningham F, Marth GT, et al. A standard variation file format for human genome sequences. Genome Biol. 2010;11:R88.
14. Horaitis O, Cotton RG. The challenge of documenting mutation across the genome: the human genome variation society approach. Hum Mutat. 2004;23:447–52.
15. An International System for Human Cytogenetic Nomenclature. Basel: S. Karger AG; 2009.
16. Flicek P, Ahmed I, Amode MR, Barrell D, Beal K, Brent S, et al. Ensembl 2013. Nucleic Acids Res. 2013;41:D48–55.
17. Lappalainen I, Lopez J, Skipper L, Hefferon T, Spalding JD, Garner J, et al. DbVar and DGVa: public archives for genomic structural variation. Nucleic Acids Res. 2013;41:D936–41.
18. Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, et al. The Bioperl toolkit: Perl modules for the life sciences. Genome Res. 2002;12:1611–8.
19. Song T, Hwang KB, Hsing M, Lee K, Bohn J, Kong SW. gSearch: a fast and flexible general search tool for whole-genome sequencing. Bioinformatics. 2012;28:2176–7.
20. Chen Y, Cunningham F, Rios D, McLaren WM, Smith J, Pritchard B, et al. Ensembl variation resources. BMC Genomics. 2010;11:293.

21. Ensembl predicted data. [http://www.ensembl.org/info/genome/variation/predicted_data.html].
22. Vihinen M. Variation ontology for annotation of variation effects and mechanisms. Genome Res. 2014;24:356–64.
23. Rios D, McLaren WM, Chen Y, Birney E, Stabenau A, Flicek P, et al. A database and API for variation, dense genotyping and resequencing data. BMC Bioinformatics. 2010;11:238.
24. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001;29:308–11.
25. Forbes SA, Bhamra G, Bamford S, Dawson E, Kok C, Clements J, et al. The Catalogue of Somatic Mutations in Cancer (COSMIC). Curr Protoc Hum Genet. 2008;Chapter 10:Unit 10 11.
26. Meyer LR, Zweig AS, Hinrichs AS, Karolchik D, Kuhn RM, Wong M, et al. The UCSC genome browser database: extensions and updates 2013. Nucleic Acids Res. 2013;41:D64–9.