




SOFTWARE TOOL ARTICLE

# ClinVar data parsing [version 1; referees: 2 approved]

Xiaolei Zhang <sup>1,2\*</sup>, Eric V. Minikel<sup>3,4\*</sup>, Anne H. O'Donnell-Luria<sup>3-5</sup>,  
Daniel G. MacArthur<sup>3,4</sup>, James S. Ware<sup>1,2,6</sup>, Ben Weisburd<sup>3,4</sup>

<sup>1</sup>National Heart and Lung Institute, Imperial College London, London, SW7 2AZ, UK

<sup>2</sup>Royal Brompton Cardiovascular Research Centre, Royal Brompton & Harefield Hospitals NHS Trust, London, SW3 6NP, UK

<sup>3</sup>Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, Massachusetts, 02114, USA

<sup>4</sup>Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, 02142, USA

<sup>5</sup>Boston Children's Hospital, Boston, Massachusetts, 02115, USA

<sup>6</sup>MRC London Institute of Medical Sciences, Imperial College London, London, W12 0NN, UK

\* Equal contributors

**v1** First published: 23 May 2017, 2:33 (doi: [10.12688/wellcomeopenres.11640.1](https://doi.org/10.12688/wellcomeopenres.11640.1))



Latest published: 23 May 2017, 2:33 (doi: [10.12688/wellcomeopenres.11640.1](https://doi.org/10.12688/wellcomeopenres.11640.1))



## Abstract

This software repository provides a pipeline for converting raw ClinVar data files into analysis-friendly tab-delimited tables, and also provides these tables for the most recent ClinVar release. Separate tables are generated for genome builds GRCh37 and GRCh38 as well as for mono-allelic variants and complex multi-allelic variants. Additionally, the tables are augmented with allele frequencies from the ExAC and gnomAD datasets as these are often consulted when analyzing ClinVar variants. Overall, this work provides ClinVar data in a format that is easier to work with and can be directly loaded into a variety of popular analysis tools such as R, python pandas, and SQL databases.

## Open Peer Review

Referee Status:  

	Invited Referees	
	1	2
<b>version 1</b> published 23 May 2017	 report	 report

- 1 **Fiona Cunningham** , European Bioinformatics Institute, UK  
**Sarah Hunt**, European Bioinformatics Institute, UK
- 2 **Steven M Harrison** , Harvard Medical School, USA

## Discuss this article

Comments (0)

**Corresponding author:** Ben Weisburd ([weisburd@broadinstitute.org](mailto:weisburd@broadinstitute.org))

**Competing interests:** No competing interests were disclosed.

**How to cite this article:** Zhang X, Minikel EV, O'Donnell-Luria AH *et al.* **ClinVar data parsing [version 1; referees: 2 approved]** Wellcome Open Research 2017, 2:33 (doi: [10.12688/wellcomeopenres.11640.1](https://doi.org/10.12688/wellcomeopenres.11640.1))

**Copyright:** © 2017 Zhang X *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution Licence](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The author(s) is/are employees of the US Government and therefore domestic copyright protection in USA does not apply to this work. The work may be protected under the copyright laws of other jurisdictions when used in those jurisdictions.

**Grant information:** This work was supported by the Wellcome Trust [107469/Z/15/Z] and the National Institute of Diabetes and Digestive and Kidney Diseases and the National Institute of General Medical Sciences of the NIH [U54DK105566, R01GM104371]. EVM is supported by the National Institutes of Health under a Ruth L. Kirschstein National Research Service Award (NRSA) NIH Individual Predoctoral Fellowship (F31) [AI122592-01A1]. AHODL is supported by an by the National Institutes of Health under an National Institute of General Medical Sciences T32 institutional training grant [T32GM007748].

*The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.*

**First published:** 23 May 2017, 2:33 (doi: [10.12688/wellcomeopenres.11640.1](https://doi.org/10.12688/wellcomeopenres.11640.1))

## Introduction

ClinVar<sup>1</sup> is a public database hosted by the National Center for Biotechnology Information (NCBI) for the purpose of collecting information on genotype-phenotype relationships in the human genome. One common use case for ClinVar is as a catalog of genetic variants that have been reported to cause disease. When interpreting variants for clinical or research use, this data is commonly paired with reference data from ExAC (Exome Aggregation Consortium) or gnomAD (Genome Aggregation Database)<sup>2</sup>, particularly for large-scale analyses of reportedly pathogenic variants.

ClinVar makes its data available via FTP in three formats: XML, TXT, and VCF. However, none of these files were ideally suited for large scale data collection and downstream analysis for variant interpretation. The VCF file only contains variants present in dbSNP, which is not a comprehensive catalog of ClinVar variants. The TXT file is organized around Allele ID and the reported clinical significance is aggregated over distinct disorders. It also lacks certain annotations such as PubMed IDs for related publications, inheritance mode, prevalence, and other disease related information. The XML file is a comprehensive representation, but is organized around unique variant-condition combinations and is large and complex, making it difficult to quickly look up a variant of interest, and many potential users considering larger scale analyses may not be familiar with tools required to parse this data format. In addition, both the XML and TXT representations contain many genomic coordinates that have been parsed from HGVS notation. Therefore these representations may be right-aligned in contrast to VCF standard (<http://samtools.github.io/hts-specs/VCFv4.1.pdf>), and may also be non-minimal, i.e. containing additional nucleotides of context to the left or right of a given variant<sup>3</sup>.

To facilitate access to accurate and comprehensive ClinVar data at scale, we developed this software tool to convert the latest raw ClinVar data into multiple data files with options specified by users.

The tool supports both genome build GRCh37 and GRCh38. For each genome build, ClinVar records are parsed into separate files for simple mono-allelic variants and complex variants (i.e. with more than one variant interpreted together such as compound heterozygous and haplotype), and are further separated by whether variant records are variant-condition specific or aggregated for distinct conditions. The variant records are also annotated with reference population data from ExAC or gnomAD. The resulting files provide a summary of the most relevant fields for a range of uses in an accessible format.

## Methods

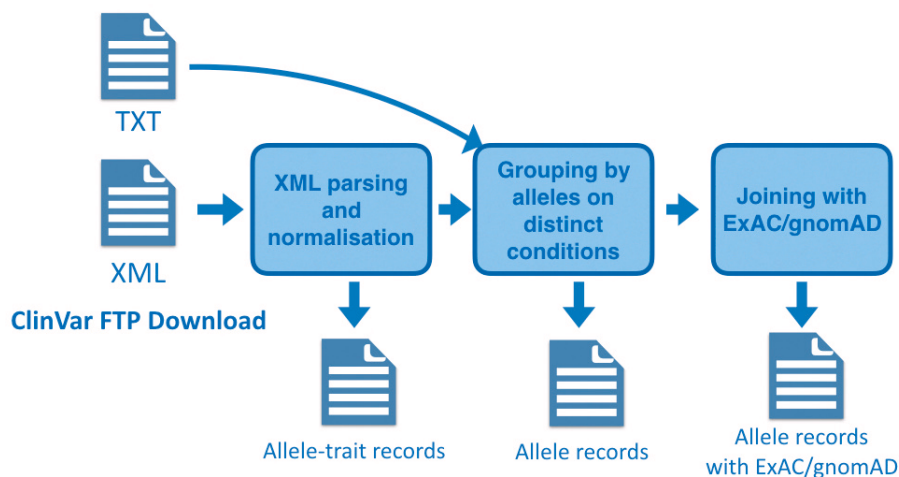
### Implementation

To create a flat representation of ClinVar data suited for general purposes, we took several steps illustrated in [Figure 1](#), which are encapsulated in the pipeline `src/master.py`:

- Download the latest XML and TXT files from ClinVar FTP.
- Parse the XML file to extract fields of interest into a flat file.
- Normalize the representation of genome coordinates using our Python implementation of `vt normalize`<sup>3</sup>.
- Group the allele-trait records by allele to aggregate record information except clinical significance from multiple submitters, independent of conditions. Join the TXT file to aggregate the clinical significances from multiple submitters and generate VCF files.
- Join with ExAC or gnomAD data and generate table files.

In the parsing process, when grouping the same allele over distinct conditions, we defined additional columns to indicate the clinical significance, since one allele may contain multiple assertions of clinical significance:

- Pathogenic is 1 if the variant has ever been asserted as "Pathogenic" or "Likely pathogenic" by any submitter for any phenotype, and 0 otherwise



**Figure 1.** The workflow to parse ClinVar data.

- Benign is 1 if the variant has ever been asserted as "Benign" or "Likely benign" by any submitter for any phenotype, and 0 otherwise
- Conflicted is 1 if the variant has ever been asserted as "Pathogenic" or "Likely pathogenic" by any submitter for any phenotype, and has also been asserted as "Benign" or "Likely benign" by any submitter for any phenotype, and 0 otherwise. A variant having one assertion of pathogenic and one of uncertain significance does not count as conflicted for this column.

## Operation

The pipeline is written in python 2.7 and R. It depends on python packages pysam (v0.11.1), pandas (v0.19.2) and pypeze (v0.1.5). Additionally, the tabix<sup>4</sup> and vt<sup>3</sup> third-party tools must be installed and executable from any directory.

To run the pipeline:

```
cd ./src
pip install -user -upgrade -r requirements.txt
python2.7 master.py -b37-genome/path/to/b37.fa -b38-genome/path/to/b38.fa \
-E/path/to/ExAC.vcf.gz -GG/path/to/gnomad.genomes.vcf.gz -GE/path/to/gnomad.exomes.vcf.gz
```

## Use case

Whiffin *et al.* (2016)<sup>5</sup> made use of the tool to efficiently extract the interpreted variants for inherited cardiovascular diseases from ClinVar, and used the variant allele frequency from ExAC to reassess the variant pathogenicity.

## Limitations

The accuracy of output files is limited by the downloadable files from the ClinVar FTP site. In the case that ClinVar releases new data with a new reporting format or an unfinished format update, our pipeline may not work for the latest release. We would recommend that users revert to the old version by specifying the input ClinVar files when executing `python master.py`.

## Summary

The software tool is developed to parse the latest ClinVar release data into analysis-friendly files to facilitate convenient data collection in variant interpretation research. The files are separated by genome build (GRCh37/GRCh38), and by whether they

represent simple mono-allelic variants, or complex multi-allelic variants such as compound-heterozygotes and haplotypes. Records are annotated with reference population data from ExAC and/or gnomAD. The representation of variant genome coordinates are normalized and the resulting output files are suitable for multi-purpose downstream variant-interpretation analysis.

## Software availability

The pipeline, its fully parsed data files and example data files are available: <https://github.com/macarthur-lab/clinvar>

Archived source code as at the time of publication: <https://doi.org/10.5281/zenodo.399052>

License: The software tool is distributed under an MIT license. ClinVar data, as a work of the United States federal government, are in the public domain and are redistributed here under the same terms as they are distributed by ClinVar itself. Importantly, note that ClinVar data are “not intended for direct diagnostic use or medical decision-making without review by a genetics professional”.

## Author contributions

EVM, DGM, and BW conceived the study. XZ, EVM, AHODL, JSW, and BW contributed to its design, developed and implemented the pipeline. XZ, EVM, AHODL, DGM, JSW, and BW prepared the manuscript.

## Competing interests

No competing interests were disclosed.

## Grant information

This work was supported by the Wellcome Trust [107469/Z/15/Z] and the National Institute of Diabetes and Digestive and Kidney Diseases and the National Institute of General Medical Sciences of the NIH [U54DK105566, R01GM104371]. EVM is supported by the National Institutes of Health under a Ruth L. Kirschstein National Research Service Award (NRSA) NIH Individual Predoc-toral Fellowship (F31) [AI122592-01A1]. AHODL is supported by an by the National Institutes of Health under a National Institute of General Medical Sciences T32 institutional training grant [T32GM007748].

*The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.*

## References

1. Landrum MJ, Lee JM, Benson M, *et al.*: **Clinvar: public archive of interpretations of clinically relevant variants.** *Nucleic Acids Res.* 2016; **44**(D1): D862–8. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
2. Lek M, Karczewski KJ, Minikel EV, *et al.*: **Analysis of protein-coding genetic variation in 60,706 humans.** *Nature.* 2016; **536**(7616): 285–291. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
3. Tan A, Abecasis GR, Kang HM: **Unified representation of genetic variants.** *Bioinformatics.* 2015; **31**(13): 2202–4. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
4. Li H: **Tabix: fast retrieval of sequence features from generic TAB-delimited files.** *Bioinformatics.* 2011; **27**(5): 718–9. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
5. Whiffin N, Minikel E, Walsh R, *et al.*: **Using high-resolution variant frequencies to empower clinical genome interpretation.** *Genet Med.* 2017. [PubMed Abstract](#) | [Publisher Full Text](#)

# Open Peer Review

Current Referee Status:  

---

## Version 1

Referee Report 12 June 2017

doi:[10.21956/wellcomeopenres.12572.r22989](https://doi.org/10.21956/wellcomeopenres.12572.r22989)



**Steven M Harrison** 

Laboratory for Molecular Medicine, Harvard Medical School, Boston, MA, USA

This paper by Zhang *et al* provides a mechanism to combine ClinVar data downloads into a format that is easier to work with and implement, as ClinVar downloads are either not fully comprehensive individually or too large and complex for the average ClinVar user to use. By including variant frequency information from gnomAD/ExAC, the authors have created a valuable resource for the genetics community. This paper is incredibly timely as more clinical laboratories and researchers utilize ClinVar and population frequency data, but incorrect implementation of data from either source can lead to incorrect filtration and classification.

To help user better understand and implement this resource, it may be helpful to change the logic for "conflicted" to instead indicate all medically significant conflicts (pathogenic/likely pathogenic versus uncertain significance/likely benign/benign) as opposed to just pathogenic/likely pathogenic versus likely benign/benign. For example, ClinVar variant <https://www.ncbi.nlm.nih.gov/clinvar/variation/186876/> has 6 submissions: 1 of Likely pathogenic and 5 of Uncertain significance. To mark the variant as "1" for "pathogenic" without an indication in the "conflicted" column could be misleading as it's not clear that the majority of submitters actually think this variant is of uncertain significance. Alternatively listing the clinical significance terms with counts would help, such as "Likely pathogenic(1);Uncertain significance(5)" but perhaps that is not available until ClinVar moves to a variant-centric XML download.

Additionally, it may be helpful to move from RCV#s to SCV#s when representing ClinVar records. Users who navigate to an RCV page may not be aware that they are only looking at interpretations regarding a variant/condition pair and not seeing all interpretations submitted for the variant. Also, by listing all SCV#s, it would be clearer to users how many total submissions there are for any given variant. Clarifying that `measureset_id = VariationID` could also be helpful as many ClinVar users are familiar with the VariationID concept but likely unaware of its name in the XML.

Lastly, if possible, it would be incredibly helpful if coverage data from gnomAD/ExAC could be added for variants in ClinVar that are absent from gnomAD/ExAC to help users determine the difference between absent due to poor coverage versus truly absent from tested alleles.

**Is the rationale for developing the new software tool clearly explained?**

Yes

**Is the description of the software tool technically sound?**

Yes

**Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?**

Yes

**Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?**

Yes

**Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?**

Yes

**Competing Interests:** No competing interests were disclosed.

**I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.**

Referee Report 05 June 2017

doi:[10.21956/wellcomeopenres.12572.r22988](https://doi.org/10.21956/wellcomeopenres.12572.r22988)



**Fiona Cunningham** <sup>1</sup>, **Sarah Hunt**<sup>2</sup>

<sup>1</sup> European Molecular Biology Laboratory, Wellcome Trust Genome Campus, European Bioinformatics Institute, Cambridge, UK

<sup>2</sup> European Molecular Biology Laboratory, European Bioinformatics Institute, Cambridge, UK

The authors provide a tool to reformat a ClinVar data release into a set of simply formatted files. They augment the ClinVar information with frequency data from the ExAC and gnomAD projects, which are the popular reference sets used in the community.

Existing tools like VEP and SNPeff allow variants to be annotated with allele frequencies from popular reference sets and ClinVar clinical significance classifications. They provide some but not full phenotype and evidence details.

ClinVar provides data in multiple file formats which are of different degrees of readability and parse-ability and contain different data types. Extracting the full information from these files can be non trivial, so this tool will be of use to groups wishing to use ClinVar data in large scale analyses.

The normalization step is essential to avoid losing potential matching records, but it would add clarity to state this step left-aligns allele representations as some groups use a different convention. Some information on the performance of the pipeline would also be useful.

All in all though, this is a useful tool for the community.

**Is the rationale for developing the new software tool clearly explained?**

Yes

**Is the description of the software tool technically sound?**

Yes

**Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?**

Yes

**Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?**

Yes

**Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?**

Yes

***Competing Interests:*** No competing interests were disclosed.

**We have read this submission. We believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.**

---