The Journal of Open Source Software

vcftoolz: a Python package for comparing and evaluating Variant Call Format files.

Steve Davis[1]

1 U.S. Food and Drug Administration

DOI: 10.21105/joss.01144

Software

- Review
- Repository
- Archive

Submitted: 27 November 2018
Published: 26 March 2019

Summary and Need Statement

The analysis of next-generation sequence data often involves variant calling – the process of identifying differences between genomes. The standard output file format of variant callers is the Variant Call Format (VCF) (https://github.com/samtools/hts-specs).

Researchers need to view and compare VCF files when comparing the behavior of different variant calling algorithms (and even the same algorithm with different parameters). Additionally, the performance of a variant calling algorithm can be evaluated by comparing against a known truth VCF dataset.

Here, we present `vcftoolz`, software to facilitate comparing and evaluating the variant calls in VCF files. The core functionality of `vcftoolz` is the capability to compare two or more VCF files, producing a report, Venn Diagrams, and a spreadsheet identifying the concordance between the VCF files. The artifacts produced by `vcftoolz` are not available from other tools.

The `vcftoolz` software is designed to work with bacterial variant files. It has been tested in a 3-way comparison of VCF files, with each file having 200 samples and 500 snps per sample for a total of 100,000 variants in each file.

Related Research

The `vcftoolz` software is being used as part of an ongoing effort to compare and evaluate the variant callers used by multiple government agencies involved in the analysis of pathogenic organisms of interest to food safety. In this effort, we use multiple variant callers to construct VCF files from food-borne pathogens. The `vcftoolz` software identifies the concordance between the VCF files produced by the alternative variant callers and facilitates algorithm improvements.

Prior Related Work

The `RTG Tools` package (Cleary et al., 2015) has advanced capabilities to compare VCF files containing complex variants, but does not support VCF files with multiple samples per file.

The `BCFtools` (Clarke et al., n.d.) package has the capability to create intersections, unions and complements of VCF files, as well as other useful tools for working with VCF files.

The `VCFtools` (Danecek et al., 2011) package has the capability to calculate differences between VCF files, among other functions.

Links

Documentation: https://vcftoolz.readthedocs.io/en/latest/readme.html

Source Code: https://github.com/CFSAN-Biostatistics/vcftoolz

PyPI Distribution: https://pypi.python.org/pypi/vcftoolz

References

Clarke, N., Collier, T., Danecek, P., Herrero, J., Kretzschmar, W., Li, H., McCarthy, S., et al. (n.d.). BCFtools. Retrieved November 26, 2018, from https://github.com/samtools/bcftools

Cleary, J. G., Braithwaite, R., Gaastra, K., Hilbush, B. S., Inglis, S., Irvine, S. A., Jackson, A., et al. (2015). Comparing variant call files for performance benchmarking of next-generation sequencing variant calling pipelines. *BioRxiv*, 023754. doi:10.1101/023754

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., et al. (2011). The variant call format and vcftools. *Bioinformatics*, *27*(15), 2156–2158. doi:10.1093/bioinformatics/btr330