The Journal of Open Source Software

VCFPy: a Python 3 library with good support for both reading and writing VCF

Manuel Holtgrewe[1] **and Dieter Beule**[1]

1 Berlin Institute of Health, Kapelle-Ufer 2, 10117 Berlin

DOI: 10.21105/joss.00085

Software

- Review
- Repository
- Archive

Summary

VCF file format (Danecek et al. 2011) is the standard file format for genetic variants, both small and structural variants. It has broad adaption in the Bioinformatics community and is used both by most projects, software, and databases these days.

There is a number of Python libraries for processing VCF, but most focus on reading VCF and not allowing for easily creating or augmenting VCF headers and records. For example, the most popular library PyVCF does not allow for built-in modification of the per-sample `FORMAT/*` records. PySAM (the wrapper for htslib) does only have very limited support for modifyin VCF records at all.

VCFPy addresses these issues and provides a well-documented, easy to use, and pythonic interface to reading and writing VCF files. It supports VCF v4.3, reading and writing of both plain-text and bgzip-compressed VCF files, as well as Tabix indices. Further, the project is well-documented and uses automatic testing as well as static code analysis for enforcing software quality standards.

References

Danecek, Petr, Adam Auton, Goncalo Abecasis, Cornelis A Albers, Eric Banks, Mark A DePristo, Robert E Handsaker, et al. 2011. "The Variant Call Format and Vcftools." *Bioinformatics* 27 (15). Oxford Univ Press: 2156–8.