




BAMnostic: an OS-agnostic toolkit for genomic sequence analysis

Marcus D Sherman¹ and Ryan E Mills^{1, 2}

¹ Department of Computational Medicine and Bioinformatics, University of Michigan ² Department of Human Genetics, University of Michigan

DOI: [10.21105/joss.00826](https://doi.org/10.21105/joss.00826)

Software

- [Review](#) 
- [Repository](#) 
- [Archive](#) 

Submitted: 15 June 2018

Published: 09 August 2018

License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC-BY](#)).

Summary

Sequencing technologies typically produce millions of plain text entries representing the genetic sequences of DNA or RNA fragments. With these data, bioinformatic pipelines give genetic context to the fragments by aligning them to larger reference sequences such as the resolved human genome. In order to handle these data structures in a standardized way, the Sequence Alignment Map (SAM, plain text), and Binary Alignment Map (BAM, byte-encoded) formats were created. As a standard, the BAM format is one of the most widely used formats for storing and processing sequencing data (Li et al., 2009). It is not uncommon to have a single file that can be 1 TB in its compressed binary-encoded form.

A high-throughput sequencing library (`htslib`) was developed to establish a standard encoding and compression schema to handle BAM files (Li et al., 2009). However, `htslib` currently does not readily support Windows environments and, due to its `htslib` dependency, the most popular Python toolset (`pysam` (2018a)) also cannot be used in a Windows environments or outside the CPython runtime (`pysam`, 2018b). Furthermore, both `pysam` and `htslib` have no intention to support Windows in the foreseeable future. This is a significant limitation as no other published Python implementation (besides `pysam`) can perform random access operation on BAM files.

To overcome the `htslib` dependency, `BAMnostic` was written from the ground-up as a fully featured, *pure Python* implementation of BAM file random access and parsing. Special care was taken to ensure `BAMnostic` had no dependencies outside of the Python standard library. As a corollary of the lack of dependencies, `BAMnostic` is not bound to a specific Python version (from 2.7 onward), or runtime (CPython and all stable versions of PyPy). Additionally, `BAMnostic` retains much of the same BAM file API as `pysam`. This allows `BAMnostic` to work as a drop-in replacement for `pysam` in everything from small web applications to full Python builds in Windows environments. As such, `BAMnostic` potentially makes genomic research and analytics available to a much greater software demographic.

`BAMnostic` is shipped with a small example BAM file for testing purposes (found under `bamnostic.example_bam`) and detailed documentation on both [Read the Docs](http://bamnostic.readthedocs.io/) at <http://bamnostic.readthedocs.io/> and within the packages docstrings. `BAMnostic` can be found on GitHub at <https://github.com/betteridiot/bamnostic>, [conda-forge](#), and the [Python Package Index \(PyPI\)](#) under the [BSD 3-Clause "New" License](#).

Acknowledgments

This work was supported by the University of Michigan [REM, MDS], the National Institutes of Health [R01HG007068 to REM], and the Rackham Merit Fellowship [MDS].

References

- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., et al. (2009). The sequence alignment/map format and samtools. *Bioinformatics*, 25(16), 2078–2079. doi:[10.1093/bioinformatics/btp352](https://doi.org/10.1093/bioinformatics/btp352)
- pysam. (2018a, May). Pysam. *GitHub*. Retrieved from <https://github.com/pysam-developers/pysam>
- pysam. (2018b, May). Pysam/issues. *GitHub*. Retrieved from <https://github.com/pysam-developers/pysam/issues/575>