St Petersburg University

13th International Multiconference on “Bioinformatics of Genome Regulation and Structure/Systems Biology”

Novosibirsk 2022

Pygenomics: manipulating genomic intervals and data files in Python

Gaik Tamazian, Nikolay Cherkasov, Alexander Kanapin and Anastasia Samsonova
Institute for Translational Biomedicine

Research Objectives

Developing a Python package with the following features:

- Pure Python implementation without third-party dependencies (e.g., modules compiled with C or C++)
- Using the Python standard library only
- Adherence to Python 3.7, thus enabling compatibility with CPython and PyPy.
- Supporting reading and writing in widely used bioinformatic data formats
- Providing operations with genomics intervals such as searching, intersecting, merging, and subtracting
- Demonstrating usability of the developed package by applying it to analysis of real-world bioinformatic datasets.

Materials and Methods

We developed the Python package *pygenomics* following the functional programming paradigm:

- Implementing most of the package functions without side effects (that is, making the functions *pure*)
- Localizing in separate entities routines that produce side effects
- Using stream-based input-output
- Making package objects immutable.

Pygenomics provides a developer-friendly framework that includes:

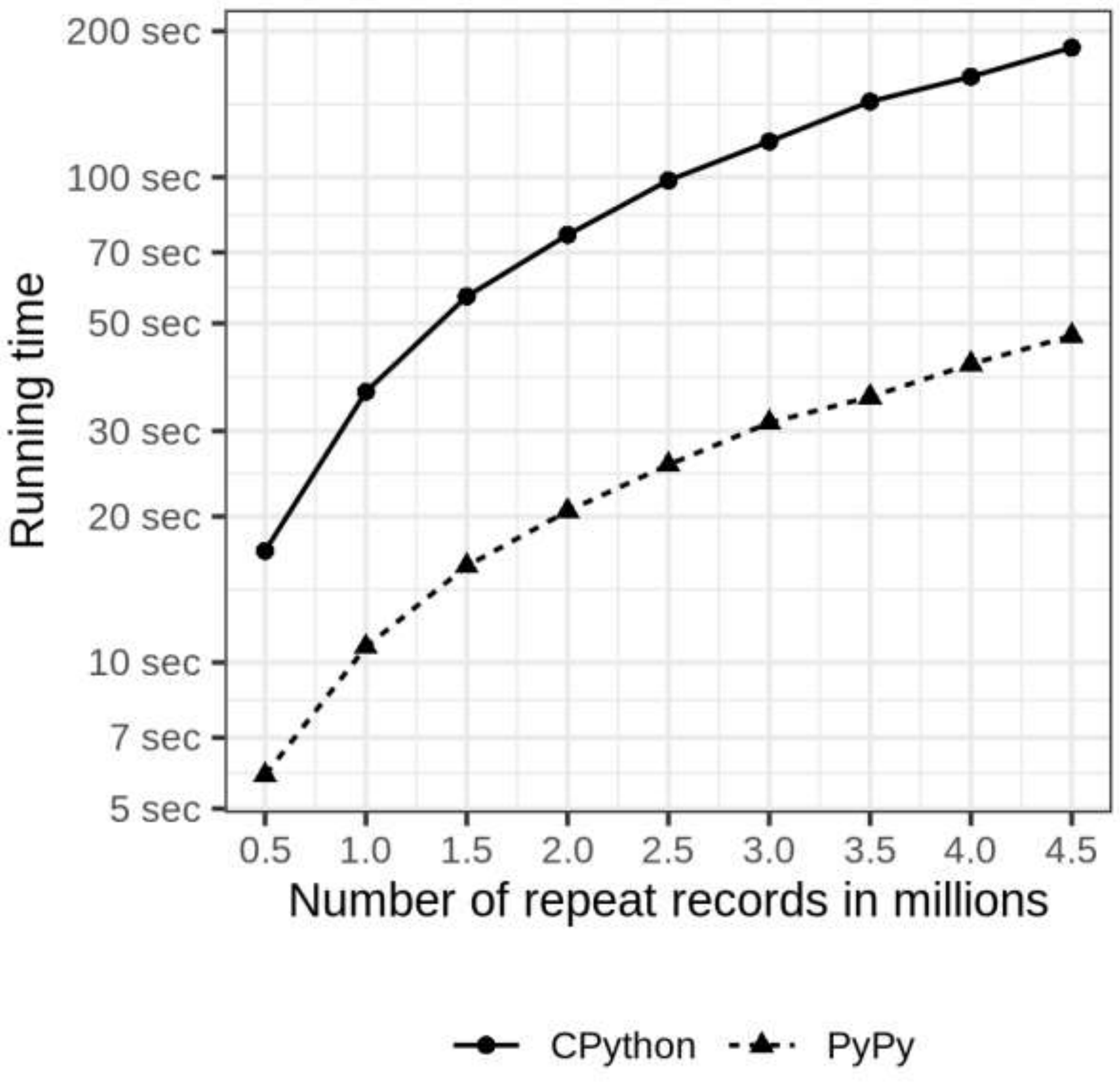
- In-source documentation
- Type annotations that enable static type checking
- Property-based and randomly generated test suites.

Results

Pygenomics supports a number of bioinformatic data formats, including

- BAM
- BED
- GFF3
- VCF.

Implementation of *pygenomics* in pure Python and according to the functional paradigm enables acceleration while running the package routines with PyPy. The figure shows the performance increase for the use case of reading and merging genomic intervals of human genome repeats annotated by RepeatMasker.



Conclusions

Pygenomics is publicly available on GitLab:

- https://gitlab.com/gtamazian/pygenomics - the main repository of the package
- https://gitlab.com/gtamazian/pygenomics_ext/ - the extra package that provides routines for working with custom data formats
- https://gitlab.com/gtamazian/pygenomics_examples - examples of using the API of *pygenomics* for developing custom scripts
- https://gitlab.com/gtamazian/pygenomics_paper - examples of incorporating *pygenomics* into bioinformatic pipelines.

Pygenomics provides a solid foundation for building bioinformatic software and data processing pipelines.

Acknowledgements

This research was supported by Grant of Russian Science Foundation (RSF) 20-14-00072 “Investigating genome integrity in oncologic and neuropsychiatric disorders”.

Thank you for your attention!