Jumpstart your genomics pipelines with genomepy

This manuscript (permalink) was automatically generated from vanheeringen-lab/genomepy_manuscript@ad8fcac on November 24, 2020.

Authors

- **Siebren Frolich**
  0000-0001-6925-8446 · siebrenf
  Department of Molecular Developmental Biology, Radboud University

- **Maarten van der Sande**
  0000-0001-7803-1526 · Maarten-vd-Sande · MaartenvdSande
  Department of Molecular Developmental Biology, Radboud University

- **Simon van Heeringen**
  0000-0002-0411-3219 · simonvh · svheeringen
  Department of Molecular Developmental Biology, Radboud University

Abstract

Summary

Analyzing genomics data, including RNA-, ATAC- and ChIP-sequencing, requires multiple types of support data, such as genome sequence and gene annotations. These resources can generally be retrieved from multiple organizations, where they exist at multiple versions, and may have been generated with varying methods. Which data to use depends on the context of the research, such as collaboration partners, data reuse or data quality. Many of the bioinformatic workflows and pipelines available to date require the user to make this informed decision and supply the support data. Obtaining this data can be a tedious and error-prone process and does not allow for full computational reproducibility.

Here we present genomepy, a quality-of-life enhancement tool, that can navigate the genome databases of Ensembl, UCSC and NCBI. Genomepy can search and install genome sequences and gene annotation data from these providers in a consistent, reproducible and documented manner. The search function retrieves genomes related to the search term, and can do so for one or all providers to allow the user to make an informed decision. The install function downloads a specified genome with sensible defaults, while providing full control to advanced features. Additionally, gene annotations can be downloaded and converted to commonly used formats, with built-in checks for compatibility with the genome. Genomepy can optionally create genome indexes for commonly used aligners, including splice-aware aligners utilizing both genome and gene annotations. Genomes and gene annotations not available on supported databases can be processed by genomepy as well, providing a consistent workflow with any genome.

Genomepy provides this functionality and more via command line interface and Python application programming interface, aimed at easy of use and integration in automated pipelines.

Availability and implementation

Genomepy can be installed using Bioconda and Pip, and is available at https://github.com/vanheeringen-lab/genomepy.

Introduction

Its big, its vague, and you just want to run your pipeline yesterday, right?

Table 1: Genome providers.

Database	Fun aspect
Ensembl	Generally seen as standard, updates infrequently, incompatible chromosome names
UCSC	multiple GTF formats, GTFs labelled incorrectly
NCBI	Different pipelines, looks like Ensembl, updates frequently

Related Work

- its not like refgenie, but they could work nicely with eachother!

- its missing in most workflows

there's a need for something that does the first step. genomepy fill that need.

genomepy

search, download, sensible defaults, reproducible, automatable. about those defaults…

Install via conda, pip or git.

basic steps in CLI

Repeat steps in API Extended steps, link to seq2science implementation?

Conclusions

Research is about making informed decisions. Choosing a reference assembly is no different. Genomepy offers an overview of the three largest genome providers, making this choice easier.

After choosing an assembly, data must be downloaded and processed for compatibility with downstream tools. Genomepy provides this functionality, while providing logging. Even if the required reference data is not available on the three largest genome providers, genomepy can process external data to provide a consistent output.

While genomepy makes choices during the processing, each of these can be tuned to the specific needs of a project using the CLI. More control can be achieved via the Python API. Combined, these features make genomepy ideal for integration in automated sequencing workflows, as demonstrated in seq2science [1].

Acknowledgements

We thank the Department of Molecular (Developmental) Biology, out github contributors, and issue posters for their patience, feedback and insight. We thank black, pytest, CodeCoverage and TravisCI for enduring our abuse and teaching us patience. And finally, we thank Manubot [2] for assisting with this manuscript.

Code availability

Genomepy can be installed using Bioconda and Pip. The code is available at https://github.com/vanheeringen-lab/genomepy.

References

1. **seq2science**
   Maarten Van Der Sande, Siebren Frölich, Jos Smits, Simon Van Heeringen
   Zenodo (2020-11-05) https://doi.org/ghktg7
   DOI: 10.5281/zenodo.3921913

2. **Open collaborative writing with Manubot**
   Daniel S. Himmelstein, Vincent Rubinetti, David R. Slochower, Dongbo Hu, Venkat S. Malladi, Casey S. Greene, Anthony Gitter
   PLOS Computational Biology (2019-06-24) https://doi.org/c7np
   DOI: 10.1371/journal.pcbi.1007128 · PMID: 31233491 · PMCID: PMC6611653