Bedtoolsr: An R package for genomic data analysis and manipulation

Mayura N. Patwardhan[1], **Craig D. Wenger**[2], **Eric S. Davis**[3], **Douglas H. Phanstiel**[1,3,4,5]

[1]Thurston Arthritis Research Center, University of North Carolina, Chapel Hill, NC 27599, USA

[2]Independent Scholar, Long Beach Township, NJ 08008

[3]Curriculum in Bioinformatics & Computational Biology, University of North Carolina, Chapel Hill, NC 27599, USA

[4]Department of Cell Biology & Physiology, University of North Carolina, Chapel Hill, NC 27599, USA

[5]Lineberger Comprehensive Cancer Research Center, University of North Carolina, Chapel Hill, NC 27599, USA

Summary

The sequencing of the human genome and subsequent advances in DNA sequencing technology have created a need for computational tools to analyze and manipulate genomic data sets. The bedtools software suite and the R programming language have emerged as indispensable tools for this purpose but have lacked integration. Here we describe bedtoolsr, an R package that provides simple and intuitive access to all bedtools functions from within the R programming environment. We provide several usability enhancements, support compatibility with past and future versions of bedtools, and include unit tests to ensure stability. bedtoolsr provides a user-focused, harmonious integration of the bedtools software suite with the R programming language that should be of great use to the genomics community.

Introduction

The sequencing of the human genome and subsequent advances in DNA sequencing technology have transformed modern biological research by producing data sets of ever-increasing size and complexity. While these technologies have led to breakthroughs in genetics research, the incredible throughput and breadth of the resulting data have spurred a reliance on computational tools and programming languages to interpret the results. In 2010, Quinlan et al. developed bedtools, a powerful suite of command-line tools for 'genome arithmetic' that has become one of the most widely used and indispensable tools for genomic data analysis (Quinlan and Hall 2010). A year later, pybedtools extended the features of bedtools to the python programming language (Dale, Pedersen, and Quinlan

2011). During that same time period, the use of the programming language R—with a rich trove of libraries for statistical analysis and data visualization—skyrocketed in the biological sciences (Tippmann 2015). While several R packages have been developed for bedtools-like genome analysis, their usage and functionality differ significantly from that of bedtools (Riemondy et al. 2017; Lawrence et al. 2013). These differences make them more difficult to use for those who are already familiar with bedtools behavior and lacks some of the capabilities of bedtools.

Here we describe bedtoolsr, an R package that allows seamless integration of bedtools functions into the R programming environment. bedtoolsr functionality, inputs, outputs, and documentation perfectly replicate those found in the command-line version of bedtools and offer new features for improved usability within the R environment.

Methods

bedtoolsr is an R package that allows access to all bedtools functions from within the R environment. To support past, current, and future versions of bedtools, we wrote bedtoolsr using a metaprogramming approach. The bedtoolsr package is built by a python script that reads function names, parameters, default settings, and documentation from a local installation of bedtools and constructs a distributable R package custom-built for that bedtools version. bedtoolsr is version controlled and freely available on the software development platform GitHub. To ensure stability, bedtoolsr supports continuous integration and includes unit tests for every function. These unit tests were implemented using the R package testthat (Wickham 2011) and can be run immediately after installation to ensure proper functionality. The continuous integration service Travis CI runs every time a code change is posted to GitHub to safeguard against any updates that might introduce flaws, faults, or failures.

Results

bedtoolsr was written with user experience in mind. To minimize the learning curve for those already familiar with bedtools we aimed to perfectly replicate the bedtools experience while adding all of the features of an R package. As such, bedtoolsr supports every currently available bedtools function and all function parameters. Parameters have the exact same names and documentation as those provided by bedtools with code autocompletion support from within RStudio.

bedtoolsr extends bedtools features for improved ease of use. Inputs for bedtoolsr functions can be provided either as file paths or as R objects (e.g. data.frames, data.tables, tibbles, etc). bedtoolsr automatically detects whether the input is a file path or R object and handles the data accordingly. To simplify usage, bedtoolsr comes preloaded with chromosome size files for commonly used genomes that are required by many bedtools functions. Results can be returned as a data frame or written directly to a file. To ensure proper installation of the package, users can run unit tests for most functions which can be executed with a single command following installation.

Discussion

bedtoolsr provides seamless integration of the bedtools software suite into the R programming environment. The package was designed to be as user-friendly as possible and should be intuitive for those already familiar with the bedtools command-line tool. The ability to handle multiple data types, the forward and backward compatibility, and the included unit tests ensure stability and ease of use. The harmonious combination of these two powerful analytical platforms should make bedtoolsr a valuable and widely used tool for genomic analysis.

Acknowledgements

References

Dale Ryan K, Pedersen Brent S, and Quinlan Aaron R. 2011 "Pybedtools: A Flexible Python Library for Manipulating Genomic Datasets and Annotations." Bioinformatics 27 (24): 3423–4. 10.1093/bioinformatics/btr539. [PubMed: 21949271]

Lawrence Michael, Huber Wolfgang, Pagès Hervé, Aboyoun Patrick, Carlson Marc, Gentleman Robert, Morgan Martin T, and Carey Vincent J. 2013 "Software for Computing and Annotating Genomic Ranges." PLoS Comput. Biol 9 (8): e1003118 10.1371/journal.pcbi.1003118. [PubMed: 23950696]

Quinlan Aaron R, and Hall Ira M. 2010 "BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features." Bioinformatics 26 (6): 841–42. 10.1093/bioinformatics/btq033. [PubMed: 20110278]

Riemondy Kent A, Sheridan Ryan M, Gillen Austin, Yu Yinni, Bennett Christopher G, and Hesselberth Jay R. 2017 "Valr: Reproducible Genome Interval Analysis in R." F1000Res. 6 (June): 1025 10.12688/f1000research.11997.1. [PubMed: 28751969]

Tippmann Sylvia. 2015 "Programming Tools: Adventures with R." Nature 517 (7532): 109–10. 10.1038/517109a. [PubMed: 25557714]

Wickham Hadley. 2011 "Testthat: Get Started with Testing." R Journal 3 10.32614/RJ-2011-002.