Streamlining data-intensive biology with workflow systems

This manuscript (permalink) was automatically generated from bluegenes/2020-gep@da0d815 on May 14, 2020.

Authors

- **Taylor Reiter**
  0000-0002-7388-421X · taylorreiter · ReiterTaylor
  Department of Population Health and Reproduction, University of California, Davis · Funded by Grant XXXXXXXX

- **C. Titus Brown**
  0000-0001-6001-2677 · ctb · ctitusbrown
  Department of Population Health and Reproduction, University of California, Davis · Funded by Moore Foundation GBMF4551

- **N. Tessa Pierce**
  0000-0002-2942-5331 · bluegenes · saltyscientist
  Department of Population Health and Reproduction, University of California, Davis · Funded by NSF 1711984

Abstract

As both sequencing technologies and data have proliferated, the bottleneck of biological sequence analysis has shifted from data generation to analysis.

The emergence of workflow systems designed for bioinformatics has altered the landscape of ...

Fortunately, analysis tools and techniques have evolved to cope with this ever-increasing flood of data. Reliable and user-friendly workflow systems and software management have emerged to facilitate interrogation of many thousands of samples. For fundamental steps such as quality control, standardized protocols are now available meaning researchers can spend less time rewriting common analyses and more time examining the biological intricacies of their data. In cases where the data are too large for even high-performance computing environments, a series of tools have emerged that are capable of using small, representative subsets of massive datasets to produce comparable results. While adoption of these tools can both facilitate and expedite reproducible data analysis, knowledge of and training in these techniques is still lacking.

Here, we provide insight on workflow systems that have emerged to fill the gap for biologists....

Here, we provide a series of tips, tools, and “good enough” practices for biologists venturing into the realm of biological sequence analysis. The guidelines and tools presented below are designed to apply to novel or publicly-available sequencing data sets and across the range of computational resource options available to researchers.

The majority of this manuscript will covers understanding how to conduct computational analyses on sequencing data, Except for data acquisition, the tools and guidelines presented below apply to either novel or publicly-available data.

Author Summary

In this paper, we present our guide for biological sequence data analysis, developed through our own teaching, training and analysis. We recognize that this is currently biased towards our own use casea nd experiences, but we hope to engage in robust discussion with the open source community in order in include the best set of practices.

Our main goal is to accelerate scientists conducting sequence analyses into organized workflow practices that benefit their own research while also facilitating open and reproducible science. %Our main goal is to accelerate biologists/bioinformaticians into organized workflow practices that benefit their own research while also facilitating open, reproducible analyses

Sequencing data are now widely available for species across the tree of life, and new sequencing data continues to be generated at a fantastic clip. %(cite sra growth?). The wealth of information present in sequencing data has the potential to revolutionize our understanding of the diversity and function of communities, building basic understanding from ecosystems to human health. However, sequence analysis remains both complex and computationally intensive, problems that are compounded during analysis of large datasets.

The magnitude of sequencing data requires a principled approach to management, analysis, and dissemination of results. As sequencing analysis has matured over the past decade, several papers have presented “best” or “good enough” practices for computational biological analyses ADD BACK CITATIONS These recommendations have both helped build consensus and fueled additional tool and workflow development. Since the latest paper in 2017 WILSONCITE, a number of important tools have greatly reduced the barrier to entry and opened the door to end-to-end reproducible analyses. % simple, shareable, etc Many of these changes owe their origin, at least in part, to the open science movement and the recognition of the importance of entry-level training, such as that provided by The Carpentries TEALCITE (open sci movement CITE).

The key advancements over the past few years have come in workflow scripting, software management, and tools that handle biological data at scale. % and sharing? Role of github/open code? The combination of workflow languages (e.g. snakemake, nextflow, common workflow language) and package installations (e.g. conda) have revolutionized bioinformatic analysis development. These tools enable researchers to build reproducible analyses that can be automatically executed in a directed fashion. With integrated installations, these workflows can work across different computational systems, and can even serve as a form of documentation for the analysis. Finally, when paired with new tools leveraging computational approximations, this suite of tools enables researchers to cope with the enormity of sequencing data. %have emerged a promising solution to coping with the enormity of sequencing data. %..provide researchers a framework/structure %Adopting workflow-based systems may be the single best step you can take to improve your analyses (here’s where to talk up workflows!) %bonus: these integrate with software installation! Also provide a bunch of other neat data-sciencey logging and benchmarking.

In this paper, we build on our experiences training researchers as part of The Carpentries and other courses and workshops. We present a roadmap for biological sequence analysis, beginning at data acquisition and providing specific recommendations for tools that ensure the integrity of your data along the way. We emphasize the importance of adopting a workflow-based approach to enhance documentation, automation and reproducibility of your science. Adopting these approaches will not only propel your own research, but will also facilitate sharing, discussion, peer review, etc. Here, we present our best advice on how to get the most out of your sequencing data and time.

%- analogy: manual approach works well once, but learning to use power tools will make it work many times? %p3ish %Via these tips, we hope to accelerate biologists/bioinformaticians into open and

reproducible workflow practices. % %- workflow tips and good approaches (e.g. making big data less big) to data analysis.

References