BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 10 2014, pages 1488–1490
doi:10.1093/bioinformatics/btu055

Genetics and population analysis Advance Access publication January 30, 2014

Mobyle SNAP Workbench: a web-based analysis portal for population genetics and evolutionary genomics

James T. Monacell[1,2] and Ignazio Carbone[1,2,*]

[1]Center for Integrated Fungal Research, Department of Plant Pathology and [2]Bioinformatics Research Center, North Carolina State University, NC 27695, USA

Associate Editor: Gunnar Ratsch

ABSTRACT

Summary: Previously we developed the stand-alone SNAP Workbench toolkit that integrated a wide array of bioinformatics tools for phylogenetic and population genetic analyses. We have now developed a web-based portal front-end, using the Mobyle portal framework, which executes all of the programs available in the stand-alone SNAP Workbench toolkit on a high-performance Linux cluster. Additionally, we have expanded the selection of programs to over 189 tools, including population genetic, genome assembly and analysis tools, as well as metagenomic and large-scale phylogenetic analyses. The Mobyle SNAP Workbench web portal allows end users to (i) execute and manage otherwise complex command-line programs, (ii) launch multiple exploratory analyses of parameter-rich and computationally intensive methods and (iii) track the sequence of steps and parameters that were used to perform a specific analysis. Analysis pipelines or workflows for population genetic, metagenomic and genome assembly provide automation of data conversion, analysis and graphical visualization for biological inference.

Availability: The Mobyle SNAP Workbench portal is freely available online at http://snap.hpc.ncsu.edu/. The XMLs can be downloaded at http://carbonelab.org/system/files/snap_xmls.tgz. Each XML provides links to help files, online documentation and sample data.

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on July 1, 2013; revised on November 27, 2013; accepted on January 23, 2014

1 INTRODUCTION

Previously we developed the SNAP Workbench toolkit that integrated phylogenetic and population genetic analyses and made complicated sets of program functions, settings and parameters accessible and transparent to the user (Aylor, *et al.*, 2006; Price and Carbone, 2005). The SNAP Workbench provided a graphical user interface for command-line programs and a workflow to guide the user on the assumptions and limitations of the methods that test for population differentiation. Although useful for methods based on summary statistics (Excoffier and Lischer, 2010) and simple evolutionary models (Fu and Li, 1993; Tajima, 1989) further insights into the magnitude and direction of evolutionary forces required the use of coalescent (Griffiths and Tavaré, 1994a, b; Kuhner, 2009) and Bayesian methods (Beerli and Palczewski, 2010) that are parameter-rich and frequently computationally intensive (Wakeley and Lessard, 2006). This translated into long computational times, often requiring several days or weeks, particularly for larger datasets and complex models that included migration and recombination. Also restrictive was the large parameter space that must be examined for complex phylogeographic models to ensure reliable and accurate parameter estimation. In practice, this requires performing multiple separate independent simulations, which is prohibitive on a single stand-alone machine. To overcome these limitations, we have migrated our stand-alone SNAP Workbench to a web-based platform that allows for execution on a high-performance Linux cluster. This eliminates the need to provide multi platform installations of the workbench and has allowed us to expand our repertoire of tools to include large-scale phylogenetic analyses, metagenomics, genome assembly and comparative analyses and an enriched population genetic toolkit (see Supplementary Table 1S).

2 SYSTEMS AND METHODS

The Mobyle SNAP Workbench portal utilizes the Mobyle XML language and workflow engine (Neron *et al.*, 2009) to interface command-line programs. The XML specifies all runtime parameters, input and output files required for successful program execution and visualization of results. Functionality is built into each XML to provide the end user with curated example files and default parameters for mock execution, as well as definitions of key program parameters and access to online documentation and relevant publications for more detailed information. Adoption of the Mobyle program description language also allows for integration into existing Mobyle servers and programs, which would be beneficial for resource sharing; however, XMLs can also be downloaded and installed locally. Mobyle's implementation of workflows allows chaining together a series of XMLs such that a single user submission will run each XML in sequence as a separate job. Alternatively, scripting can be used such that all programs are executed consecutively in a single XML. A novelty in our implementation of workflows is the automation of intermediate data processing steps that would normally require user input. For example, in a typical genetree (Griffiths and Tavaré, 1994a, b) analysis many files, representing possible rooted trees, are generated that are the starting trees for coalescent simulations to determine the most likely rooted gene genealogy. We have automated the process of (i) selecting the rooted tree with the highest likelihood, (ii) performing simulations on the best tree to estimate ages of mutations and time to the most recent common ancestor and (iii) drawing the tree showing mutation probabilities, haplotype distributions and time scale in coalescent or real time units. Workflows are particularly important in processing

*To whom correspondence should be addressed.

next generation sequencing (NGS) reads or metagenomic data because these files are typically multiple gigabytes in size and require several consecutive pre-processing steps. In this case, each workflow can be scripted into a single XML, eliminating the transfer of intermediate metadata over the network. A formal description of the XML grammar (https://projets.pasteur.fr/projects/mobyle/wiki) is available on the Mobyle project website. Although there is currently no mechanism in mobyle for browsing files on the server, the path to files can be specified in the XML. We provide a tool for uploading large data files using the Velocity service (https://velocity.ncsu.edu/) hosted at NC State University. Users that set up a local instance of the portal can also upload NGS data files via sftp using clients such as Cyberduck for Mac (http://cyberduck.ch/) and WinSCP for Windows (http://winscp.net/eng/index.php). The workflows track all metadata from original source data to final results, thereby ensuring that data recording and analysis are explicitly traceable.

3 FRAMEWORK

Several integrative computational analysis frameworks have been developed that bring together tools, parameter settings and metadata for streamlining computational analyses and making results transparent and reproducible (Nekrutenko and Taylor, 2012). Galaxy (Goecks *et al.*, 2010) and Mobyle (Neron *et al.*, 2009) are two open, widely used web-based platforms being actively developed to bring together diverse computational tools and to provide a framework for adding new tools and workflows. We developed new program XMLs and workflows that can be deployed on the Mobyle web-based framework (Supplementary Tables 1S and 2S). These include XMLs and workflows for (i) large-scale phylogenetic and metagenomics analyses, (ii) computationally intensive simulations in population genetics and (iii) routine summary statistics for data exploration and *a priori* inferences on biological processes. The Mobyle SNAP Workbench web-portal allows researchers to seamlessly manage and execute complex command-line programs with multiple input files and parameters on a high-performance Linux cluster; these tools are parameter-rich or memory-intensive, often requiring several days or weeks to run. Our optimization includes selecting appropriate number of compute nodes and machine architecture for MPI programs and when possible, parallelization of computational tasks to execute on multiple machines. A unique feature of the portal is the implementation of workflows that link together several programs sequentially. This greatly facilitates exploratory population genetic and genomic analyses for the novice user but also allows for efficient use of cluster resources. The current version of our portal utilizes version 1.0.7 of Mobyle and is deployed in the NC State University Virtual Computing Lab (VCL). Users can also download program XMLs (http://carbonelab.org/system/files/snap_xmls.tgz) and the most recent version of Mobyle to install a local instance of the portal (https://projets.pasteur.fr/projects/mobyle/wiki/download). Programs and associated workflows are organized into fourteen subject areas (Supplementary Table S1). The XMLs are distributed to work on batch submission systems that are supported by Mobyle (https://projets.pasteur.fr/projects/mobyle/wiki). In addition to providing curated sample datasets and parameter values in XMLs, the Mobyle SNAP Workbench framework extends Mobyle's interactive tutorial pages with embedded YouTube videos. Tutorials are designed to guide the user through several sample analyses illustrating the portal data-entry-interface design and user-interactive web applets.

3.1 Implementation

The Mobyle SNAP Workbench portal is publicly available at http://snap.hpc.ncsu.edu/. We also distribute the program descriptions in XML format, which can be used in combination with Mobyle to host a web server. The web server can be hosted on a single unix/linux-based machine capable of hosting a web server, but many programs are optimized to take advantage of Message Passing Interface (MPI) and multiple processing resources available in a computing cluster for increased computational power. Mobyle SNAP Workbench has been useful for teaching by providing students with a convenient way to learn to use bioinformatics tools without requiring any familiarity with unix and program-specific command-line parameters. The implementation of the portal on a linux cluster has greatly facilitated phylogenetic/metagenomic analyses and genome assemblies as part of an ongoing interdisciplinary study of fungal endophytes and their function in boreal forests (http://www.endobiodiversity.org/). We anticipate that the portal will facilitate genome-scale coalescent-based inferences from large resequencing projects (Wei *et al.*, 2013) and enhance exploratory comparative analyses spanning the population/species interface and higher taxonomic scales.

ACKNOWLEDGEMENTS

We thank Hervé Ménager, Bertrand Néron, Eric Sills, Jianwei Dian and Doug Brown for helping with the installation and optimization of Mobyle and XMLs on the NC State High Performance Computing (HPC) cluster. We thank Alex Griffing and Stephen Tourigny for their work in developing PCA and genome assembly XMLs, respectively; Ross Whetten for guidance in developing the filter and assembly workflows; Kisurb Choe for developing the STRUCTURE workflow; and Rakhi Singh for help in debugging tools.

Funding: The NC State High Performance Computing (HPC) resources provided by the NC State Information Technology Division with support from the Office of the Provost and Office of Research and Graduate Studies (in part). Bioinformatics Training Grant (NIEHS 5 T32 ES007329 to J.T.M.); National Science Foundation's Dimensions of Biodiversity (DoB) program (Award Number 1046167 to I.C.).

Conflict of Interest: none declared.

REFERENCES

Aylor,D.L. *et al.* (2006) SNAP: combine and Map modules for multilocus population genetic analysis. *Bioinformatics*, **22**, 1399–1401.

Beerli,P. and Palczewski,M. (2010) Unified framework to evaluate panmixia and migration direction among multiple sampling locations. *Genetics*, **185**, 313–326.

Excoffier,L. and Lischer,H.E. (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol. Ecol. Resour.*, **10**, 564–567.

Fu,Y.X. and Li,W.H. (1993) Statistical tests of neutrality of mutations. *Genetics*, **133**, 693–709.

Goecks,J. *et al.* (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.*, **11**, R86.

Griffiths,R.C. and Tavaré,S. (1994a) Ancestral inference in population genetics. *Statistical science*, **9**, 307–319.

Griffiths,R.C. and Tavaré,S. (1994b) Simulating probability distributions in the coalescent. *Theor. Popul. Biol.*, **46**, 131–159.

Kuhner,M.K. (2009) Coalescent genealogy samplers: windows into population history. *Trends Ecol. Evol.*, **24**, 86–93.

Nekrutenko,A. and Taylor,J. (2012) Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. *Nat. Rev. Genet.*, **13**, 667–672.

Neron,B. *et al.* (2009) Mobyle: a new full web bioinformatics framework. *Bioinformatics*, **25**, 3005–3011.

Price,E.W. and Carbone,I. (2005) SNAP: workbench management tool for evolutionary population genetic analysis. *Bioinformatics*, **21**, 402–404.

Tajima,F. (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, **123**, 585–595.

Wakeley,J. and Lessard,S. (2006) Corridors for migration between large subdivided populations, and the structured coalescent. *Theor. Popul. Biol.*, **70**, 412–420.

Wei,W. *et al.* (2013) A calibrated human Y-chromosomal phylogeny based on resequencing. *Genome Res.*, **23**, 388–395.