Title: Deploying Galaxy on the Cloud

Authors: Enis Afgan[1], Dannon Baker[1], Nate Coraor[3], The Galaxy Team[2], Anton Nekrutenko[3], James Taylor[1]

Author affiliations:

[1]Department of Biology and Department of Mathematics & Computer Science, Emory University {E.A. email: eafgan@emory.edu}

[2]http://galaxyproject.org

[3]Huck Institutes of the Life Sciences and Department of Biochemistry and Molecular Biology, The Pennsylvania State University

Project website: http://usegalaxy.org/cloud

Project source code: http://bitbucket.org/afgane/galaxy-central-gc2/

Open Source License used: Academic Free License

Abstract:

DNA sequencing has become one of the most transformative high-throughput techniques in life sciences. Novel sequence-based assays have made sequencing an indispensable tool for studying gene expression and regulation, chromatin structure, and sequence variation. What is most transformative however, is the wide availability of “next-generation” DNA sequencing (NGS) instruments, enabling any investigator, for a modest cost, to produce enormous amounts of DNA sequence data. However, working with raw data generated by next-generation sequencers and transforming it into biologically meaningful information requires significant computing infrastructure and informatics support. For the majority of experimentalists that lack needed computational support, just storing and managing the vast amount of data produced by these technologies presents a significant informatics burden As a result, for an experimental group with no computational expertise, simply running a data analysis program is a barrier, let alone building a compute and data storage infrastructure capable of dealing with volume and processing requirements of NGS data.

As a first step in combating the NGS data deluge, there exists an open-source system, Galaxy. Galaxy provides an integrated analysis environment where domain scientists can, without informatics expertise, interactively construct multi-step analyses, with outputs from one step feeding seamlessly to the next. In order to utilize Galaxy and ease NGS analyses, there is a need for availability of computational infrastructure. Fortunately, a computational model – cloud computing – has recently emerged and is well suited to the analysis of large-scale sequence data. However, cloud computing resources are not yet suitable for immediate “as is” use by experimental biologists. As a step in the direction of enabling seamless NGS analyses on the cloud, and thus removing limitations of local or publicly offered computational services and contentions that arise in such environments, we have developed Galaxy Cloud (GC). GC is a comprehensive manager for enabling, running, and scaling the Galaxy application on cloud computing infrastructures. It offers a simple web-based interface that allows anyone to acquire the desired computational and storage resources on a cloud infrastructure, and perform NGS analysis through the familiar Galaxy interface and corresponding tools (complete Galaxy functionality is supported). GC automatically handles all aspects of resource acquisition, configuration, and data persistence, thus entirely insulating a user from the low-level computational details. This talk will focus on the motivation, use cases, and available functionality within GC.