

**Title:** Enabling NGS Analysis with(out) the Infrastructure

**Authors:** [Enis Afgan](#)<sup>1</sup>, Dannon Baker<sup>1</sup>, Nate Coraor<sup>2</sup>, Anton Nekrutenko<sup>2</sup>, James Taylor<sup>1</sup>

**Author affiliations:**

<sup>1</sup>Department of Biology and Department of Mathematics & Computer Science, Emory University {E.A. email: [efagan@emory.edu](mailto:efagan@emory.edu)}

<sup>2</sup>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/galaxy/cloudman>

**Open Source License used:** Academic Free License

### **Abstract:**

Running tools and performing analyses to transform sequence data into biologically meaningful information requires sophisticated computational infrastructure and support. The size of the required computational infrastructure is outpacing what individual researchers, many labs, and even universities are able to support. In addition, the setup and maintenance associated with a computational infrastructure presents significant problems for individual investigators and small labs that may not have the necessary informatics support. Fortunately, cloud computing provides unique capabilities for transparent scaling and sharing of computational infrastructures. Built on the Galaxy CloudMan platform, we have enabled the entire Galaxy application - completely configured with a range of tools and reference genomes - to transparently utilize AWS cloud resources. The presented solution delivers a fully functional infrastructure capable of performing complex genomic analyses in a matter of minutes.

This talk presents key new features of Galaxy CloudMan that focus around extension, transparency, and automation. Namely, we have **automated the process of deploying CloudMan** on a cloud infrastructure with the accompanying data, tools, and applications, making it completely transparent, reproducible, and accessible. **Any individual instance of CloudMan is now self-contained**, meaning that it does not require an external broker or service to operate. Moreover, this enables each instance of CloudMan to be customized by deploying new or alternative tools, configurations, and data, thus supporting the widely varied needs of individual investigators and labs. **CloudMan now supports setup of different cluster modes**, allowing one to utilize all of the CloudMan's infrastructure management features (e.g., cluster setup, NFS setup, data persistence, (automatically) adding/removing instances, sharing) but without setting up Galaxy. Coupled with the CloudBioLinux AMI that CloudMan builds upon, this feature allows any of the tools in NERC BioLinux to be run on a cluster managed by CloudMan without any additional setup. Additionally, any tool or application that can utilize a general purpose cluster can be installed on the deployed cluster while allowing CloudMan to manage the infrastructure. **CloudMan now supports sharing of cloud cluster instances**. This functionality allows an analysis to remain in the cloud (i.e., no need to download results and make available elsewhere) while minimizing the expense incurred by resources that need to be provided by the analysis owner. In addition to enabling publishing and sharing of data analyses, this feature allows sharing of customized instances of CloudMan where tools and/or data have been modified. This functionality minimizes repeat effort and offers tool developers a platform for easily distributing their tools while minimizing any otherwise required setup (for both developers and users). Lastly, continuing the automation effort, **CloudMan now supports the notion of infrastructure autoscaling**. This feature allows a user to specify bounds for the size of their cluster while letting CloudMan automatically adjust the current number of the compute resources to match the current system load, thus taking maximum advantage of the elastic infrastructure underlying the computation. This feature supports the *set-it-and-forget-it* paradigm of providing a compute infrastructure for users without requiring them to manage it. This talk will highlight each of these major advancements in CloudMan and showcase their impact on user experience when using Galaxy and cloud computing resources.