Reproducible Research in the Cloud with the Refinery Platform

Nils Gehlenborg[1], *Shannan J Ho Sui*[2], *Ilya Sytchev*[2], *Stefan Luger*[3], *Fritz Lekschas*[1], *Richard W Park*[1], *Jennifer Marx*[1], *Scott Ouellette*[1], *David R Jones*[4], *Anton Xue*[1], *Psalm Haseley*[1], *Marc Streit*[3], *Winston Hide*[2,4], *Peter J Park*[1]

[1] Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA
[2] Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA
[3] Department of Computer Science, Johannes Kepler University Linz, Linz, Austria
[4] Sheffield Institute for Translational Neuroscience, University of Sheffield, Sheffield, UK

Website: *http://refinery-platform.org*
Repository: *https://www.github.com/parklab/refinery-platform*

Correspondence: *nils@hms.harvard.edu*

The Refinery Platform is a data analysis environment for reproducible research that links a data repository with analysis pipelines and visualization tools within a single user interface. The goal of Refinery is to facilitate analysis and interpretation of genomic and epigenomic data in a reproducible fashion. To support this, the data repository is built around the ISA-Tab data model (http://isa-tools.org) and analyses are executed in Galaxy (http://usegalaxy.org). Workflows are configured, launched, and monitored through the Refinery user interface, which offers a sophisticated file browser that operates on data set sample annotations. Among other efforts, we have created an instance of the Stem Cell Commons based on Refinery for the Harvard Stem Cell Institute with over 200 stem cell related data sets.

Originally designed to run on institutional or lab compute clusters, we have recently extended Refinery to be able to execute analyses on the Amazon Web Services (AWS) platform and to overcome some of the limitations of typical research computing environments. Refinery relies on Galaxy Cloudman (http://cloudman.irb.hr) to provide a Galaxy cluster on AWS. To reduce the effort required to deploy Refinery instances on AWS, we are creating custom Cloudman machine images that include the tools that we need for specific instances. By deploying our own compute cluster on AWS rather than relying on the availability of local infrastructure, the platform will also be more attractive to outside developers.

A second challenge that we recently addressed to make analyses more reproducible is the visualization of the provenance graphs that result from the execution of workflows on data sets with dozens of files. While node-link diagrams are useful to convey the flow of data from the input through the tools of workflows to the outputs, this representation does not scale to more than a handful of typical workflow executions. We have designed a new approach that employs dynamic graph aggregation and expansion based on user interest. Users can expand a subset of highly compressed graph to view details of a particular analysis without losing the overall context. Additionally, filters can be applied to hide part of the provenance graph—for example analyses conducted before a particular date—on demand.

The combination of cloud-based workflow execution, automated tracking of provenance information, and tools to visualize data provenance allows Refinery users to conduct reproducible computational research without any additional effort.

The Refinery Platform project is funded by NIH (R00 HG007583) and the Harvard Stem Cell Institute. Additional support is provided by an AWS in Education Research Grant.