Refinery: a data management, analysis, and visualization platform utilizing the Galaxy workbench

Nils Gehlenborg¹,², Shannan J Ho Sui²,³, David R Jones⁴, Jennifer K Marx¹, Chuck McCallum¹, Sarah Morgan⁴, Geoffrey M Nelson¹, Scott Ouellette¹, Ilya Sytchev², Winston Hide²,³,⁴, Peter J Park¹,³

¹Department of Biomedical Informatics, Harvard Medical School, ²Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA, ³Harvard Stem Cell Institute, Cambridge, MA, USA, ⁴Sheffield Institute for Translational Neuroscience, University of Sheffield, Sheffield, UK

Having a reproducible approach to managing data, running analysis workflows, and visualizing data is a common challenge across the bioinformatics community. These three tasks are typically isolated from one another which makes comprehensive tracking of data lineage, or data provenance, particularly difficult. Here we provide an update on the Refinery Platform, which addresses these challenges with a centralized data repository and tools to explore, analyze and visualize data. Our platform enables users to: 1) annotate and store all their data in a private, searchable database, 2) share data sets with other members via simple-to-use collaboration features, 3) run Galaxy Workflows on their data in a scalable fashion, and 4) launch interactive Docker-based visualizations on their raw data and derived results. We have recently improved capabilities for performing genomic analyses by leveraging Galaxy’s Dataset Collections, and by exposing Galaxy Workflow parameters through our Tool Launching interface. There has also been a large effort within the visualization tool space that allows users to launch interactive visualizations on their data from within our platform, and for prospective tool developers to wrap arbitrary web-based visualization tools in Docker containers with minimal overhead. Our vision is that the Refinery Platform will be a central resource for a community of researchers keen on tracking their data provenance and conducting reproducible science.

Acknowledgements
This project is funded by the Harvard Stem Cell Institute and the NIH/NHGRI (R00 HG007583).