A federated Galaxy for user-friendly, large-scale cancer genomics research

Jeremy Goecks and James Taylor
A federated Galaxy platform for analyses of cancer datasets that are widely distributed across private institutional repositories, commercial computing clouds, and public NIH computing clouds with applications to precision oncology.
The Galaxy Project
http://galaxyproject.org
>6500 analysis tools and >100 visualizations available
Web UI for accessibility
programmatic API for automation and scalability

Runs on laptop, computing clusters, and many clouds

Main public Galaxy server at https://usegalaxy.org
>300,000 analysis user jobs run each month
>100,000 registered users with >2PB of user data

>6500 citations across all areas of biomedical science

>100 public servers across the world
genomics, proteomics, image analysis, metabolomics, flow cytometry, drug discovery, public health, many more…

Funded by NHGRI, NSF, and now NCI and is/has been used by many ITCR projects
Team

Galaxy PIs:
- Jeremy Goecks, Oregon Health and Science University (contact PI)
  - Also CompBio lead for precision oncology at OHSU
- James Taylor, Johns Hopkins University (co-PI)
  - Several active cancer projects
- Anton Nekrutenko, Penn State University
- Dan Blankenberg, Cleveland Clinic

Kyle Ellrott, Oregon Health and Science University
- TCGA/GDAN and GA4GH contributor
Aim 1: Distributed data

Galaxy currently requires (a) all data to be on a local filesystem and (b) complete control over all data.

Goal: Extend Galaxy for working with distributed cancer genomics and phenotypic data.

Ideal scenario: joint analysis of genomics and imaging data on NCI clouds together with private datasets.
Aim 1 Subaims

Robust authentication and authorization in Galaxy

Extend Galaxy Data Libraries for remote access to cancer genomics datasets

Extend Galaxy Data Sources for remote access to cancer genomics datasets

Enable Galaxy users to access their own data in the cloud

Implement community schema (e.g., GA4GH) and provide access to major cancer datasets
Aim 2: Distributed (data-local) compute

Galaxy cannot run analyses across different computing clouds using cross-platform technologies

Goal: Distributed analyses in Galaxy using shared workflow representations

Ideal scenario: in-place data analysis for datasets on multiple clouds using software containers and cross-platform workflows
Aim 2 Subaims

Multi-cloud computing and container orchestration via CloudVE

Complete implementation of the Common Workflow Language in Galaxy

Implement community schemas (e.g., GA4GH) in Galaxy for running analysis jobs and workflows

Implement context-aware computing to improve analysis efficiency

Deploy a Federated Galaxy onto major cancer and public clouds
Aim 3

Precision oncology requires integrating public datasets with private patient data to improve therapeutic predictions and understand adaptation.

Goal: Apply Federated Galaxy to precision oncology research.

Ideal scenario: our Federated Galaxy includes machine learning workflows to meet common precision oncology needs.
Aim 3 Subaims

Establish Galaxy workflows and create baseline using existing datasets and methods

Extend workflows for new methods, datasets, and combinations of datasets

Automated hyperparameter optimization using cross-fold validation

Interactive patient-centric reports using ITCR visualizations
Precision Oncology at OHSU

Drs. Gordon Mills, Joe Gray, and Ray Bergan

Unique contributions: adaptive, combinatorial, and deep profiling in clinical trials

Breast, prostate, AML, and pancreas

Human Tumor Atlas Network data generation center (Goecks—co PI)

Initial workflows planned:

- Signatures for single Tx response
- Signatures for combination Tx response
- Adaptive response prediction
Acknowledgements

Galaxy core team and community

Goecks Lab @ OHSU
Taylor Lab @ Johns Hopkins
Nekrutenko Lab @ Penn State
Blankenberg Lab @ Cleveland Clinic

… and thousands of developers, trainers, and scientists around the world

Precision Oncology at OHSU

Gordon Mills
Joe Gray
Ray Bergan

Brett Johnson
Annette Kolodzie
Swapnil Parmar

… hundreds more