The Galaxy Platform for Accessible, Reproducible, and Scalable Biomedical Data Science

Jeremy Goecks
Associate Professor, Department of Biomedical Engineering
Section Head for Cancer Data Science, The Knight Cancer Institute
Oregon Health & Science University

OHSU KNIGHT CANCER Institute

Two Upcoming CWIG Galaxy Talks

Today's talk:

- What is Galaxy?
- Applications of Galaxy in Cancer Informatics

Next Month:

- Dr. Enis Afgan will present a demonstration of using Galaxy with software containers and workflows

Outline

The Galaxy Platform

Machine Learning Applications for Cancer

An Interactive Hub for Multiplex Tissue Image Analysis

How to Maximize the Value of Data-Intensive Science

Accessibility — Empower scientists regardless of informatics expertise

- Data-intensive science requires use of large datasets, computational resources, and analysis methods

Reproducibility — Ensure that data-intensive analyses are high-quality

- Critical for advancing science, including peer review, validation, and extension

Communication — Clearly sharing what has been done with others

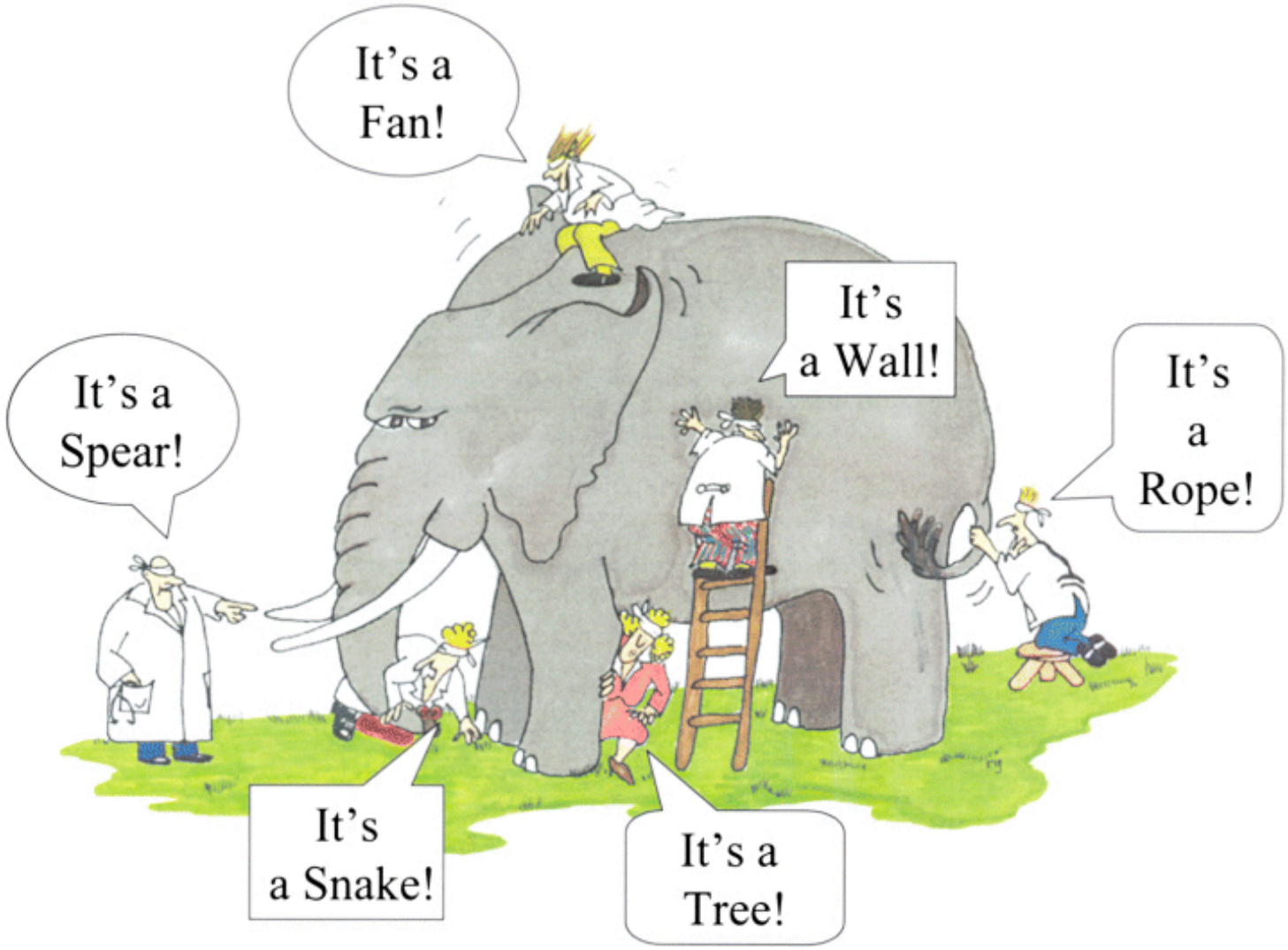
- Multiple levels of information are needed from broad overview down to essential details

What is Galaxy?

Different things to different communities

Galaxy user communities

- Scientists
- Tool Developers
- Educators
- Service Providers



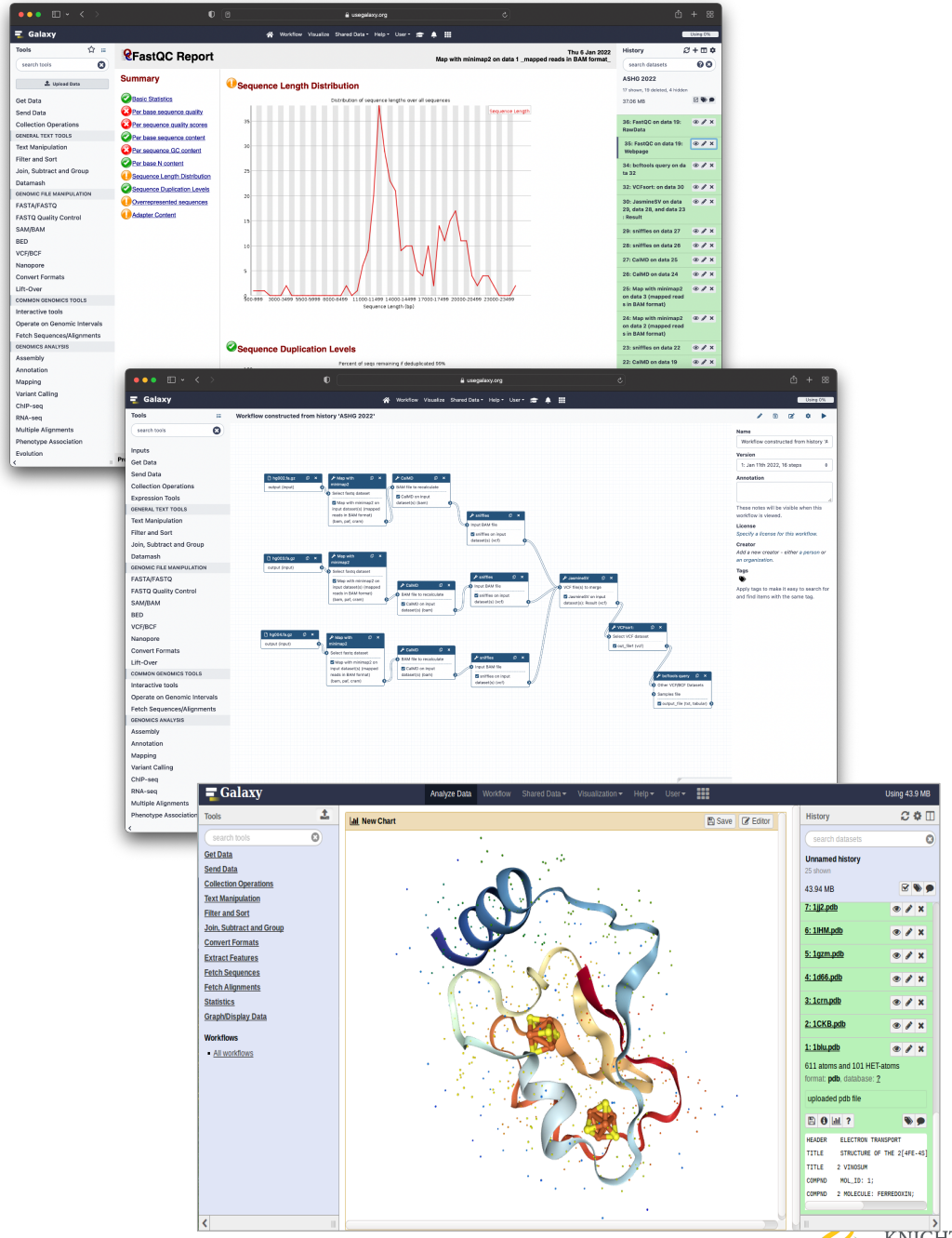
Key Features of Galaxy

- GUI for interactively running analysis tools on biomedical datasets
- Toolshed with 1,000s of tools ready to run
- Full featured workflow functionality
- Graphical interface for handling >1,000 samples
- Run Jupyter, RStudio, and other Interactive Tools for custom analyses
- Extensive training tutorials and infrastructure
- 6TB of latest, curated reference data

…all accessible via a Web browser or the command line and can be used with:

- Your own laptop
- Public high performance computational infrastructure
- Institutional computing resources
- Commercial cloud platforms

…created and supported by a large international community of scientists, developers, service providers, and educators



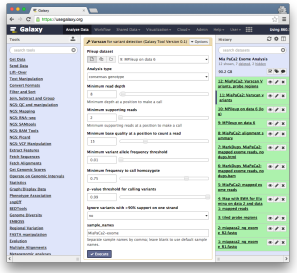
OHSU KNIGHT CANCER Institute

http://galaxyproject.org

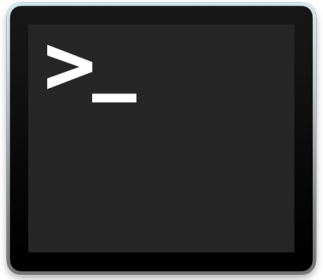
Integrate datasets, analysis tools, visualizations, and computing resources for large-scale biomedical data science

Interfaces

Web UI



Programmatic API



Datasets

Galaxy

API

Computing Resources

slurm workload manager

Condor High Throughput Computing

Google Cloud Platform

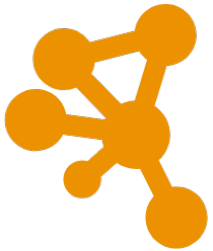
amazon web services

Analysis Tools and Visualizations

Filter by columns

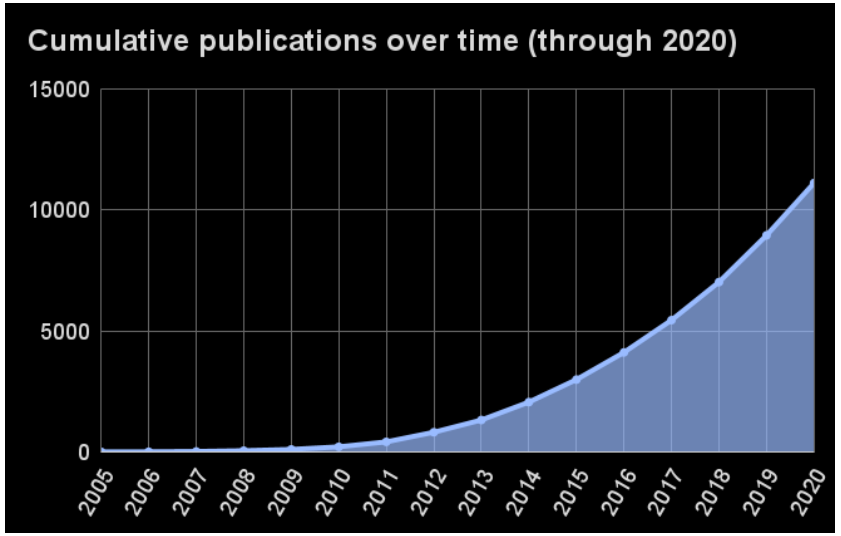
Differential expression

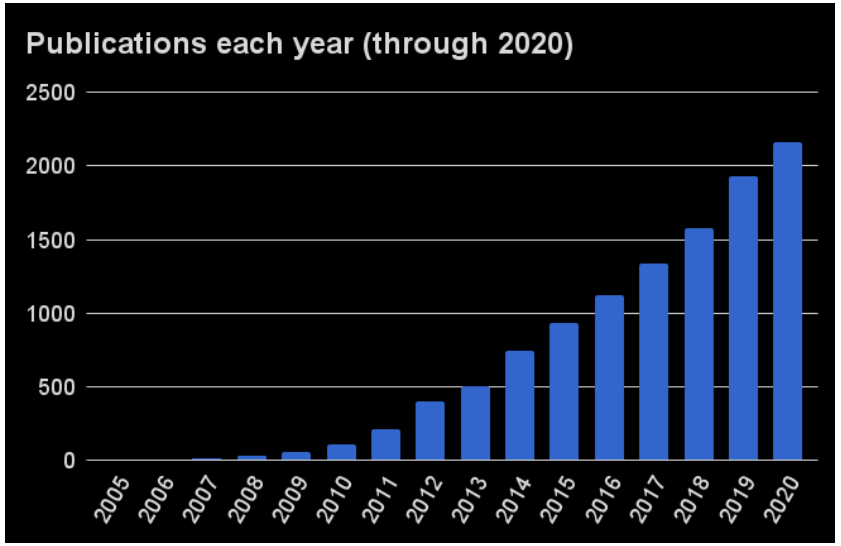
Drug response predictor



KNIGHT CANCER Institute OHSU

Galaxy is Widely Cited



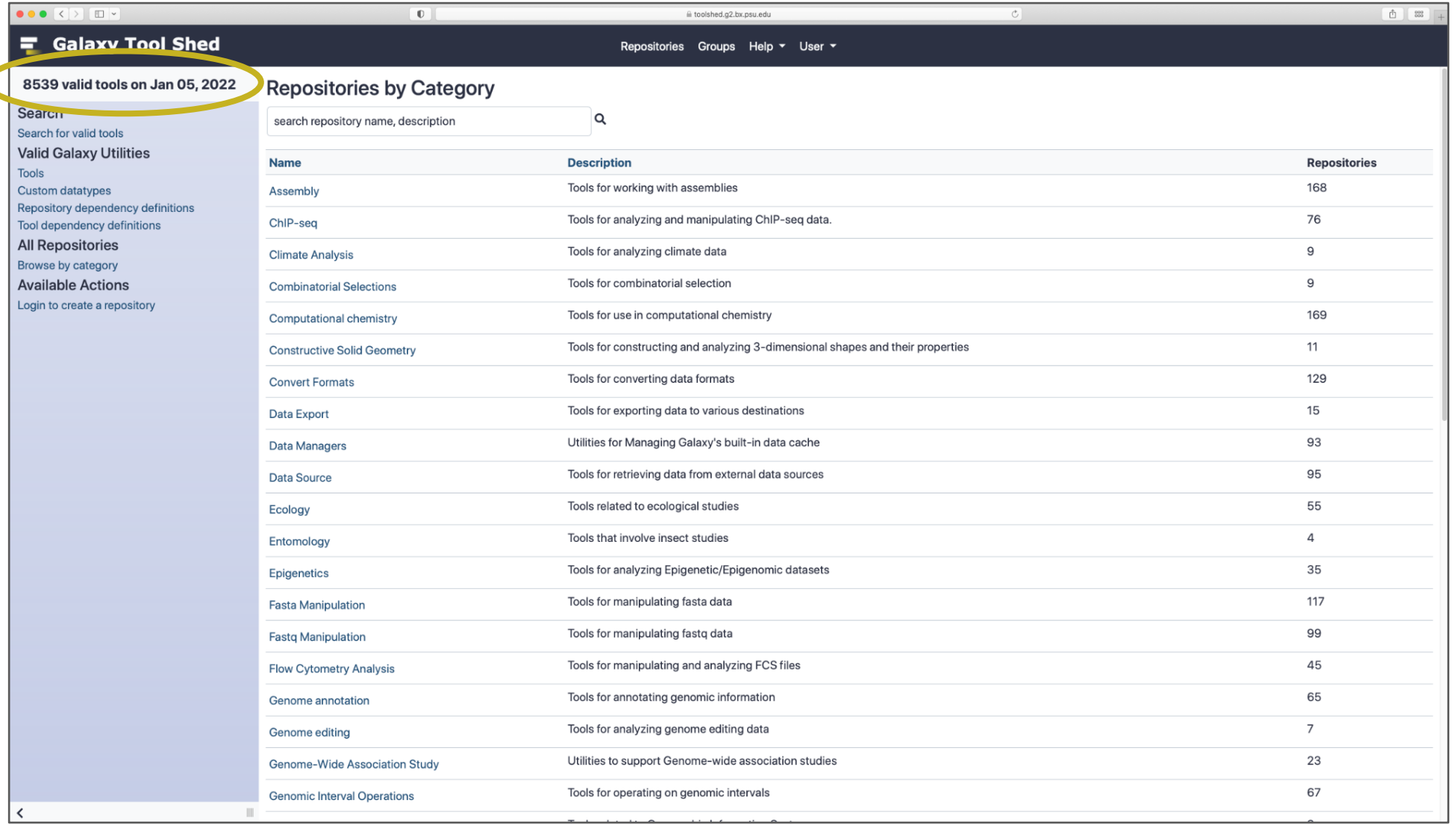


Year	# Pubs	Methods	UsePublic	Workbench	UseMain	RefPublic	UseLocal	Tools	IsGalaxy	Reproducibility	Cloud	Other	Shared	Unknown	HowTo	Project	Education	Visualization	UseCloud
2005	1															1			
2006	4			3											1				
2007	12	2		7							1				2	2			
2008	32	15		12	1				2					2		1		1	
2009	53	27		18				3	2			1	1	4	1	1			
2010	107	50		36	1			1	5			1	1	7	4	5			
2011	205	93		69	1			8	16	6	3		8	3	4	6	1	1	
2012	398	197	1	128	3		3	30	15	7	14	9	12	10	12	10	1	2	
2013	506	264	16	149	92	12	29	37	29	9	22	9	22	13	7	6	3	3	2
2014	741	331	60	226	98	30	43	67	48	25	40	39	23	7	12	7	2	8	1
2015	930	474	140	233	116	52	58	68	49	26	48	33	23	14	8	11	1	7	3
2016	1125	575	213	246	115	116	73	73	49	46	37	47	19	13	20	7	2	9	4
2017	1334	760	279	239	138	110	98	76	72	72	35	26	24	23	8	6	5	7	1
2018	1576	1027	375	235	185	137	109	51	46	76	27	18	26	25	16	8	6	6	3
2019	1933	1315	585	237	205	168	140	63	51	85	22	17	25	5	12	6	12	4	2
2020	2164	1549	704	217	265	153	134	79	60	64	11	17	21	4	10	11	27	10	3
2021	575	412	222	63	57	43	24	22	19	7	4	2	14	2	5	8	8		
2022	1					1													
Total	11697	7091	2595	2118	1277	822	711	578	463	423	264	219	219	132	122	96	68	58	19

11697

KNIGHT CANCER Institute
OHSU

Galaxy Toolshed



Name	Description	Repositories
Assembly	Tools for working with assemblies	168
ChIP-seq	Tools for analyzing and manipulating ChIP-seq data.	76
Climate Analysis	Tools for analyzing climate data	9
Combinatorial Selections	Tools for combinatorial selection	9
Computational chemistry	Tools for use in computational chemistry	169
Constructive Solid Geometry	Tools for constructing and analyzing 3-dimensional shapes and their properties	11
Convert Formats	Tools for converting data formats	129
Data Export	Tools for exporting data to various destinations	15
Data Managers	Utilities for Managing Galaxy's built-in data cache	93
Data Source	Tools for retrieving data from external data sources	95
Ecology	Tools related to ecological studies	55
Entomology	Tools that involve insect studies	4
Epigenetics	Tools for analyzing Epigenetic/Epigenomic datasets	35
Fasta Manipulation	Tools for manipulating fasta data	117
Fastq Manipulation	Tools for manipulating fastq data	99
Flow Cytometry Analysis	Tools for manipulating and analyzing FCS files	45
Genome annotation	Tools for annotating genomic information	65
Genome editing	Tools for analyzing genome editing data	7
Genome-Wide Association Study	Utilities to support Genome-wide association studies	23
Genomic Interval Operations	Tools for operating on genomic intervals	67

https://toolshed.g2.bx.psu.edu/

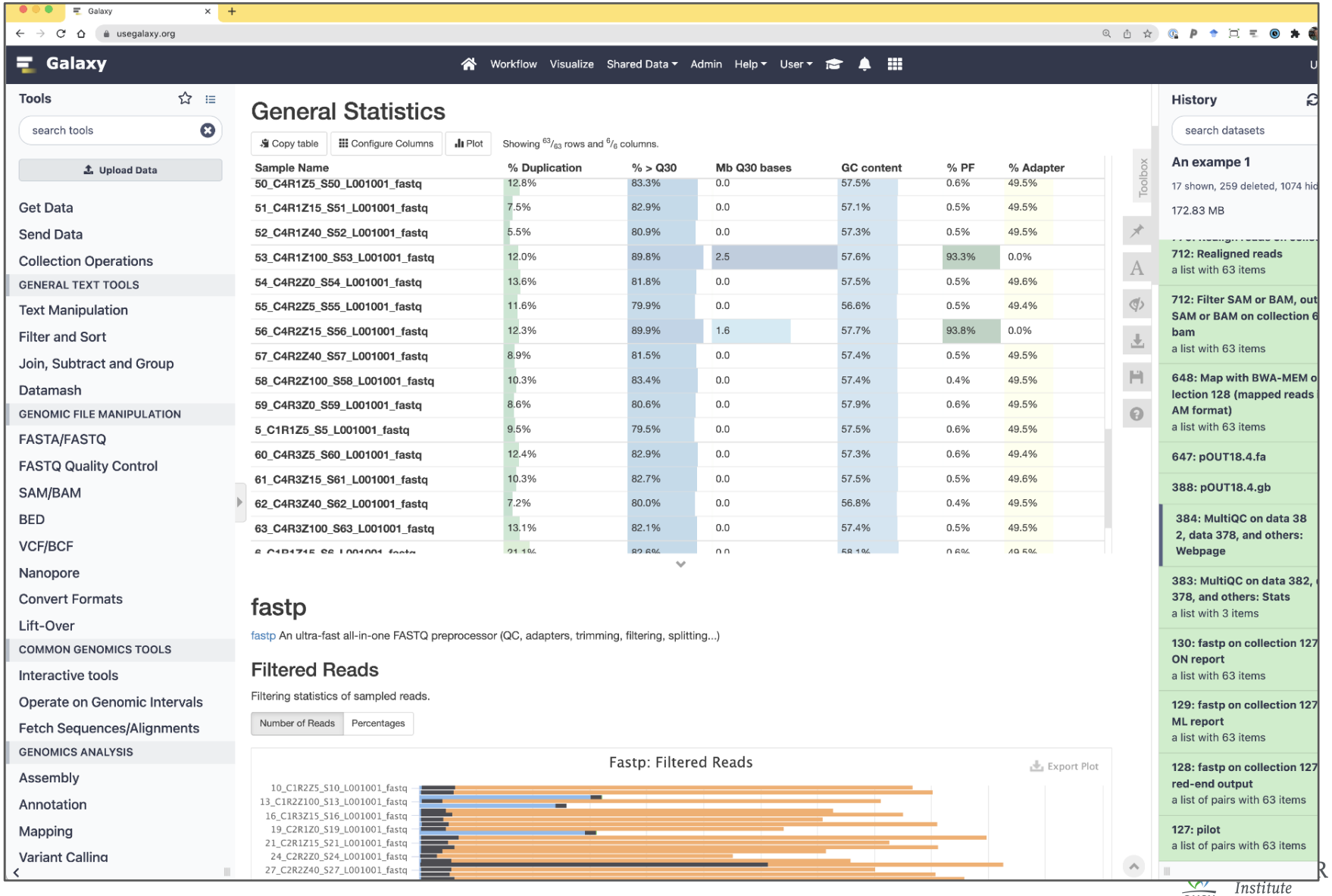
The Galaxy User Interface Makes Everything Accessible

Accessible yet powerful

All analysis functionality is available from the Web

Many recent advances that increase the power and flexibility of the user interface:

- Collections
- Workflow reports

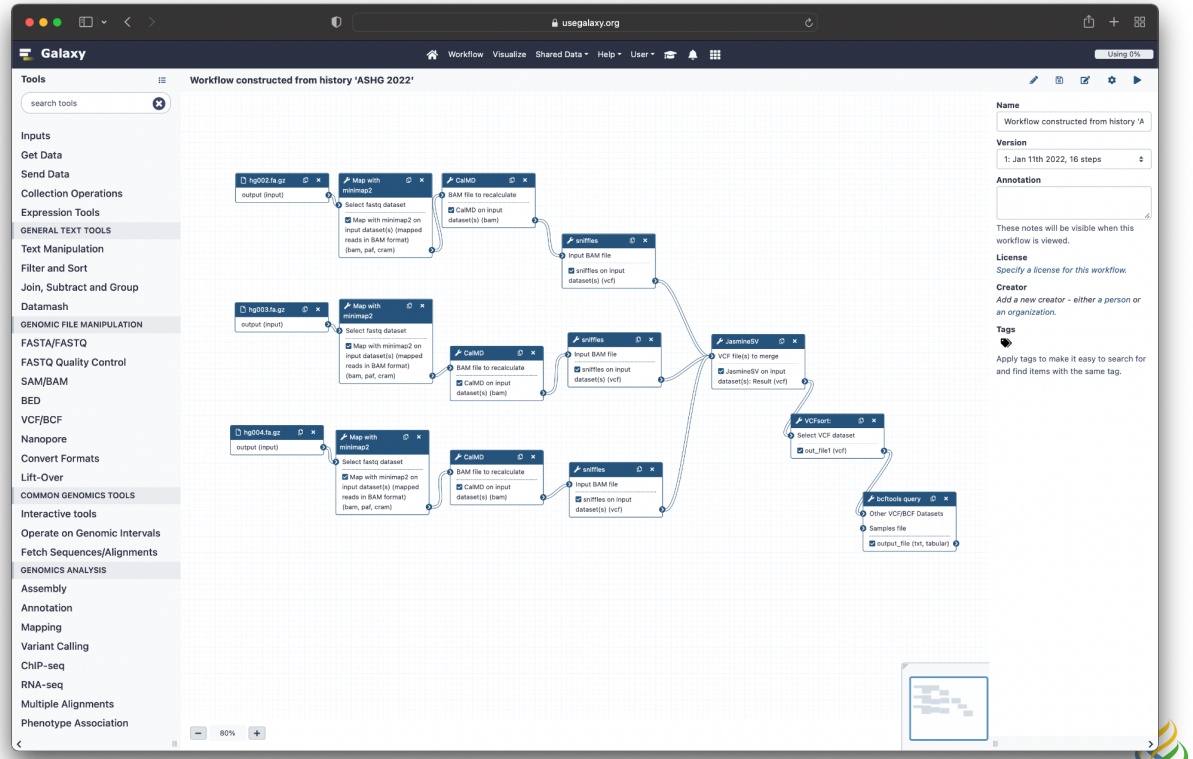


Reproducibility is Central to the Galaxy Framework

The software framework, tools, and utilities are all open-source

All parameters are recorded for all analyses and stored in Galaxy's database

This is true for both individual tool executions and multi-step workflows

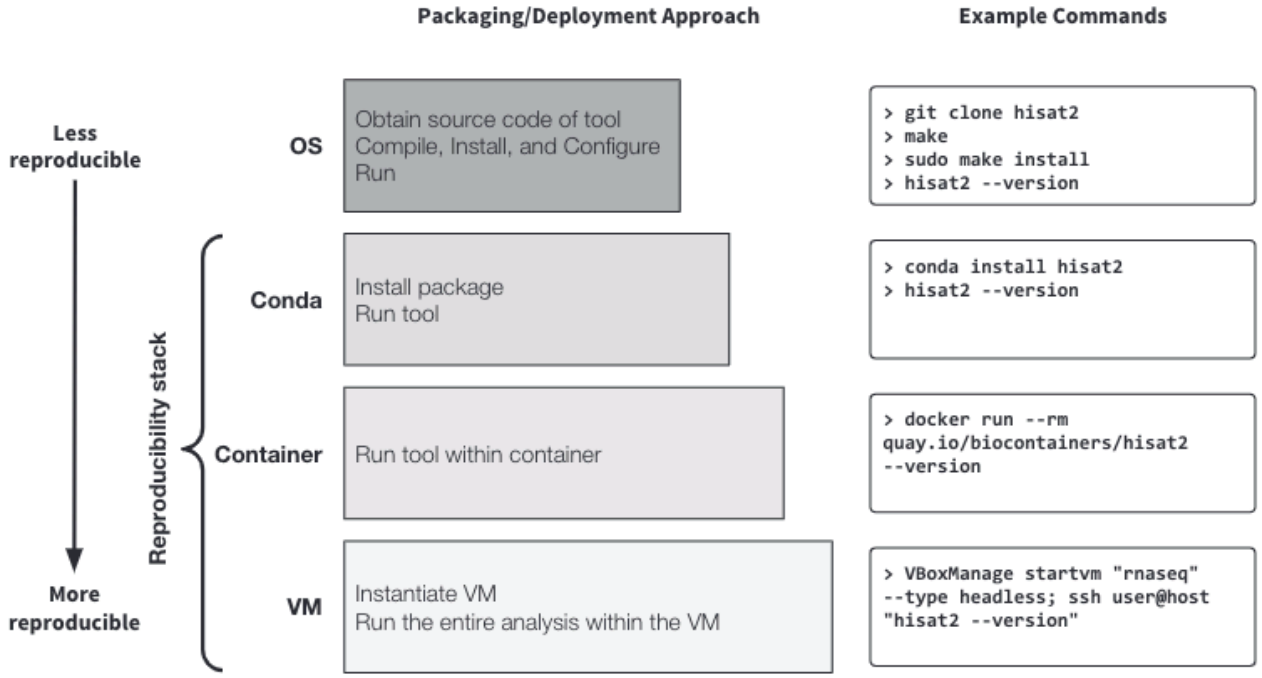


KNIGHT CANCER Institute
OHSU

The Galaxy Reproducibility Stack

Layers of reproducibility built on virtualization technologies

Used for automated dependency resolution in Galaxy



Grüning et al. (2018) *Cell Systems*

KNIGHT CANCER Institute
OHSU

There are Many Ways to Communicate in Galaxy

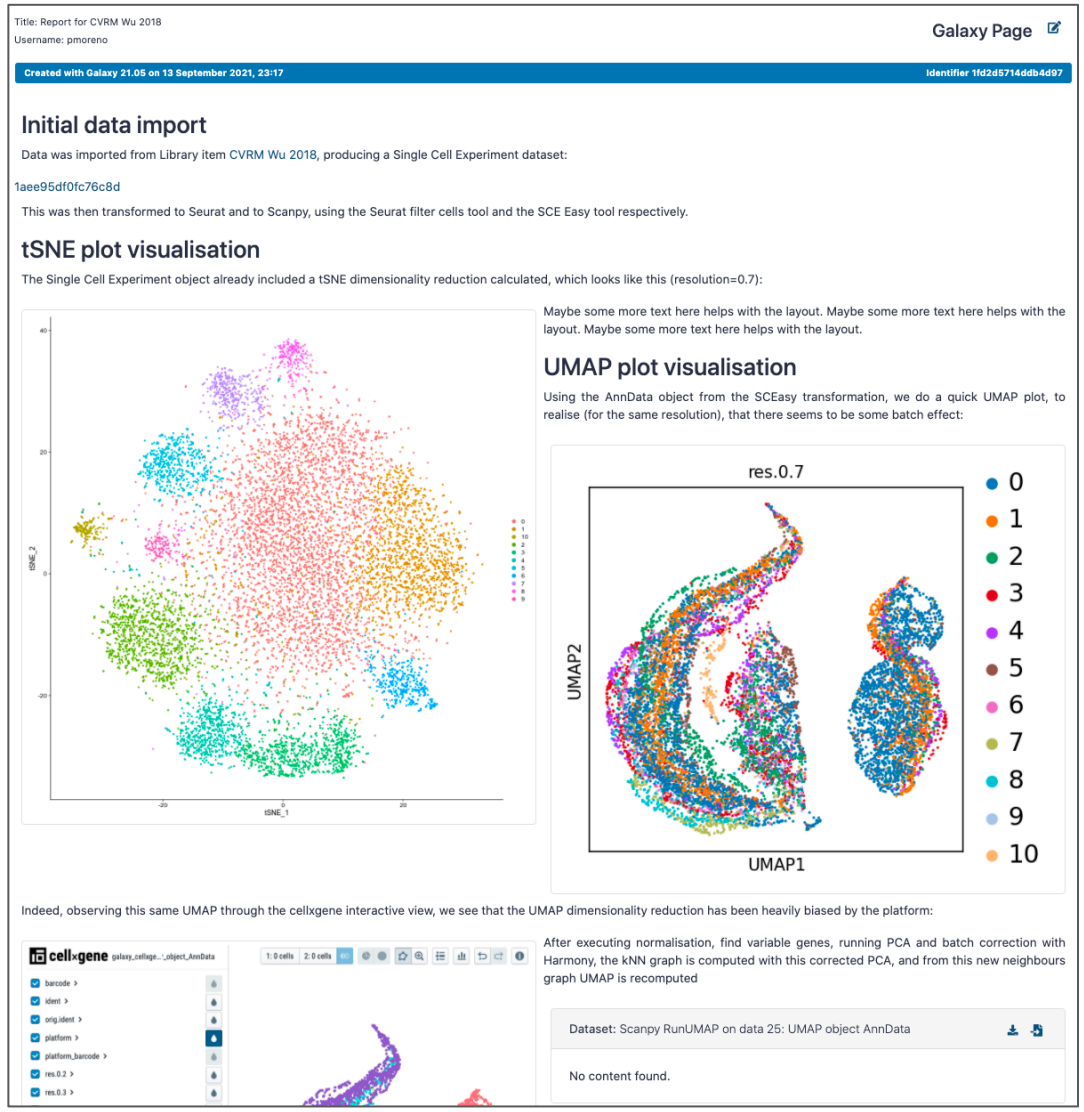
All Galaxy tools, histories, workflows, and visualizations can be shared via a web link

- Can share with everyone or particular users
- Can include in publications

Workflow reports make it possible to generate summaries of complex analyses

Importing/exporting:

- Galaxy histories can be imported/exported for archival
- Galaxy workflows can be exported and shared for archival
- Increasing support for Common Workflow Language (CWL)

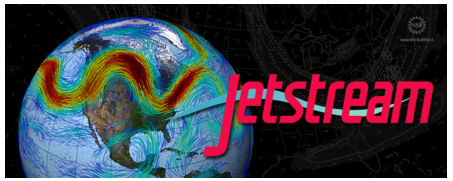


KNIGHT CANCER Institute
OHSU

Four ways to use Galaxy

1. Public servers such as http://usegalaxy.org and many more at https://galaxyproject.org/use/
2. Your laptop or local computer
3. Install locally with many compute environments

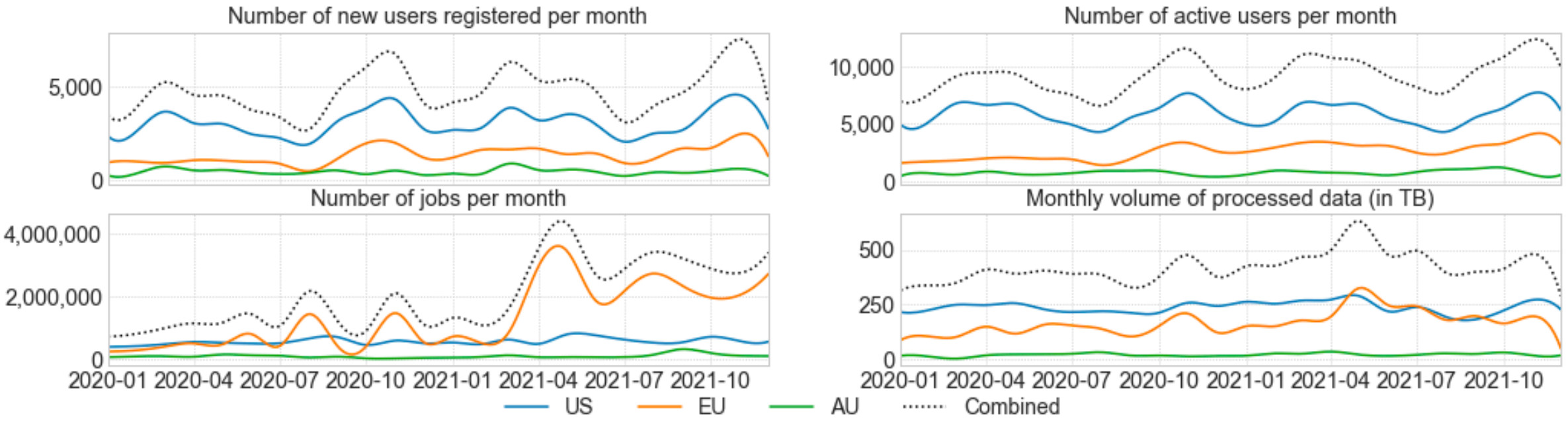
4. Deploy on a cloud



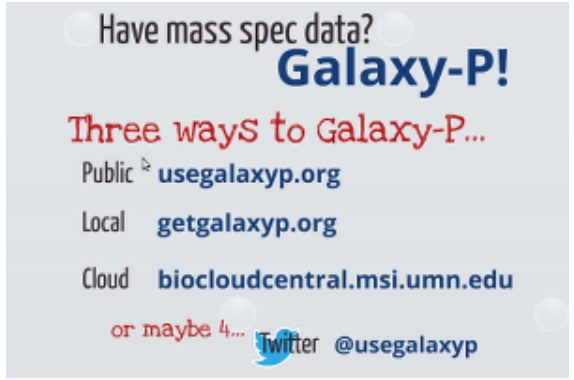
The universe has many Galaxies

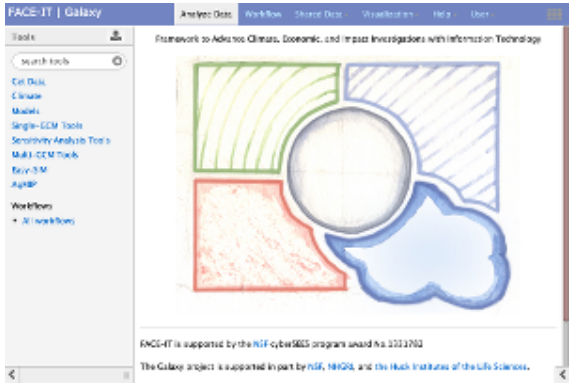
usegalaxy.*: the big three

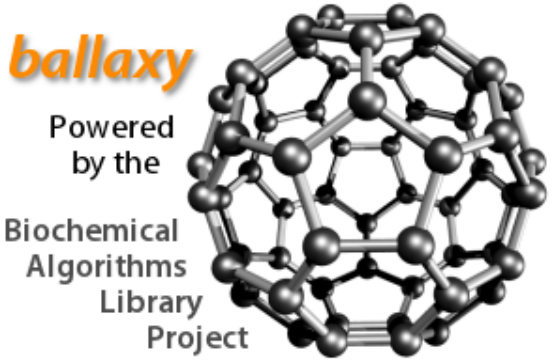
Usage of Three Main Public Galaxy Servers

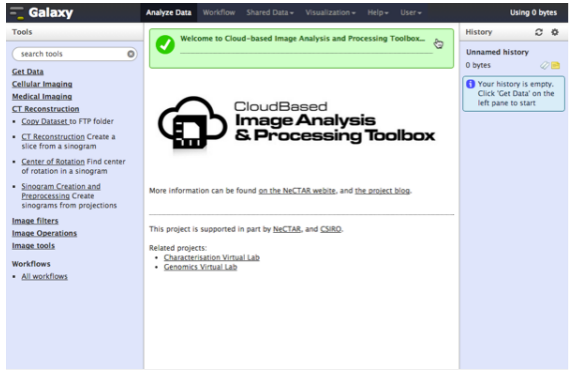


KNIGHT CANCER Institute
OHSU









Climate Change

Proteomics
Metabolomics
Drug Discovery
Cosmology
Image Analysis
Flow Cytometry

Natural Language

https://galaxyproject.org/use/

KNIGHT CANCER Institute
OHSU

Galaxy meets key scientific needs

Scientists get:

- Web-based GUI with thousands of tools to create complex analysis workflows
- Single, shared platform for computational and non-computational users

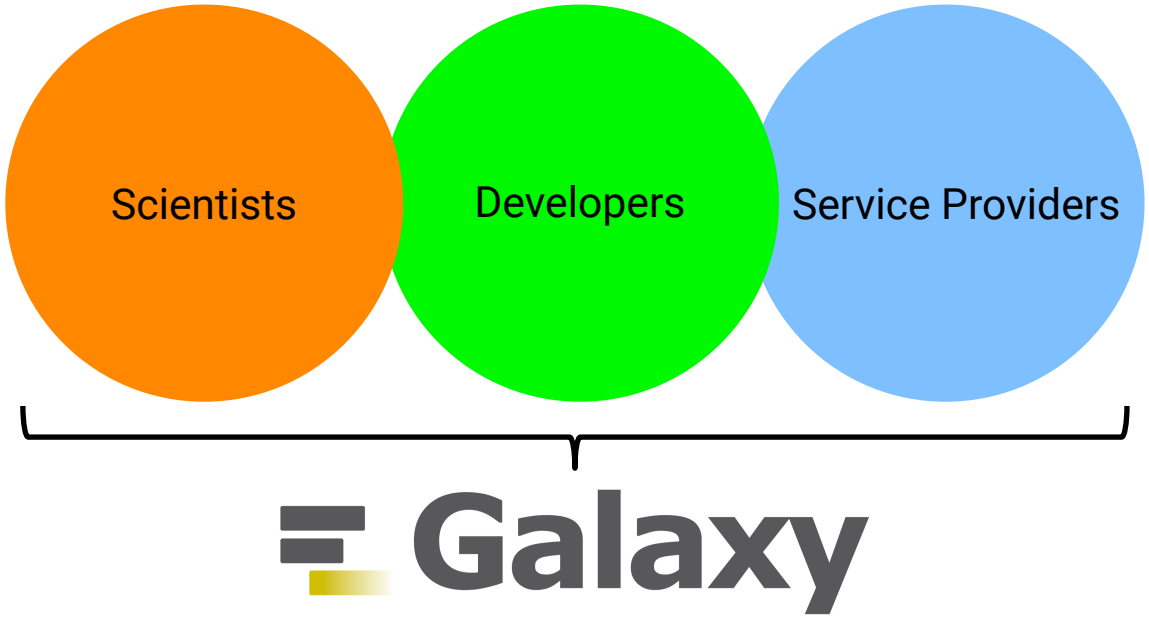
Developers get:

- Access to thousands of users
- Easy to connect new tools/visualizations with other tools/ visualizations

Service providers get:

- Efficient use of large hardware allocations via API for local service integration
- Automated tool/dependency management

Everyone gets community support, training and advice



OHSU

Galaxy Interactive Tools

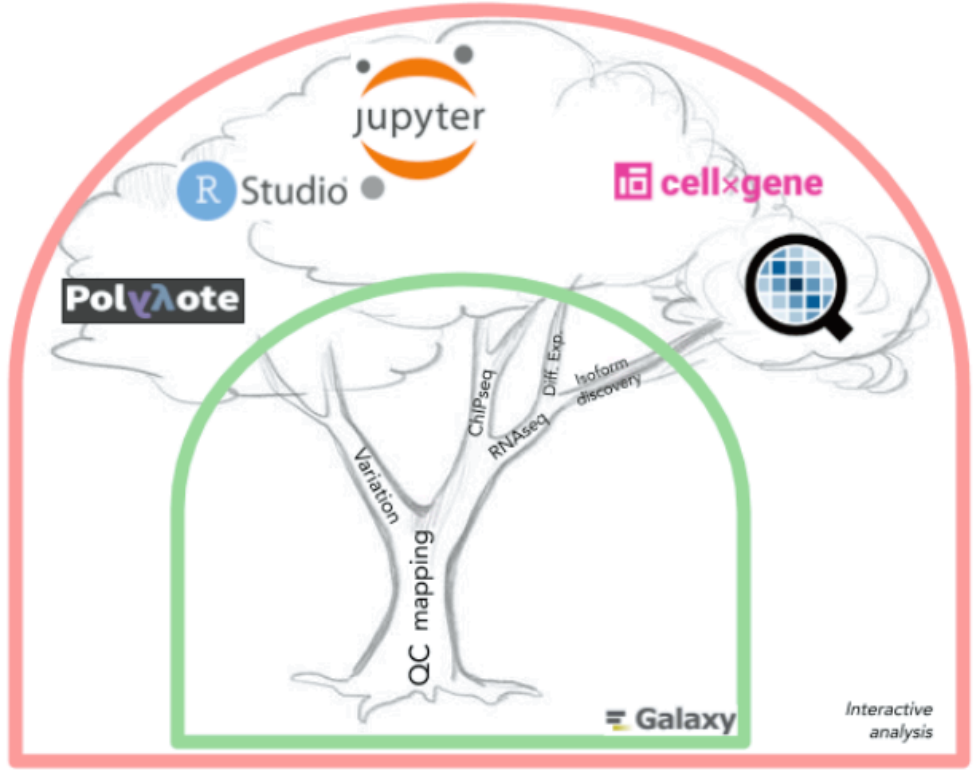
Makes interactive web tools available in Galaxy

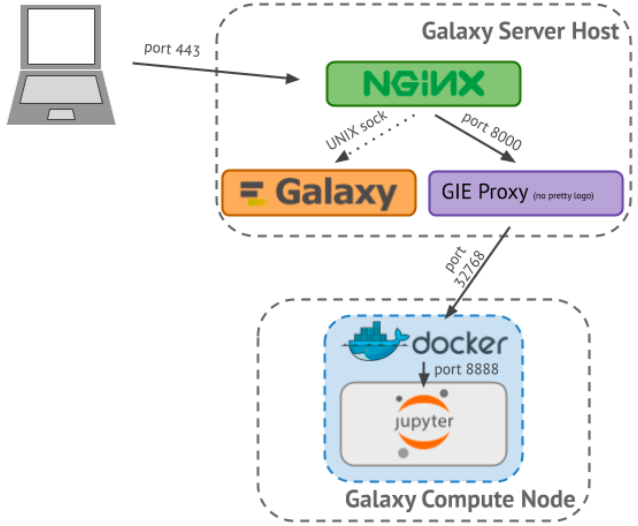
Some examples:

- Jupyter and RStudio for programmatic analyses
- CellxGene for dynamic visualization

Technical approach

- Software containers are managed by Galaxy
- Galaxy datasets can be imported and exported
- Can be included in workflows for QC and dynamic interactions





KNIGHT CANCER Institute
OHSU

Galaxy Dataset Collections

Today's analyses often involve many samples

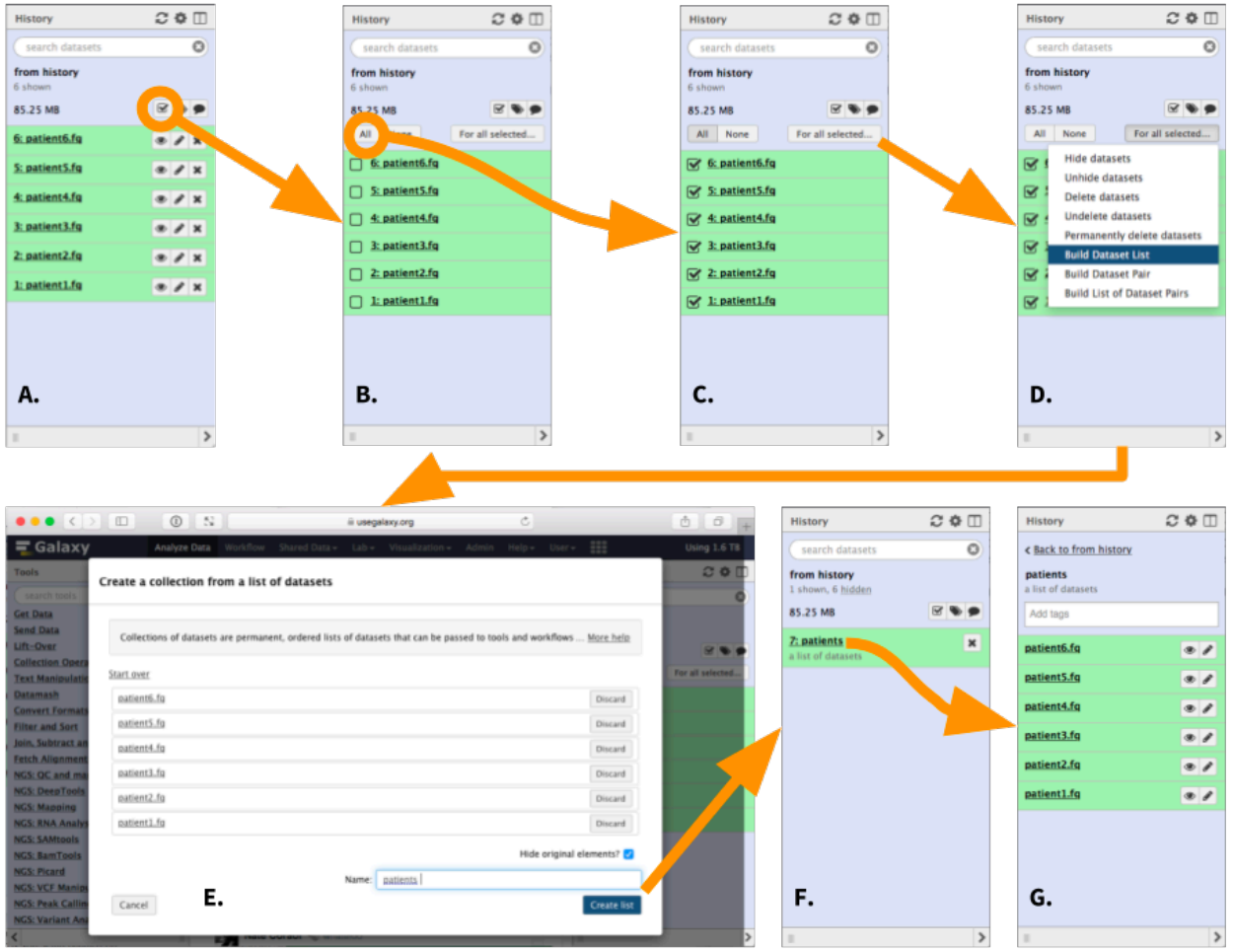
- Want to process all samples in the same way

Collections enable processing of datasets in the same way (map)

- Galaxy transparently runs a tool on each dataset in a collection

Also can combine many dataset into one (reduce)

- Often this is datatype/analysis specific



KNIGHT CANCER Institute
OHSU

Galaxy as part of NIH Data Science Infrastructure

NHGRI

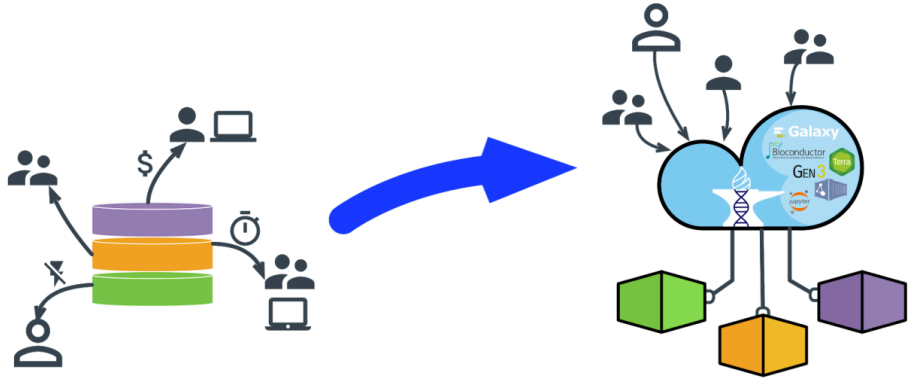
- A Component of the AnVIL, the NHGRI Data Commons

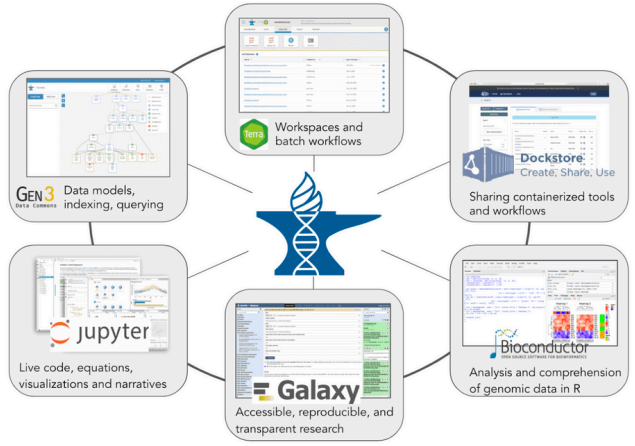
NCI

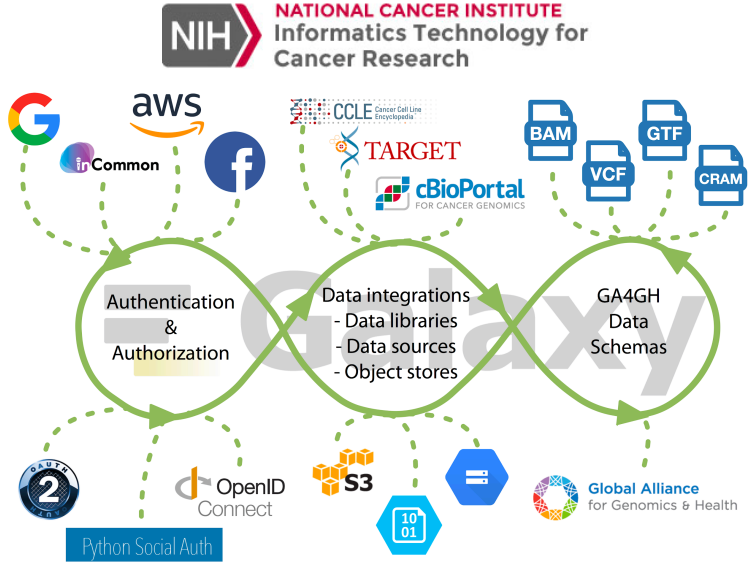
- Developing connections to the NCI Cancer Research Data Commons

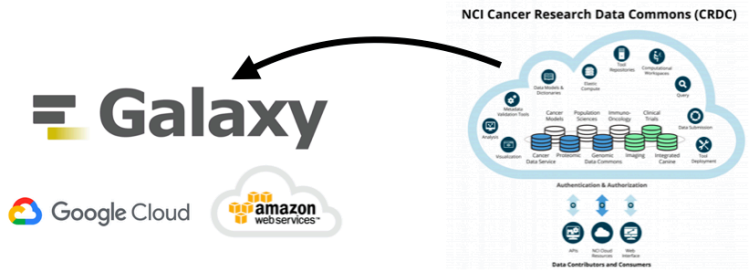
Key Advances

- Data-local computing is possible when Galaxy runs on commercial clouds, so no egress fees
- Substantial use of software containers and cloud service for deployment and tool execution



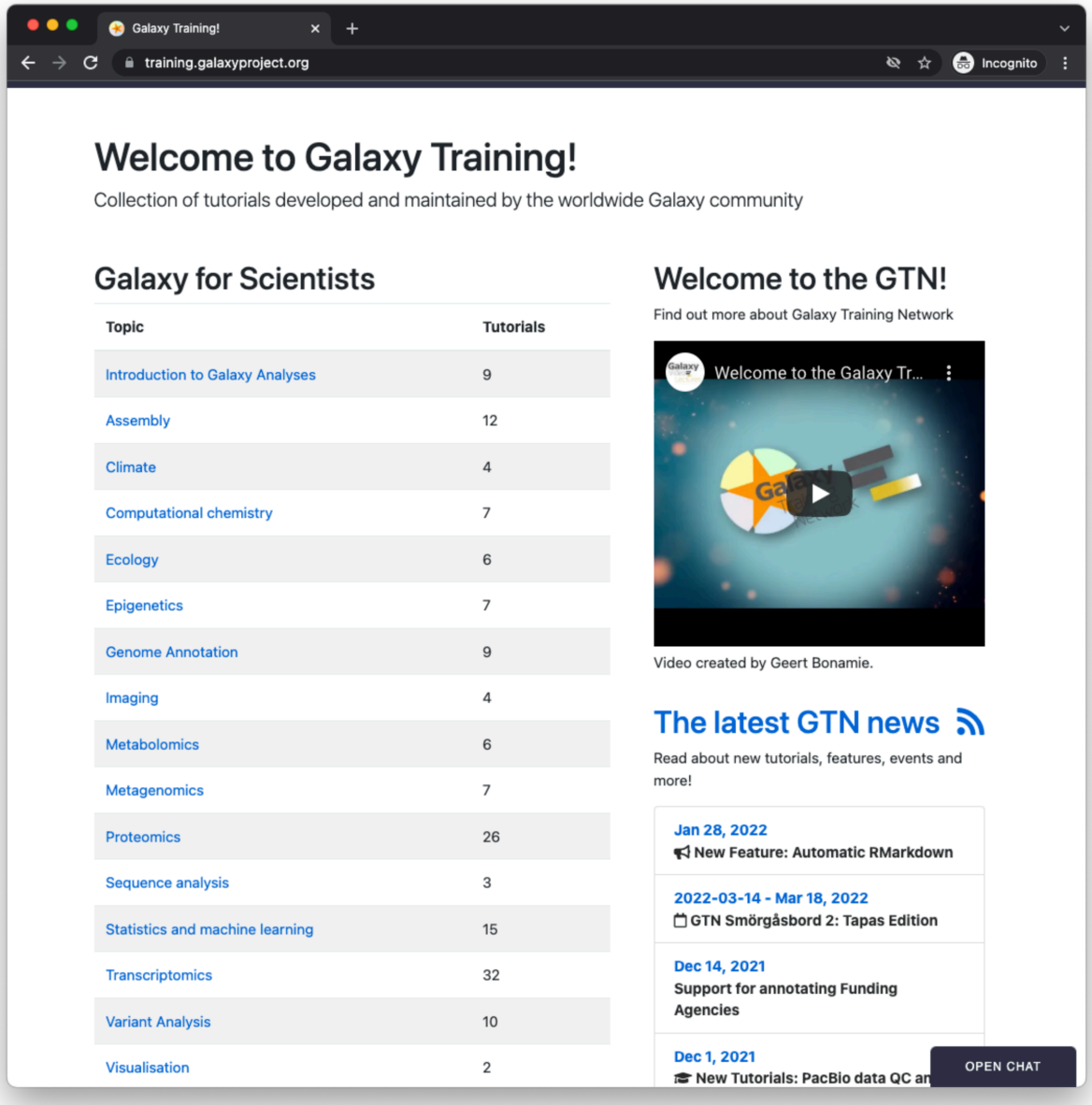


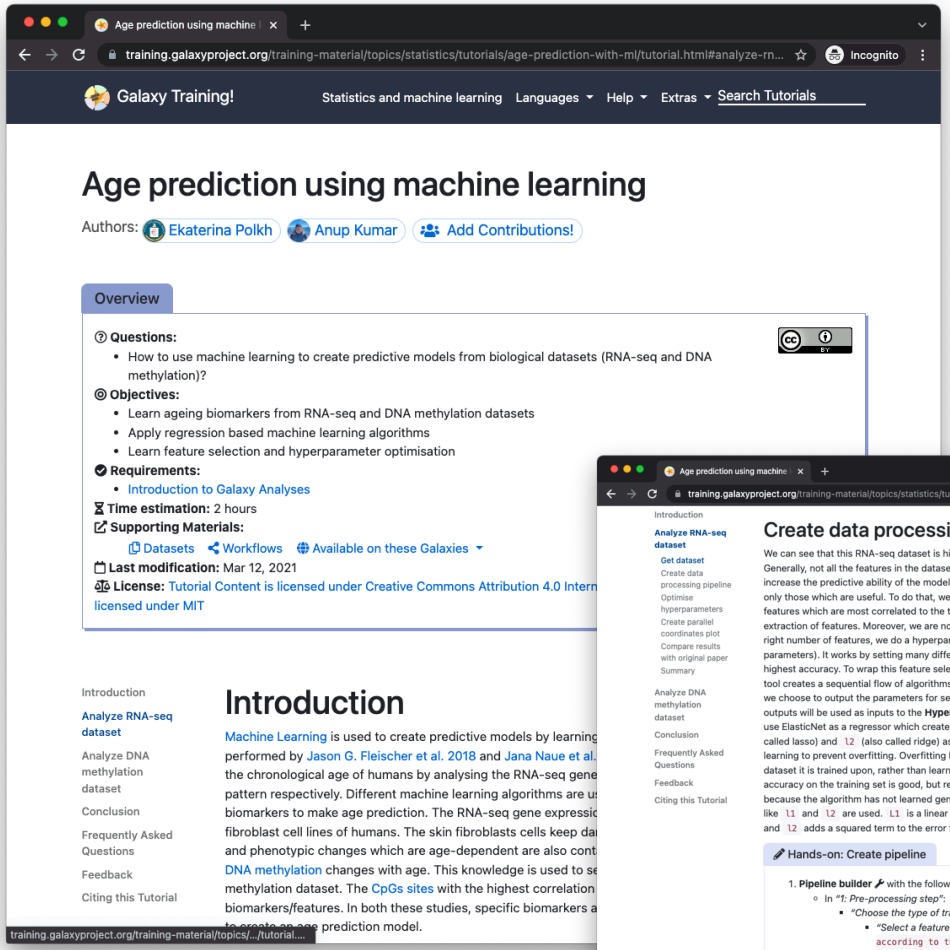




KNIGHT CANCER Institute
OHSU

The Galaxy Training Network is a Fantastic Resource





OHSU Institute

The Gallantries, Galaxy Training Network &
Galaxy Community are happy to announce

GTN Smörgåsbord 2

14-18 March 2022

Save the date!
bit.ly/smorgasbord2

Join a **free**, **global**, week-long Galaxy Training event covering everything from RNA-Seq, Single Cell, Proteomics, SARS-CoV-2 *and more!* This year will include Galaxy Admin Training.

@gxytraining @Gallantries_EU

With the support of the European Union

Galaxy COMMUNITY HUB | Use | Learn | Community | Deploy & Develop | Support | @jxtx | Search Galaxy | Edit

← Back to Events

Key Dates | Conduct | Abstracts | Schedule | Training | CoFest | Sponsors | Travel | Register | Childcare | Organizers

2022 Galaxy Community Conference (GCC2022)

July 16-23, 2022

University of Minnesota, Twin Cities

Minneapolis, Minnesota, United States

#UseGalaxy2022

https://galaxyproject.org/events/gcc2022/

@galaxyproject

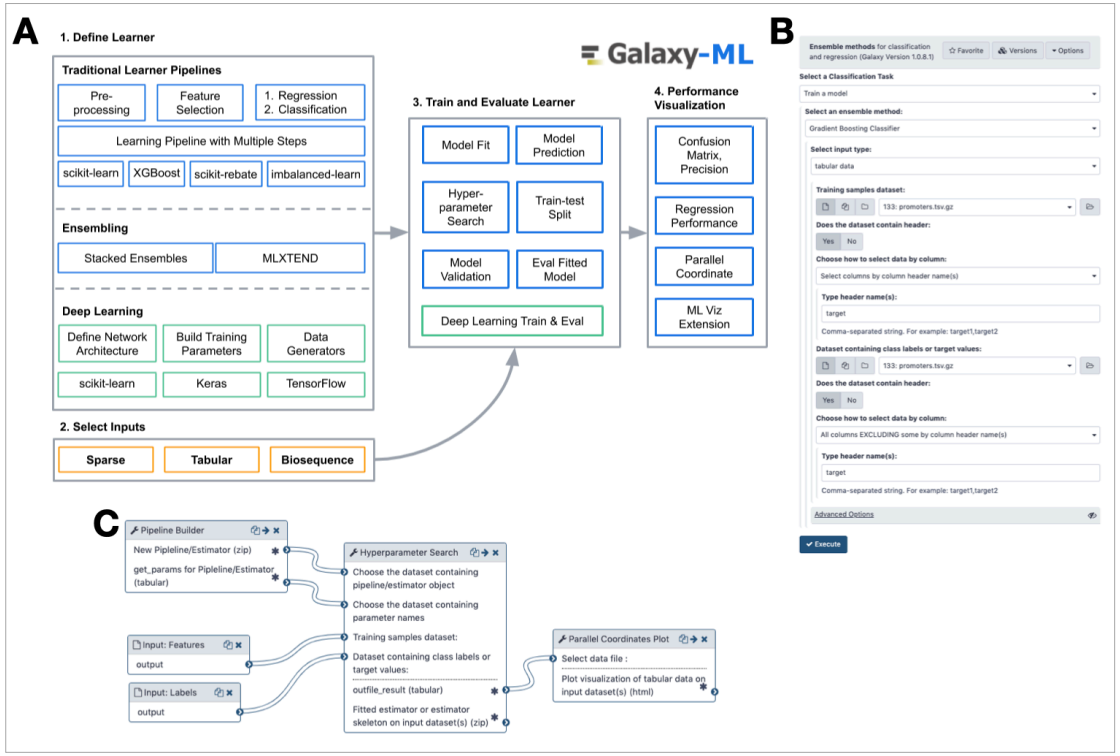
Outline

The Galaxy Platform

Machine Learning Applications for Cancer

An Interactive Hub for Multiplex Tissue Image Analysis

Galaxy-ML: A general purpose machine learning toolkit for Galaxy



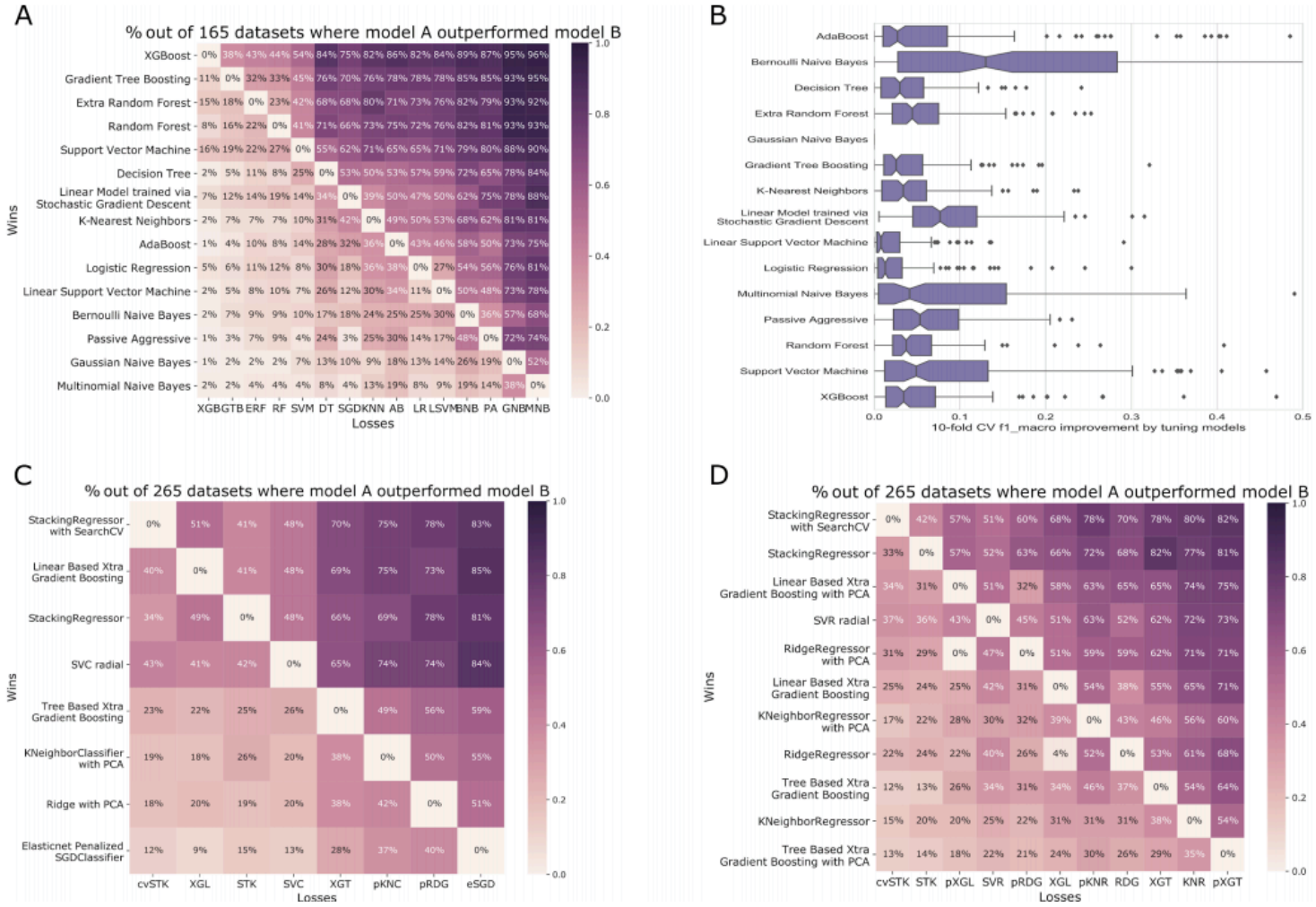
Software Library	Applications
Scikit-learn [18]	Various approaches for preprocessing, modeling, ensembling, and evaluation
Scikit-rebate [19]	Feature selection
Imbalanced-learn [20]	Approaches for working with imbalanced datasets
XGBoost [21]	Modeling using high-performance gradient boosting with decision trees or linear models
Keras [22]	Modeling using deep learning
Mlxtend [23]	Modeling using meta-ensembles
LightGBM [28]	Modeling using gradient boosting with the LightGBM algorithm

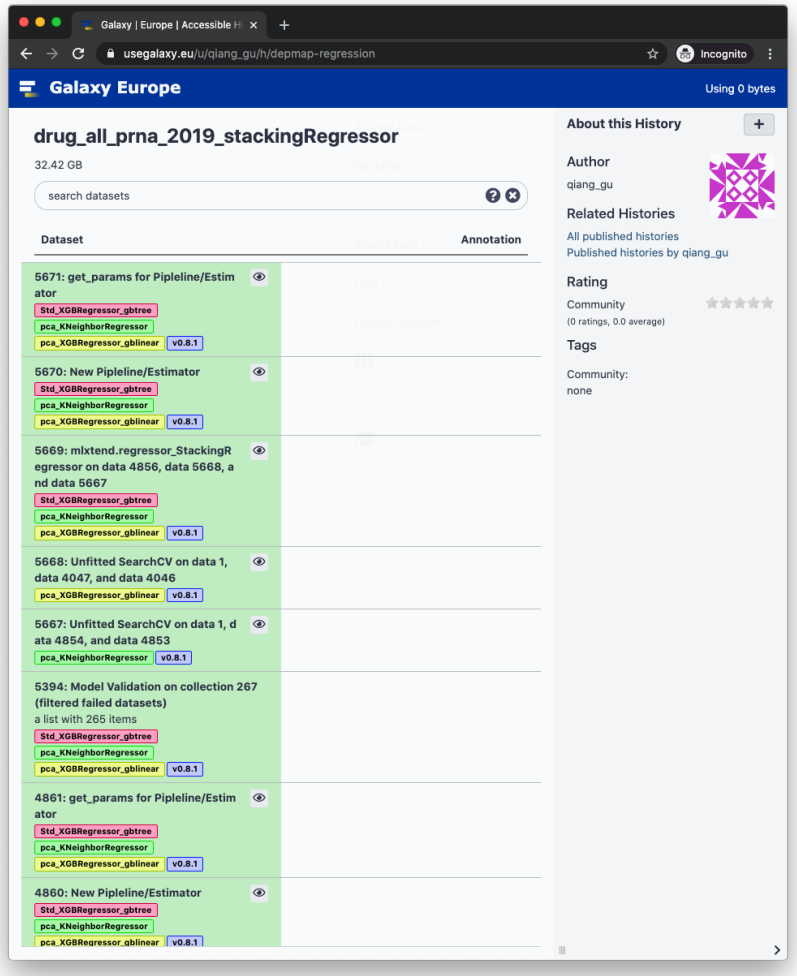
https://doi.org/10.1371/journal.pcbi.1009014.t001

Galaxy-ML: An accessible, reproducible, and scalable machine learning toolkit for biomedicine
Gu et al. (2021) *PLOS Comp Biology*. https://doi.org/10.1371/journal.pcbi.1009014

KNIGHT CANCER Institute
OHSU

Scalable and Reproducible Machine Learning

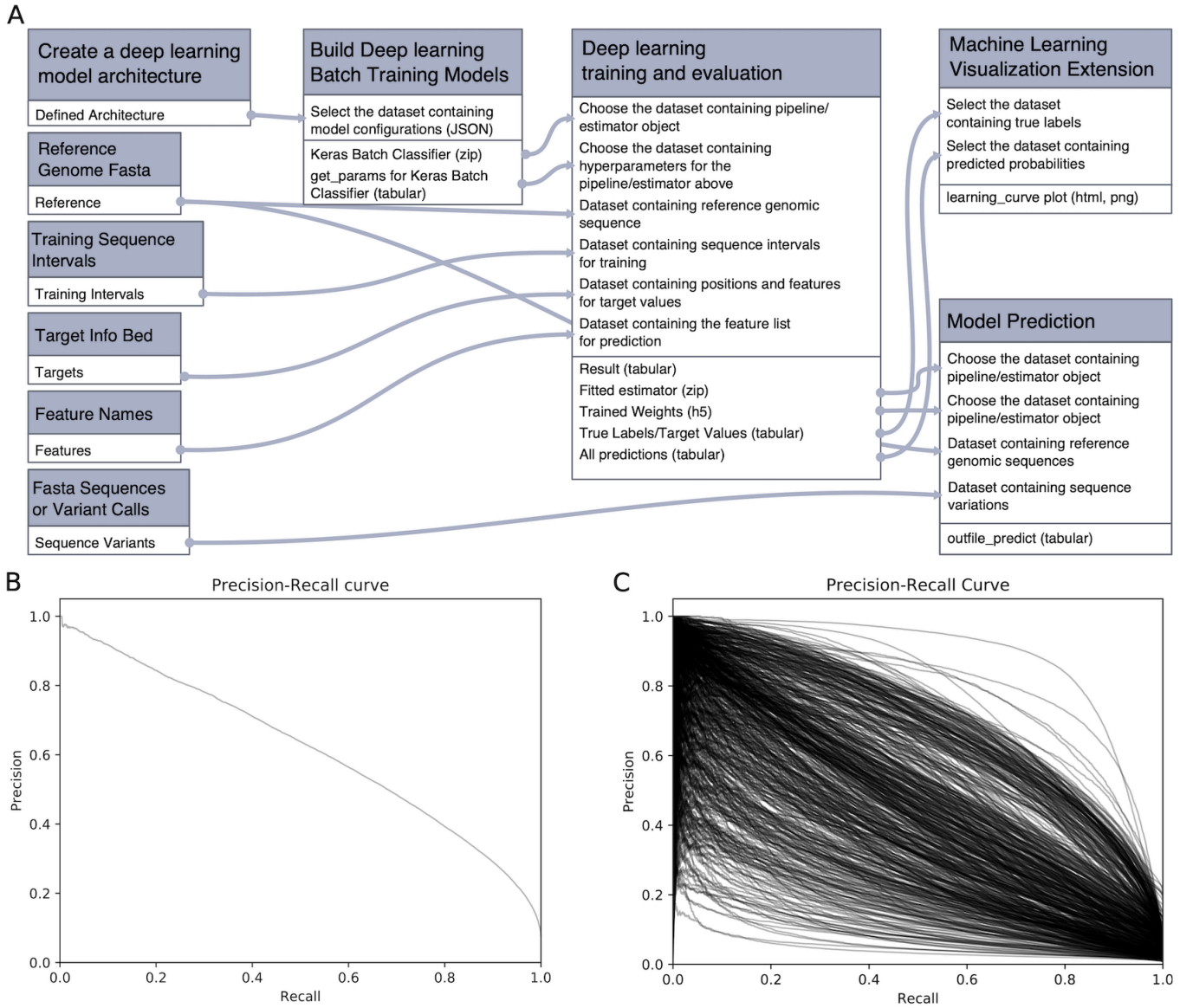




- Thousands of model trained automatically on hundreds of datasets
- Collections are used extensively

KNIGHT CANCER Institute
OHSU

Deep Learning in Galaxy

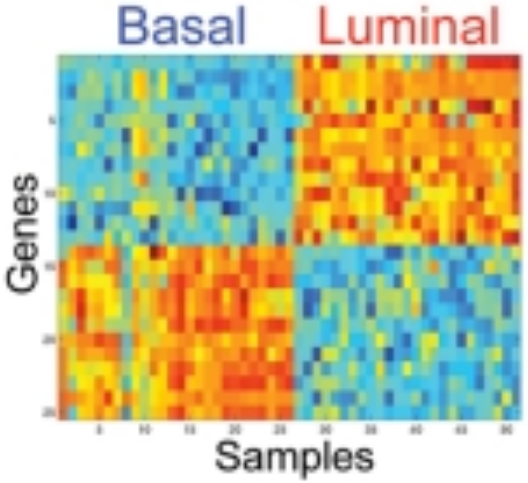


KNIGHT CANCER Institute
OHSU

Building and Evaluating Transcriptional Signatures

A single gene or group of genes with a unique pattern of expression that occurs as a result of normal function, perturbation, or disease

DNA alterations alone not sufficient for precision cancer therapy



Four approaches to building signatures from bulk TCGA RNA-seq

- Single genes / differential gene expression
- Gene sets + GSVA (Hänzelmann et al., 2013)
- VIPER (Alvarez et al., 2016)
- Learned signatures resulting from DNA alterations (Way and Greene, 2018)

Evaluation framework is independent of signature type

- Metric #1: Accuracy on TCGA and transfer to new cohorts
- Metric #2: Association with response to therapy

Learning Transcriptional Signatures Associated with DNA Alterations

Use machine learning to identify the transcriptional signature that a DNA alteration induces

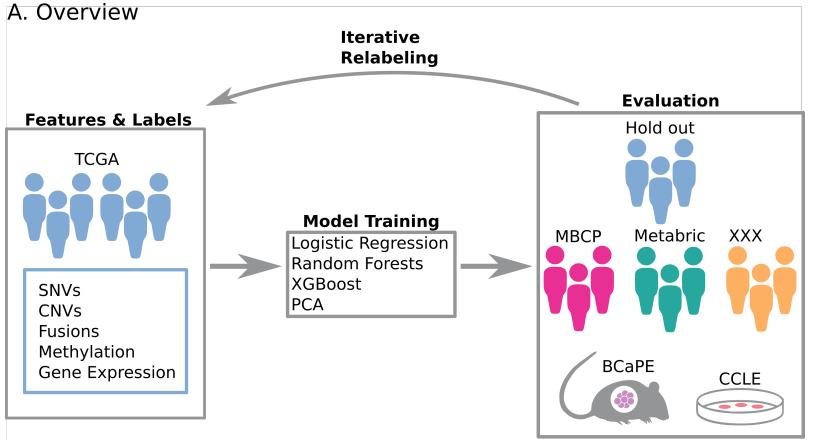
- *"What does a TP53 mutation or CDKN2A loss look like at the transcriptional level?"*
- Inputs: RNA-seq, Labels/predictions: DNA alterations

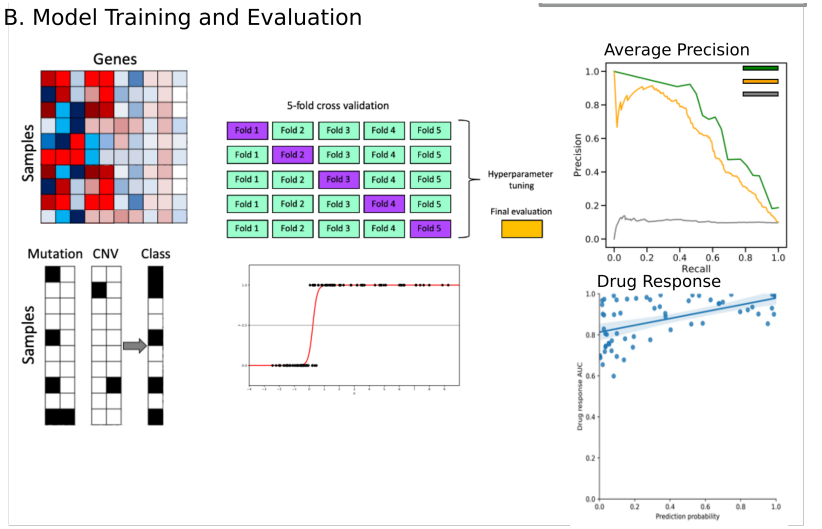
Very useful:

- Transcriptional state is dynamic and more closely reflects tumor activity than DNA (e.g., "hidden responders")
- Interpretable

Can learn using large cohorts without drug response data

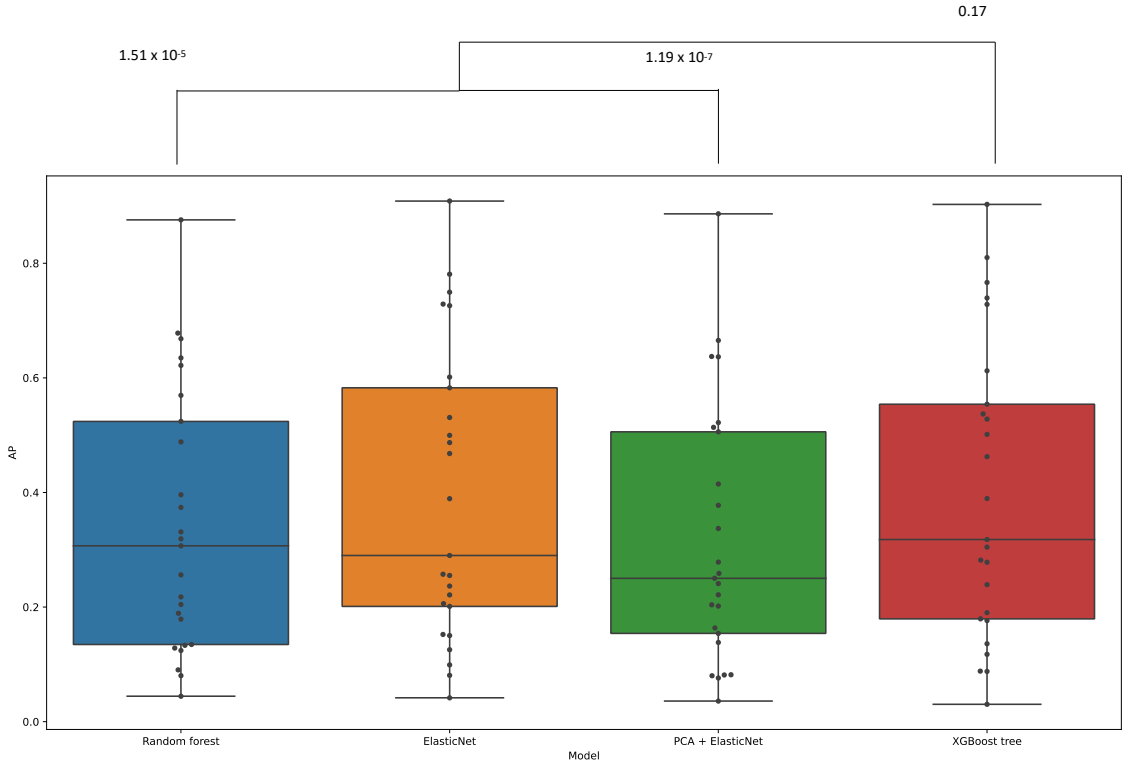
- self/unsupervised learning very likely to be highly useful in the future

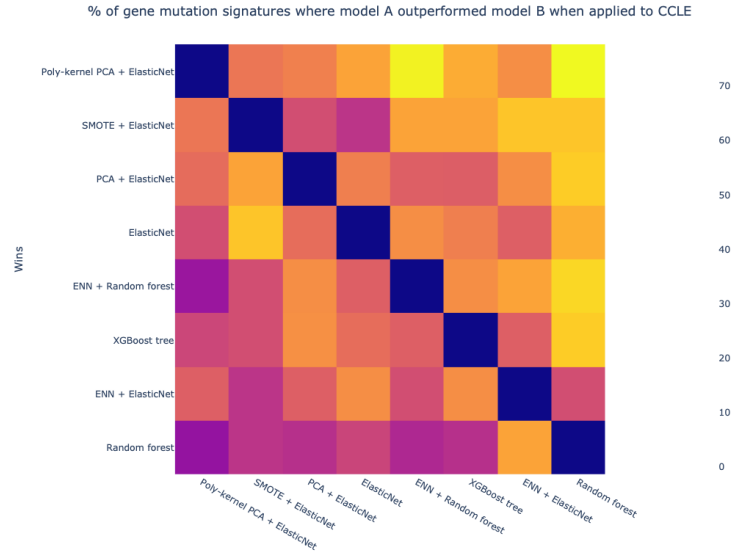


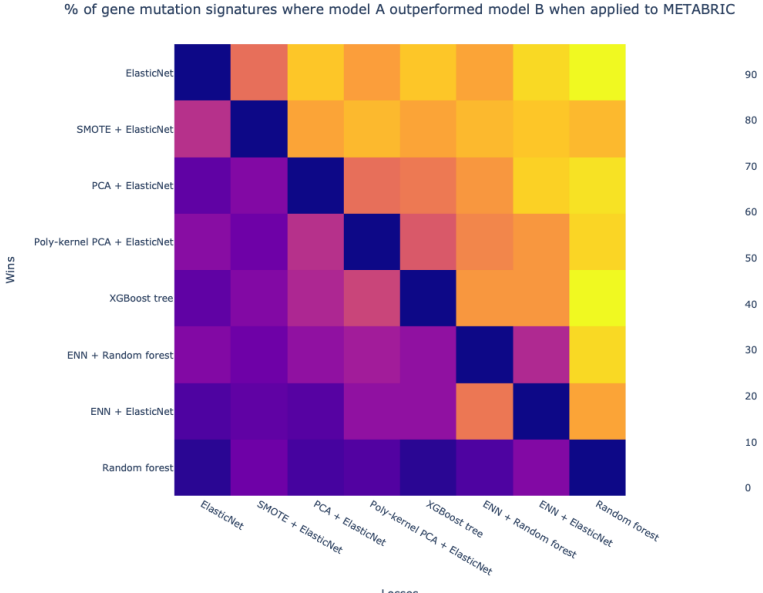


KNIGHT CANCER Institute
OHSU

Multiple Modeling Approaches Work and Models Transfer to Other Cohorts







KNIGHT CANCER Institute

Learned Signatures are associated with Drug Response in Preclinical Cohorts

Associations are consistent across preclinical cohorts (cell lines + pdxs)

Learned signatures outperform others transcriptional signatures for many drugs

KNIGHT CANCER Institute
OHSU

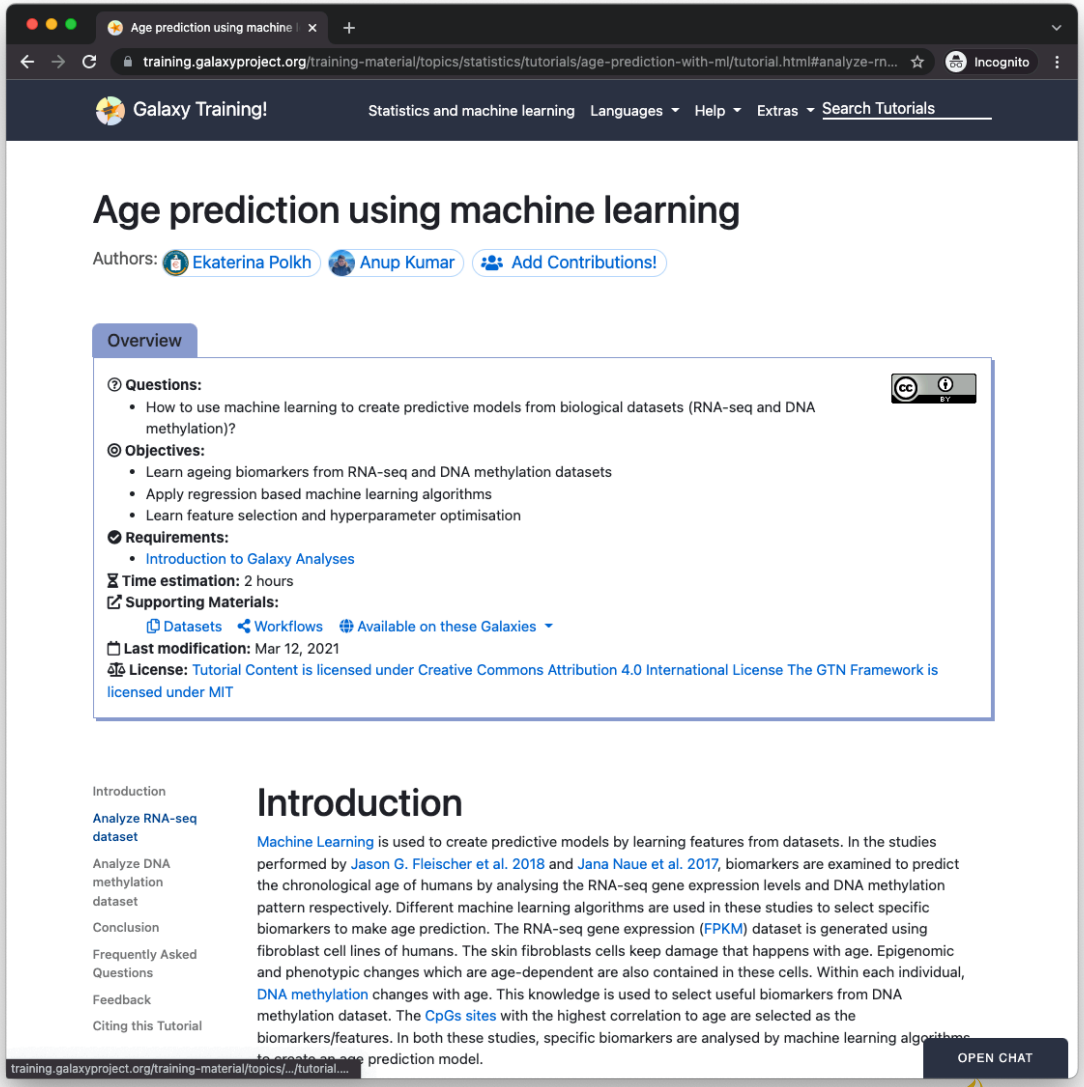
The Power of Adding Tools to the Galaxy Ecosystem

When a tool is integrated into Galaxy, it can be connected to all other Galaxy tools

For ML tools, this means that they can be used after primary analyses to enable end-to-end analysis

- Primary analysis: quantify features from data (e.g. variants or gene expression levels)
- ML analysis: use extracted features to build predictive model

Example: predicting age from RNA-seq data and identifying aging biomarkers



KNIGHT CANCER Institute
OHSU

Outline

The Galaxy Platform

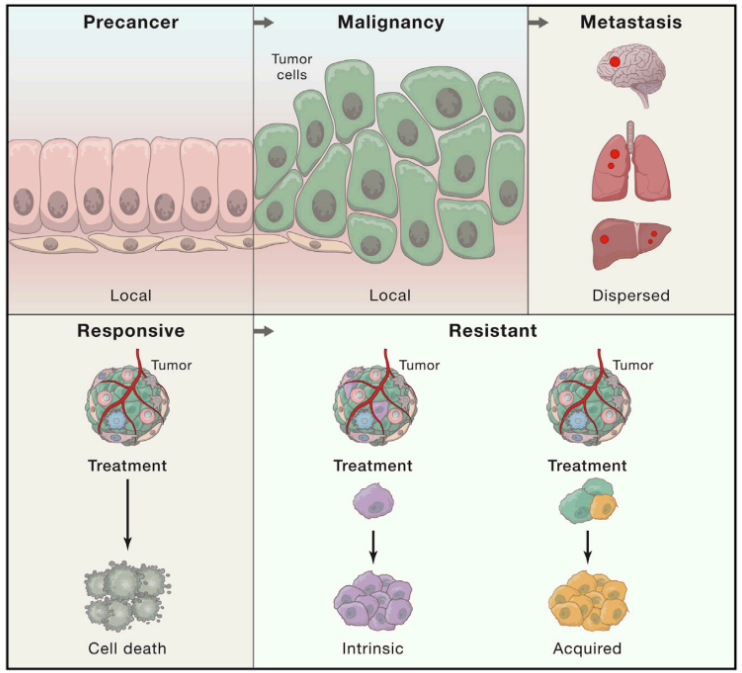
Machine Learning Applications for Cancer

An Interactive Hub for Multiplex Tissue Image Analysis

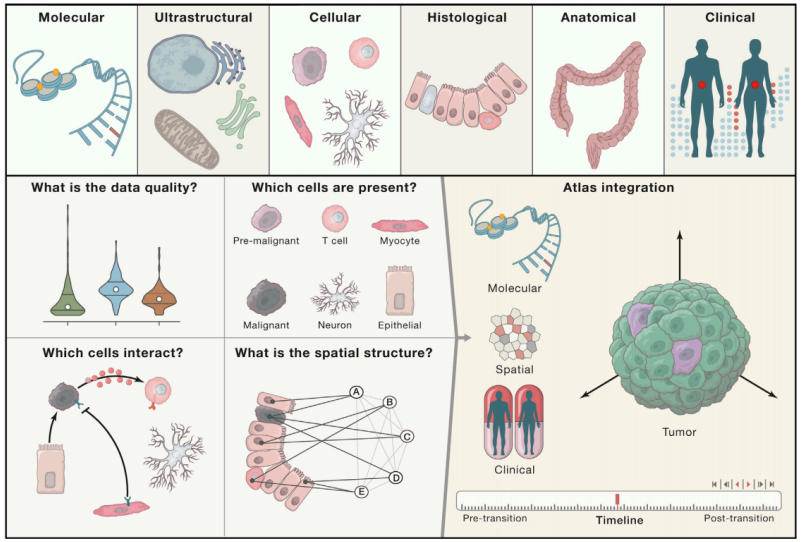
Cancer Moonshot NCI Human Tumor Atlas Network

https://humantumoratlas.org/

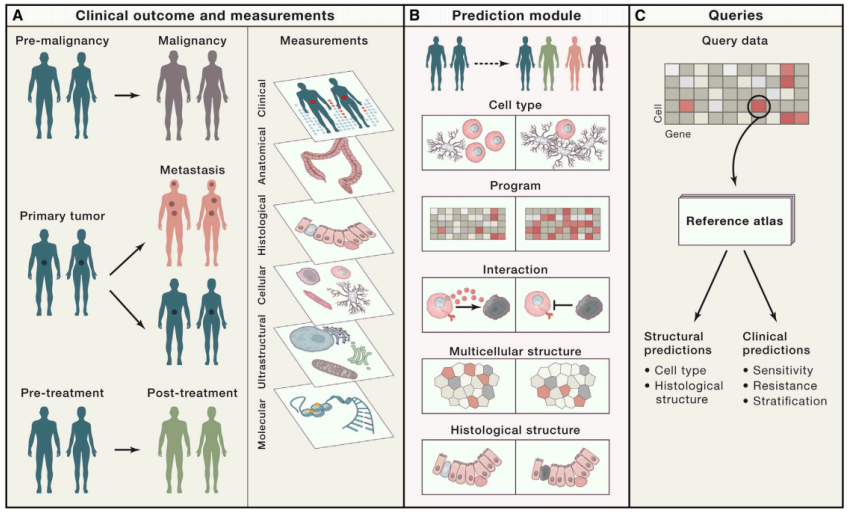
Studying key transitions



Deep Profiling



Systems Biology Understanding and Clinical Predictors



Rozenblatt-Rosen et al., https://doi.org/10.1016/j.cell.2020.03.053

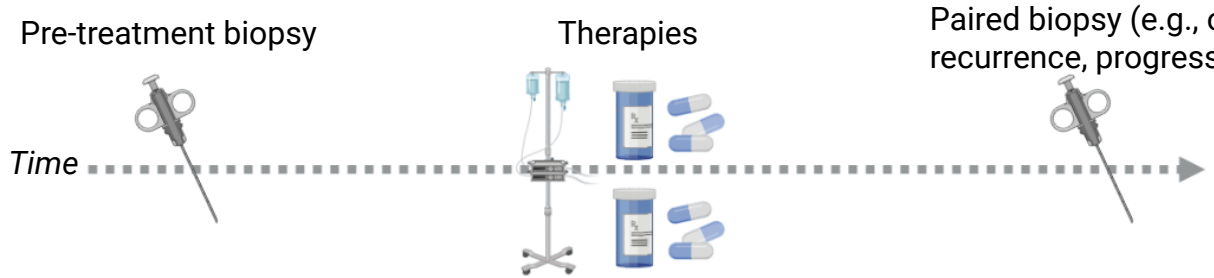
Omics and Multidimensional Spatial Atlas

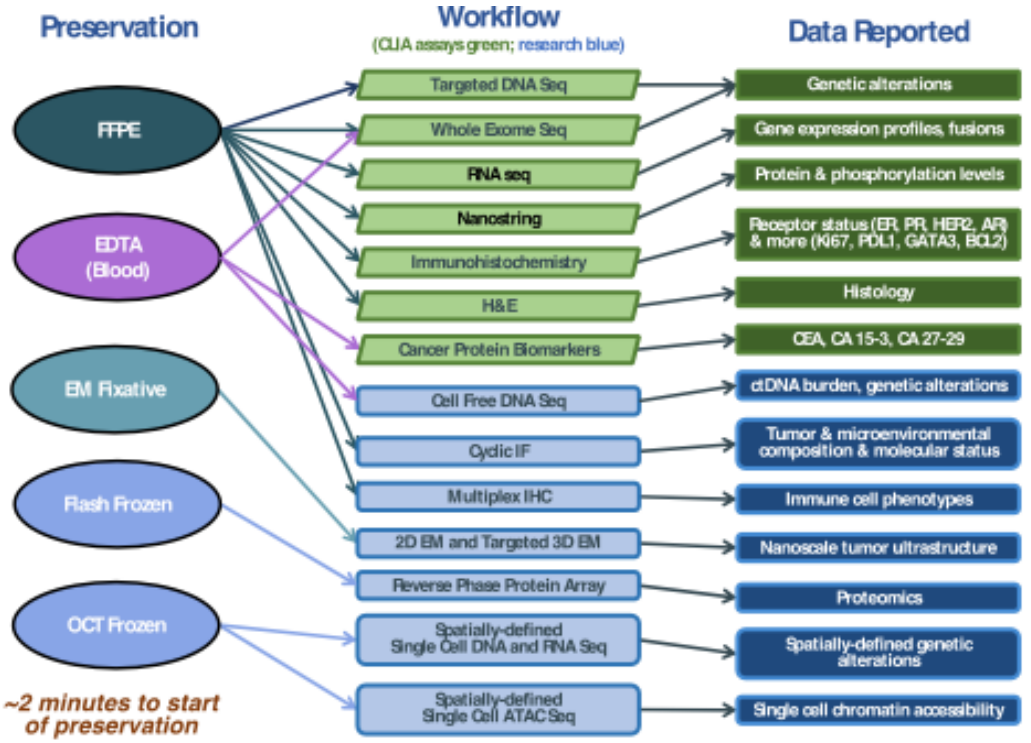
Collection of longitudinal and paired biopsies for *same patients*

A large suite of omics and imaging assays is applied to each biopsy

Omics and imaging data is connected to clinical attributes to:

- Characterize how tumor is adapting to therapy
- Identify potential mechanisms of resistance to therapy





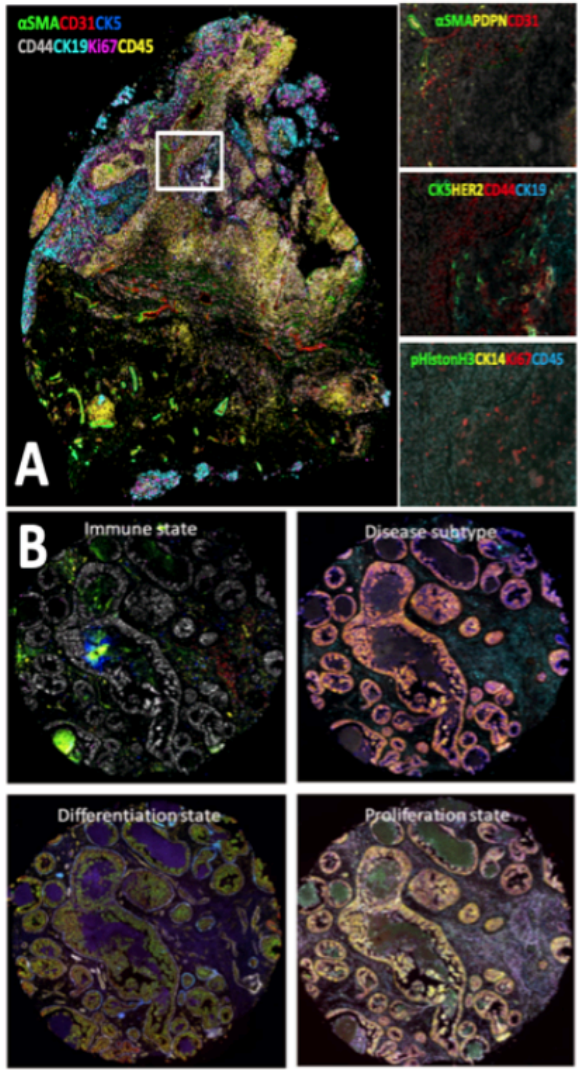
Multiplex Tissue Imaging

Spatial omics assays that assay FFPE create single-cell 2D/3D tumor maps

All sorts of interesting biology that can be investigated:

- Compositional information
- Spatial information such as microstructures (e.g., TLS, tumor-immune borders)

Examples: Cyclic Immunofluorescence, CODEX, IMC, Multiplex Immunohistochemistry



		Biomarker	Target
Cell State	Differentiation	CK5*	Cytoskeleton
		CK7*	Cytoskeleton
		CK8+18*	Cytoskeleton
		CK14*	Cytoskeleton
		CK17*	Cytoskeleton
		CK19*	Cytoskeleton
		E-Cadherin*	Epithelial cell junction
		Chromogranin A#	Neuro-endocrine cell
		CD133	Stem cell
	Proliferation	HER2*	Oncoprotein
		ER*	Hormonal receptor
		PgR	Hormonal receptor
		Ki67*	Cell cycle
		PCNA*	Cell cycle
		Phospho-histone H3*	Mitosis
		Cleaved PARP	Apoptosis
		Cleaved Caspase3	Apoptosis
		TP53#	Cell cycle check point
		pS6RP	protein translation
	Architecture	CoxIV*	Mitochondira
		Tubulin*	Cytoskeleton
		Lamin A/C*	Nuclear envelope
		Fibrillarin	Nucleolus
		Gamma Tubulin	Centrosome
Microenvironment	Immune Component	CD3*	T-cell
		CD4*	T-cell
		CD8*	T-cell
		CD20*	B-cell
		CD45*	Immune cell
		CD68*	Macrophage
		FoxP3*	Regulatory T-cell
		PD1*	Immune check point
		PD-L1	Immune check point
	Stroma/Vascular	β-Catenin	Cell adhesion
		Vimentin*	Cytoskeleton
		CD44*	Basal/stem cell
		α-SMA*	Cytoskeleton
		CD31*	Endothelial cell
		VE-Cadherin	Endothelial cell junction
		Podoplanin*	Endothelial cell
		Fibrinogen	blodd vessel
		Collagen IV	basement membrane
		Laminin	basement membrane
		Desmin	Pericyte cytoskeleton
		PDGFR beta	Pericyte receptor

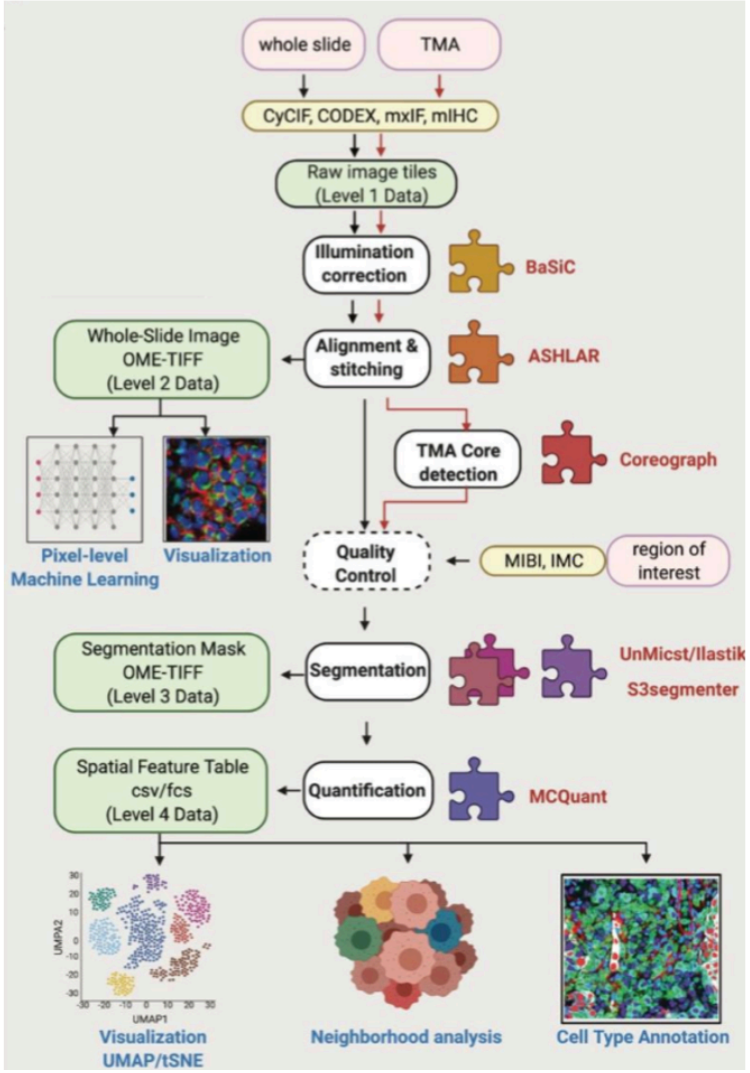
Analysis of Multiple Tissue Imaging Datasets

Datasets are image stacks that are tens-hundreds of gigabytes

10-20 analysis steps are needed for fully processing image stacks

Two general areas:

- Primary image analysis to create single-cell datasets
- Secondary analysis of single-cell datasets



An Interactive Hub for Multiplex Tissue Imaging Analysis

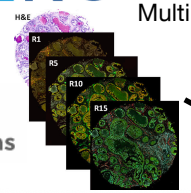
HTAN
HUMAN TUMOR ATLAS NETWORK

OMERO

SYNAPSE

NIH NATIONAL CANCER INSTITUTE
Cancer Research Data Commons

Multiplex Tissue Imaging
Raw Dataset



A Web-based portal for interactive data analysis and visualization

GUI that **anyone** can use to run analyses and visualize data

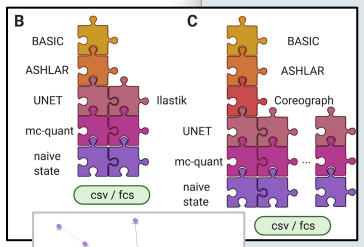
Workflow engine for programmatic and scalable access

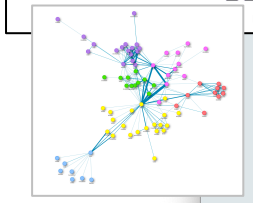
Extensible to incorporate a wide set of tools and visualizations

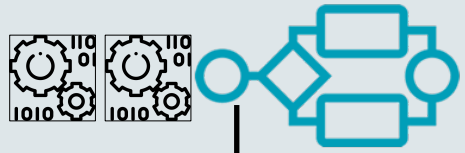
Run on both local computing resources and commercial clouds

Galaxy

Primary Image Analysis Tools and Workflows





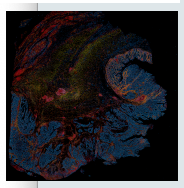


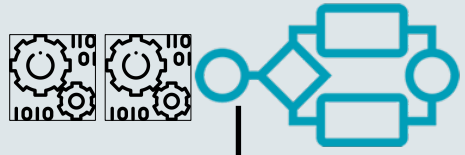
1. Segmented nuclei and cell
2. Single-cell marker intensities and spatial locations
3. Visualizations / Visual analytics

Cell Phenotyping, Spatial Analysis Workflows

HuBMAP CCF Portal

cellxgene





1. Single-cell Phenotypes
2. Cell Population Levels
3. Spatial Characterization
4. Visualizations / Visual analytics

Google Cloud

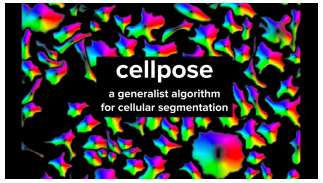
aws

Goal: “End-to-end” Multiplex Tissue Imaging Analysis and Visualization

1. Primary analysis—from image stack to cell feature tables

MCMICRO

Schapiro, Santagata, and Sorger labs



Pachitariu lab

2. Downstream analysis—cell phenotyping and spatial analyses

SCIMAP
scimap.xyz
SPATIAL SINGLE-CELL ANALYSIS TOOLKIT

Dr. Ajit Johnson Nirmal, Sorger lab

squidpy

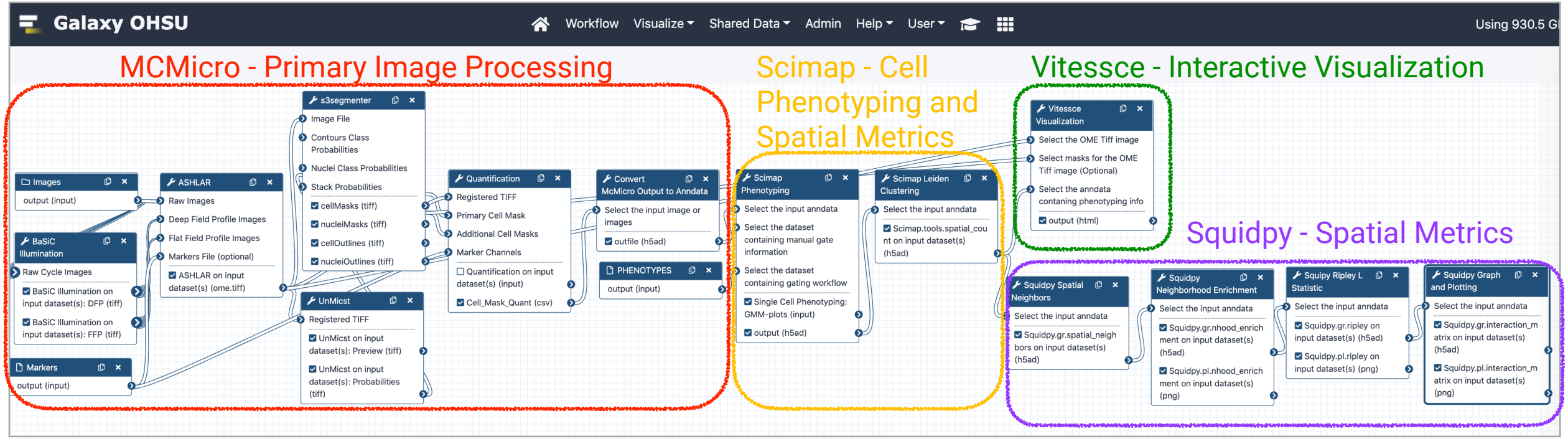
Theis lab

3. Visualization

ITESSCE

Gehlenborg lab / HuBMAP

End-to-end Multiplex Tissue Imaging Workflow

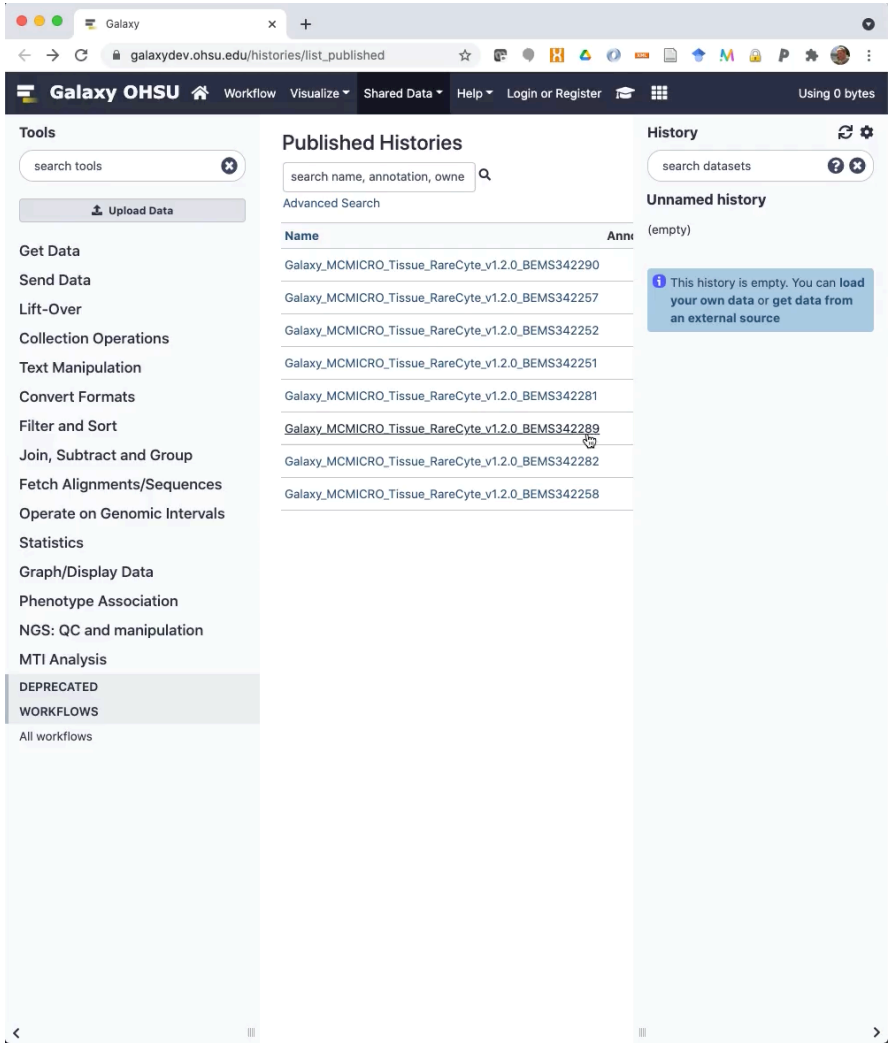


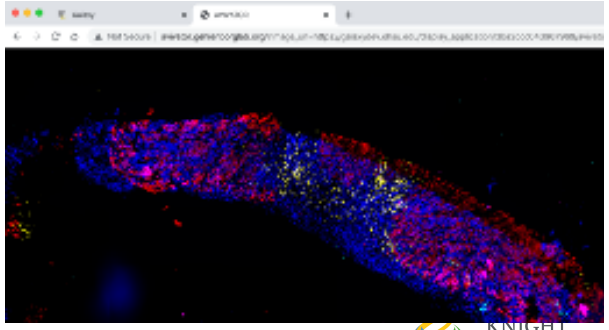
Registered Image Viewing with Viv

View image immediately in web browser and running analysis

No downloads, easy and fast even for large datasets

Review for visual QC such as registration errors





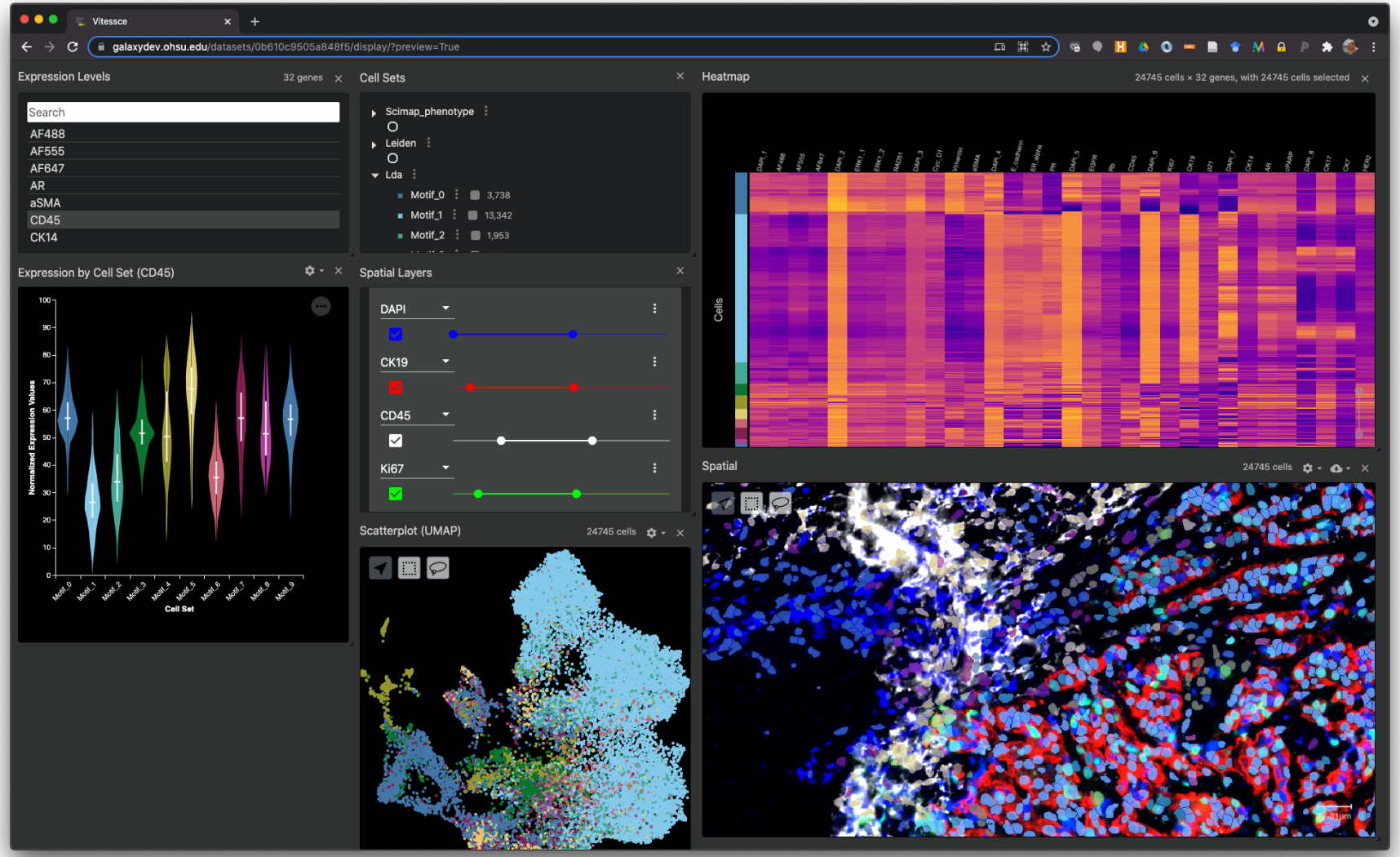
Viv: Manz et al., https://doi.org/10.31219/osf.io/wd2gu
http://viv.gehlenborglab.org/

Phenotypes + Spatial Layout with Vitessce

Three ways to identify phenotypes: gated, leiden, and LDA

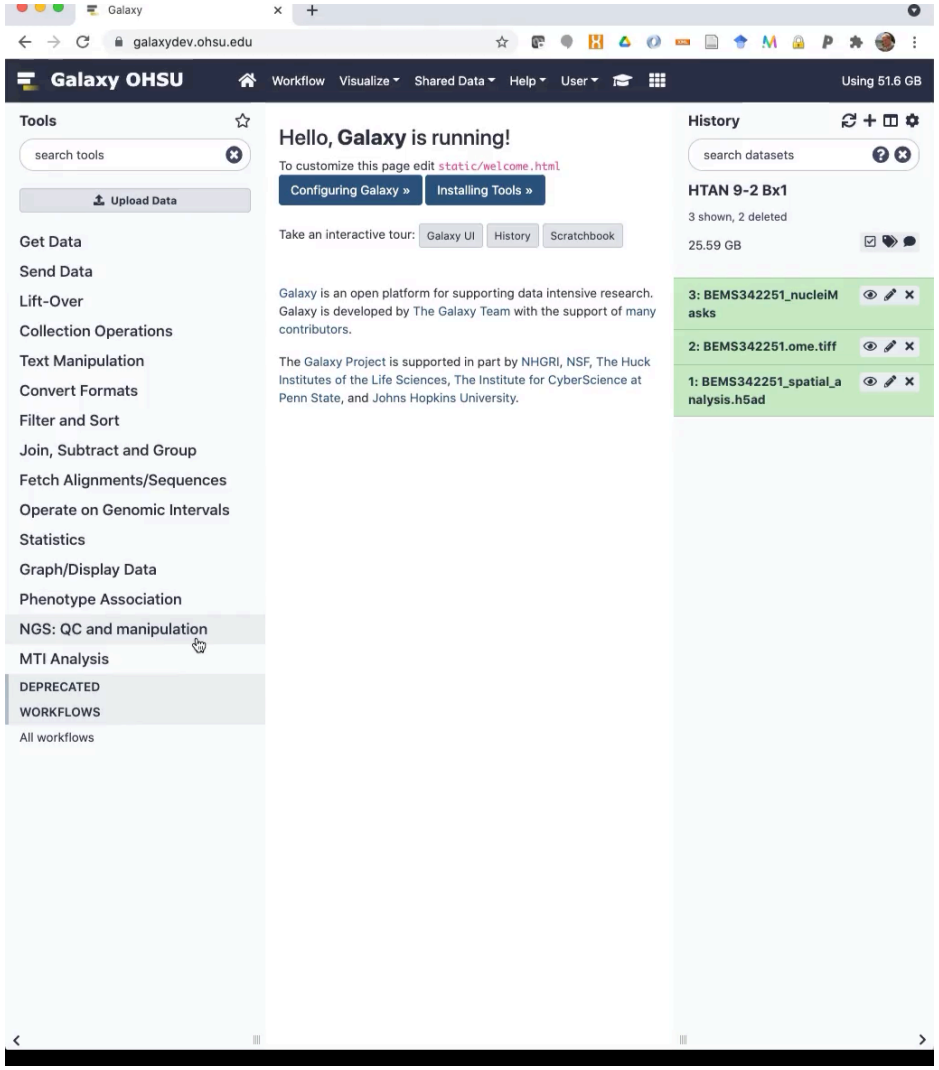
Overlay cell states with raw image + channels

Lower right: immune aggregation (white) adjacent to luminal tumor cells (red and green)



Vitessce: http://beta.vitessce.io/

Creating a Vitessce Dashboard with Galaxy

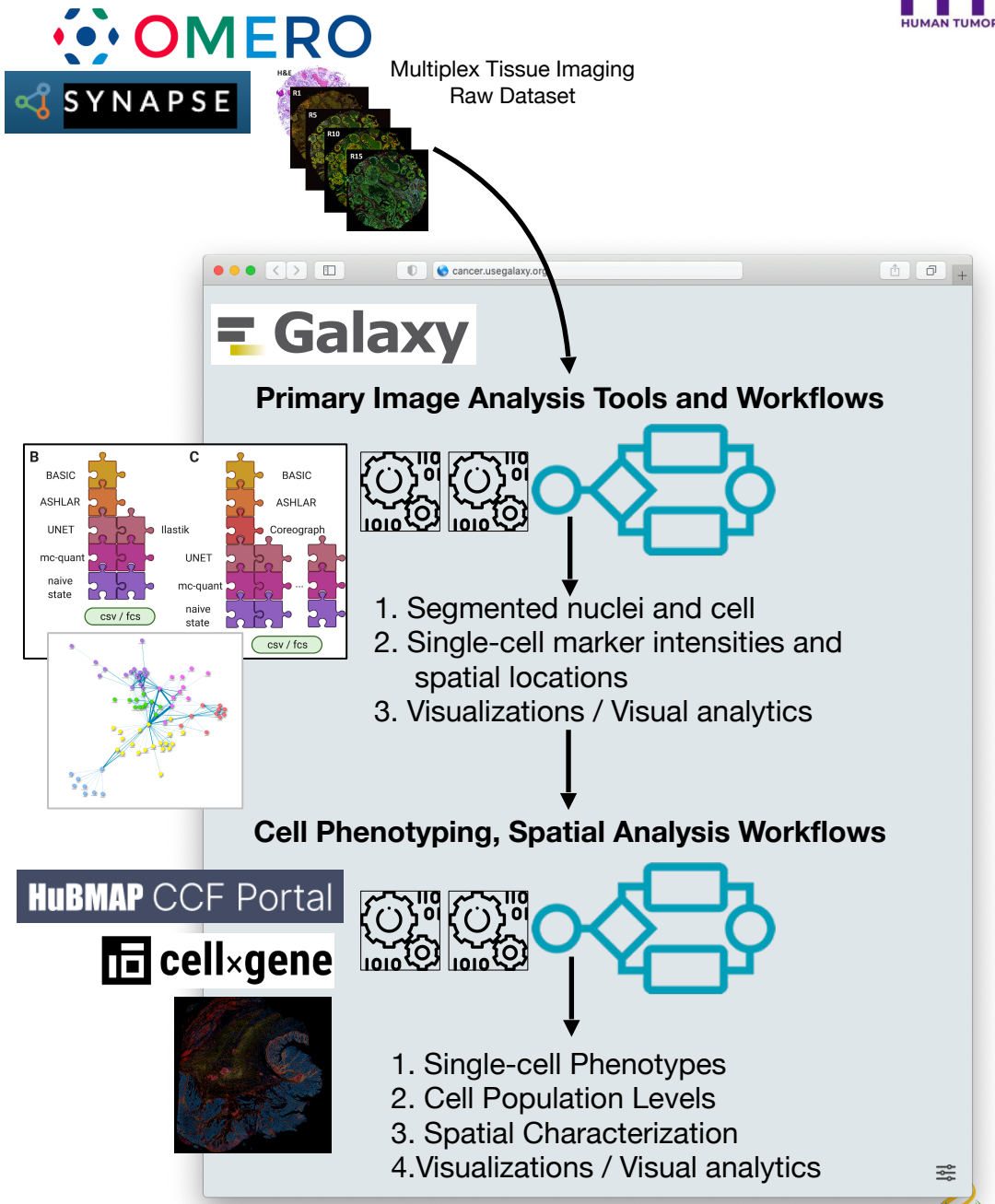


Vitessce: http://vitessce.io/

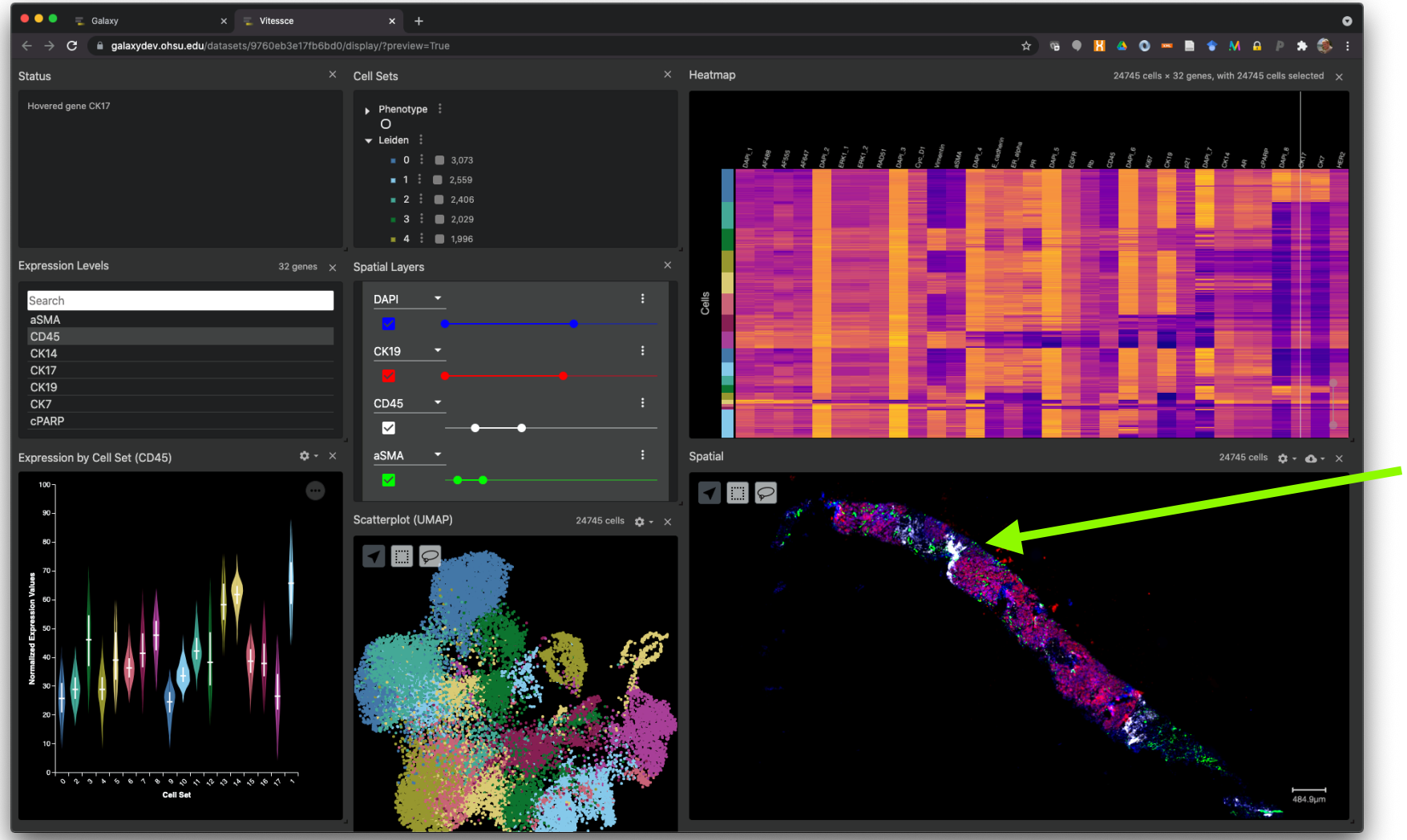
Enabling Fast and Collaborative Computational Science

Fast: When we fixed a registration issue yesterday, all analyses were rerun in a matter of hours

Collaborative: Results are available to everyone via a Web browser with no downloads, which is especially valuable for distributed teams



Observing an Immune Aggregate

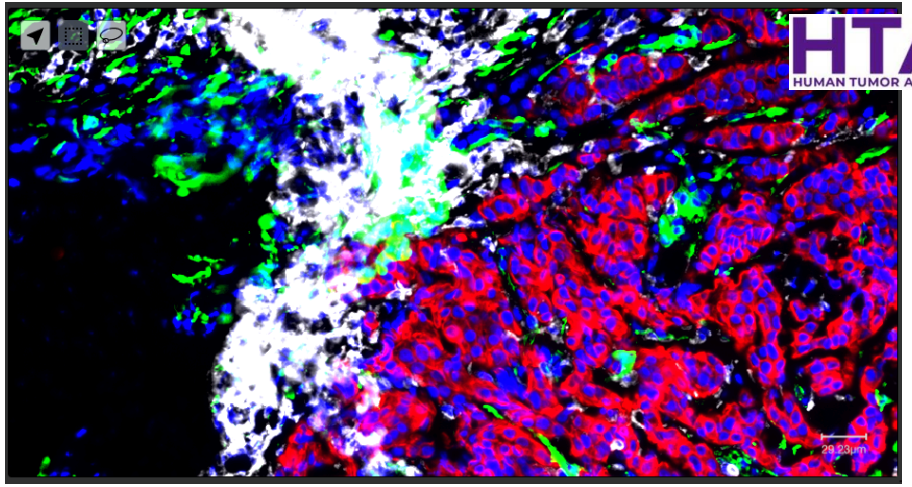


Comparing Cell State Calling Approaches

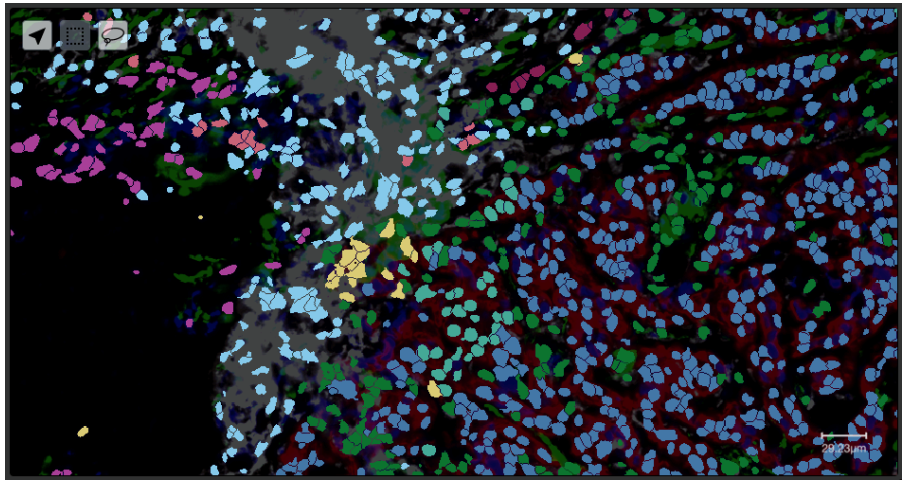
Visually similar cell states calls with Leiden and Gated approaches

Appears to be difficult to segment immune cells accurately in the immune aggregate

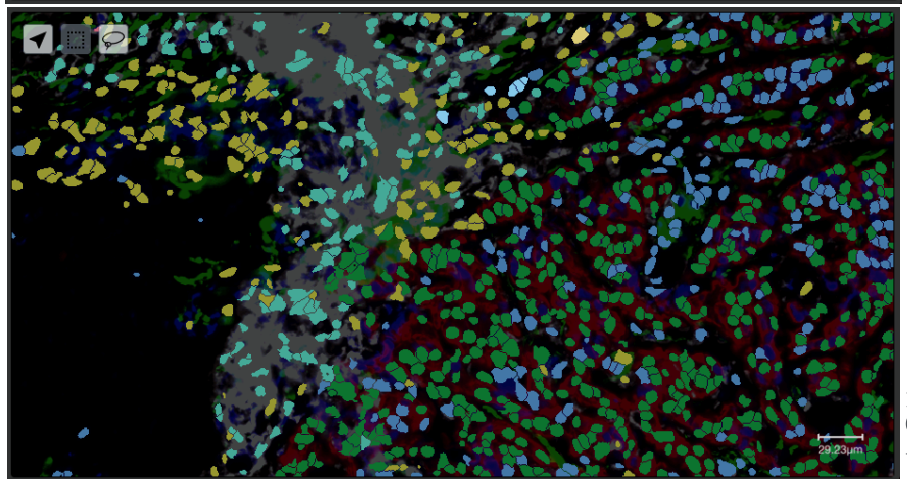
Stain



Leiden



Gated



HTAN
HUMAN TUMOR ATLAS NETWORK

KNIGHT CANCER Institute

Concluding Thoughts

Galaxy's unique aspects in my opinion:

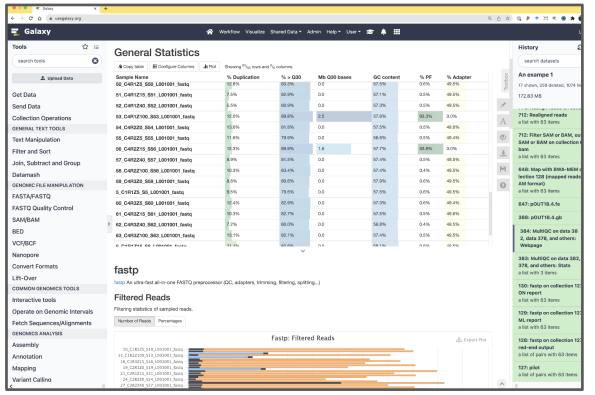
- Web-based UI, *but this is changing*
- Integrative framework for putting things together—the whole is greater than the sum of the parts
- World-wide community

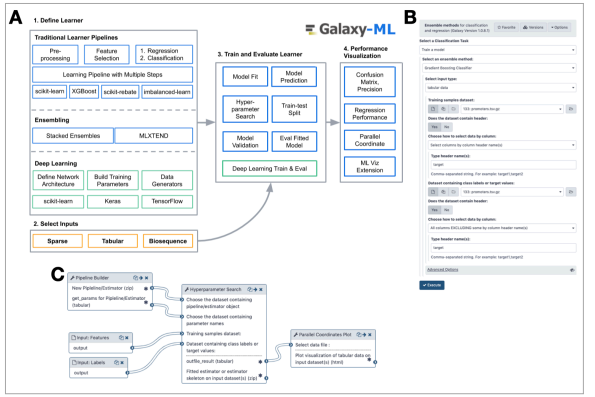
Galaxy can enable tremendously powerful cancer analyses

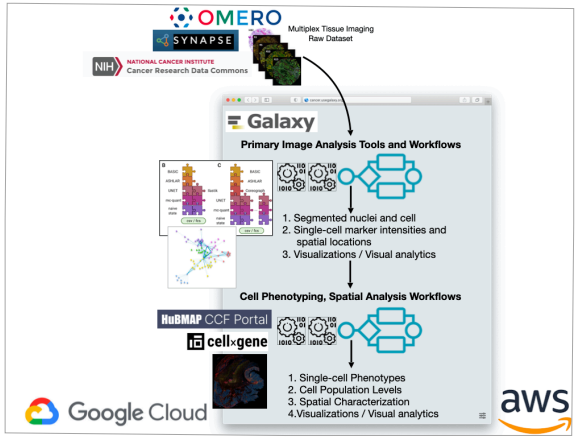
- Self-supervised and interpretable ML models for predicting response to therapy
- Analysis of multiplexed tissue imaging datasets to create 2D tumor maps

Opportunities and challenges

- Extend Galaxy with new UIs that provide both higher and lower levels of interactions
- Integrating Galaxy with *systems* rather than tools, e.g. data management systems, model zoos, and visualization hubs like cBioPortal
- Deploying a cancer-specific Galaxy server that maximizes efficiency via use of native cloud resources







KNIGHT CANCER Institute
OHSU

The Galaxy Community

All things Galaxy:
https://galaxyproject.org/

Public Galaxy servers:
https://galaxyproject.org/use/

Download and run Galaxy locally:
https://getgalaxy.org

Galaxy training network:
https://training.galaxyproject.org/

HTAN OMS Atlas Center

Oregon Health & Science University

Jeremy Goecks
Joe Gray
Gordon Mills
George Thomas

Andrew Adey
Courtney Betts
Katie Blise
Erik Burlingame
Elmar Bucher
Young Hwan Chang
Koei Chin
Hyeyoung Cho
Lisa Coussens
Allison Creason
Emek Demir
Jenny Eng
Trevor Enright
Heidi Feiler
Andrew Fields
Danielle Galipeau
Giovanney Gonzalez
Qiang Gu
Alexander Guimaraes
Zhi Hu
Brett Johnson
Annette Kolodzie
Terence Lo
Hannah Manning
Shannon McWeeney
Souraya Mitri
Zahi Mitri
Jessica Riesterer
Luke Sargent
Brendan O'Connell
Byung Park
Daniel Persson
Rosalie Sears
Xubo Song
Kiara Siex
Sam Sivagnanam
Luke Ternes
Guillaume Thibault
Nicholas Van Marter-Sanders

Harvard Medical School/ Brigham Women's Hospital

Peter K. Sorger
Sandro Santagata

Jia-Ren Lin
Yu-An Chen
Denis Schapiro
Artem Sokolov
Clarence Yapp
Jeremy L. Muhlich
Maulik Nariya
Gregory J. Baker
Juha Ruokonen
Zoltan Maliga
Connor Jacobson
Alyce A. Chen
Madison Tyler
Jennifer Guerriero

MD Anderson
Nick Navin
Yiyun Lin
Emi Sei

Quantitative Imaging Systems (Qi)
Damir Sudar
Michel Nederlof

Thank You!