



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

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

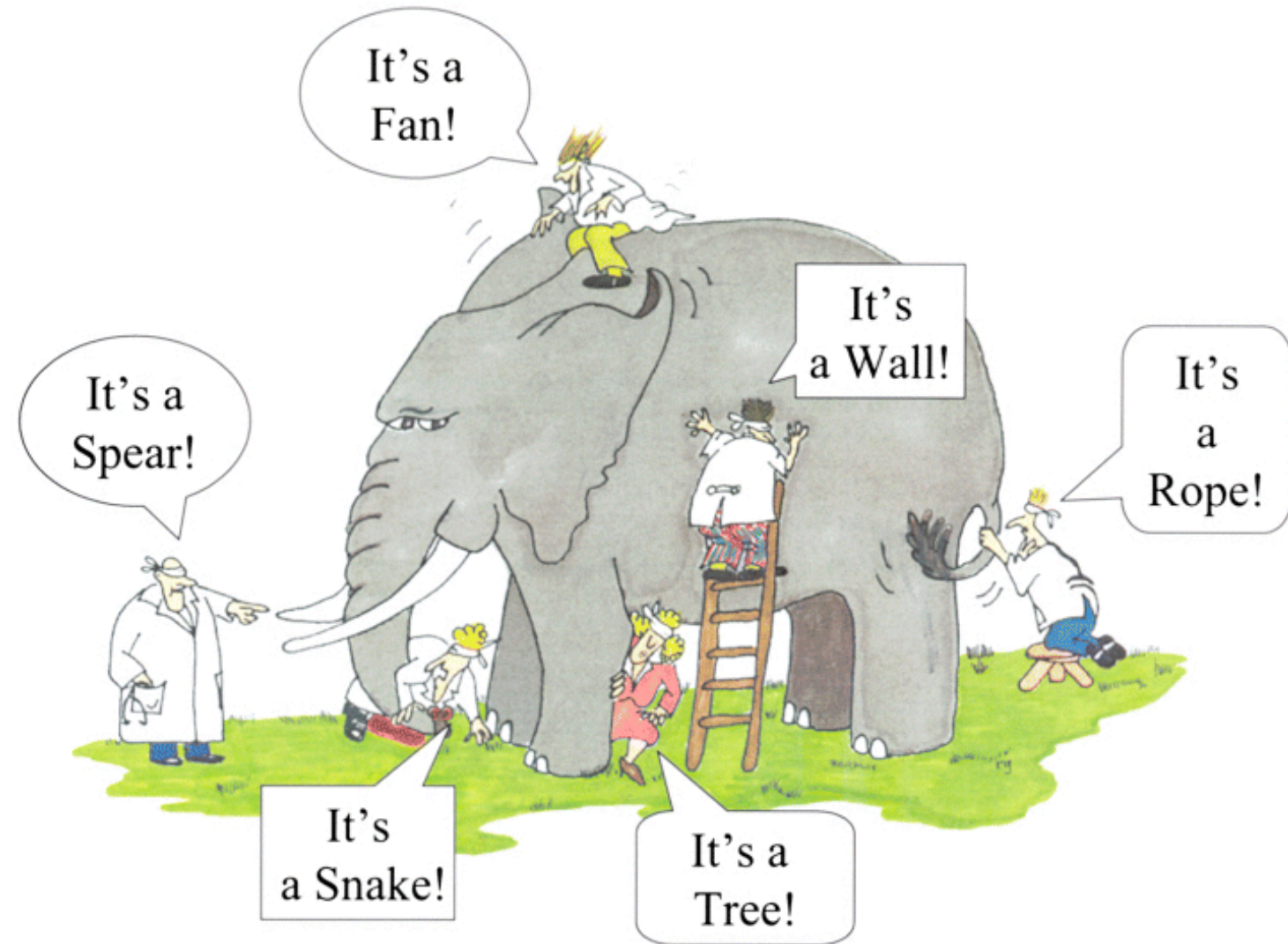
- Multiple levels of information are needed from broad overview down to essential details

What is Galaxy?

Different things to
different communities

Galaxy user
communities

- Scientists
- Tool Developers
- Educators
- Service Providers



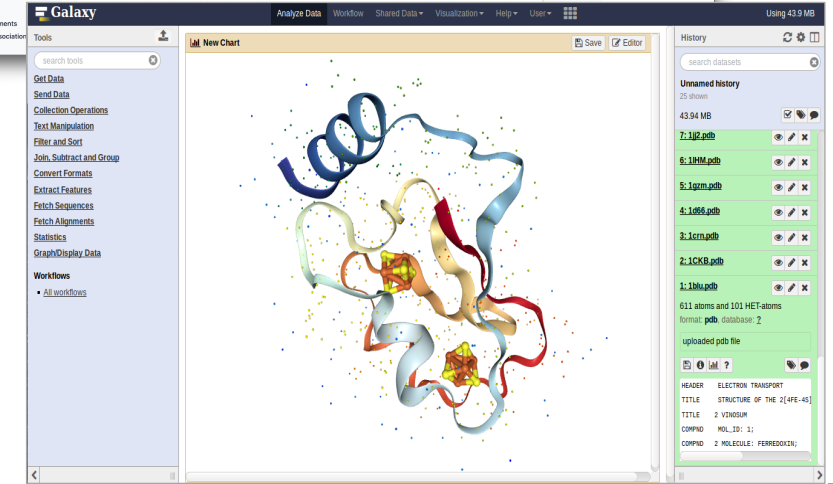
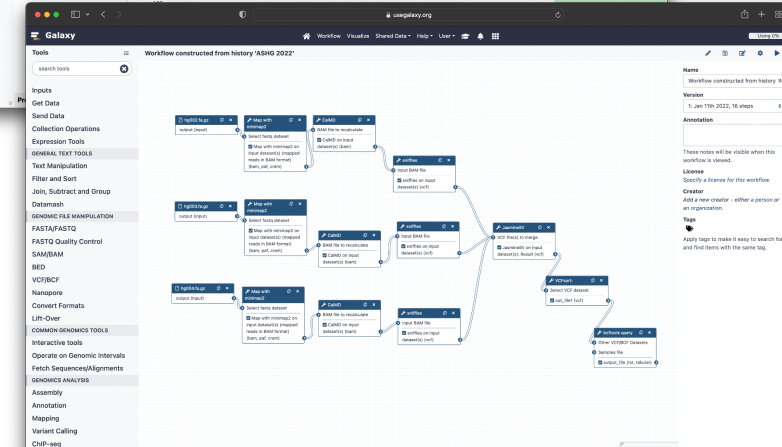
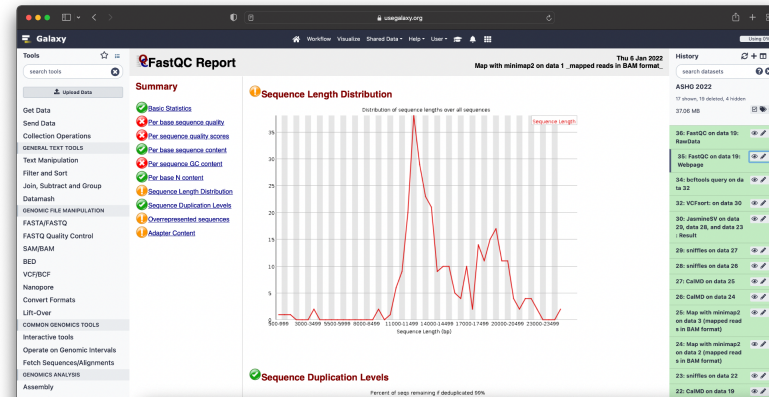
Key Features of Galaxy

- ▶ 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

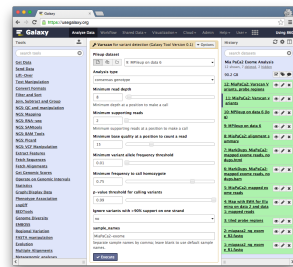
...created and supported by a large international community of scientists, developers, service providers, and educators



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

Interfaces

Web UI



Programmatic API



Datasets



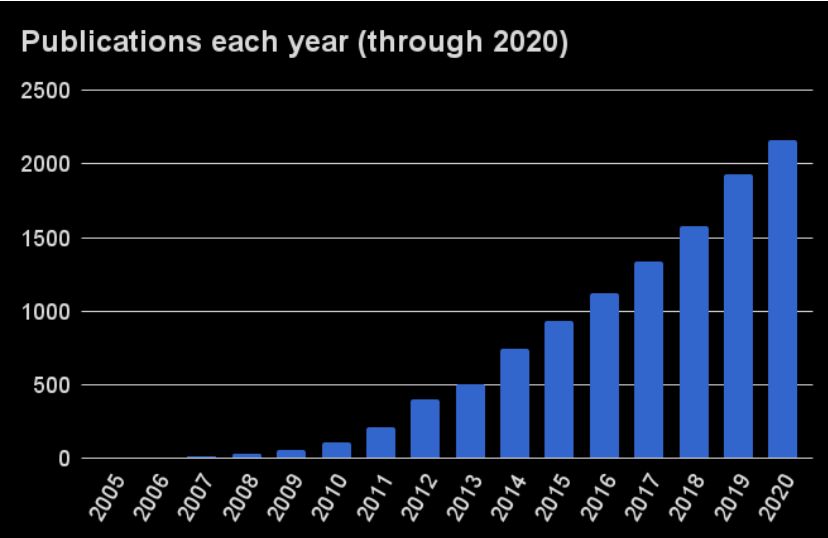
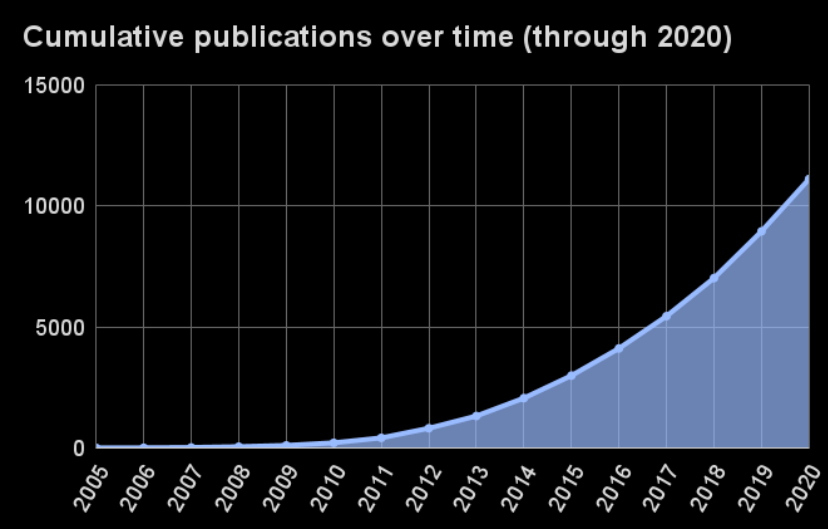
Computing Resources



Analysis Tools and Visualizations



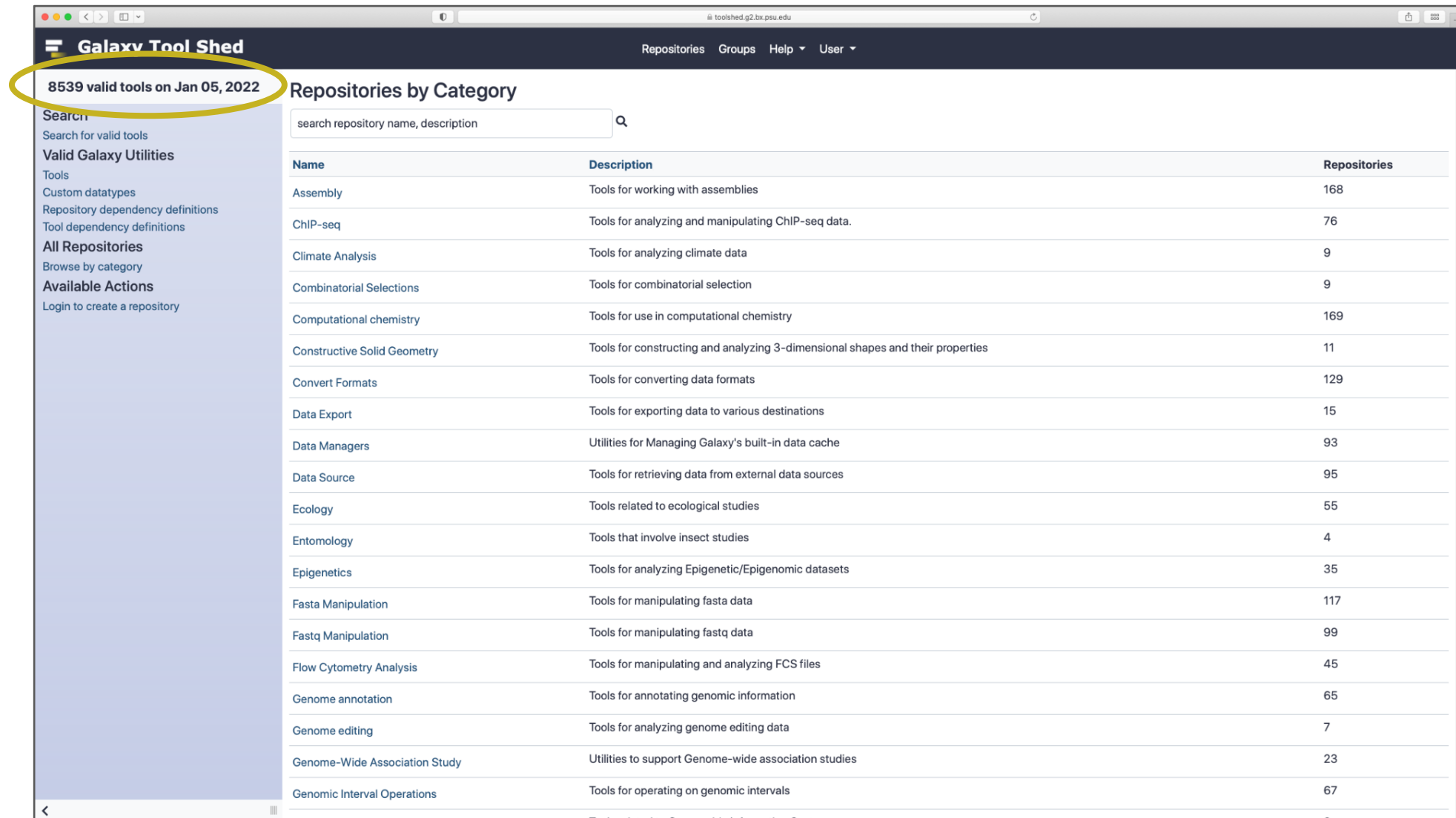
Galaxy is Widely Cited



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

Galaxy Toolshed



The screenshot shows the Galaxy Toolshed interface. At the top, the text "8539 valid tools on Jan 05, 2022" is circled in yellow. Below this, the "Repositories by Category" section is visible, featuring a search bar and a table of categories and their respective tool counts.

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

The screenshot displays the Galaxy web interface. On the left is a sidebar with tool categories: 'GENERAL TEXT TOOLS', 'GENOMIC FILE MANIPULATION', 'COMMON GENOMICS TOOLS', and 'GENOMICS ANALYSIS'. The main area shows a 'General Statistics' table with columns for Sample Name, % Duplication, % > Q30, Mb Q30 bases, GC content, % PF, and % Adapter. Below the table is a 'fastp' section with a 'Filtered Reads' plot showing read lengths for various samples.

Sample Name	% Duplication	% > Q30	Mb Q30 bases	GC content	% PF	% Adapter
50_C4R1Z5_S50_L001001_fastq	12.8%	83.3%	0.0	57.5%	0.6%	49.5%
51_C4R1Z15_S51_L001001_fastq	7.5%	82.9%	0.0	57.1%	0.5%	49.5%
52_C4R1Z40_S52_L001001_fastq	5.5%	80.9%	0.0	57.3%	0.5%	49.5%
53_C4R1Z100_S53_L001001_fastq	12.0%	89.8%	2.5	57.6%	93.3%	0.0%
54_C4R2Z0_S54_L001001_fastq	13.6%	81.8%	0.0	57.5%	0.5%	49.6%
55_C4R2Z5_S55_L001001_fastq	11.6%	79.9%	0.0	56.6%	0.5%	49.4%
56_C4R2Z15_S56_L001001_fastq	12.3%	89.9%	1.6	57.7%	93.8%	0.0%
57_C4R2Z40_S57_L001001_fastq	8.9%	81.5%	0.0	57.4%	0.5%	49.5%
58_C4R2Z100_S58_L001001_fastq	10.3%	83.4%	0.0	57.4%	0.4%	49.5%
59_C4R3Z0_S59_L001001_fastq	8.6%	80.6%	0.0	57.9%	0.6%	49.5%
5_C1R1Z5_S5_L001001_fastq	9.5%	79.5%	0.0	57.5%	0.6%	49.5%
60_C4R3Z5_S60_L001001_fastq	12.4%	82.9%	0.0	57.3%	0.6%	49.4%
61_C4R3Z15_S61_L001001_fastq	10.3%	82.7%	0.0	57.5%	0.5%	49.6%
62_C4R3Z40_S62_L001001_fastq	7.2%	80.0%	0.0	56.8%	0.4%	49.5%
63_C4R3Z100_S63_L001001_fastq	13.1%	82.1%	0.0	57.4%	0.5%	49.5%

Fastp: Filtered Reads

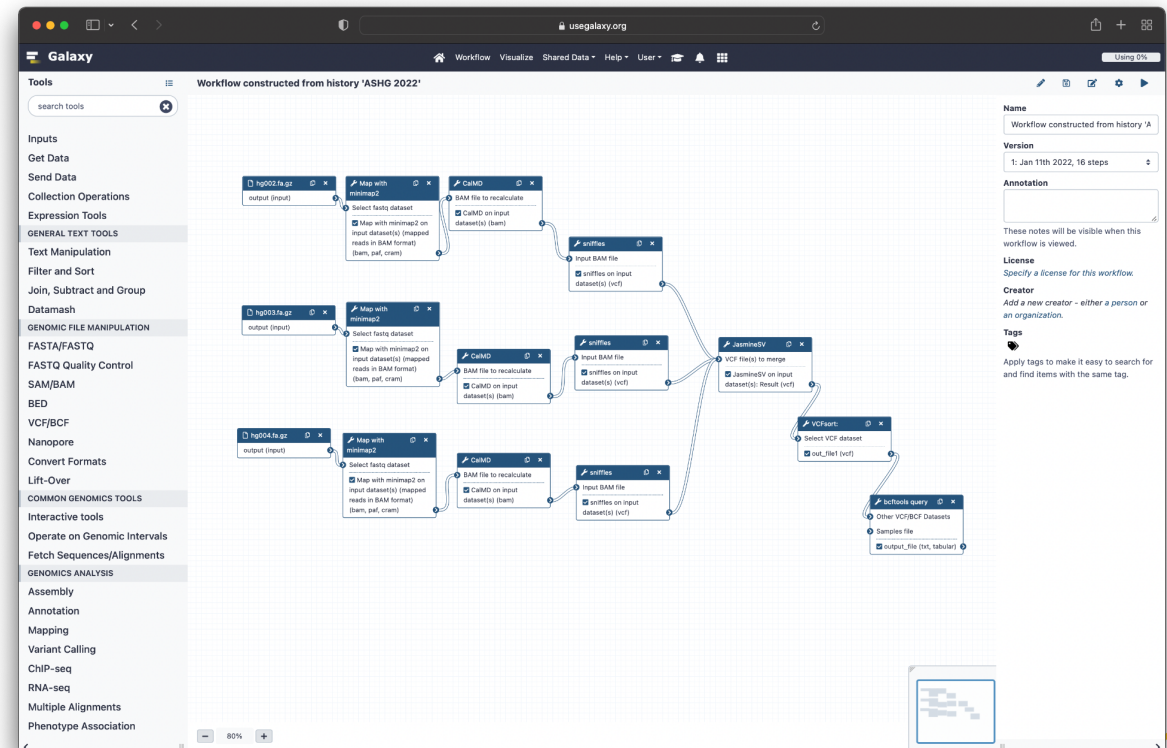
10_C1R2Z5_S10_L001001_fastq
13_C1R2Z100_S13_L001001_fastq
16_C1R3Z15_S16_L001001_fastq
19_C2R1Z0_S19_L001001_fastq
21_C2R1Z15_S21_L001001_fastq
24_C2R2Z0_S24_L001001_fastq
27_C2R2Z40_S27_L001001_fastq

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

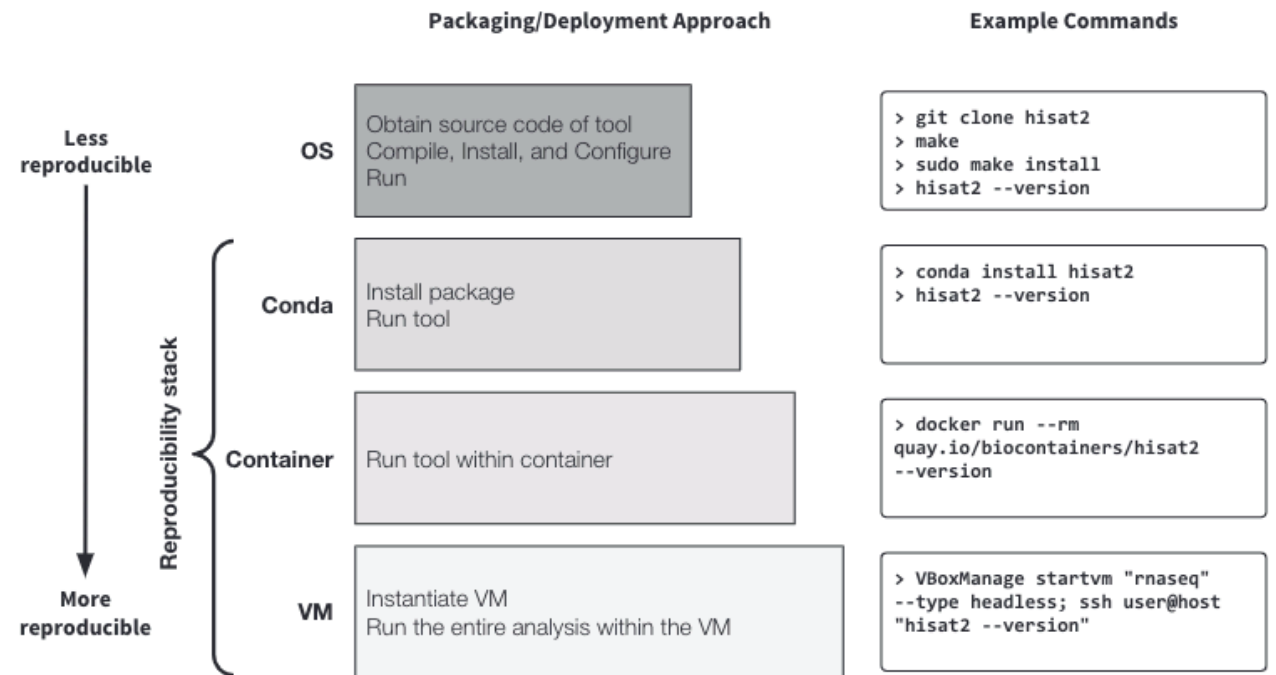
This is true for both individual tool executions and multi-step workflows



The Galaxy Reproducibility Stack

Layers of reproducibility built on virtualization technologies

Used for automated dependency resolution in Galaxy



Grüning et al. (2018) *Cell Systems*

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)

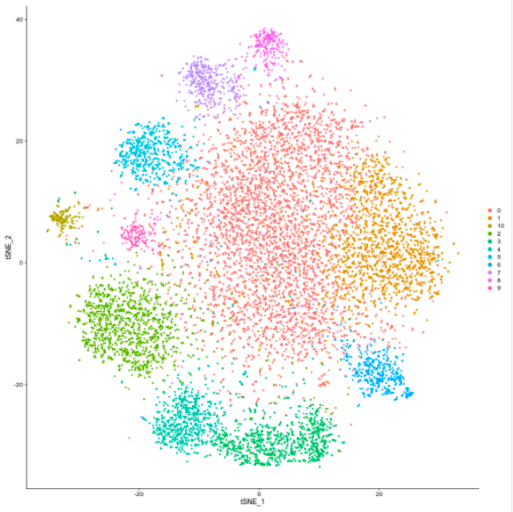
Title: Report for CVRM Wu 2018
Username: pmoreno
Created with Galaxy 21.05 on 13 September 2021, 23:17
Identifier: 1fd2d5714ddb4d97

Initial data import

Data was imported from Library item CVRM Wu 2018, producing a Single Cell Experiment dataset:
1aae95df0fc76c8d
This was then transformed to Seurat and to Scanpy, using the Seurat filter cells tool and the SCE Easy tool respectively.

tSNE plot visualisation

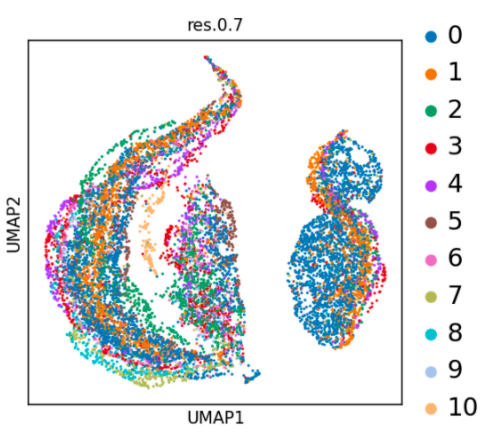
The Single Cell Experiment object already included a tSNE dimensionality reduction calculated, which looks like this (resolution=0.7):




Maybe some more text here helps with the layout. Maybe some more text here helps with the layout.

UMAP plot visualisation

Using the AnnData object from the SCEasy transformation, we do a quick UMAP plot, to realise (for the same resolution), that there seems to be some batch effect:



Indeed, observing this same UMAP through the cellxgene interactive view, we see that the UMAP dimensionality reduction has been heavily biased by the platform:



After executing normalisation, find variable genes, running PCA and batch correction with Harmony, the kNN graph is computed with this corrected PCA, and from this new neighbours graph UMAP is recomputed

Dataset: Scanpy RunUMAP on data 25: UMAP object AnnData

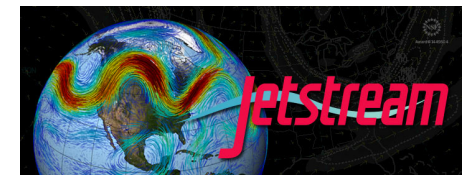
No content found.

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

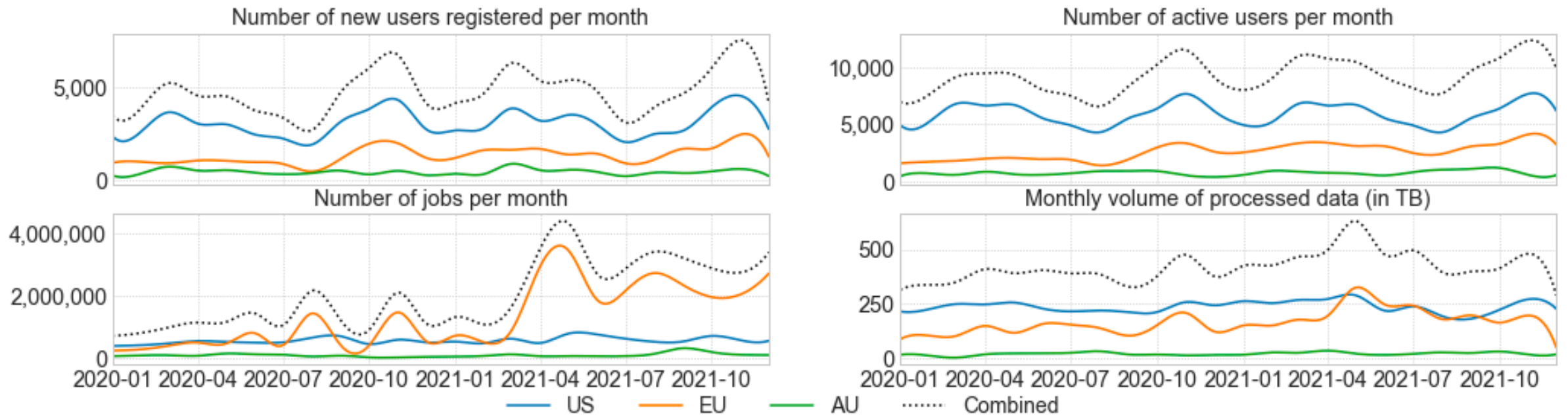
Galaxy
Main
usegalaxy.org

Galaxy
EUROPE
usegalaxy.eu

Galaxy
AUSTRALIA
usegalaxy.org.au




Usage of Three Main Public Galaxy Servers

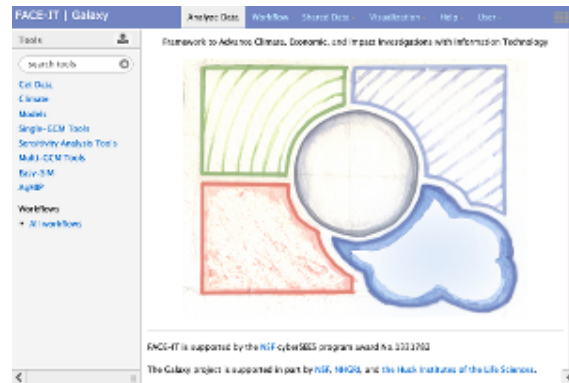


Have mass spec data?
Galaxy-P!

Three ways to Galaxy-P...

Public usegalaxy.org
 Local getgalaxy.org
 Cloud biocloudcentral.msi.umn.edu

or maybe 4...  [@usegalaxy](https://twitter.com/usegalaxy)

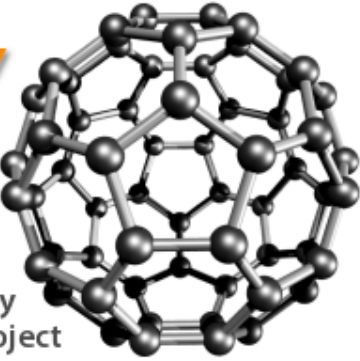


**OPEN SOURCE
 DRUG DISCOVERY**



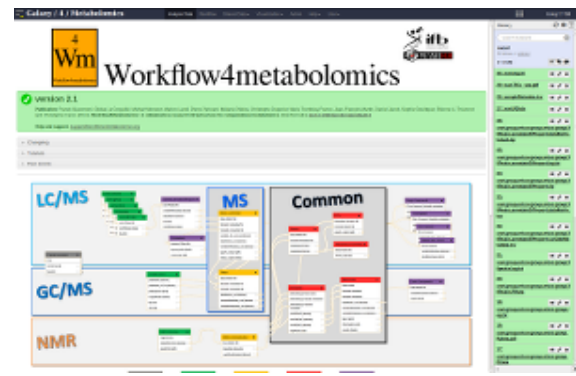
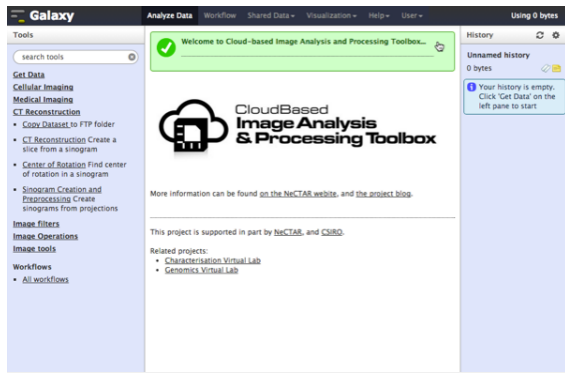

ballaxy

Powered by the
 Biochemical Algorithms Library Project




**SYM
 WXS**

Got symmetry?
 Find out.

Climate Change

**Proteomics
 Metabolomics
 Drug Discovery
 Cosmology
 Image Analysis
 Flow Cytometry**

Natural Language

<https://galaxyproject.org/use/>

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

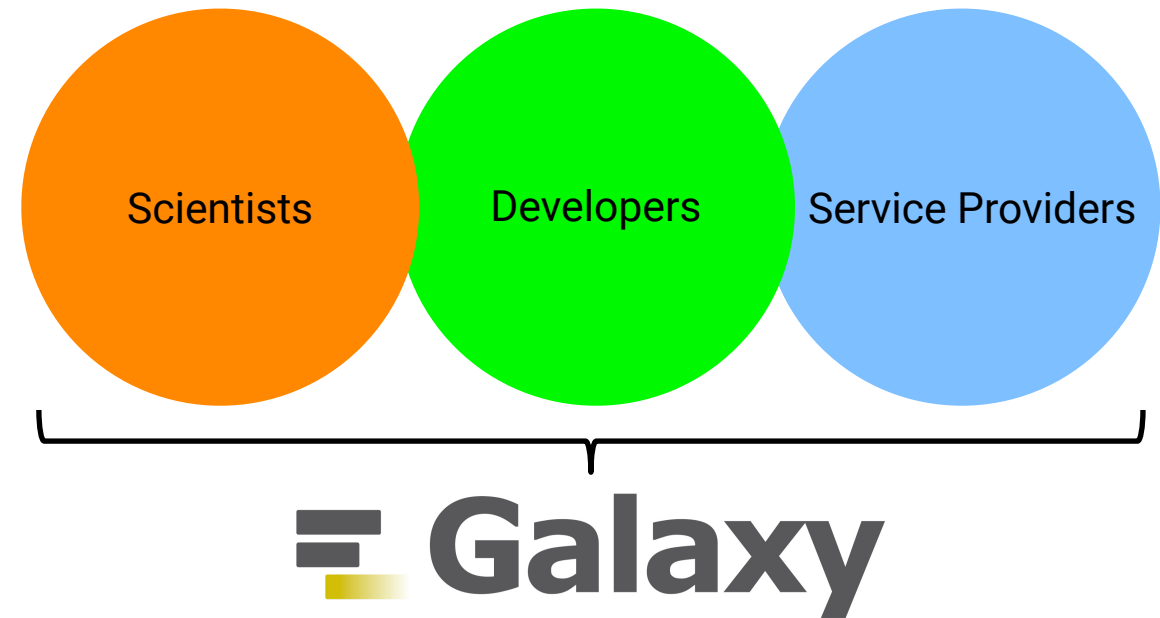
Developers get:

- 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



Galaxy Interactive Tools

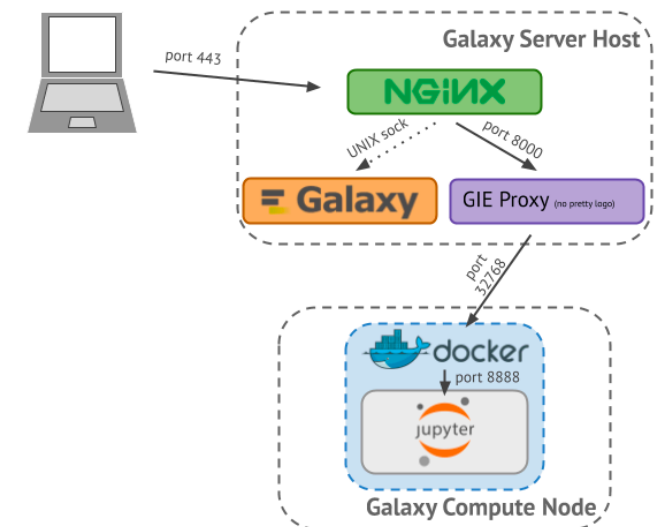
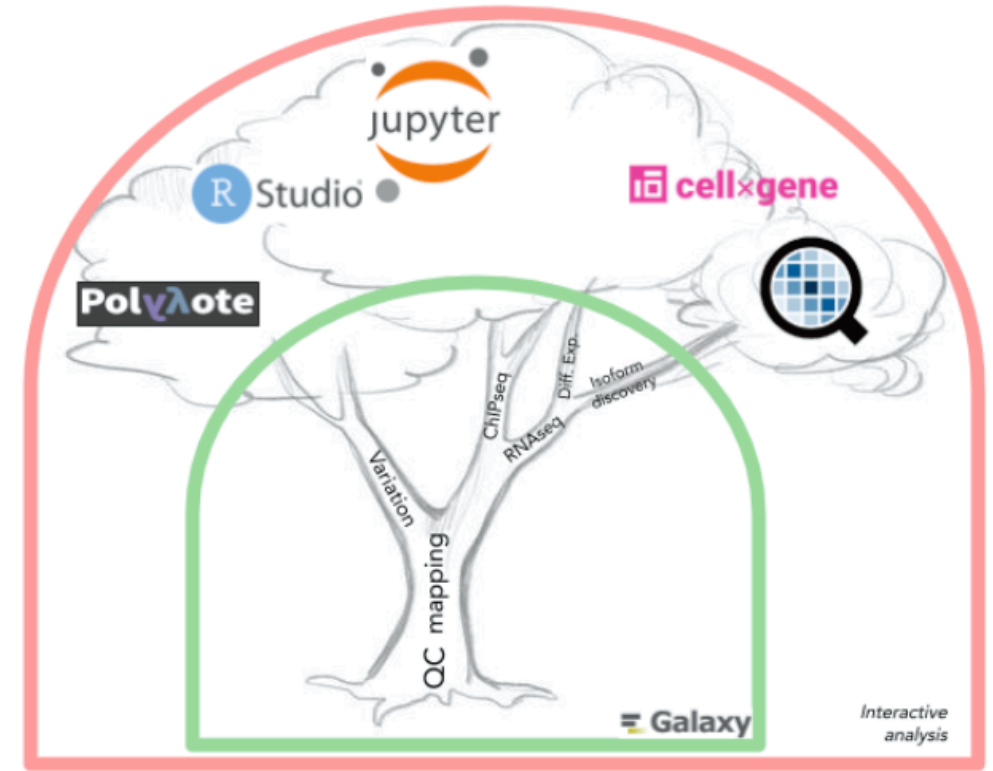
Makes interactive web tools available in Galaxy

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



Galaxy Dataset Collections

Today's analyses often involve many samples

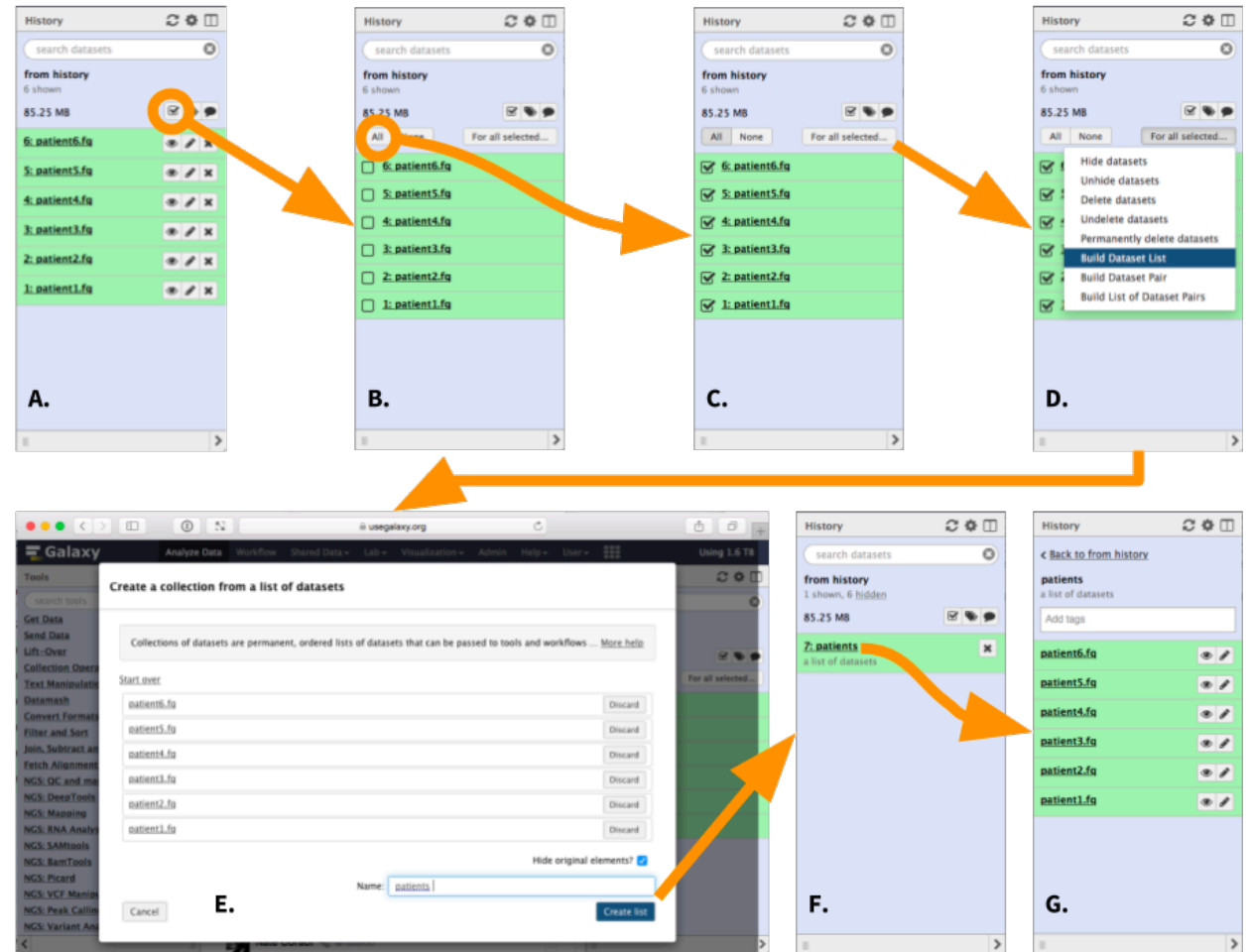
- ▶ Want to process all samples in the same way

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



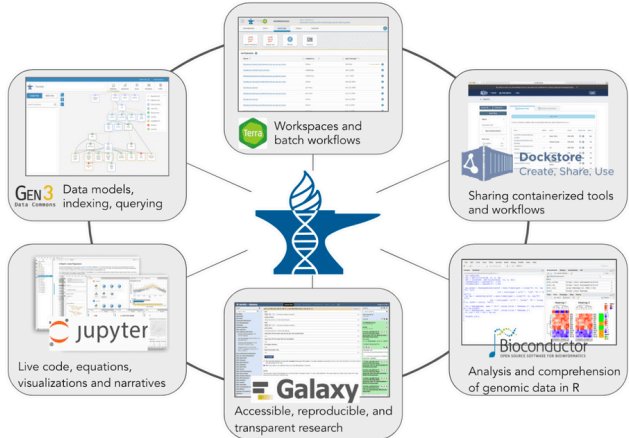
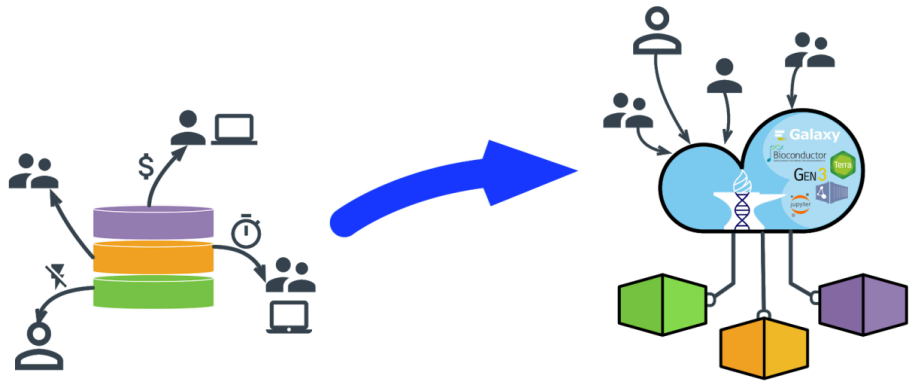
Galaxy as part of NIH Data Science Infrastructure

NHGRI

- ▶ A Component of the AnVIL, the NHGRI Data Commons

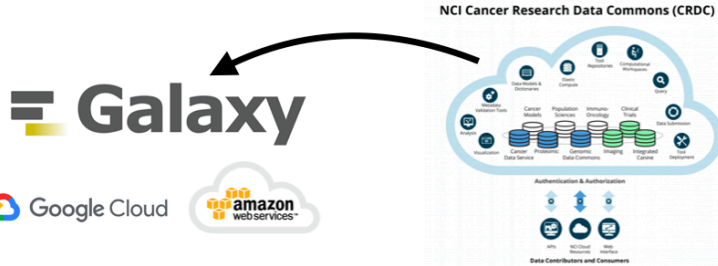
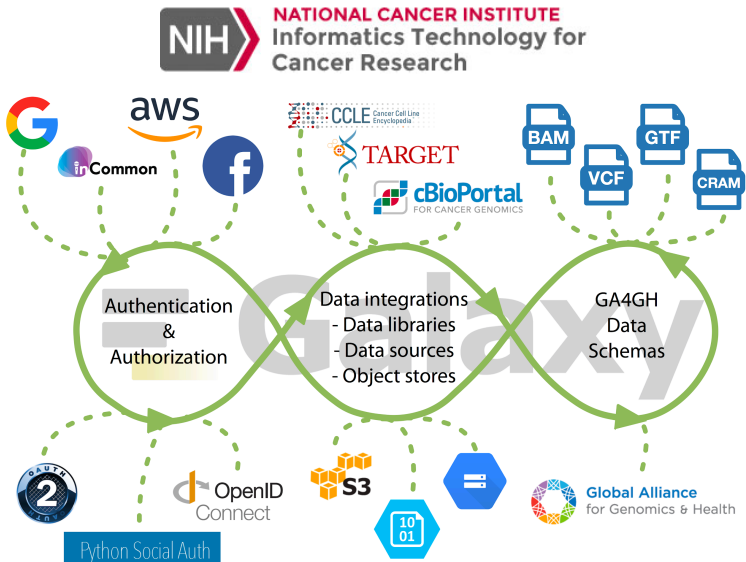
NCI

- ▶ Developing connections to the NCI Cancer Research Data Commons

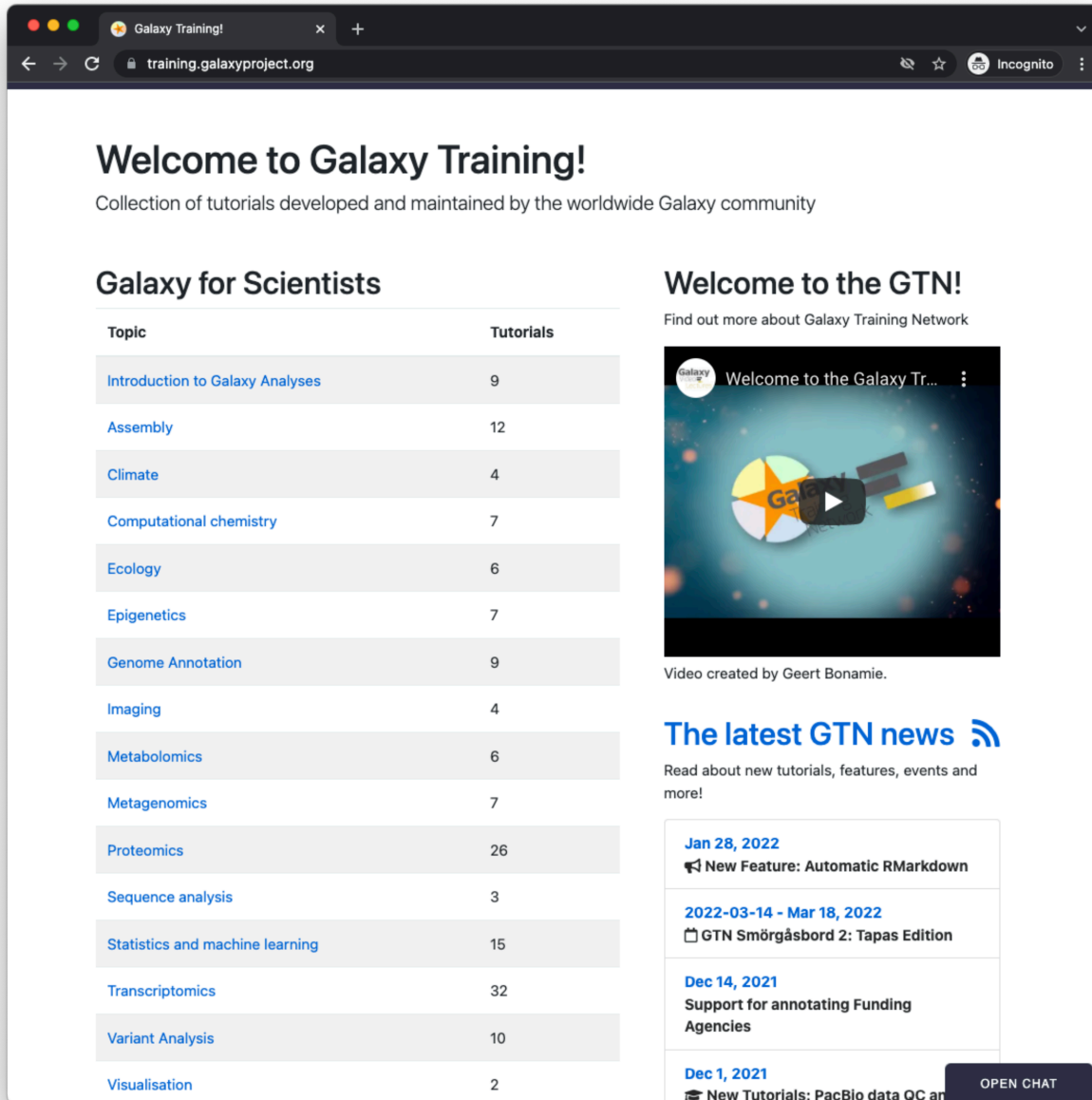


Key Advances

- ▶ 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



The Galaxy Training Network is a Fantastic Resource



Galaxy Training!
training.galaxyproject.org

Welcome to Galaxy Training!


Collection of tutorials developed and maintained by the worldwide Galaxy community

Galaxy for Scientists

Topic	Tutorials
Introduction to Galaxy Analyses	9
Assembly	12
Climate	4
Computational chemistry	7
Ecology	6
Epigenetics	7
Genome Annotation	9
Imaging	4
Metabolomics	6
Metagenomics	7
Proteomics	26
Sequence analysis	3
Statistics and machine learning	15
Transcriptomics	32
Variant Analysis	10
Visualisation	2

Welcome to the GTN!

Find out more about Galaxy Training Network



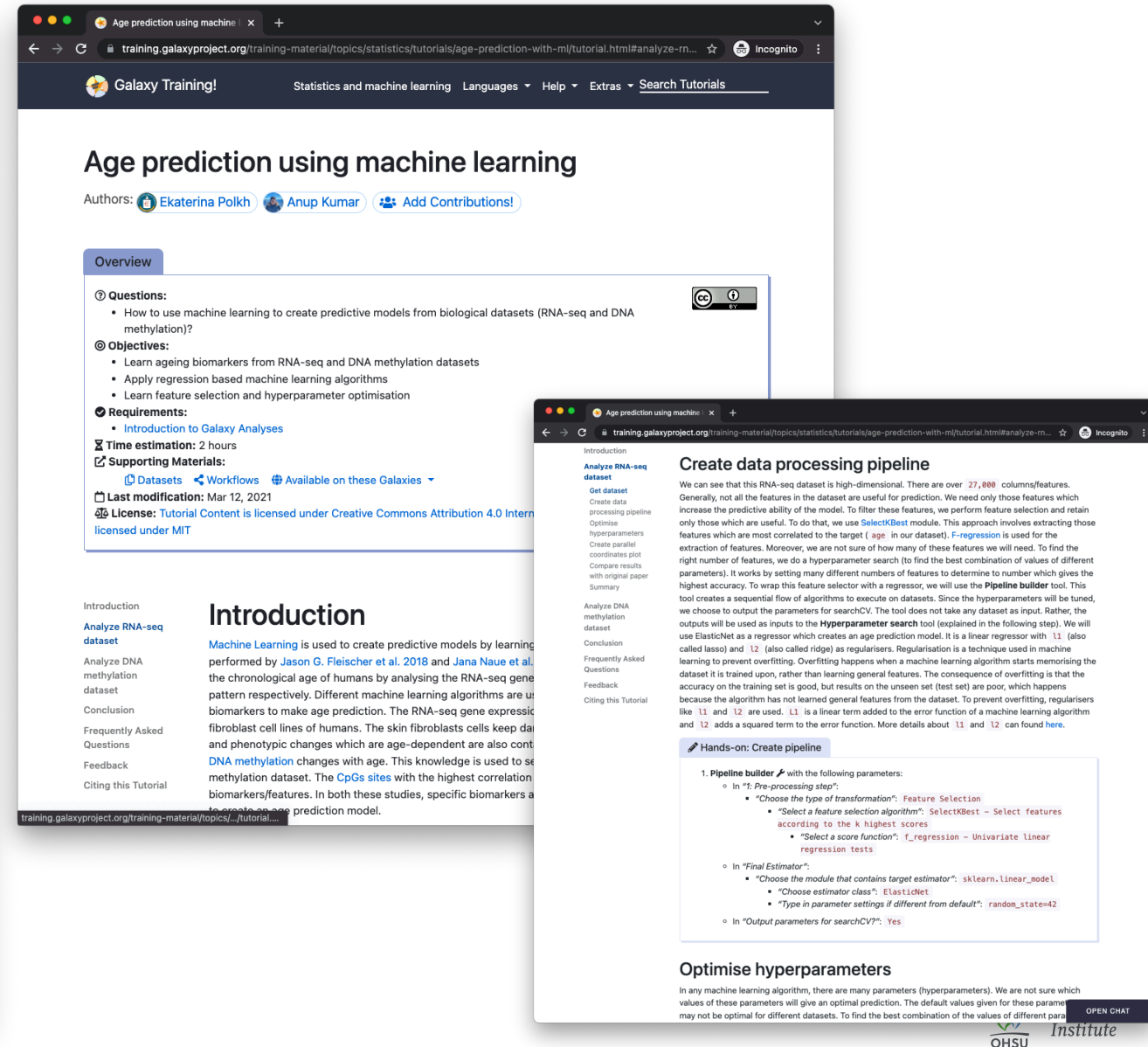
Video created by Geert Bonamie.

The latest GTN news

Read about new tutorials, features, events and more!

- Jan 28, 2022**
New Feature: Automatic RMarkdown
- 2022-03-14 - Mar 18, 2022**
GTN Smörgåsbord 2: Tapas Edition
- Dec 14, 2021**
Support for annotating Funding Agencies
- Dec 1, 2021**
New Tutorials: PacBio data QC analysis

OPEN CHAT



Age prediction using machine learning
training.galaxyproject.org/training-material/topics/statistics/tutorials/age-prediction-with-ml/tutorial.html#analyze-rn...

Authors: [Ekaterina Polkh](#) [Anup Kumar](#) [Add Contributions!](#)

Overview

- Questions:**
 - How to use machine learning to create predictive models from biological datasets (RNA-seq and DNA methylation)?
- Objectives:**
 - Learn ageing biomarkers from RNA-seq and DNA methylation datasets
 - Apply regression based machine learning algorithms
 - Learn feature selection and hyperparameter optimisation
- Requirements:**
 - [Introduction to Galaxy Analyses](#)
- Time estimation:** 2 hours
- Supporting Materials:**
 - [Datasets](#) [Workflows](#) [Available on these Galaxies](#)
- Last modification:** Mar 12, 2021
- License:** Tutorial Content is licensed under Creative Commons Attribution 4.0 International license under MIT

Introduction

Machine Learning is used to create predictive models by learning performed by Jason G. Fleischer et al. 2018 and Jana Naue et al. the chronological age of humans by analysing the RNA-seq gene pattern respectively. Different machine learning algorithms are used to create predictive models from biological datasets (RNA-seq and DNA methylation) to make age prediction. The RNA-seq gene expression data from fibroblast cell lines of humans. The skin fibroblasts cells keep data and phenotypic changes which are age-dependent are also contained in DNA methylation changes with age. This knowledge is used to create predictive models from biological datasets (RNA-seq and DNA methylation) to make age prediction. The CpGs sites with the highest correlation between CpGs sites and age are used to create predictive models.

Create data processing pipeline

We can see that this RNA-seq dataset is high-dimensional. There are over 27,000 columns/features. Generally, not all the features in the dataset are useful for prediction. We need only those features which increase the predictive ability of the model. To filter these features, we perform feature selection and retain only those which are useful. To do that, we use **SelectBest** module. This approach involves extracting those features which are most correlated to the target (age in our dataset). **F-regression** is used for the extraction of features. Moreover, we are not sure of how many of these features we will need. To find the right number of features, we do a hyperparameter search (to find the best combination of values of different parameters). It works by setting many different numbers of features to determine to number which gives the highest accuracy. To wrap this feature selector with a regressor, we will use the **Pipeline builder** tool. This tool creates a sequential flow of algorithms to execute on datasets. Since the hyperparameters will be tuned, we choose to output the parameters for searchCV. The tool does not take any dataset as input. Rather, the outputs will be used as inputs to the **Hyperparameter search** tool (explained in the following step). We will use **ElasticNet** as a regressor which creates an age prediction model. It is a linear regressor with **l1** (also called lasso) and **l2** (also called ridge) as regularisers. Regularisation is a technique used in machine learning to prevent overfitting. Overfitting happens when a machine learning algorithm starts memorising the dataset it is trained upon, rather than learning general features. The consequence of overfitting is that the accuracy on the training set is good, but results on the unseen set (test set) are poor, which happens because the algorithm has not learned general features from the dataset. To prevent overfitting, regularisers like **l1** and **l2** are used. **l1** is a linear term added to the error function of a machine learning algorithm and **l2** adds a squared term to the error function. More details about **l1** and **l2** can found [here](#).

Hands-on: Create pipeline

- Pipeline builder** with the following parameters:
 - In "Pre-processing step":
 - "Choose the type of transformation": Feature Selection
 - "Select a feature selection algorithm": SelectBest - Select features according to the k highest scores
 - "Select a score function": f_regression - Univariate linear regression tests
 - In "Final Estimator":
 - "Choose the module that contains target estimator": sklearn.linear_model
 - "Choose estimator class": ElasticNet
 - "Type in parameter settings if different from default": random_state=42
 - In "Output parameters for searchCV?": Yes

Optimise hyperparameters

In any machine learning algorithm, there are many parameters (hyperparameters). We are not sure which values of these parameters will give an optimal prediction. The default values given for these parameters may not be optimal for different datasets. To find the best combination of the values of different parameters, we use the **Hyperparameter search** tool. This tool searches for the best combination of parameters for a given dataset. We will use **ElasticNet** as a regressor which creates an age prediction model. It is a linear regressor with **l1** (also called lasso) and **l2** (also called ridge) as regularisers. Regularisation is a technique used in machine learning to prevent overfitting. Overfitting happens when a machine learning algorithm starts memorising the dataset it is trained upon, rather than learning general features. The consequence of overfitting is that the accuracy on the training set is good, but results on the unseen set (test set) are poor, which happens because the algorithm has not learned general features from the dataset. To prevent overfitting, regularisers like **l1** and **l2** are used. **l1** is a linear term added to the error function of a machine learning algorithm and **l2** adds a squared term to the error function. More details about **l1** and **l2** can found [here](#).

OPEN CHAT

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

← 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

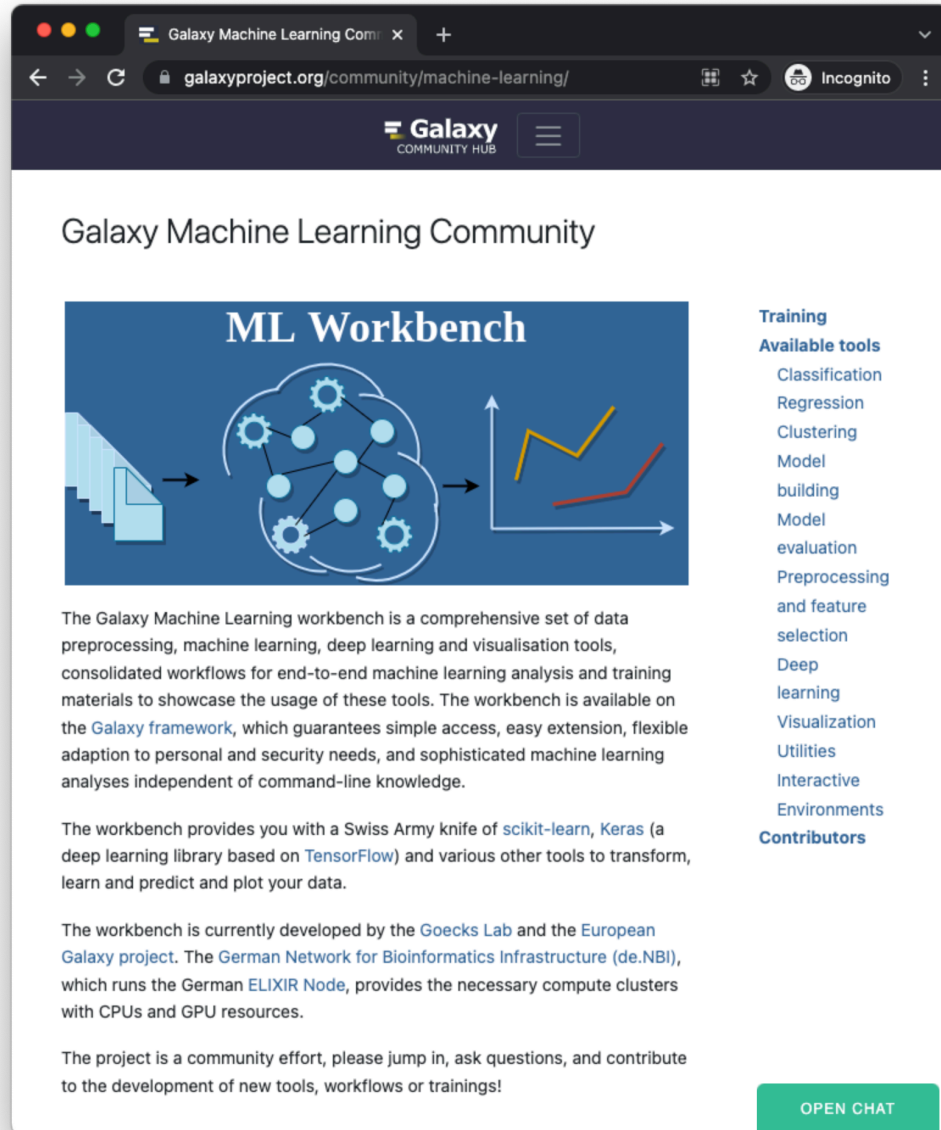
Outline

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



Galaxy Machine Learning Community

ML Workbench

The Galaxy Machine Learning workbench is a comprehensive set of data preprocessing, machine learning, deep learning and visualisation tools, consolidated workflows for end-to-end machine learning analysis and training materials to showcase the usage of these tools. The workbench is available on the **Galaxy framework**, which guarantees simple access, easy extension, flexible adaption to personal and security needs, and sophisticated machine learning analyses independent of command-line knowledge.

The workbench provides you with a Swiss Army knife of **scikit-learn**, **Keras** (a deep learning library based on **TensorFlow**) and various other tools to transform, learn and predict and plot your data.

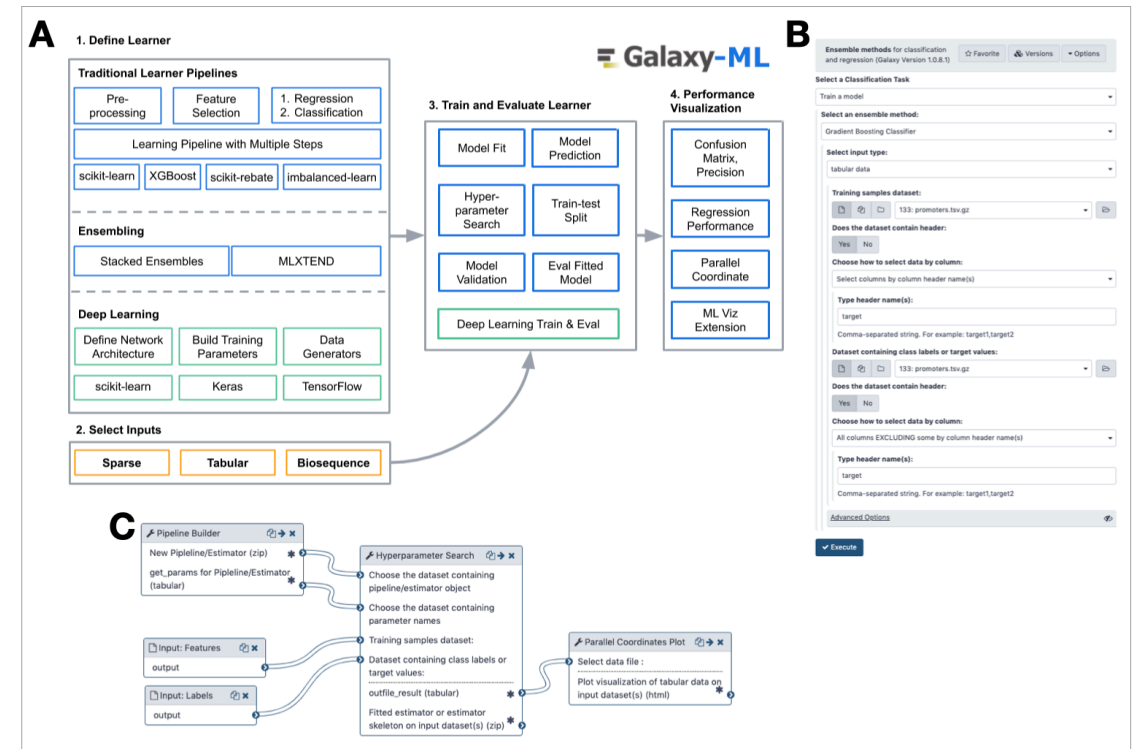
The workbench is currently developed by the **Goecks Lab** and the **European Galaxy project**. The **German Network for Bioinformatics Infrastructure (de.NBI)**, which runs the **German ELIXIR Node**, provides the necessary compute clusters with CPUs and GPU resources.

The project is a community effort, please jump in, ask questions, and contribute to the development of new tools, workflows or trainings!

Training Available tools

- Classification
- Regression
- Clustering
- Model building
- Model evaluation
- Preprocessing and feature selection
- Deep learning
- Visualization
- Utilities
- Interactive Environments
- Contributors

[OPEN CHAT](#)

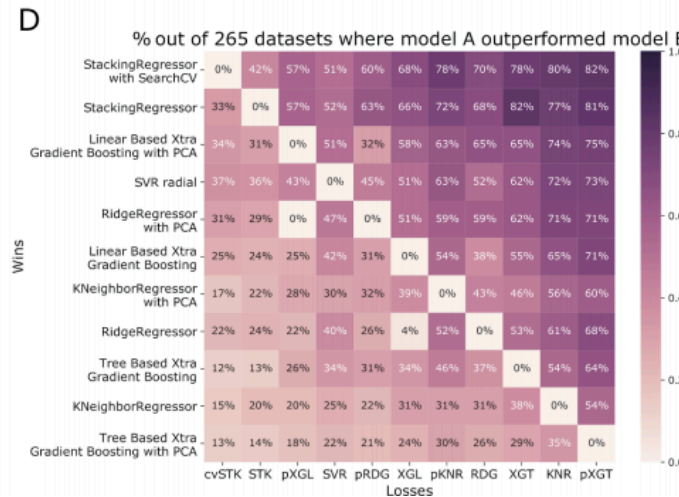
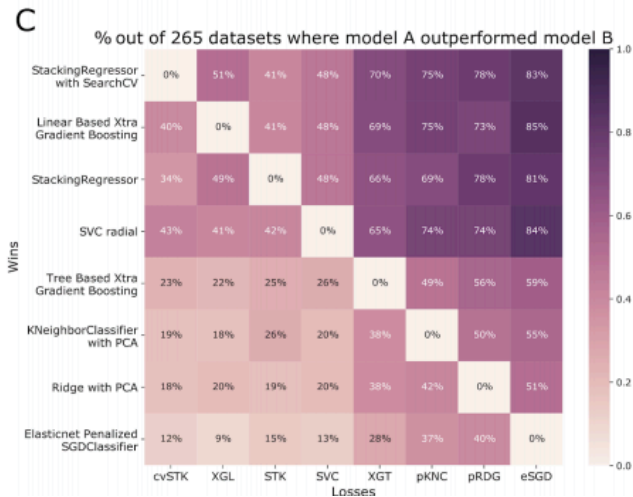
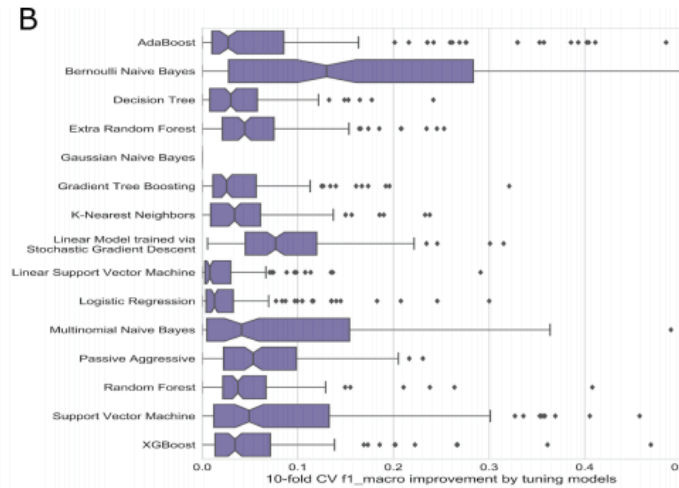
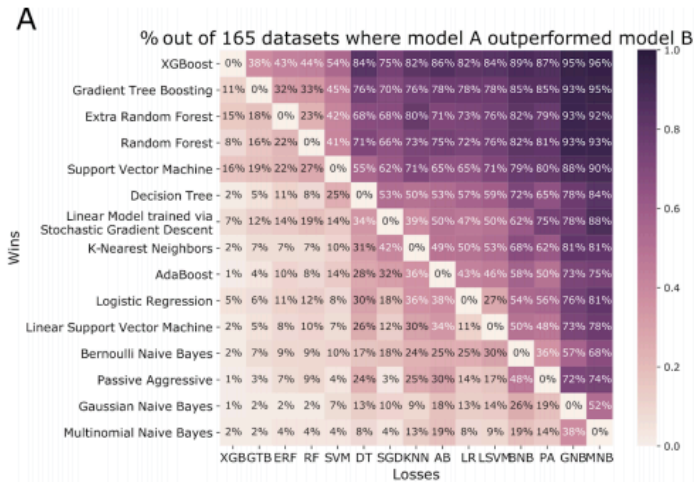


Software Library	Applications
Scikit-learn [18]	Various approaches for preprocessing, modeling, ensembling, and evaluation
Scikit-rebate [19]	Feature selection
Imbalanced-learn [20]	Approaches for working with imbalanced datasets
XGBoost [21]	Modeling using high-performance gradient boosting with decision 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>

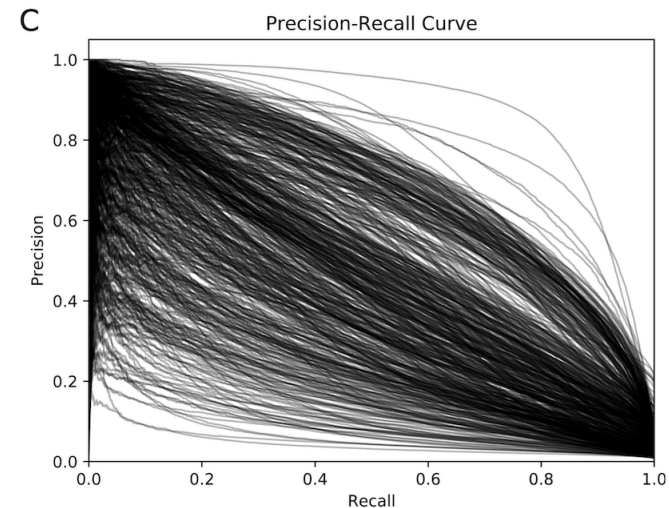
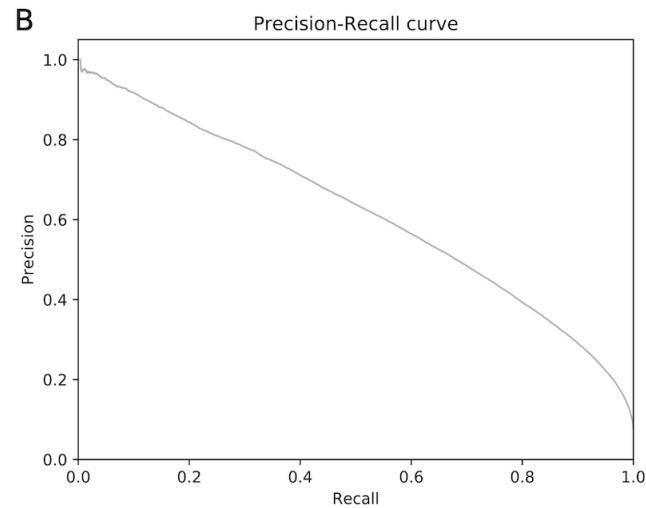
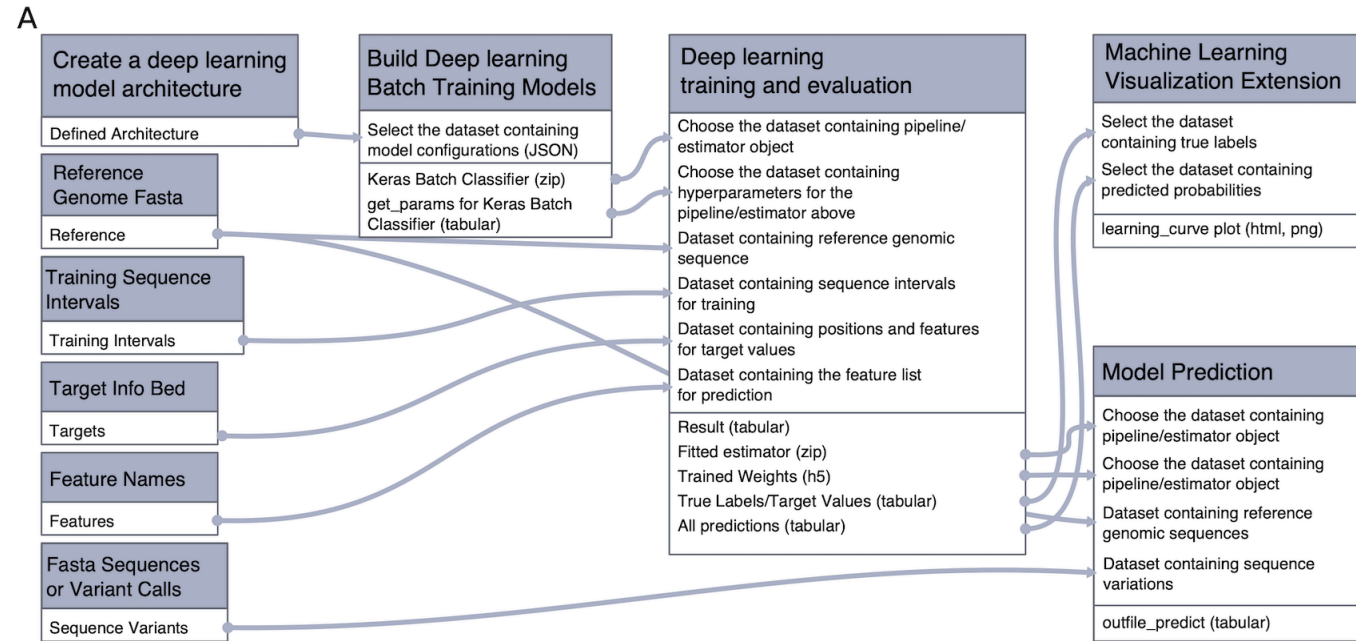
Galaxy-ML: An accessible, reproducible, and scalable machine learning toolkit for biomedicine
 Gu et al. (2021) *PLOS Comp Biology*. <https://doi.org/10.1371/journal.pcbi.1009014>

Scalable and Reproducible Machine Learning



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

Deep Learning in Galaxy



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

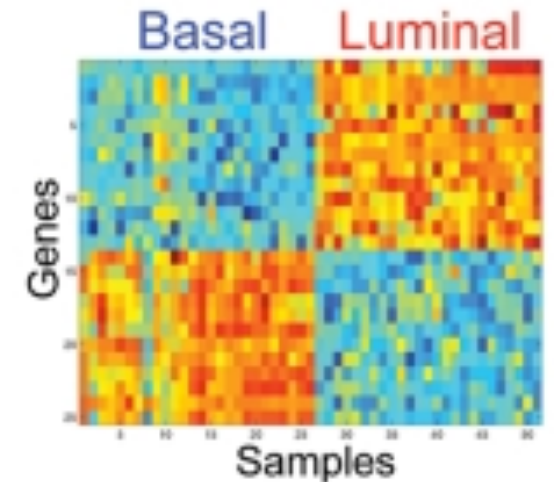
DNA alterations alone not sufficient for precision cancer therapy

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

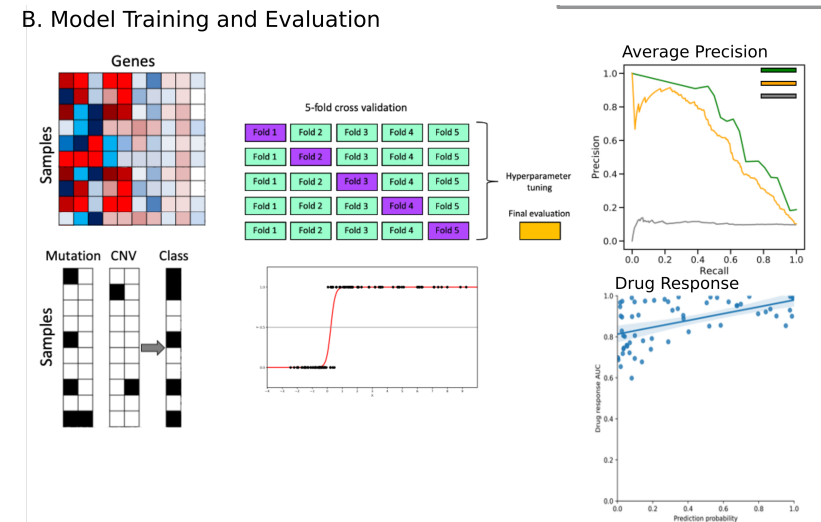
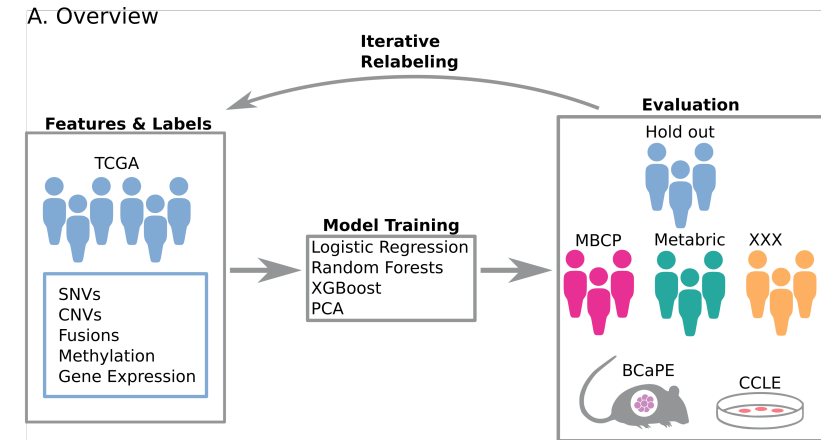
- “What does a *TP53* mutation or *CDKN2A* loss look like at the transcriptional level?”
- Inputs: RNA-seq, Labels/predictions: DNA alterations

Very useful:

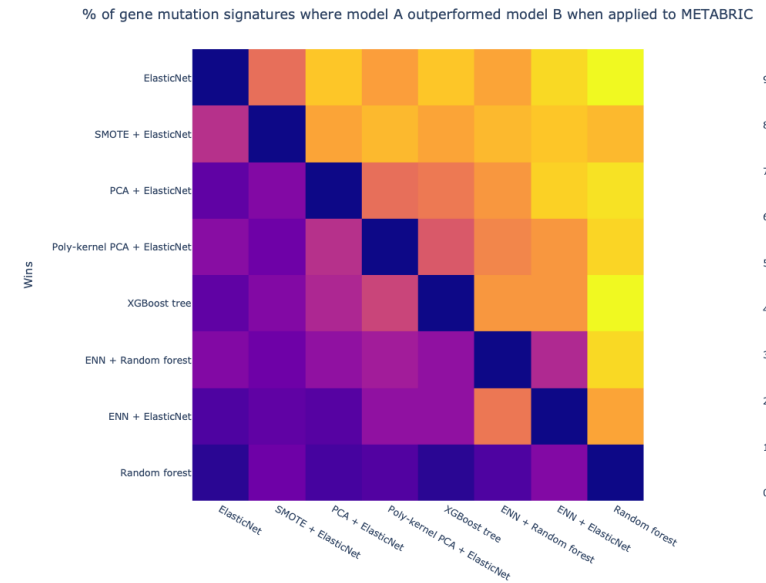
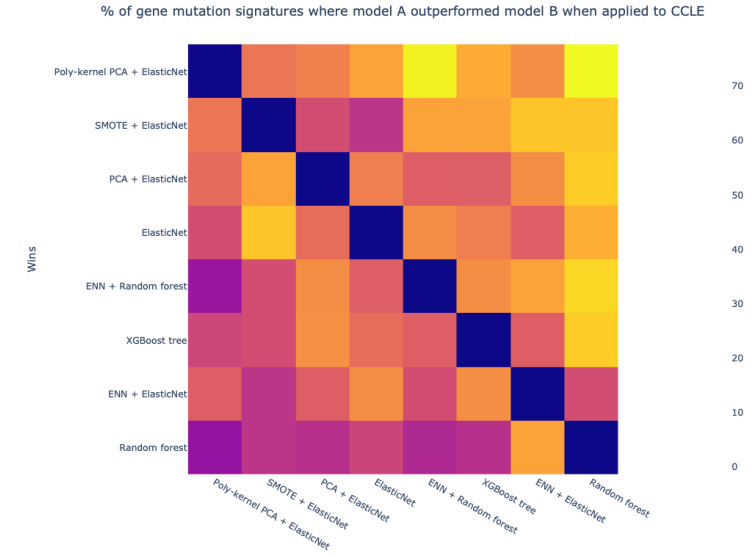
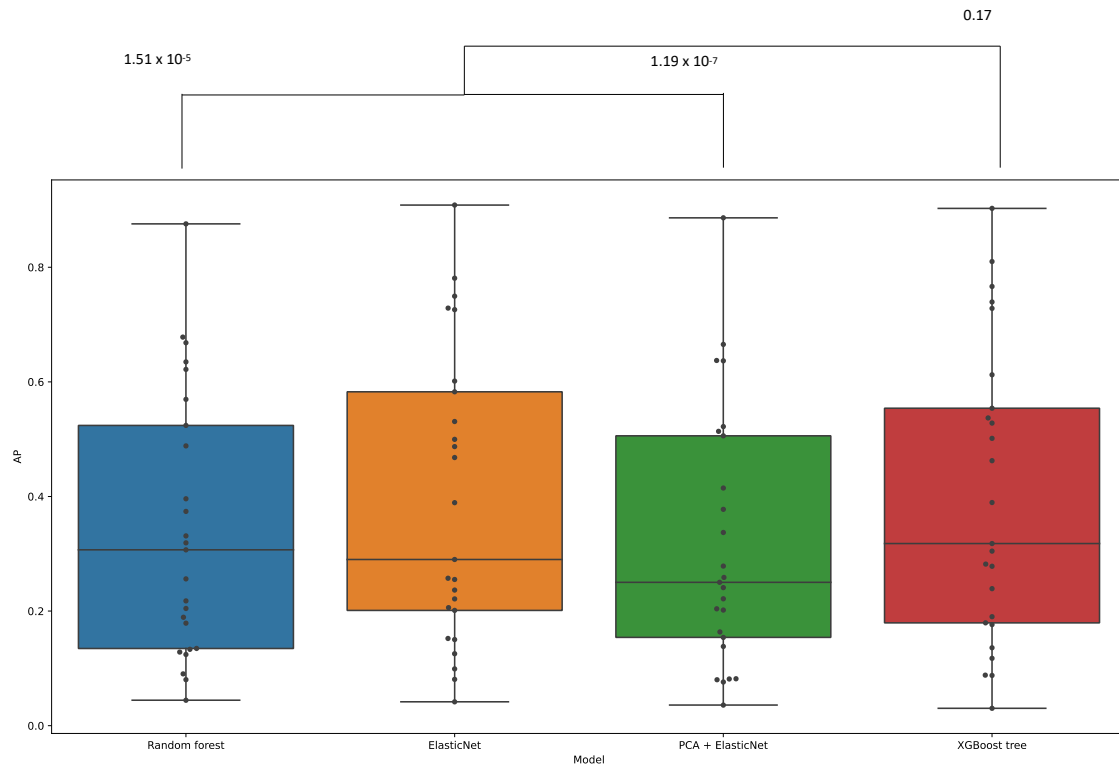
- Transcriptional state is dynamic and more closely reflects tumor activity than DNA (e.g., “hidden responders”)
- Interpretable

Can learn using large cohorts without drug response data

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



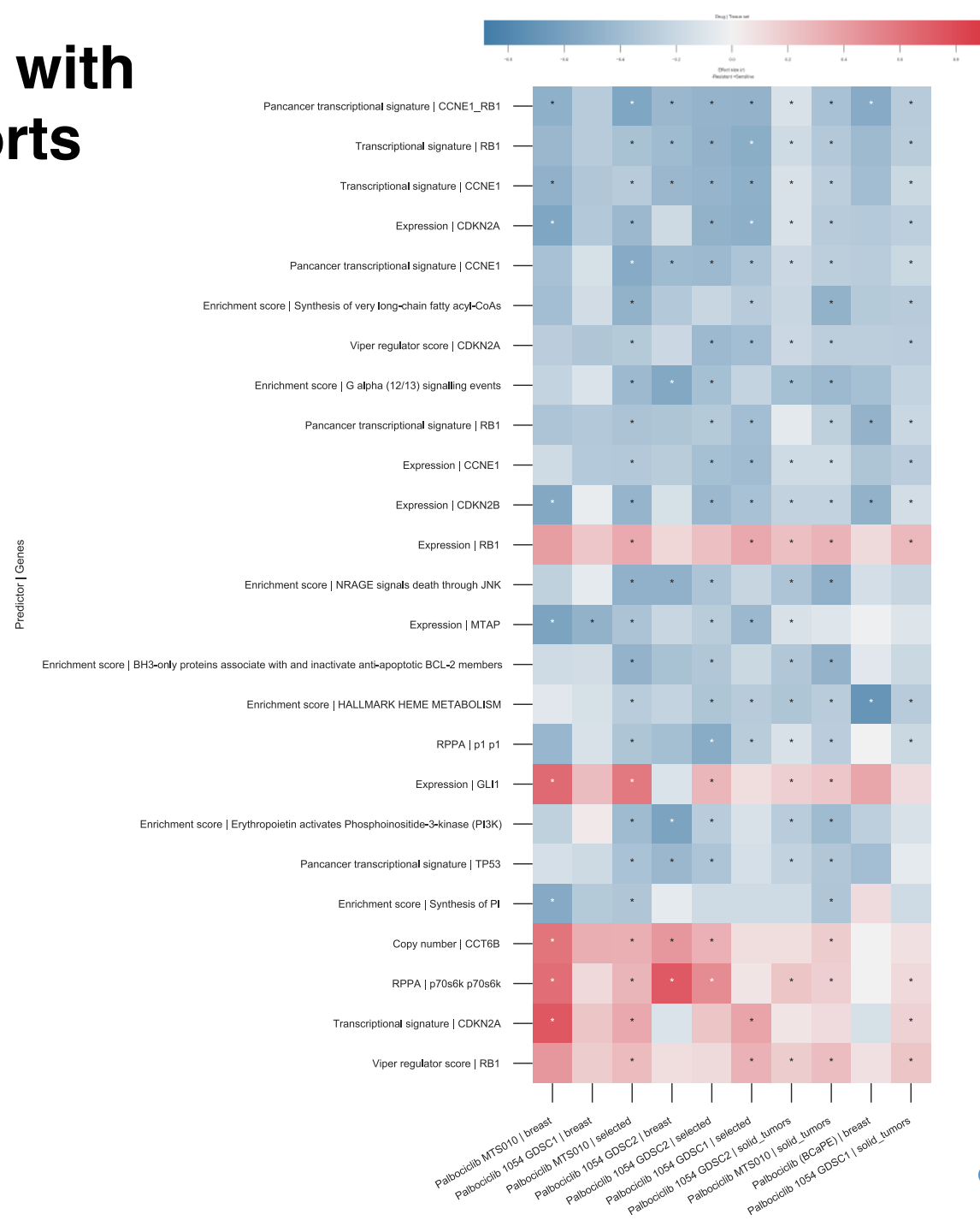
Multiple Modeling Approaches Work and Models Transfer to Other Cohorts



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



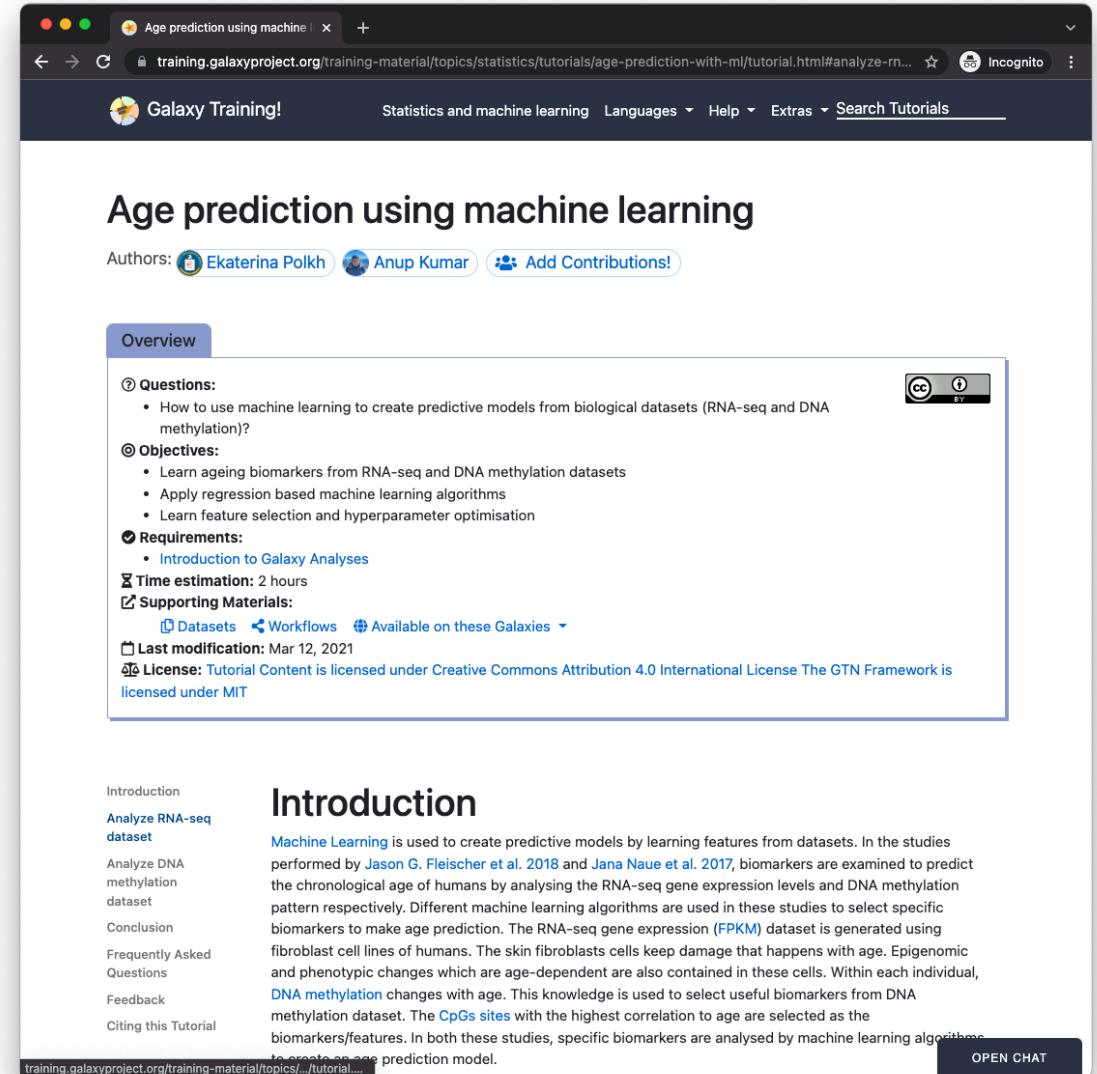
The Power of Adding Tools to the Galaxy Ecosystem

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



The screenshot shows a web browser window displaying a Galaxy Training tutorial. The page title is "Age prediction using machine learning" and it lists authors Ekaterina Polkh and Anup Kumar. The "Overview" section includes a list of questions, objectives, and requirements. The "Introduction" section discusses the use of machine learning for age prediction based on RNA-seq and DNA methylation data. The page also includes a table of contents on the left and an "OPEN CHAT" button at the bottom right.

Age prediction using machine learning

Authors: Ekaterina Polkh Anup Kumar Add Contributions!

Overview

Questions:

- How to use machine learning to create predictive models from biological datasets (RNA-seq and DNA methylation)?

Objectives:

- Learn ageing biomarkers from RNA-seq and DNA methylation datasets
- Apply regression based machine learning algorithms
- Learn feature selection and hyperparameter optimisation

Requirements:

- [Introduction to Galaxy Analyses](#)

Time estimation: 2 hours

Supporting Materials:

- [Datasets](#) [Workflows](#) [Available on these Galaxies](#)

Last modification: Mar 12, 2021

License: Tutorial Content is licensed under Creative Commons Attribution 4.0 International License The GTN Framework is licensed under MIT

Introduction

Introduction

Machine Learning is used to create predictive models by learning features from datasets. In the studies performed by Jason G. Fleischer et al. 2018 and Jana Naue et al. 2017, biomarkers are examined to predict the chronological age of humans by analysing the RNA-seq gene expression levels and DNA methylation pattern respectively. Different machine learning algorithms are used in these studies to select specific biomarkers to make age prediction. The RNA-seq gene expression (FPKM) dataset is generated using fibroblast cell lines of humans. The skin fibroblasts cells keep damage that happens with age. Epigenomic and phenotypic changes which are age-dependent are also contained in these cells. Within each individual, DNA methylation changes with age. This knowledge is used to select useful biomarkers from DNA methylation dataset. The CpGs sites with the highest correlation to age are selected as the biomarkers/features. In both these studies, specific biomarkers are analysed by machine learning algorithms to create an age prediction model.

OPEN CHAT

Outline

The Galaxy Platform

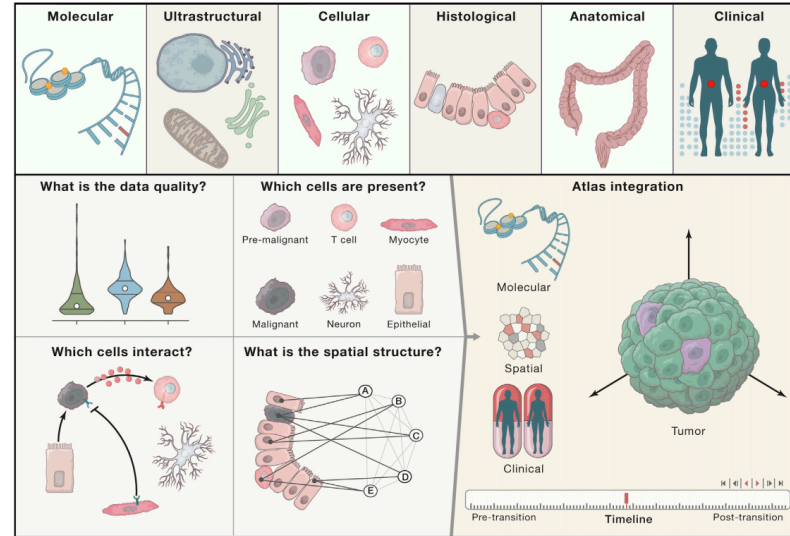
Machine Learning Applications for Cancer

An Interactive Hub for Multiplex Tissue Image Analysis

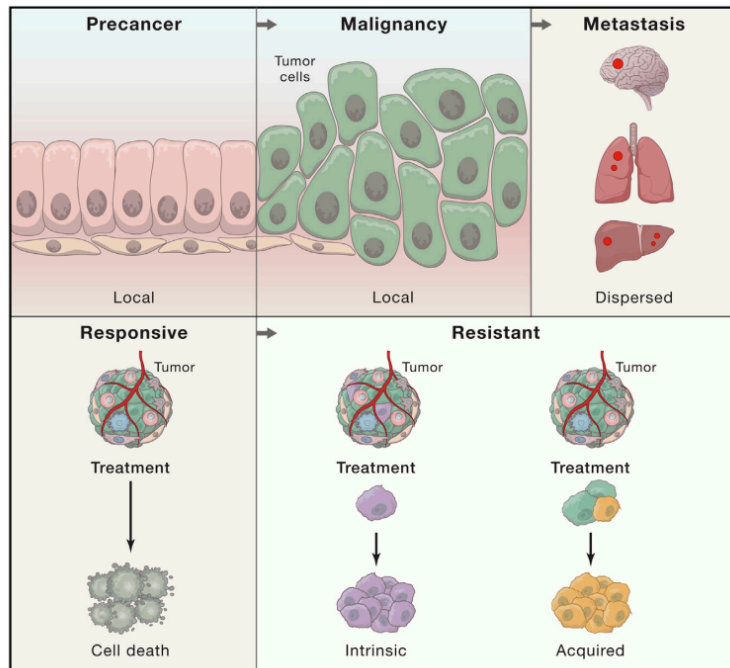
Cancer Moonshot NCI Human Tumor Atlas Network

<https://humantumoratlas.org/>

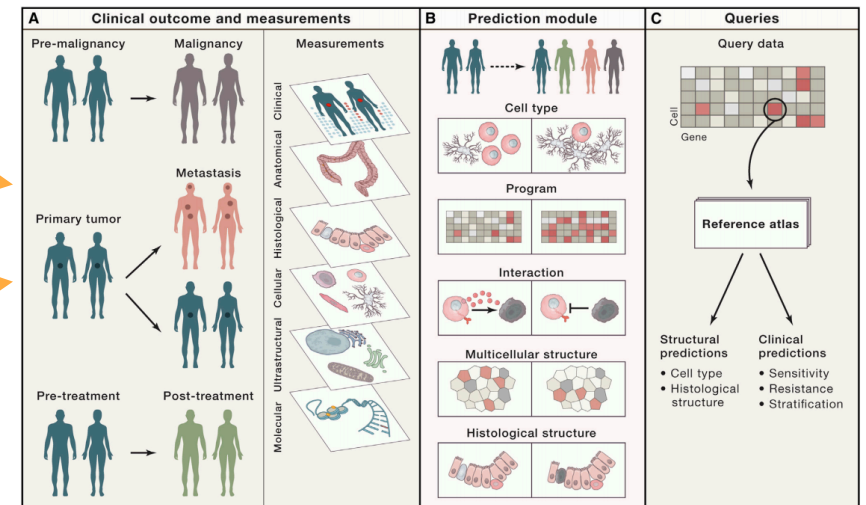
Deep Profiling



Studying key transitions

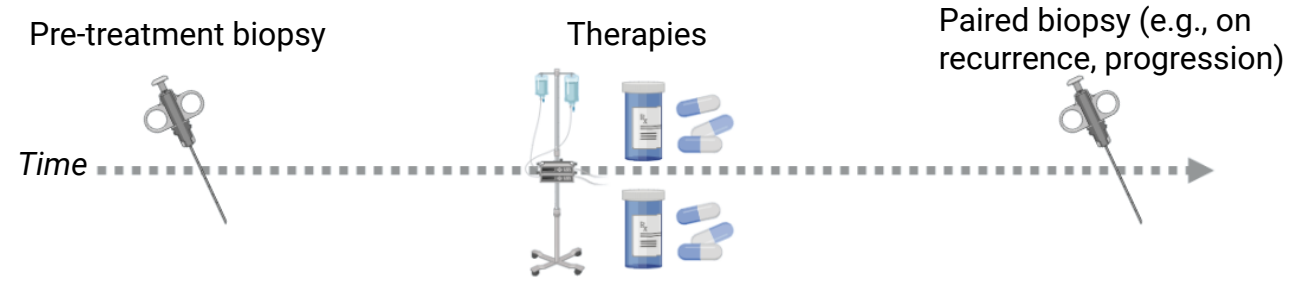


Systems Biology Understanding and Clinical Predictors



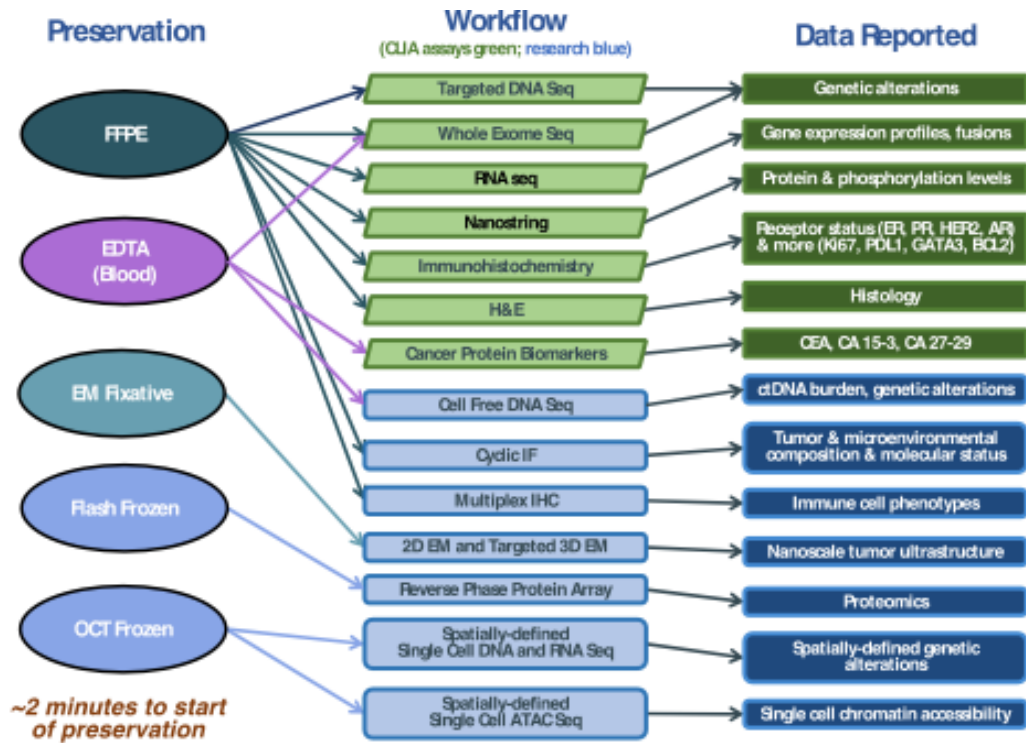
Omics and Multidimensional Spatial Atlas

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



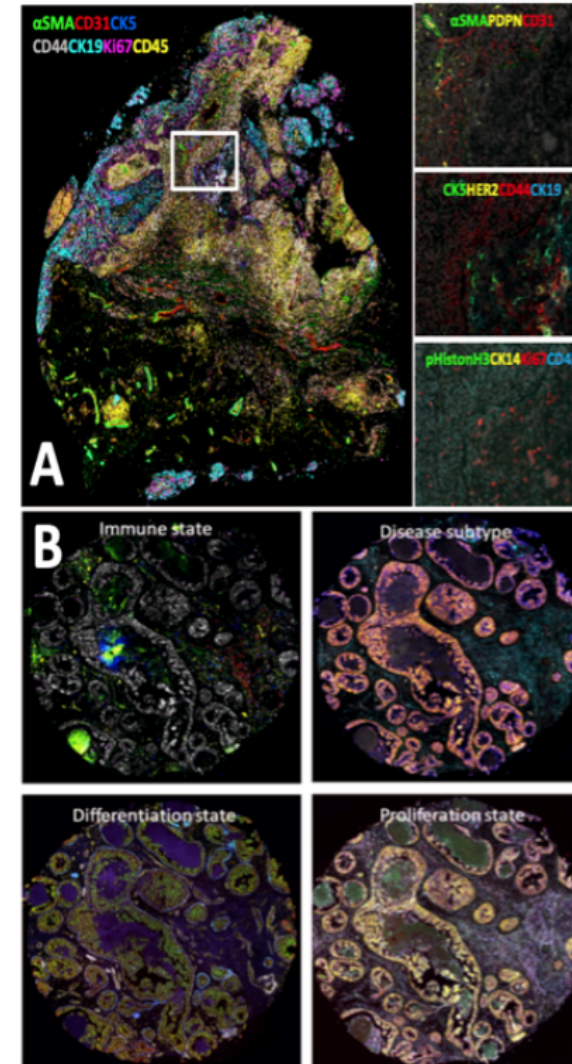
Multiplex Tissue Imaging

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

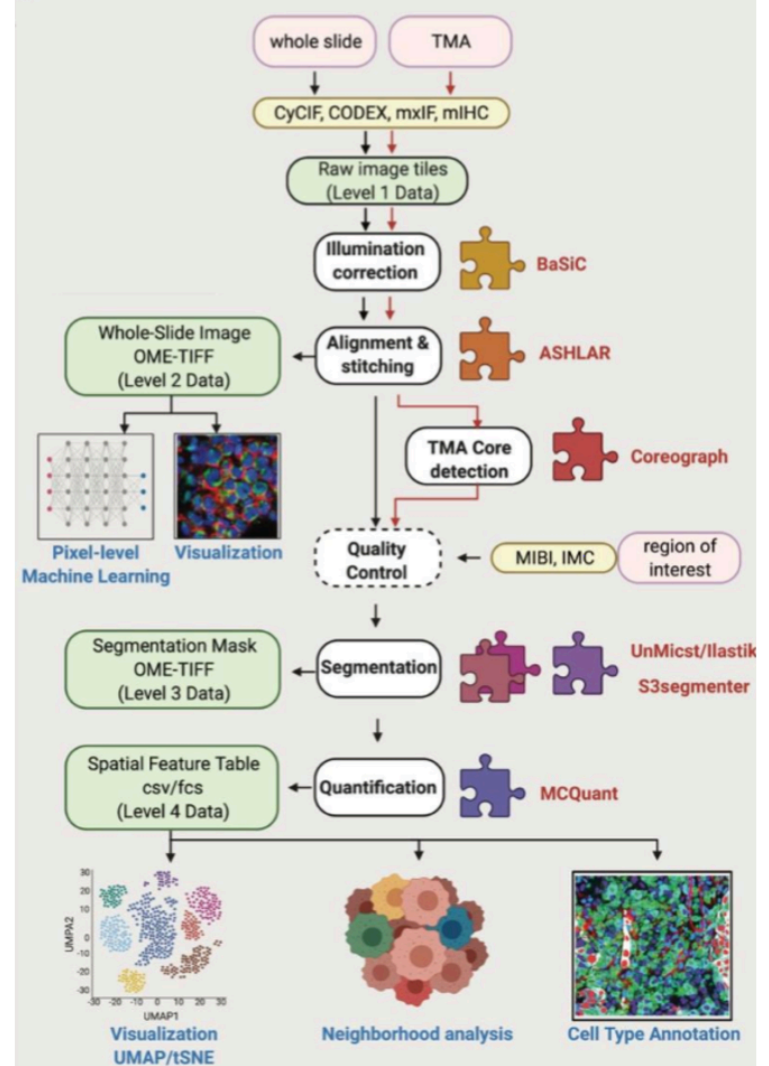
Analysis of Multiple Tissue Imaging Datasets

Datasets are image stacks that are tens-hundreds of gigabytes

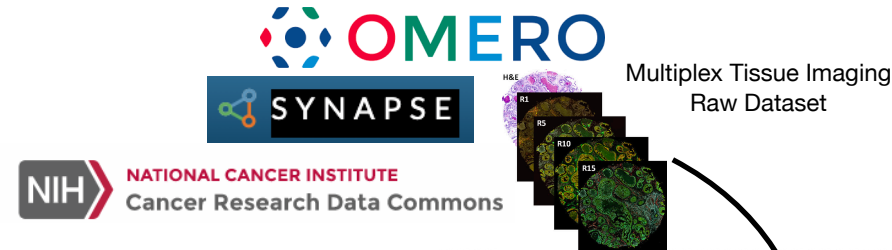
10-20 analysis steps are needed for fully processing image stacks

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



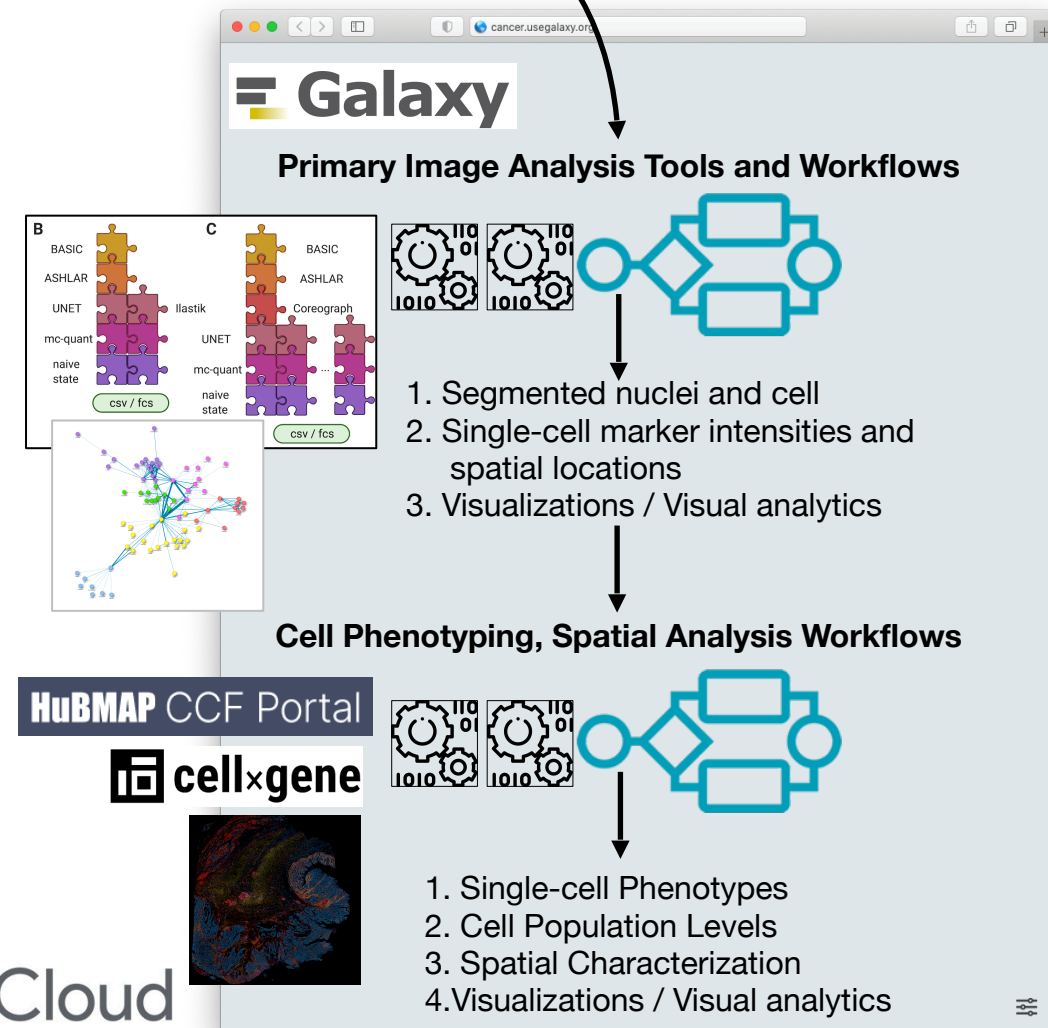
A Web-based portal for interactive data analysis and visualization

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

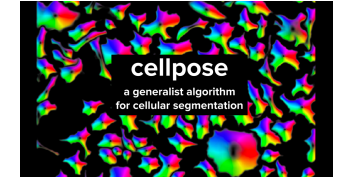


Goal: “End-to-end” Multiplex Tissue Imaging Analysis and Visualization

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



Schapiro, Santagata, and Sorger labs



Pachitariu lab

2. Downstream analysis—cell phenotyping and spatial analyses



Dr. Ajit Johnson Nirmal, Sorger lab



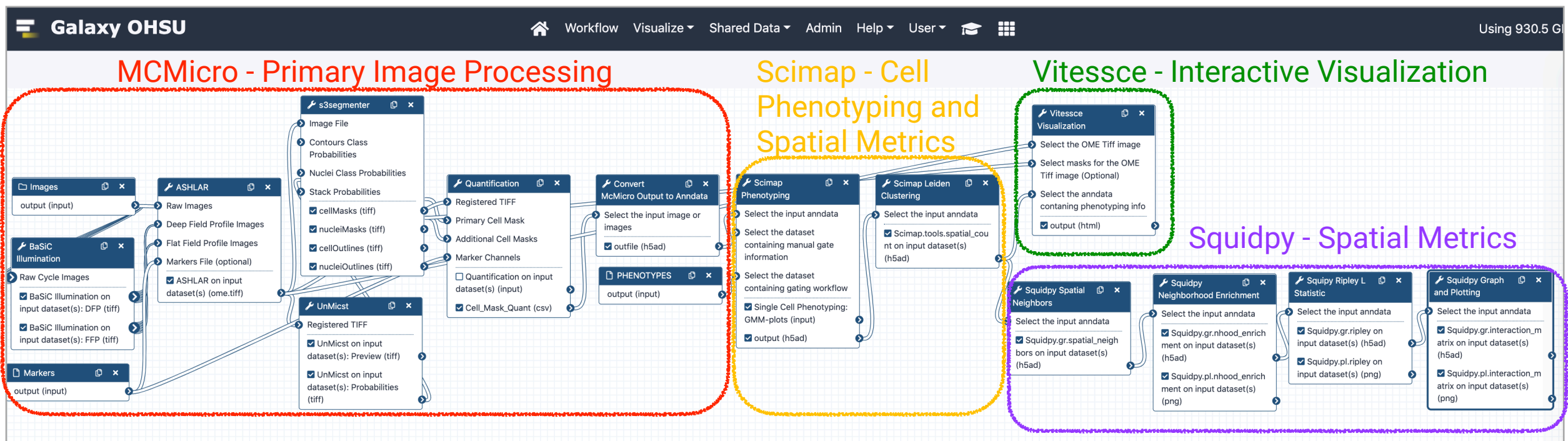
This lab

3. Visualization



Gehlenborg lab / HuBMAP

End-to-end Multiplex Tissue Imaging Workflow

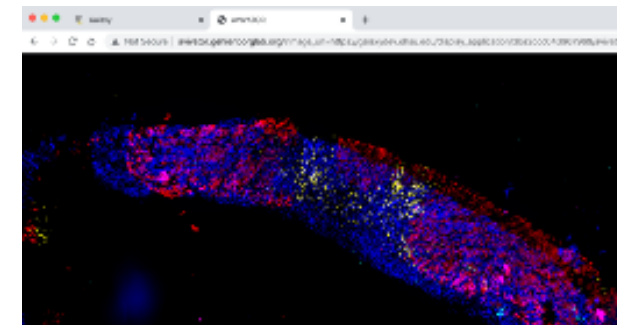
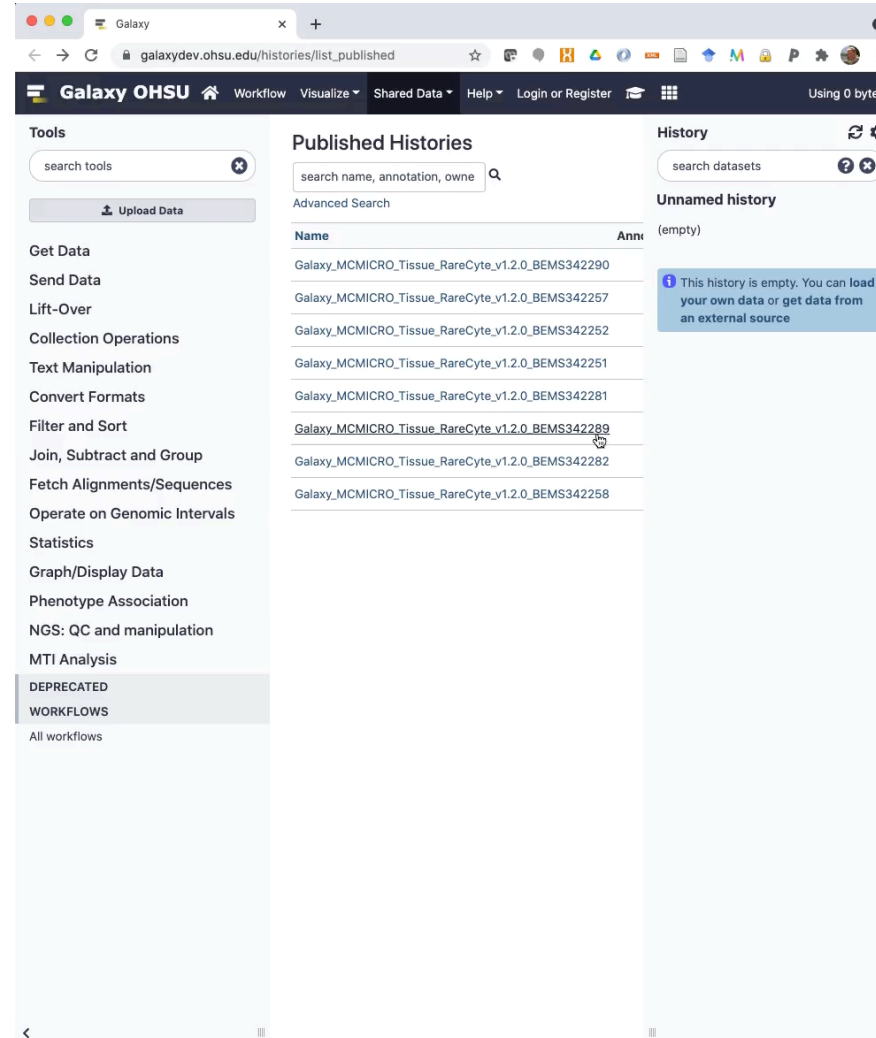


Registered Image Viewing with Viv

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



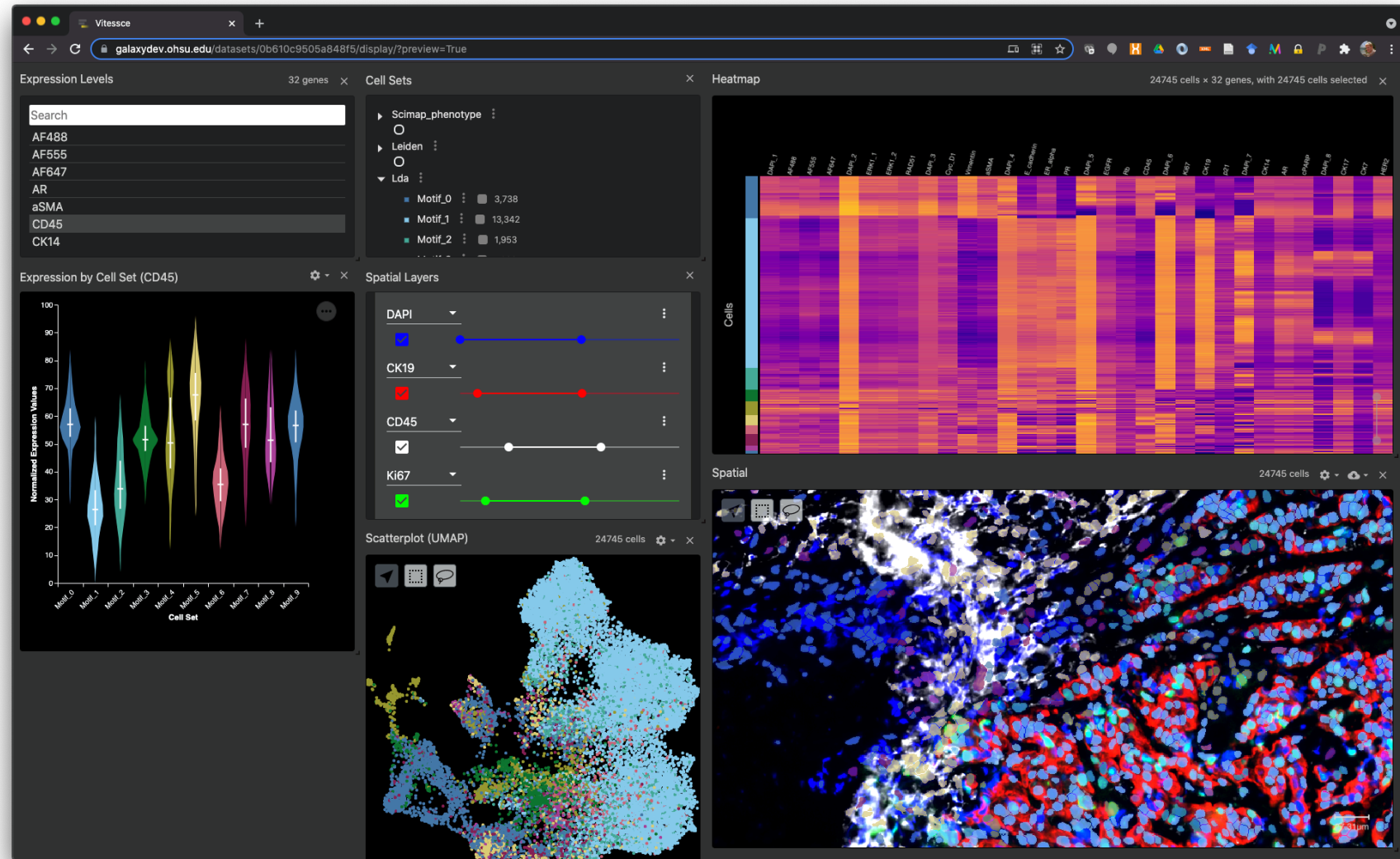
Viv: Manz et al., <https://doi.org/10.31219/osf.io/wd2gu>
<http://viv.gehlenborglab.org/>

Phenotypes + Spatial Layout with Vitesse

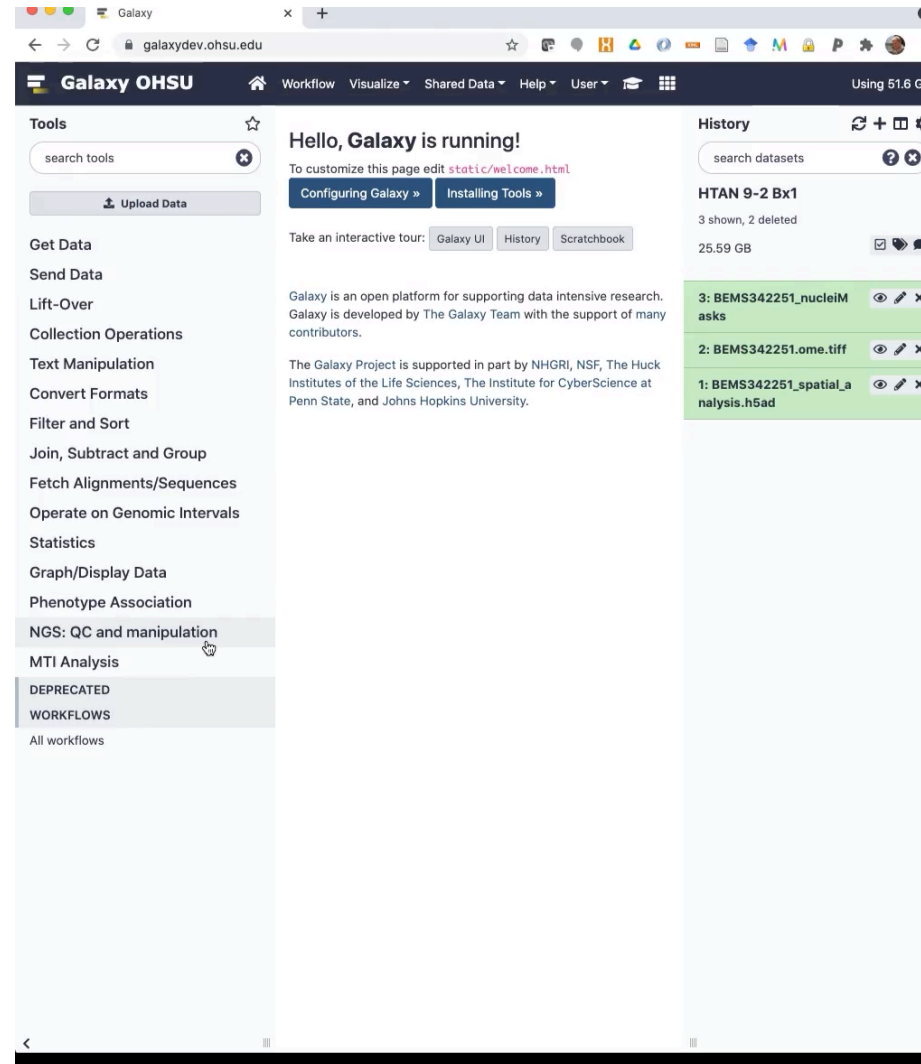
Three ways to identify phenotypes: gated, leiden, and LDA

Overlay cell states with raw image + channels

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



Creating a Vitessece Dashboard with Galaxy



Vitessece: <http://vitessece.io/>

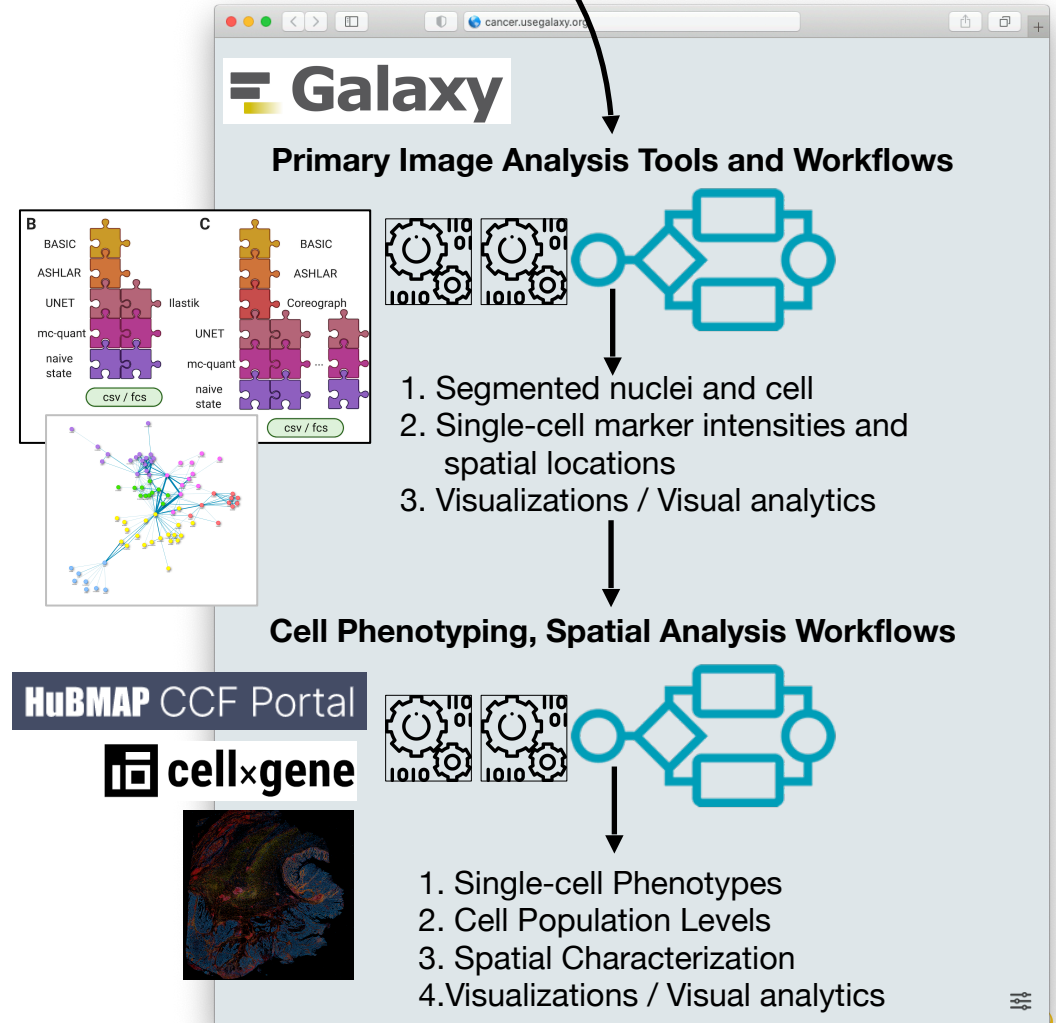
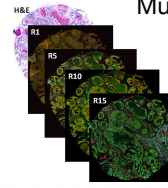
Enabling Fast and Collaborative Computational Science

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

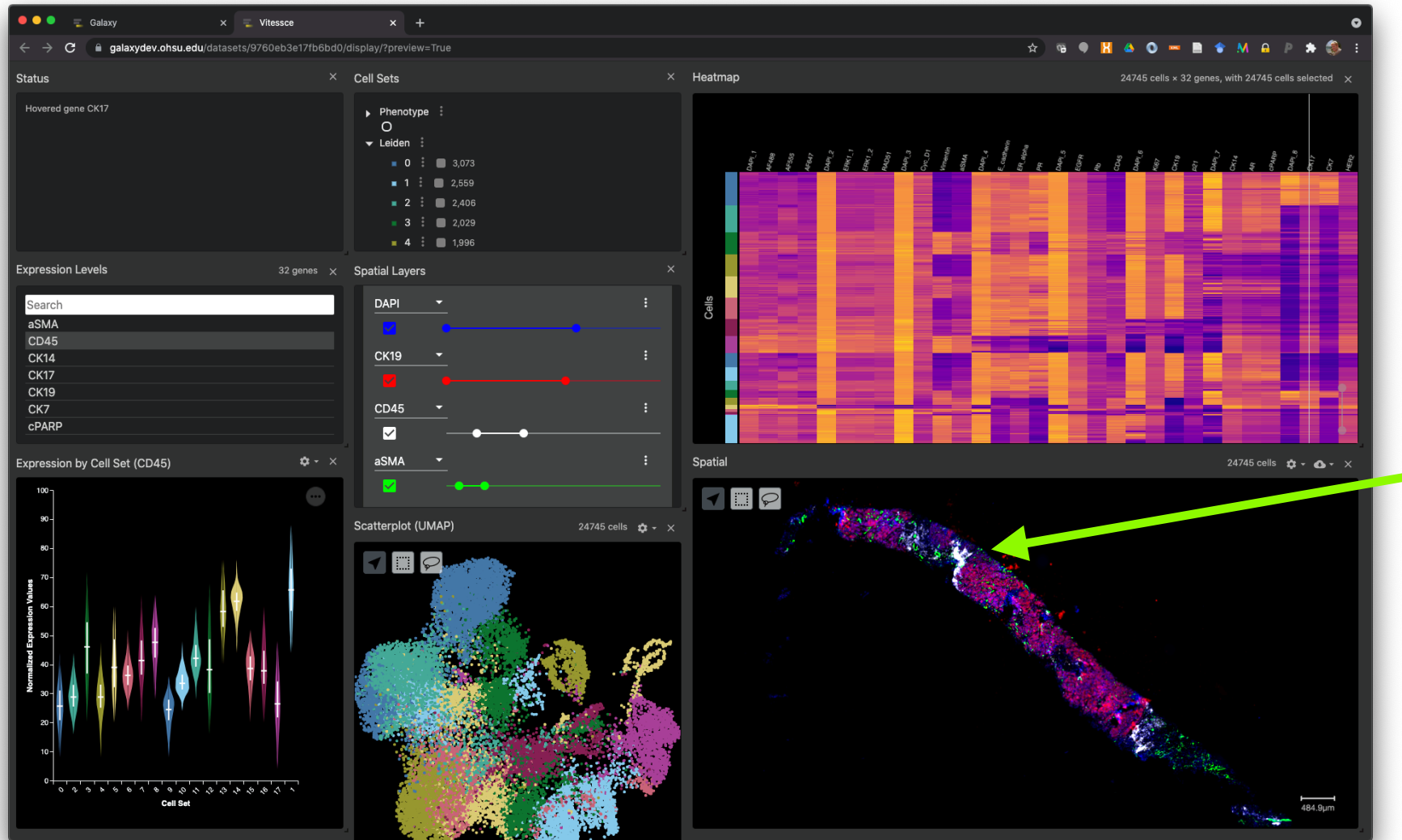
Collaborative: Results are available to everyone via a Web browser with no downloads, which is especially valuable for distributed teams



Multiplex Tissue Imaging
Raw Dataset



Observing an Immune Aggregate

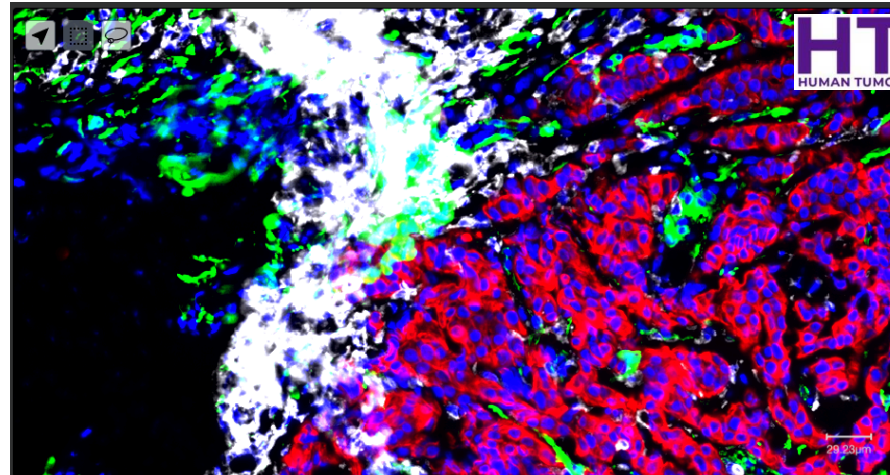


Comparing Cell State Calling Approaches

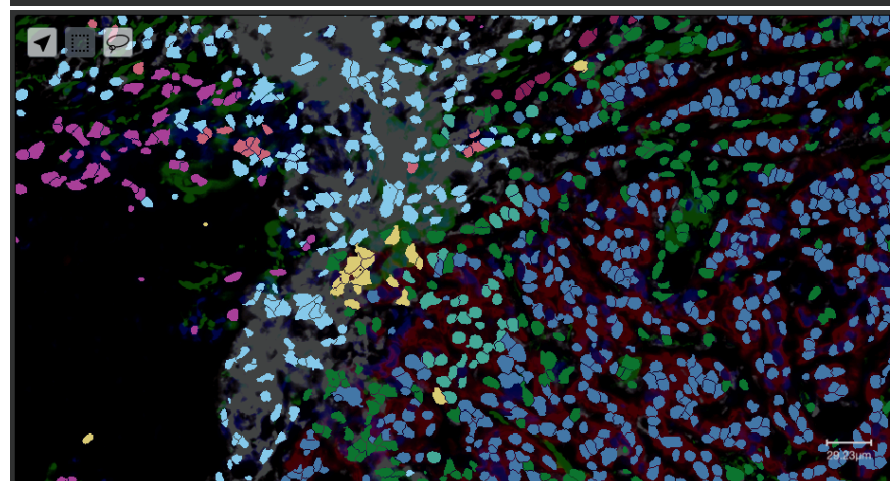
Visually similar cell states calls with Leiden and Gated approaches

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

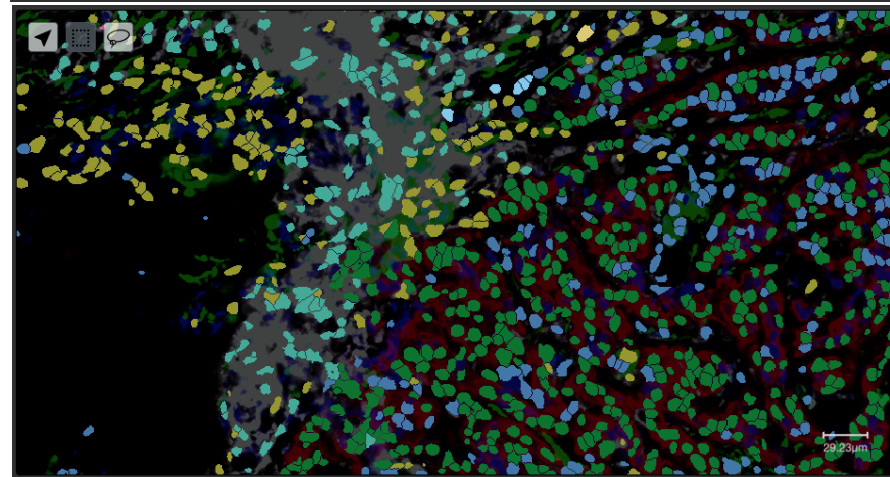
Stain



Leiden



Gated



Concluding Thoughts

Galaxy's unique aspects in my opinion:

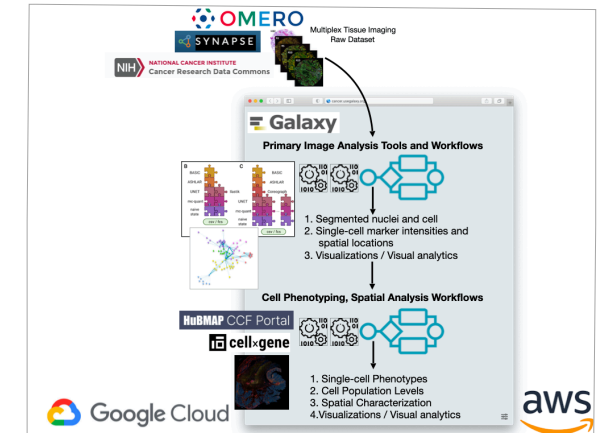
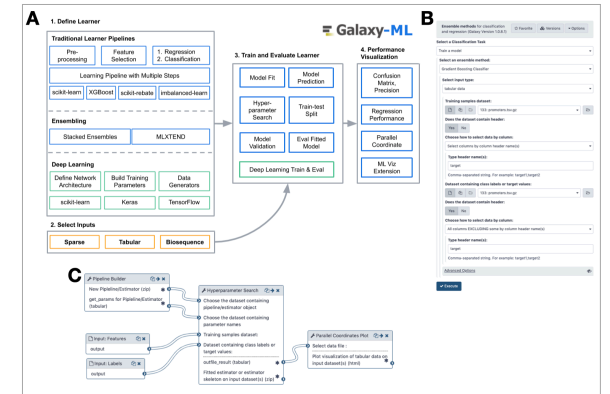
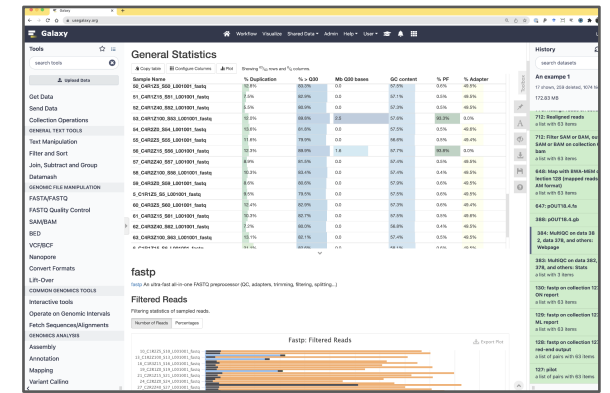
- ▶ Web-based UI, *but this is changing*
- ▶ Integrative framework for putting things together—the whole is greater than the sum of the parts
- ▶ World-wide community

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



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



SMMART
Prospect Creek Foundation



NATIONAL CANCER INSTITUTE
Informatics Technology for
Cancer Research



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 Ruukonen
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



KNIGHT
CANCER
Institute



THE UNIVERSITY OF TEXAS
MDAnderson
Cancer Center



HTAN
HUMAN TUMOR ATLAS NETWORK

Thank You!