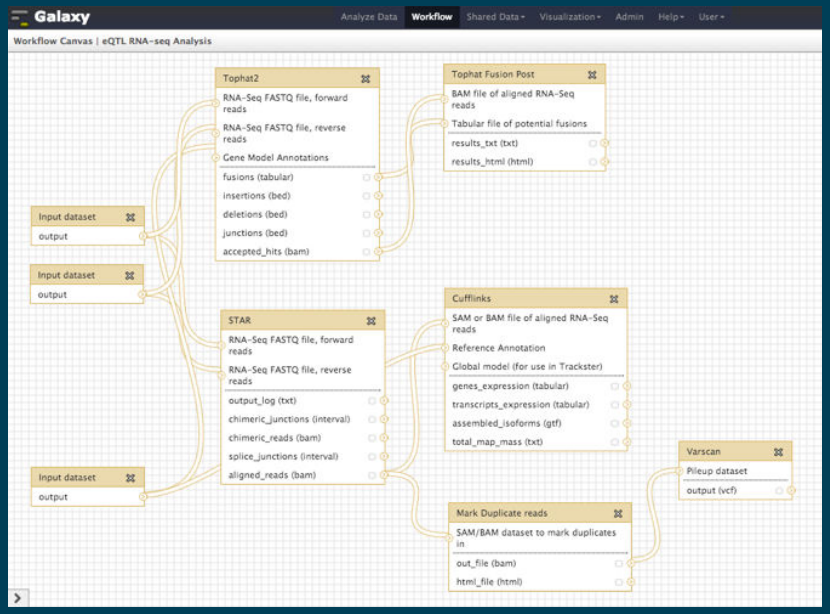
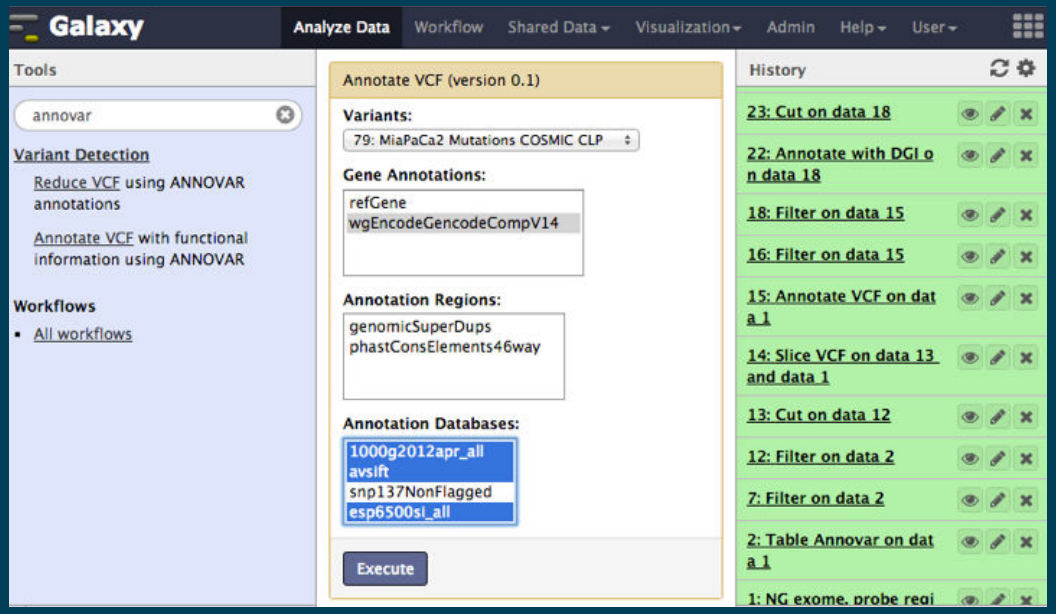




Creating and Using Tumor Genome Profiles with Galaxy

Jeremy Goecks

THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC

Topics

Galaxy

Cancer Genomics with Galaxy

My Perspective

I do computational biology/bioinformatics

I analyze all sorts of high-throughput sequencing data for many different purposes

- model, non-model
- molecular evolution, systematics, cancer

Analyzing sequence data is hard and difficult to automate

However, anyone can do it with the proper computational systems

What is the Cost of Analyzing NGS Data?

Musings

Highly accessed

The $1,000 genome, the $100,000 analysis?

Elaine R Mardis

Correspondence: Elaine R Mardis emardis@wustl.edu

Author Affiliations

The Genome Center at Washington University School of Medicine, 4444 Forest Park Blvd, St Louis, MO 63108, USA

Genome Medicine 2010, **2**:84 doi:10.1186/gm205

The electronic version of this article is the complete one and can be found online at: http://genomemedicine.com/content/2/11/84

Published: 26 November 2010

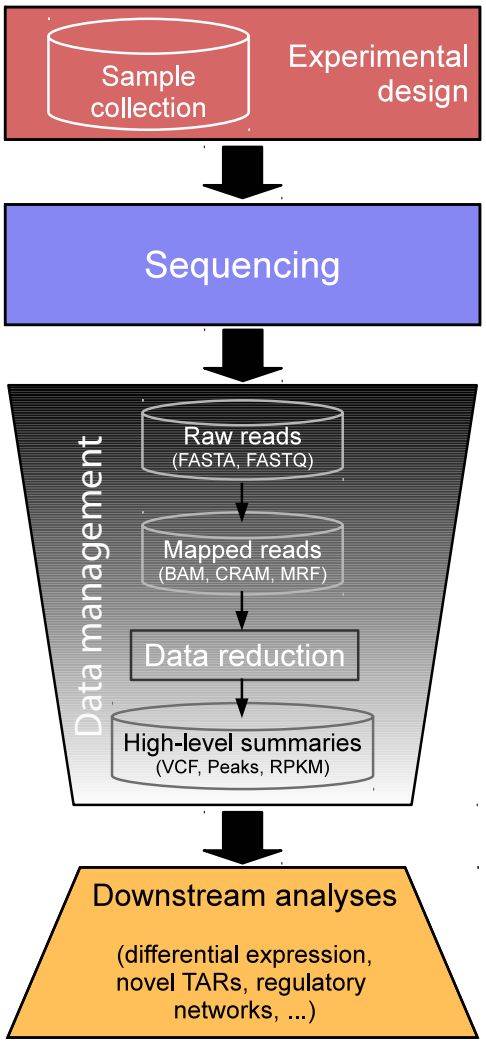
Sboner et al. Genome Biology 2011, 12:125
http://genomebiology.com/2011/12/8/125

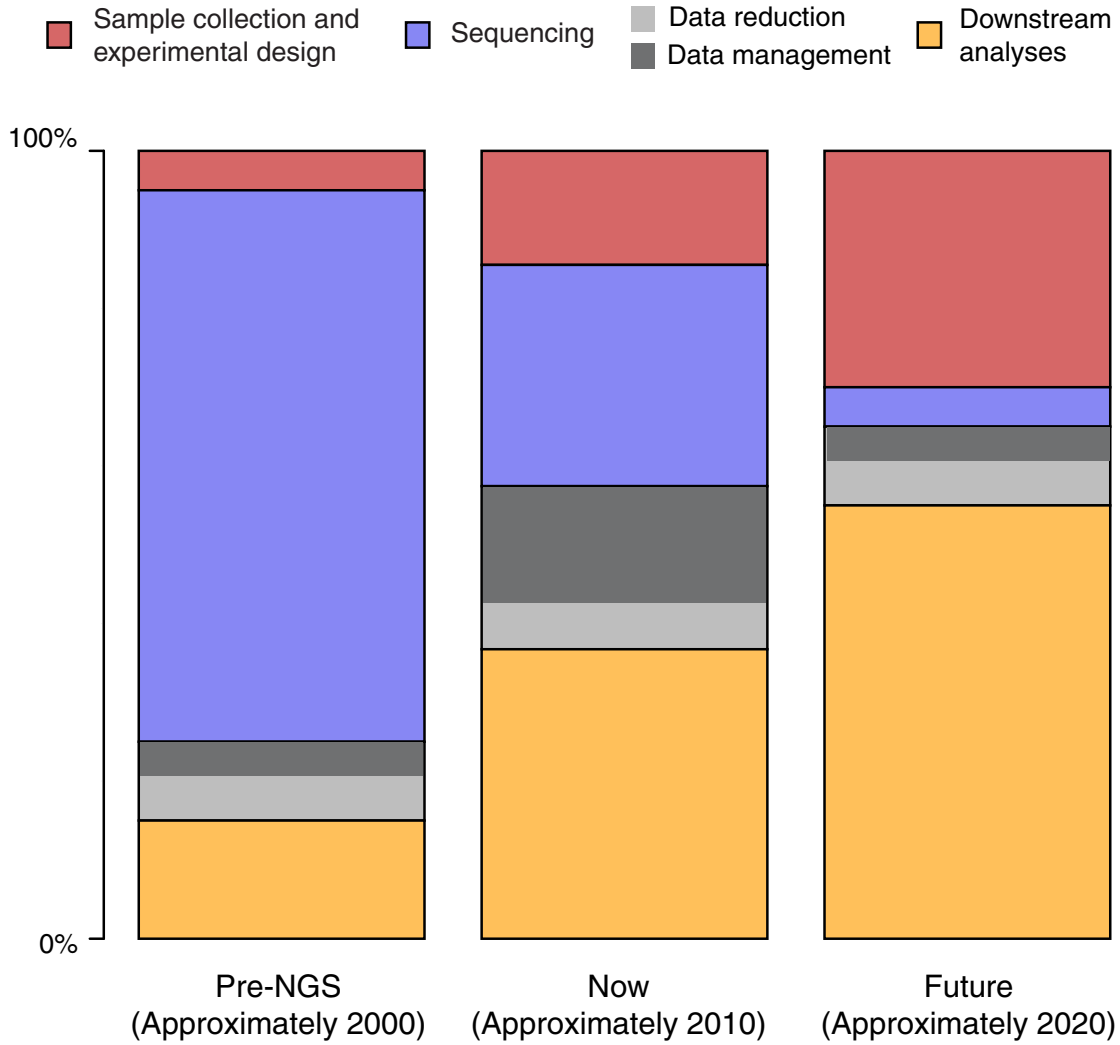
Genome Biology

OPINION

The real cost of sequencing: higher than you think!

Andrea Sboner[1,2], Xinmeng Jasmine Mu[1], Dov Greenbaum[1,2,3,4,5], Raymond K Auerbach[1] and Mark B Gerstein*[1,2,6]





Galaxy Project: Fundamental Questions

When genomics (or any biomedical science) becomes dependent on computational methods, how to:

- make tools and workflows **accessible** to scientists?
- ensure that analyses are **reproducible**?
- enable transparent **communication and reuse** of analyses?

Vision

Galaxy is an **open, Web-based platform** for accessible, reproducible, and collaborative computational genomics

Galaxy Demo

Galaxy

tachylite01.bx.mathcs.emory.edu/g/jeremy/root

Galaxy | Analyze Data | Workflow | Shared Data | Visualization | Admin | Help | User | Using 2.1 TB

Tools

search tools

Get Data
Send Data
ENCODE Tools
Lift-Over
Text Manipulation
Filter and Sort
Join, Subtract and Group
Convert Formats
Extract Features
Fetch Sequences
Fetch Alignments
Get Genomic Scores
Operate on Genomic Intervals
Statistics
Wavelet Analysis
Graph/Display Data
Regional Variation
Multiple regression
Multivariate Analysis
Evolution
Motif Tools
Multiple Alignments
Metagenomic analyses
FASTA manipulation
NGS: QC and manipulation
NGS: GATK Tools (beta)
NGS: Mapping
NGS: Indel Analysis
NGS: RNA Analysis

Hello world! It's running...

To customize this page edit `static/welcome.html`

This project is supported in part by NSF, NHGRI, and the Huck Institutes of the Life Sciences.

History

PanCan P3
9.3 GB

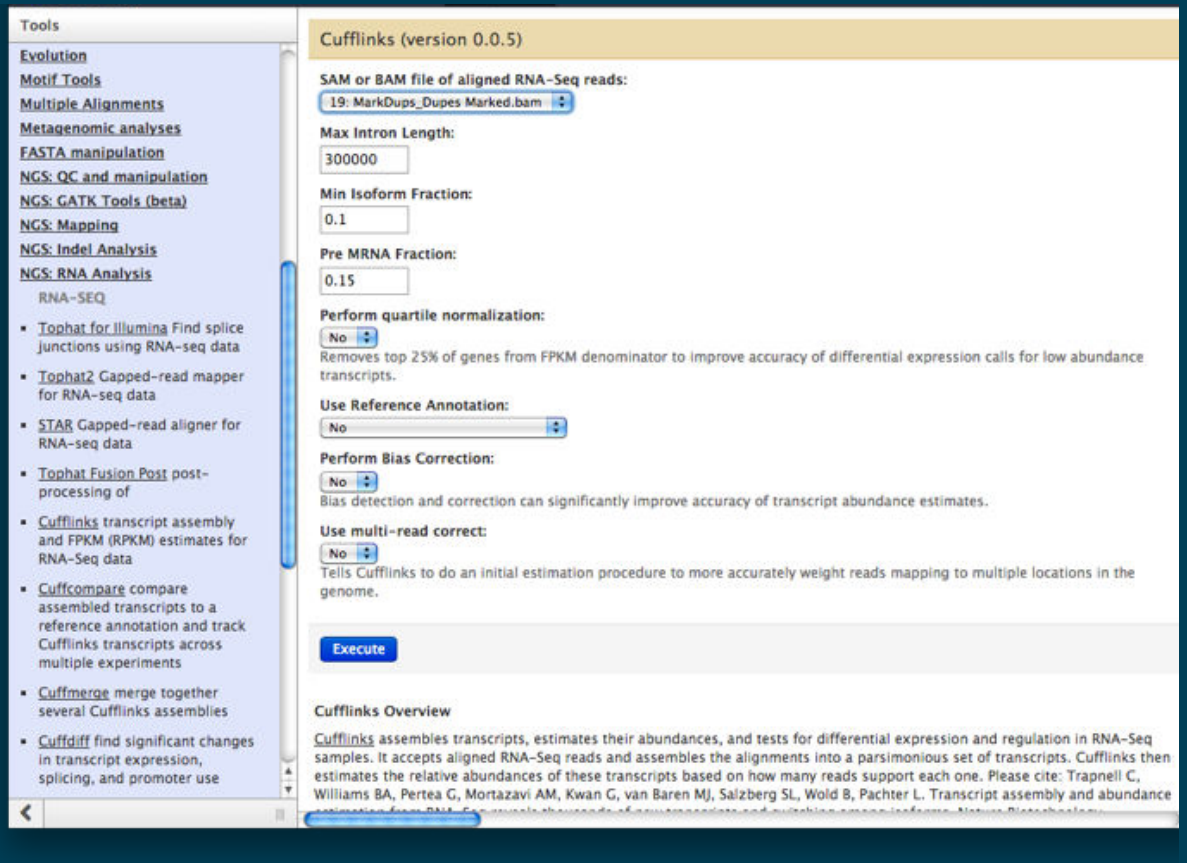
22: Tophat Fusion Post on data 10 and data 14: html results
21: Tophat Fusion Post on data 10 and data 14: text results
20: MarkDups_Dupes Marked.html
19: MarkDups_Dupes Marked.bam
17: Cufflinks on data 9 and data 4: assembled transcripts
16: Cufflinks on data 9 and data 4: transcript expression
15: Cufflinks on data 9 and data 4: gene expression
14: Tophat2 on data 3, data 4, and data 2: accepted_hits
13: Tophat2 on data 3, data 4, and data 2: splice junctions
12: Tophat2 on data 3, data 4, and data 2: deletions
11: Tophat2 on data 3, data 4, and data 2: insertions

Accessibility

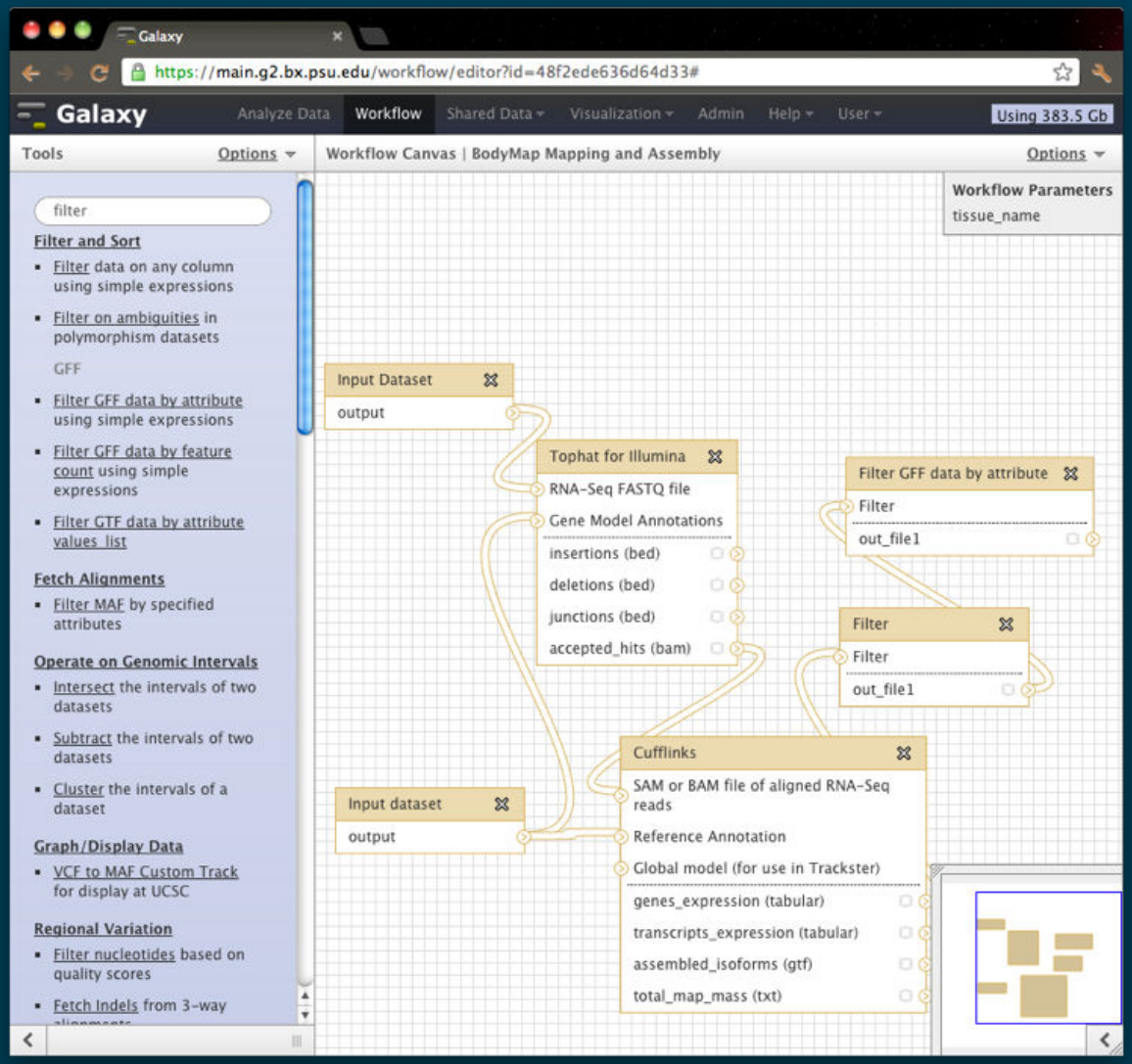
All tools looks the same

No command line or programming

Easy to chain tools together into larger analyses



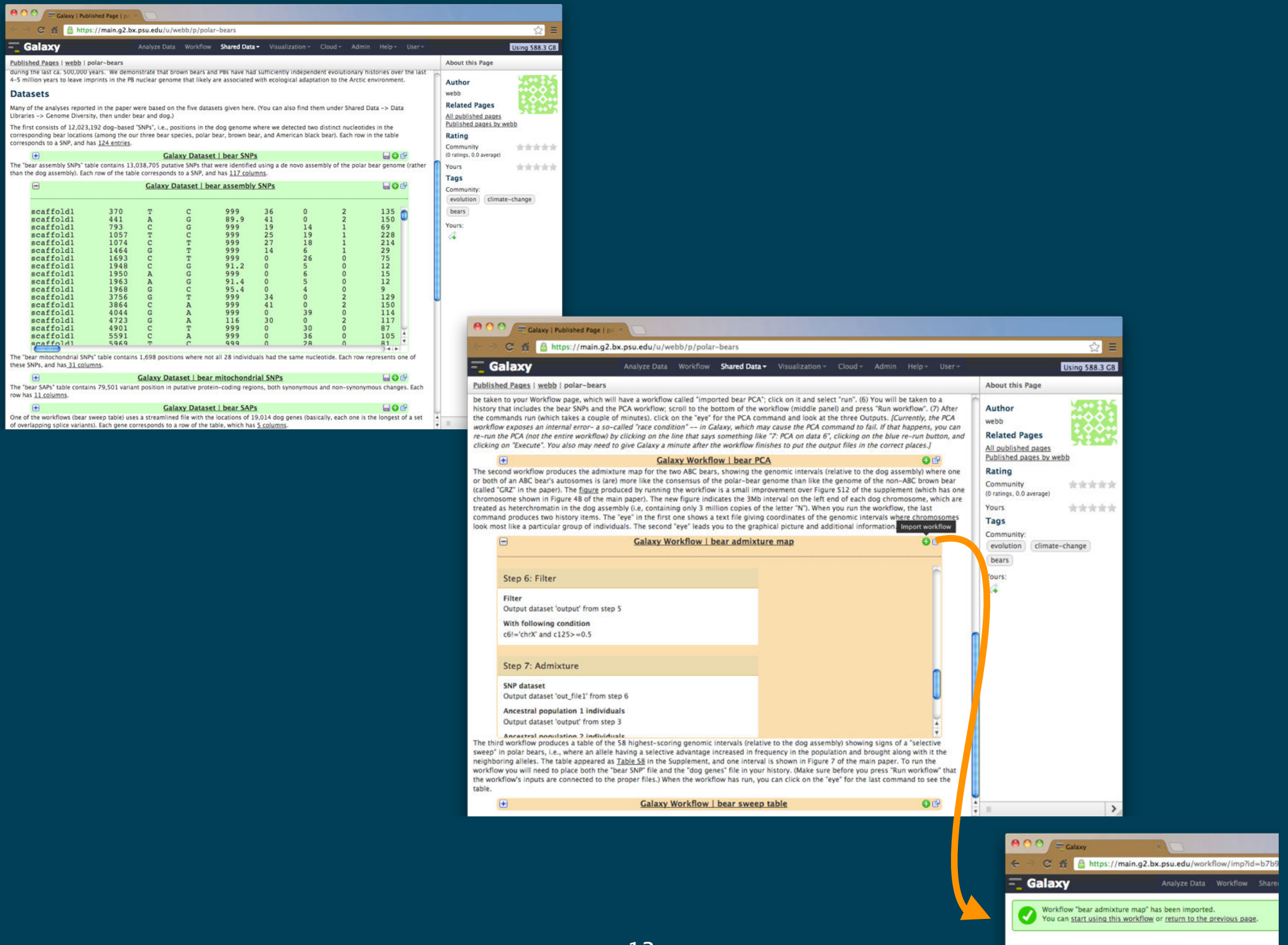
Reproducibility



Workflows enable reuse and precise reproducibility

Users can add tags and annotations for additional context

Communication and Reuse


https://main.g2.bx.psu.edu/u/webb/p/polar-bears
Galaxy
Published Pages | webb | polar-bears
Datasets
Galaxy Dataset | bear SNPs
Galaxy Dataset | bear assembly SNPs
Galaxy Dataset | bear mitochondrial SNPs
Galaxy Dataset | bear SAPs
About this Page
Author
webb
Related Pages
Rating
Tags
Galaxy Workflow | bear PCA
Galaxy Workflow | bear admixture map
Step 6: Filter
Step 7: Admixture
Galaxy Workflow | bear sweep table
Import workflow
Workflow "bear admixture map" has been imported.
You can start using this workflow or return to the previous page.

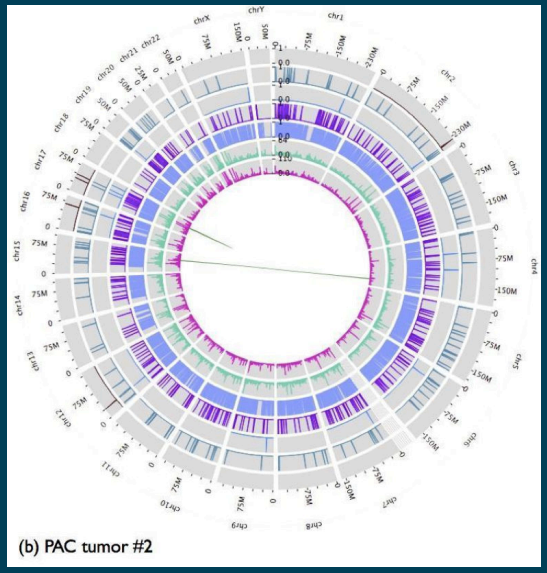
Galaxy Visualizations

Many visualizations

- biological: genome browser, Circos plot, phylogenetic tree
- numerical: bar charts, histograms, etc.

Developers can add new visualizations

Completely Web-base: no data or software downloads



What is Galaxy?

Platform for high-throughput genomics

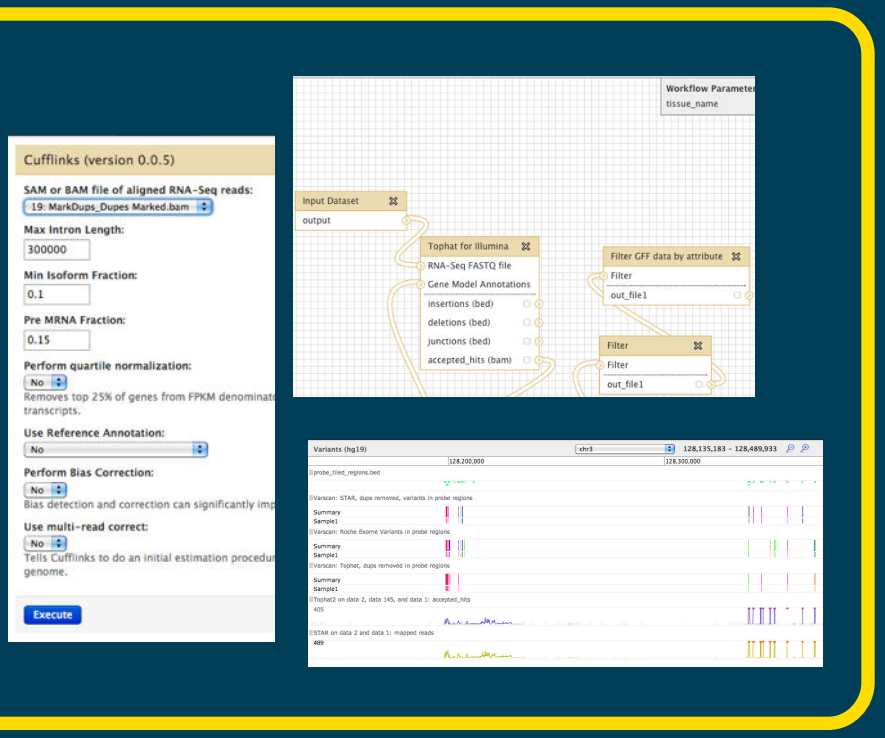
1. get and integrate public, private data
2. analyze data and create workflows
3. visualization, sharing, publication

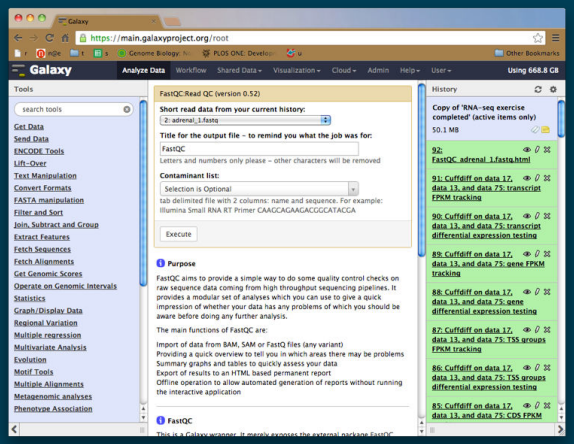
Customizable open-source software for various HPC resources

- public website — http://usegalaxy.org
- local instance
- on the cloud

Galaxy platform

- run tools, workflows on HPC resources
- minimizes data movement
- create workflows, visualizations, pages
- share *everything*





Galaxy is Very Popular

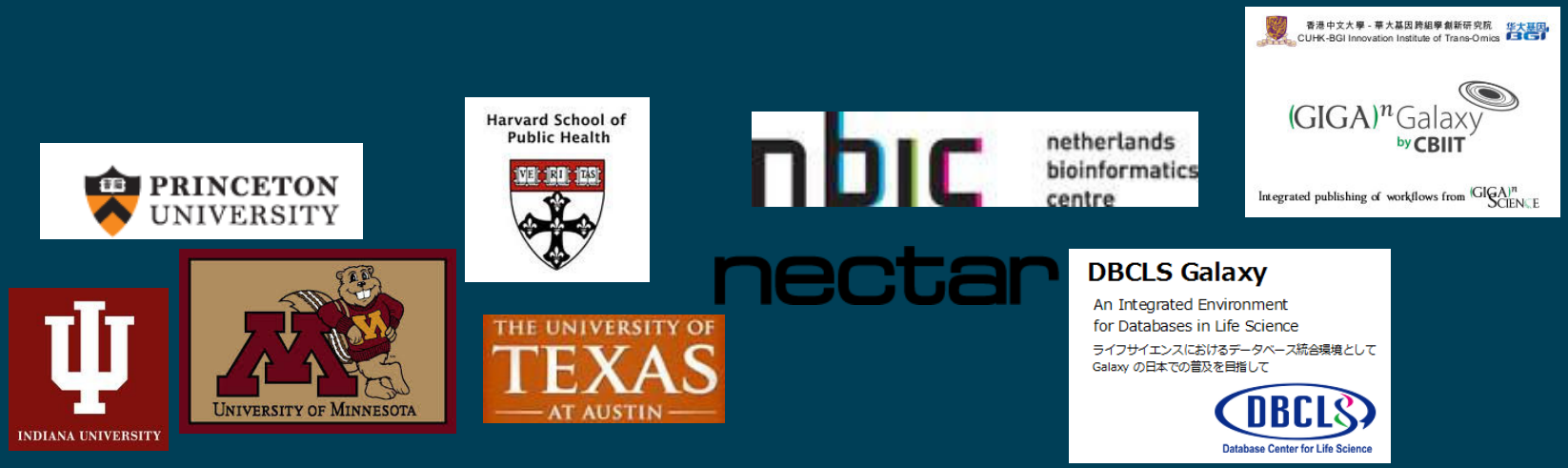
Public Website (http://usegalaxy.org), anybody can use:

- ~500 new users per month, ~200 TB of user data, ~130,000 analysis jobs per month

Used and cited in more than 1500 publications

Galaxy is Very Popular

60+ local installations all over the world



http://bit.ly/gxyservers

Topics

Galaxy

Cancer Genomics with Galaxy

Using Galaxy for Cancer Genomics

New tools

- complement existing high-throughput sequencing analysis tools

New workflows

- workflows are understandable, extendable, sharable

New visualizations and visual analysis applications

- e.g. interactively visualize and call variants in a Web browser

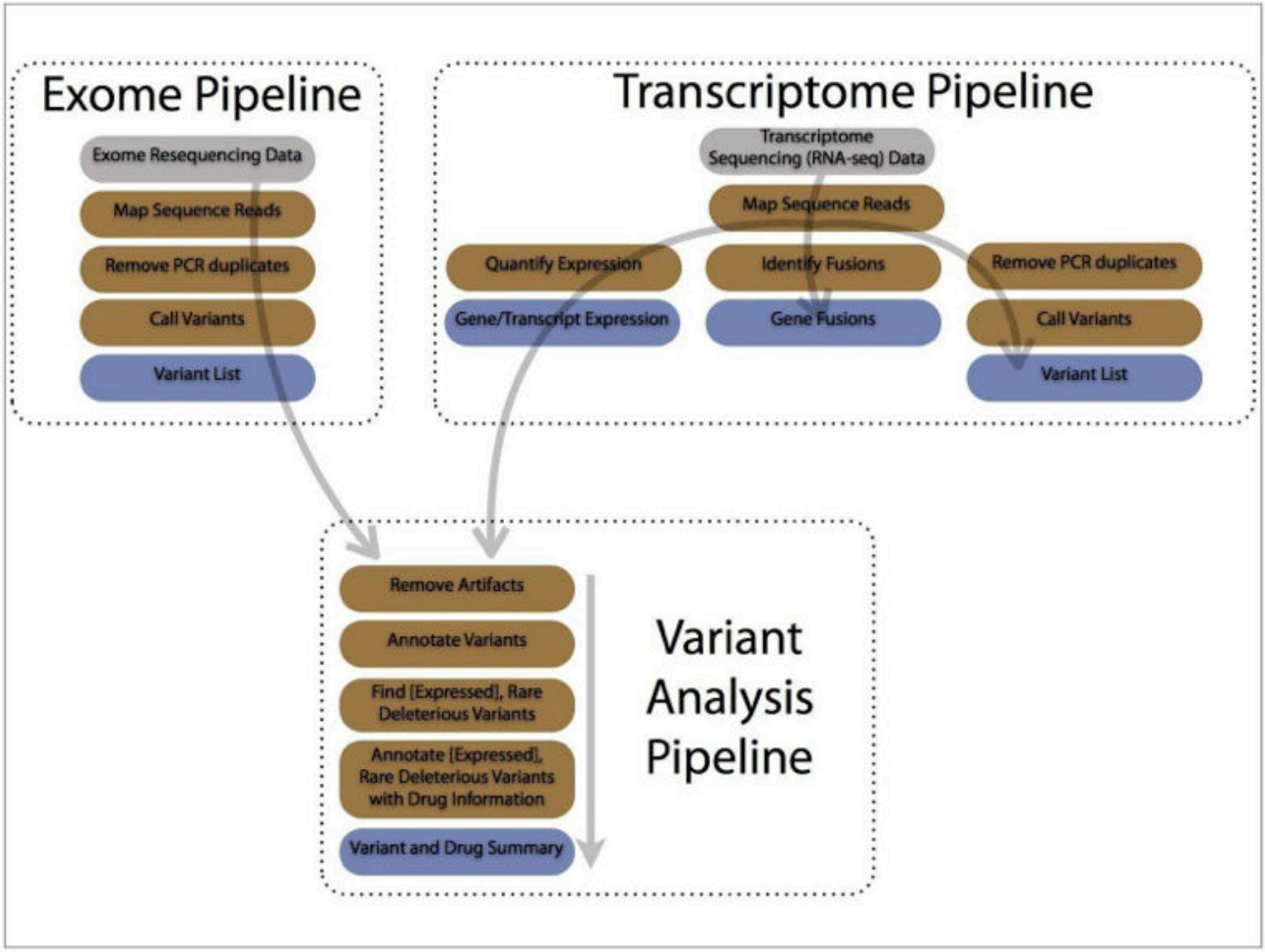
Motivating Applications

Tumor-cell line matching

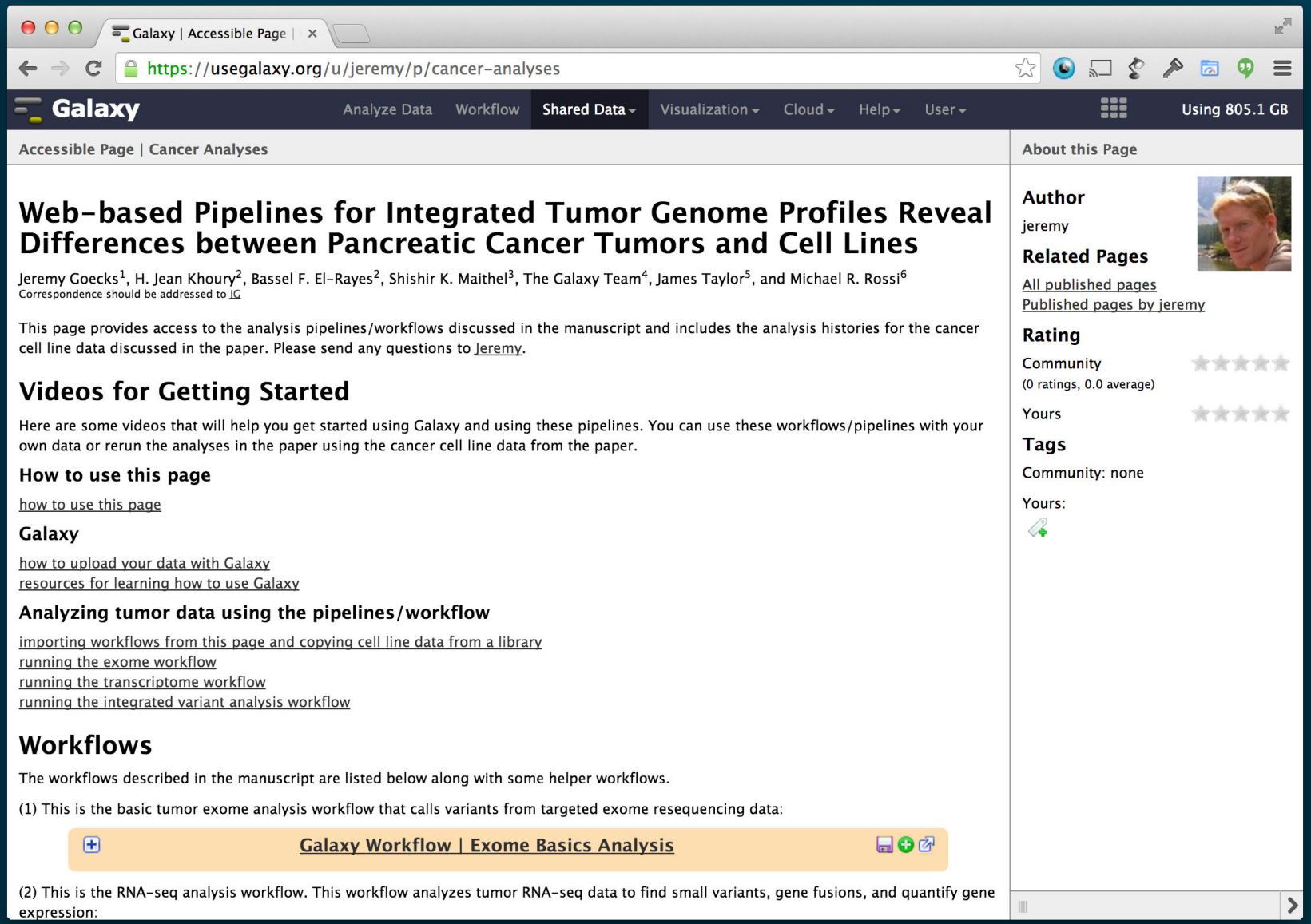
- primary pancreatic adenocarcinoma tumors vs. pancreatic cancer cell lines
- variants + gene expression

Deep targeted tumor sequencing

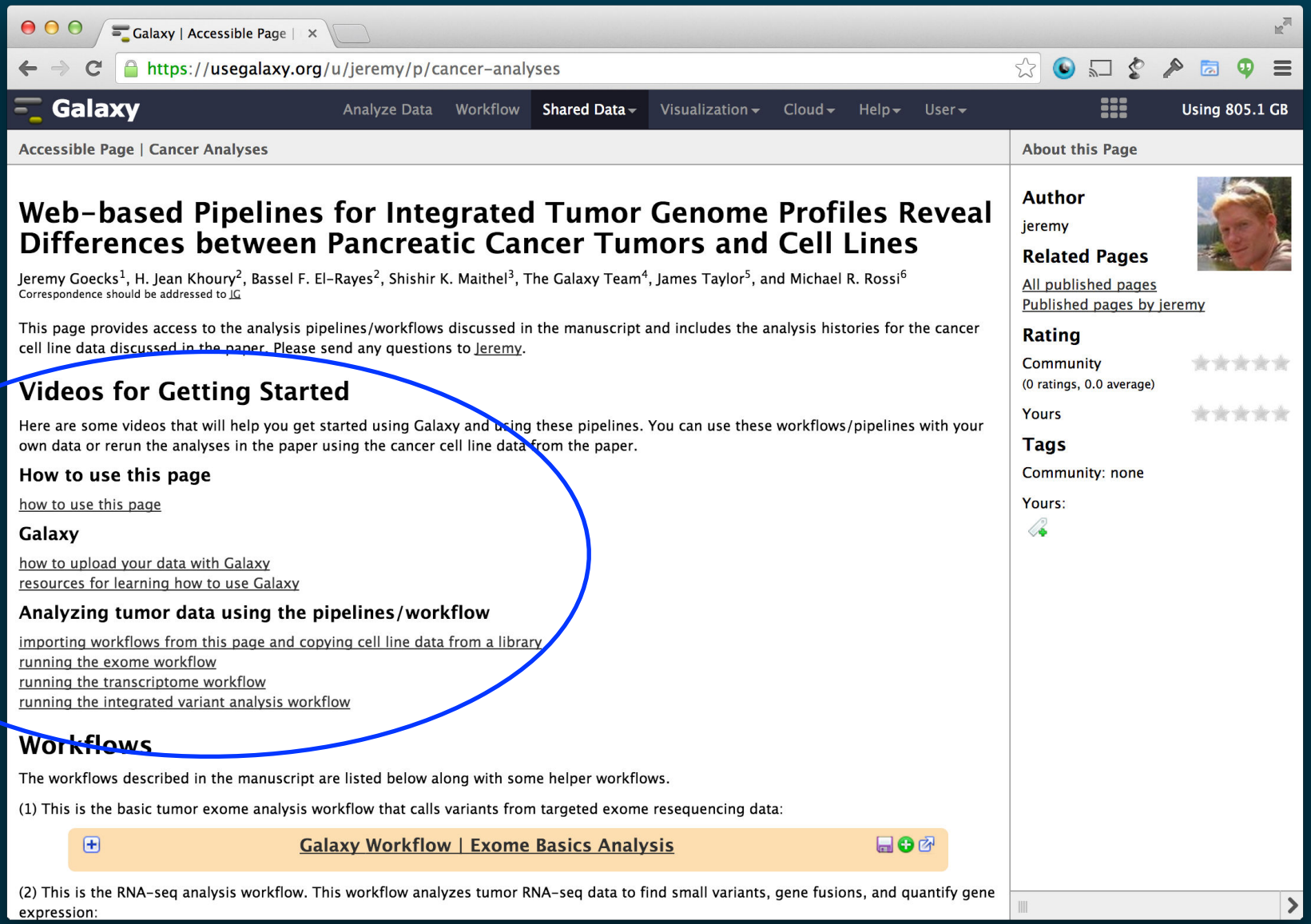
- small, inexpensive panel of ~26 genes sequenced at 5-10k coverage
- focus on clinically actionable genomic markers


Exome Pipeline
Exome Resequencing Data
Map Sequence Reads
Remove PCR duplicates
Call Variants
Variant List
Transcriptome Pipeline
Transcriptome Sequencing (RNA-seq) Data
Map Sequence Reads
Quantify Expression
Identify Fusions
Remove PCR duplicates
Gene/Transcript Expression
Gene Fusions
Call Variants
Variant List
Remove Artifacts
Annotate Variants
Find [Expressed], Rare Deleterious Variants
Annotate [Expressed], Rare Deleterious Variants with Drug Information
Variant and Drug Summary
Variant Analysis Pipeline

http://usegalaxy.org/cancer



Videos to Get Started



Galaxy | Accessible Page

https://usegalaxy.org/u/jeremy/p/cancer-analyses

Galaxy | Analyze Data | Workflow | Shared Data | Visualization | Cloud | Help | User | Using 805.1 GB

Accessible Page | Cancer Analyses

(1) This is the basic tumor exome analysis workflow that calls variants from targeted exome resequencing data:

Galaxy Workflow | Exome Basics Analysis

(2) This is the RNA-seq analysis workflow. This workflow analyzes tumor RNA-seq data to find small variants, gene fusions, and quantify gene expression:

Galaxy Workflow | Tumor RNA-seq Analysis

(3) This is the integrated variant analysis workflow. To use this workflow, two datasets in the same history are needed: (a) a variants dataset from either the exome or transcriptome analysis workflow) (b) Cufflinks Gene Expression dataset. This workflow then identifies:

deleterious variants
deleterious and druggable variants
deleterious variants in highly-expressed genes
deleterious and druggable variants in highly-expressed genes

Galaxy Workflow | Integrated Variant Analysis: Expression/Functional/Drug Annotation + Filtering

(4) This is an extended workflow for use when only a tumor exome is available. Starting with tumor exome sequencing data, it identifies deleterious variants and druggable variants:

Galaxy Workflow | Tumor Exome Analysis

(5) VCF Variant recovery. Use this workflow to obtain a list of variants in VCF format from a ANNOVAR table of variants. Variants in VCF format are useful for visualization.

Galaxy Workflow | Recover Variants from ANNOVAR

(6) Workflow to convert Tophat-fusion-post results to chrint format, which can be used to visualization fusion in Circster:

Galaxy Workflow | Tophat fusion post output to chrint

To use these workflows on a Galaxy instance other than this one, take these steps:

1. As an admin user, download the workflows that you want to use and follow the prompts to install needed tools. Here is more explanation on installing tools needed for workflows.
2. Download and install ANNOVAR (no automatic installation is possible due to ANNOVAR's licensing):

About this Page

Author

jeremy

Related Pages

All published pages
Published pages by jeremy

Rating

Community
(0 ratings, 0.0 average)

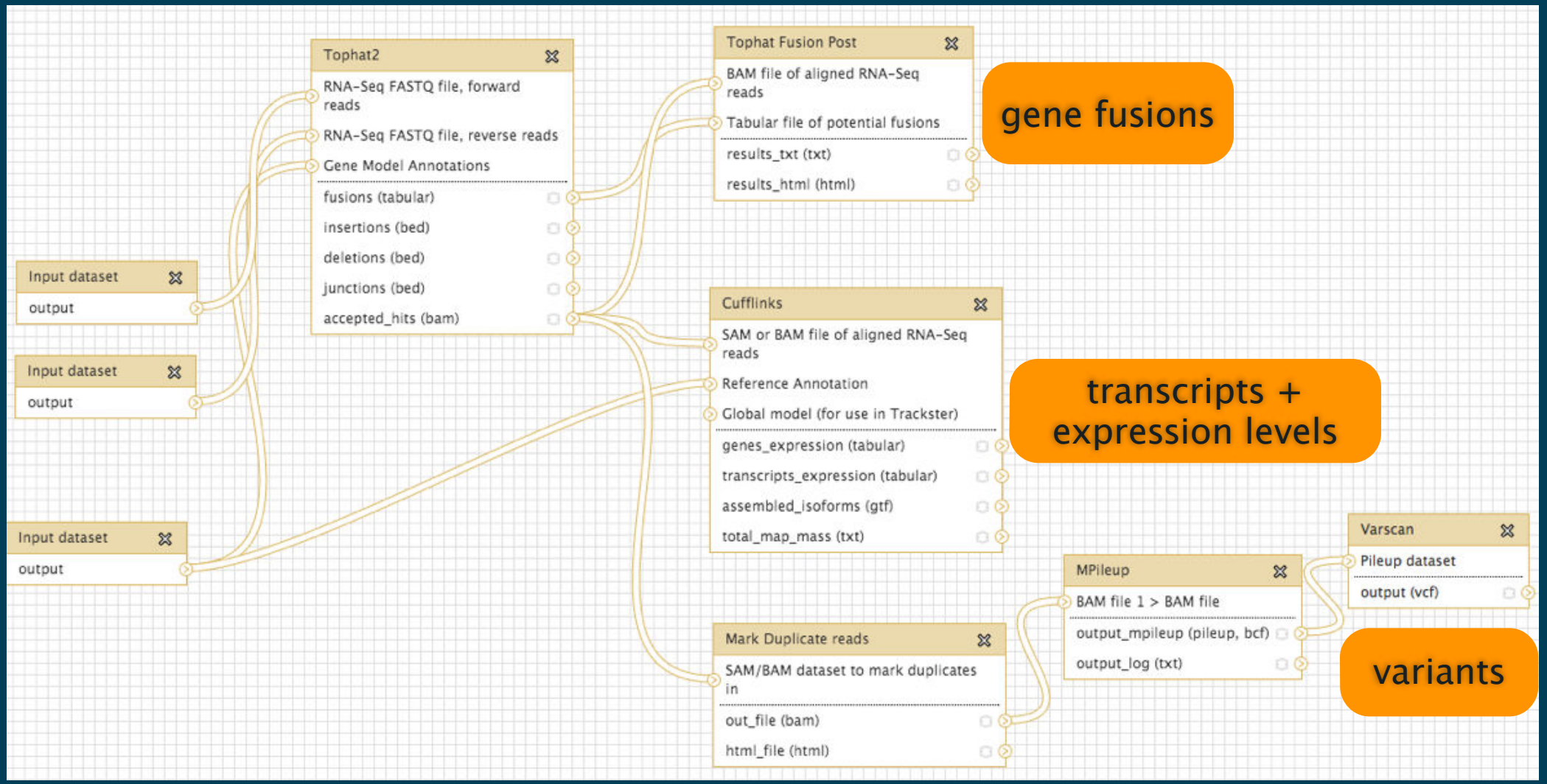
Yours

Tags

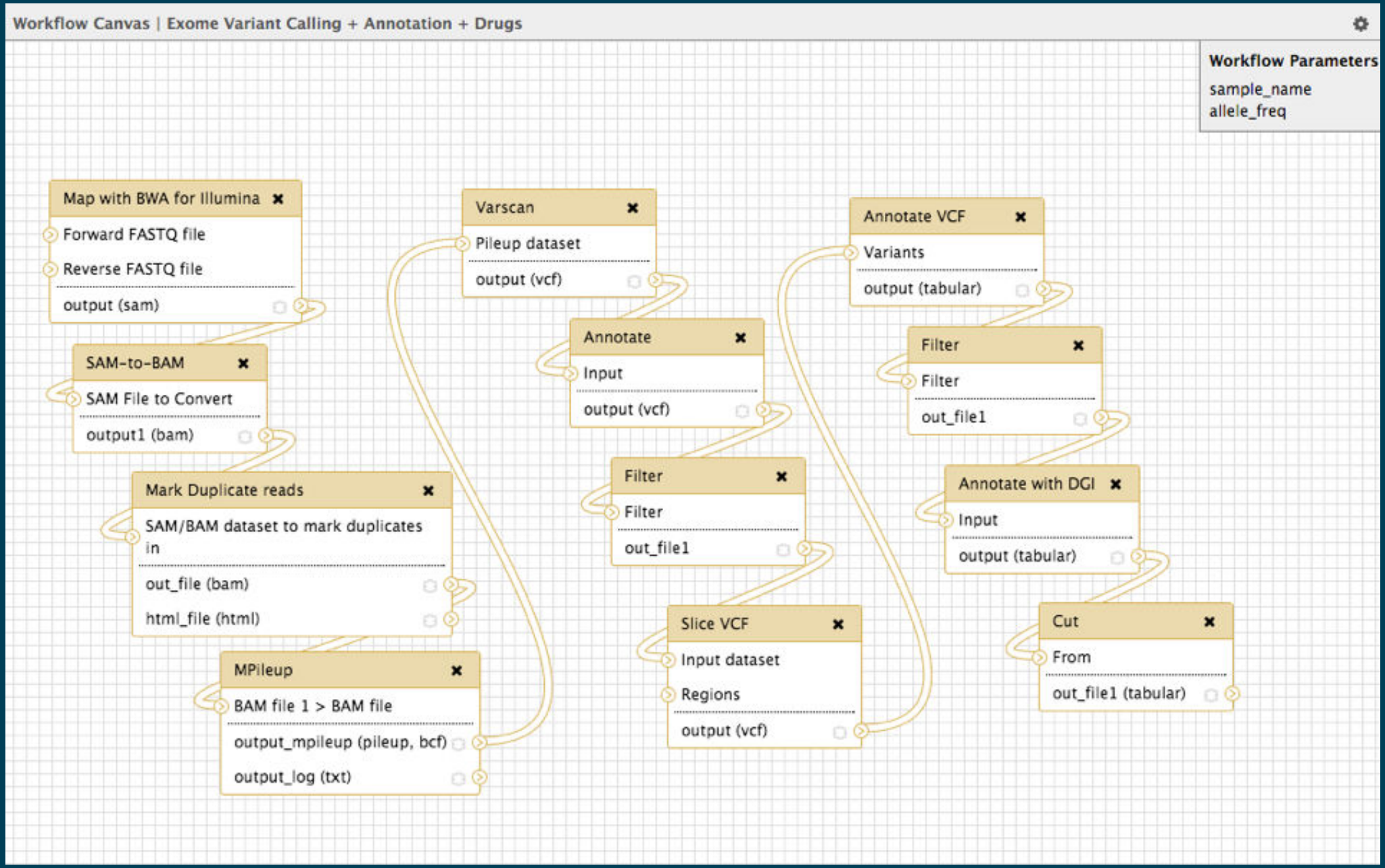
Community: none

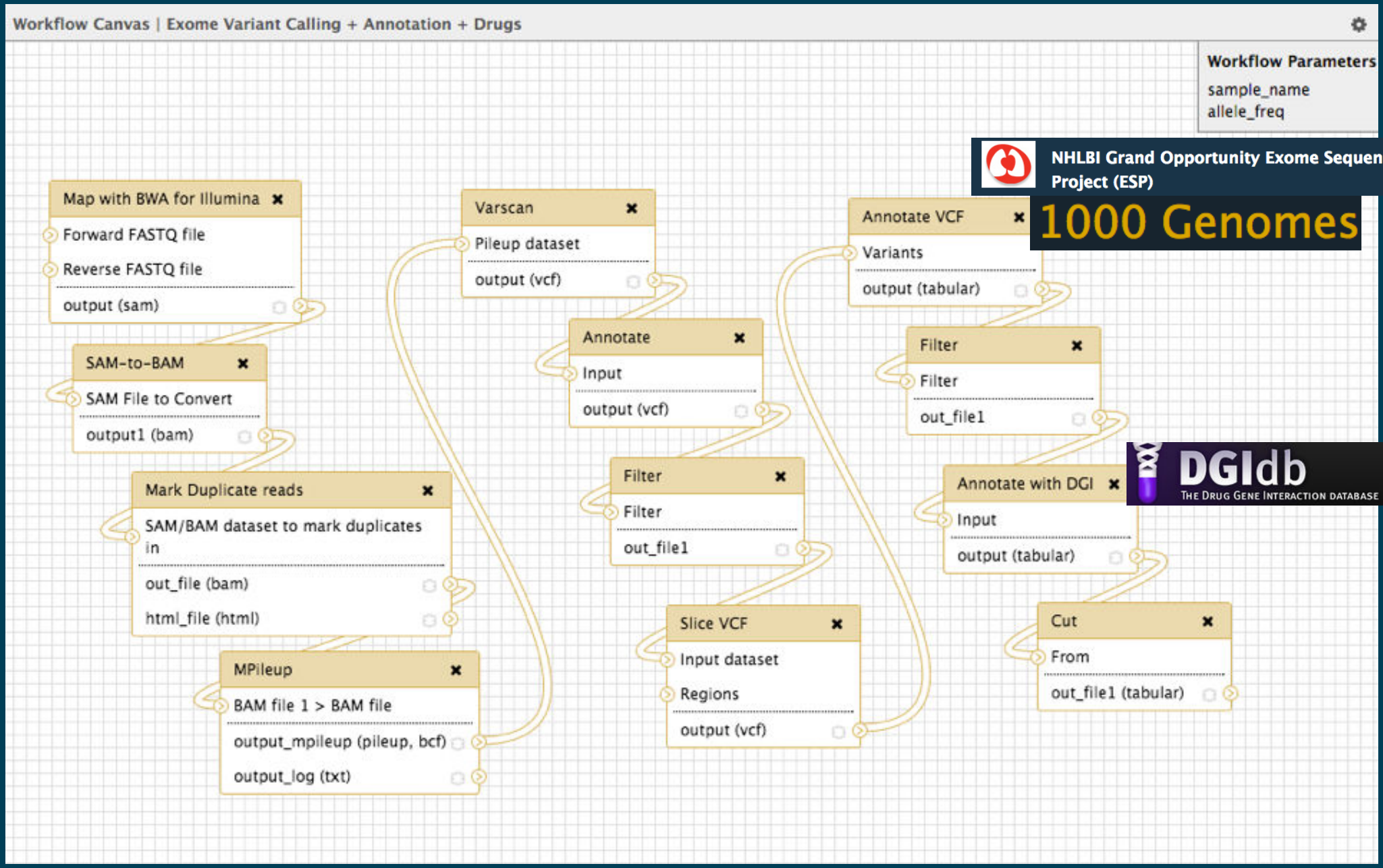
Yours:

Single Sample Transcriptome Analysis

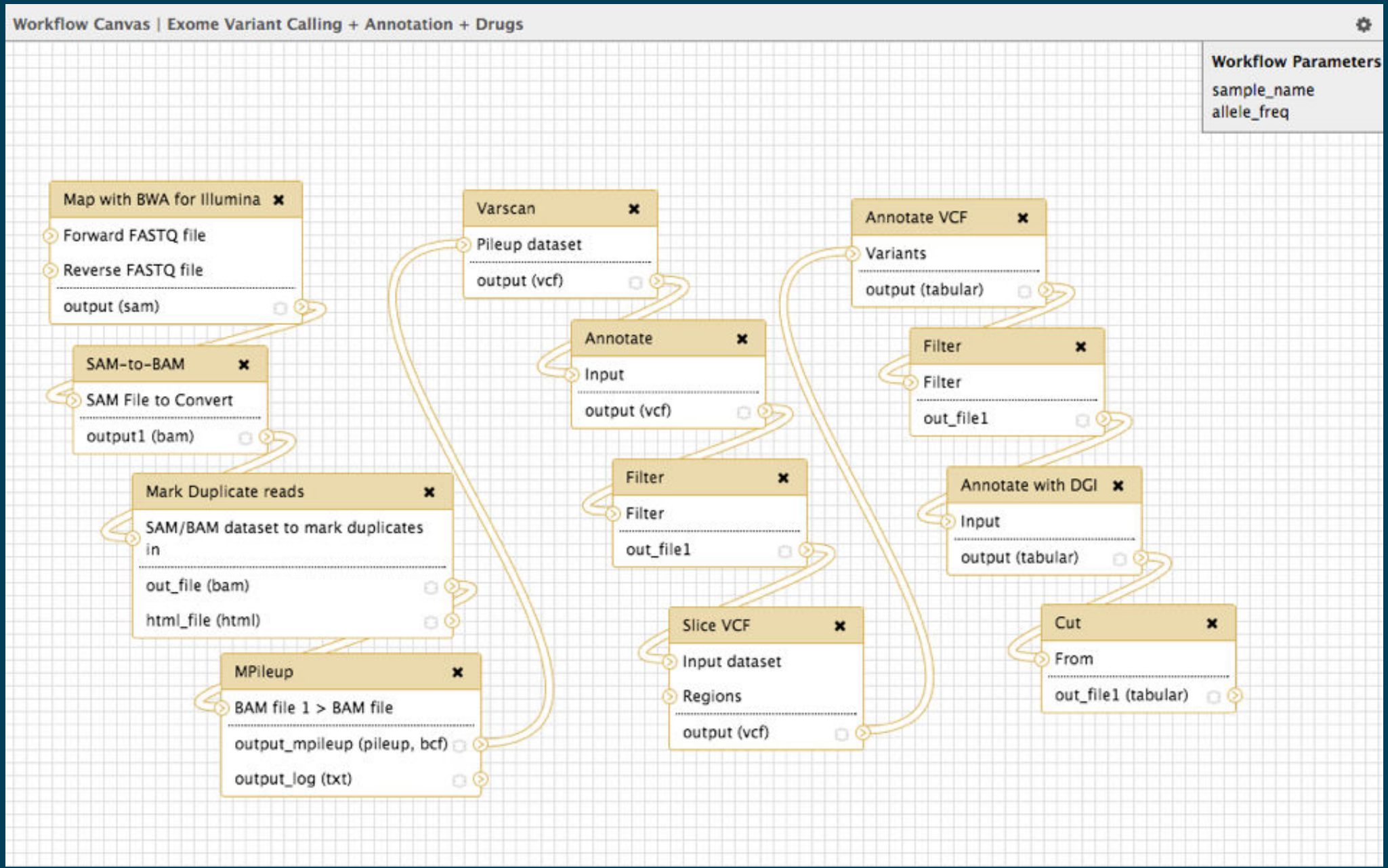


Exome: From Sequence to Drugs

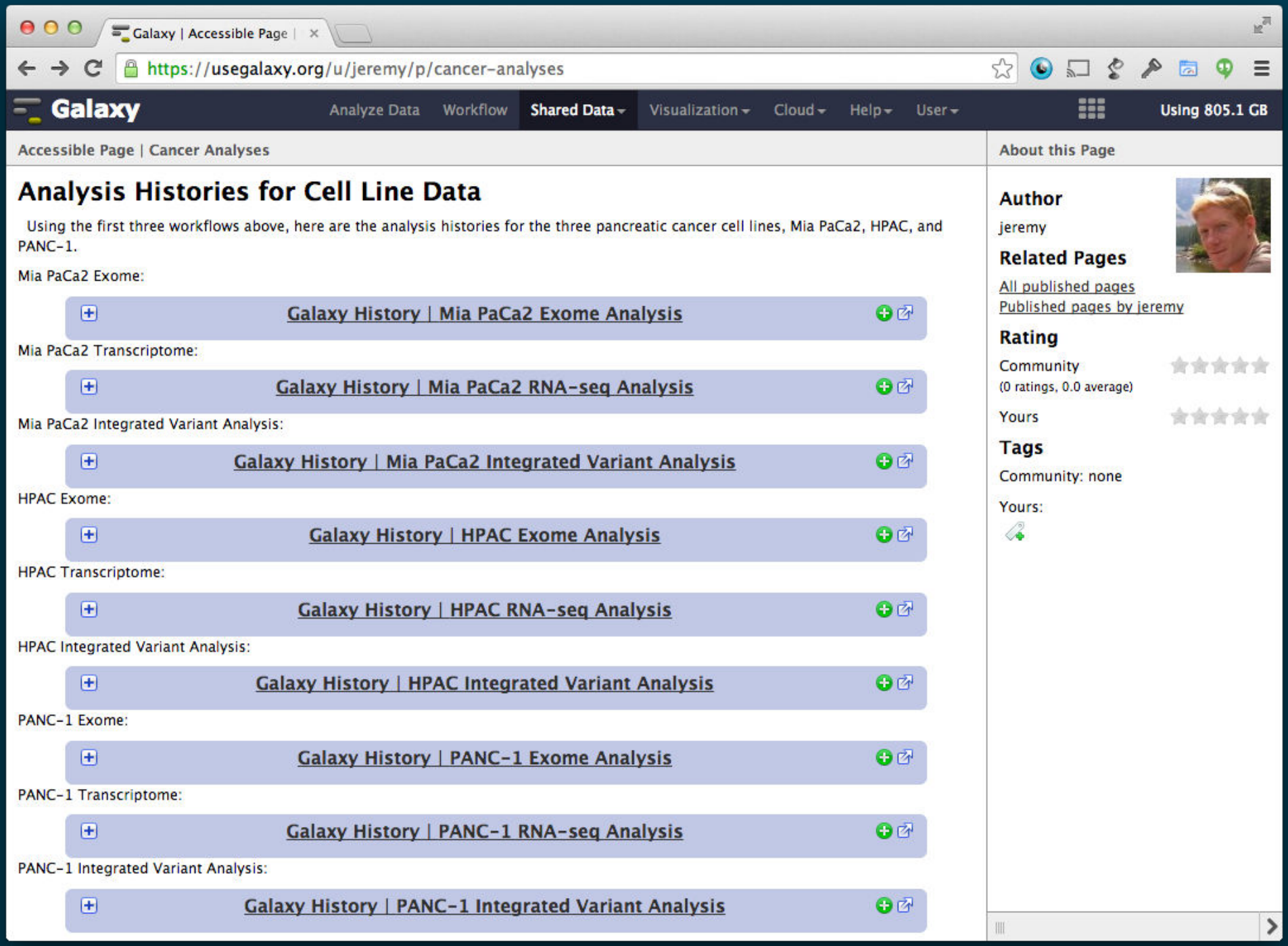


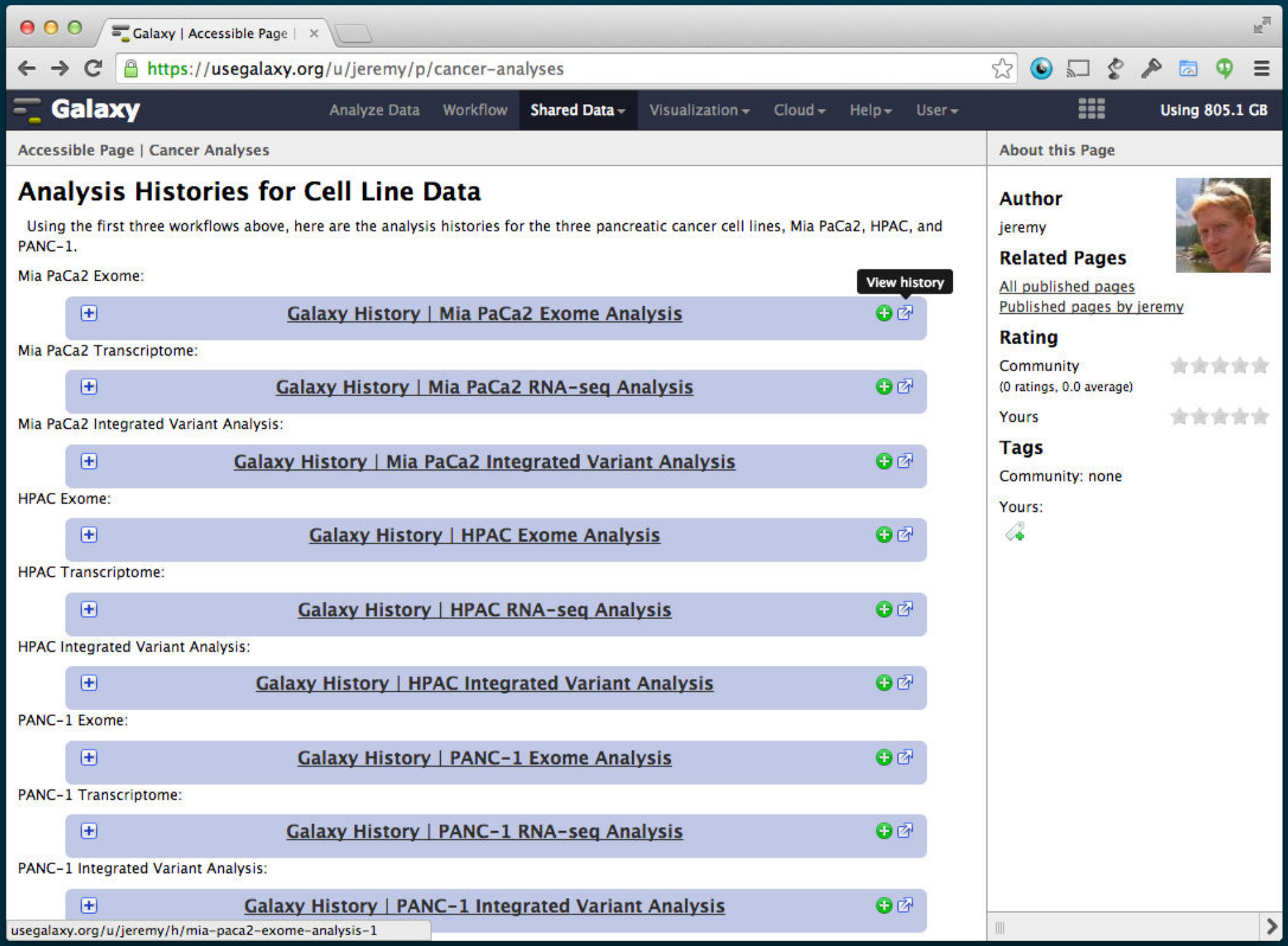


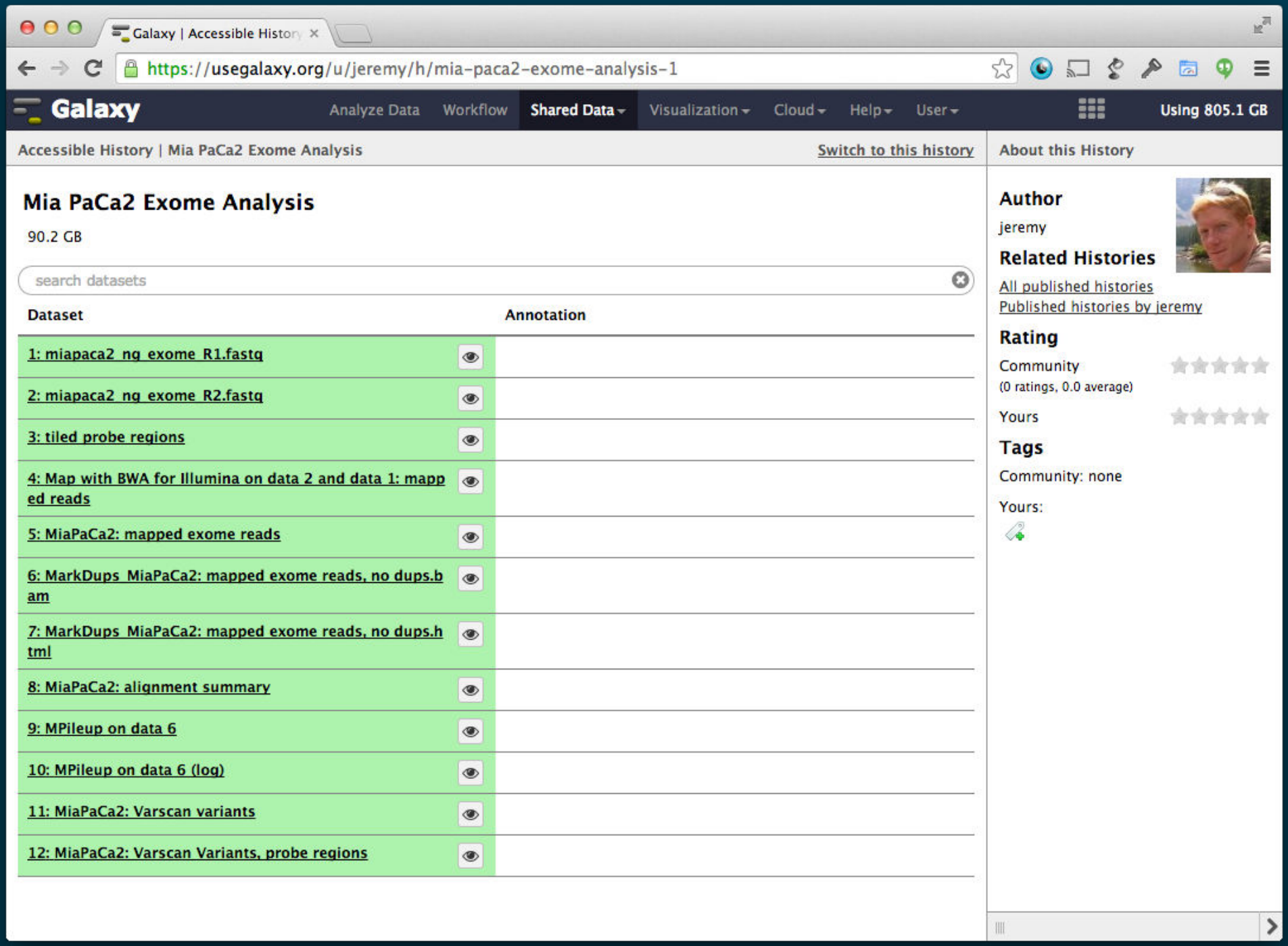
Integrates private and public data



Understandable, editable, sharable


Galaxy | Accessible Page
https://usegalaxy.org/u/jeremy/p/cancer-analyses
Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Cloud
Help
User
Using 805.1 GB
Accessible Page | Cancer Analyses
Analysis Histories for Cell Line Data
Using the first three workflows above, here are the analysis histories for the three pancreatic cancer cell lines, Mia PaCa2, HPAC, and PANC-1.
Mia PaCa2 Exome:
Galaxy History | Mia PaCa2 Exome Analysis
Mia PaCa2 Transcriptome:
Galaxy History | Mia PaCa2 RNA-seq Analysis
Mia PaCa2 Integrated Variant Analysis:
Galaxy History | Mia PaCa2 Integrated Variant Analysis
HPAC Exome:
Galaxy History | HPAC Exome Analysis
HPAC Transcriptome:
Galaxy History | HPAC RNA-seq Analysis
HPAC Integrated Variant Analysis:
Galaxy History | HPAC Integrated Variant Analysis
PANC-1 Exome:
Galaxy History | PANC-1 Exome Analysis
PANC-1 Transcriptome:
Galaxy History | PANC-1 RNA-seq Analysis
PANC-1 Integrated Variant Analysis:
Galaxy History | PANC-1 Integrated Variant Analysis
About this Page
Author
jeremy
Related Pages
All published pages
Published pages by jeremy
Rating
Community
(0 ratings, 0.0 average)
Yours
Tags
Community: none
Yours:


Galaxy | Accessible Page
https://usegalaxy.org/u/jeremy/p/cancer-analyses
Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Cloud
Help
User
Using 805.1 GB
Accessible Page | Cancer Analyses
About this Page
Analysis Histories for Cell Line Data
Using the first three workflows above, here are the analysis histories for the three pancreatic cancer cell lines, Mia PaCa2, HPAC, and PANC-1.
Mia PaCa2 Exome:
View history
Galaxy History | Mia PaCa2 Exome Analysis
Mia PaCa2 Transcriptome:
Galaxy History | Mia PaCa2 RNA-seq Analysis
Mia PaCa2 Integrated Variant Analysis:
Galaxy History | Mia PaCa2 Integrated Variant Analysis
HPAC Exome:
Galaxy History | HPAC Exome Analysis
HPAC Transcriptome:
Galaxy History | HPAC RNA-seq Analysis
HPAC Integrated Variant Analysis:
Galaxy History | HPAC Integrated Variant Analysis
PANC-1 Exome:
Galaxy History | PANC-1 Exome Analysis
PANC-1 Transcriptome:
Galaxy History | PANC-1 RNA-seq Analysis
PANC-1 Integrated Variant Analysis:
Galaxy History | PANC-1 Integrated Variant Analysis
usegalaxy.org/u/jeremy/h/mia-paca2-exome-analysis-1
Author
jeremy
Related Pages
All published pages
Published pages by jeremy
Rating
Community
(0 ratings, 0.0 average)
Yours
Tags
Community: none
Yours:

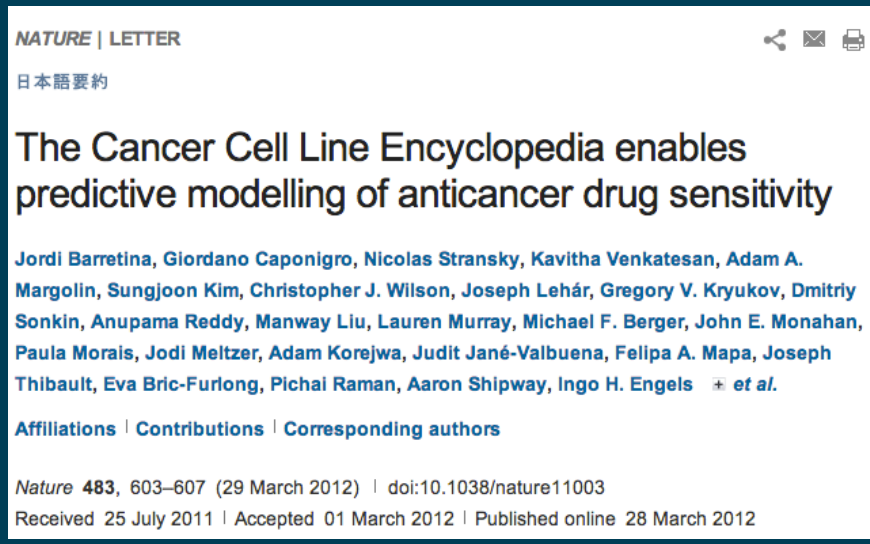

Galaxy | Accessible History
https://usegalaxy.org/u/jeremy/h/mia-paca2-exome-analysis-1
Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Cloud
Help
User
Using 805.1 GB
Accessible History | Mia PaCa2 Exome Analysis
Switch to this history
About this History
Mia PaCa2 Exome Analysis
90.2 GB
search datasets
Dataset
Annotation
1: miapaca2_ng_exome_R1.fastq
2: miapaca2_ng_exome_R2.fastq
3: tiled probe regions
4: Map with BWA for Illumina on data 2 and data 1: mapped reads
5: MiaPaCa2: mapped exome reads
6: MarkDups_MiaPaCa2: mapped exome reads, no dups.bam
7: MarkDups_MiaPaCa2: mapped exome reads, no dups.html
8: MiaPaCa2: alignment summary
9: MPileup on data 6
10: MPileup on data 6 (log)
11: MiaPaCa2: Varscan variants
12: MiaPaCa2: Varscan Variants, probe regions
Author
jeremy
Related Histories
All published histories
Published histories by jeremy
Rating
Community
(0 ratings, 0.0 average)
Yours
Tags
Community: none
Yours:

Preliminary Validation using Cell Lines

3 cell lines (MiaPaCa2, HPAC, PANC-1): exome + transcriptome sequencing

CCLE

- Highly targeted exome sequencing of 500+ cancer cell lines
- Drug response curves



NATURE | LETTER

日本語要約

The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity

Jordi Barretina, Giordano Caponigro, Nicolas Stransky, Kavitha Venkatesan, Adam A. Margolin, Sungjoon Kim, Christopher J. Wilson, Joseph Lehár, Gregory V. Kryukov, Dmitriy Sonkin, Anupama Reddy, Manway Liu, Lauren Murray, Michael F. Berger, John E. Monahan, Paula Morais, Jodi Meltzer, Adam Korejwa, Judit Jané-Valbuena, Felipa A. Mapa, Joseph Thibault, Eva Bric-Furlong, Pichai Raman, Aaron Shipway, Ingo H. Engels + *et al.*

Affiliations | Contributions | Corresponding authors

Nature **483**, 603–607 (29 March 2012) | doi:10.1038/nature11003

Received 25 July 2011 | Accepted 01 March 2012 | Published online 28 March 2012

Preliminary Validation Results

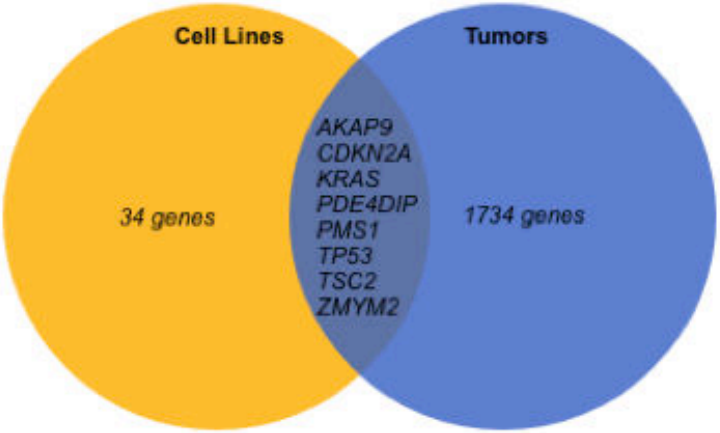
All CCLE variants (84 total) validated in cell lines

MEK inhibitor validation

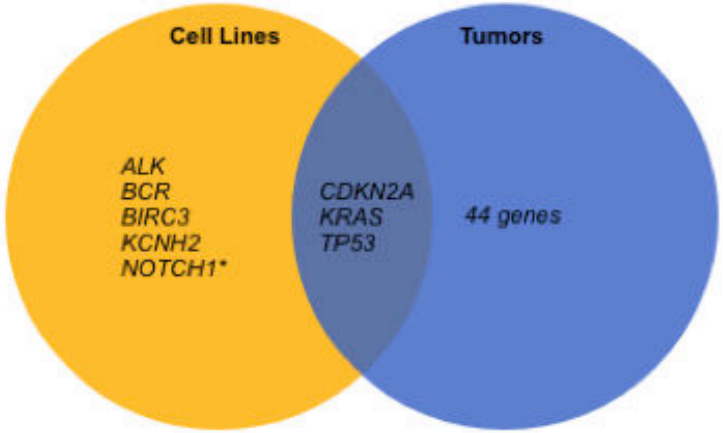
- pipelines: predicted cell lines would be responsive
- CCLE: drug response data shows cell line are responsive

PANC-1: fewer rare and deleterious mutations in cell line leads to fewer applicable drugs

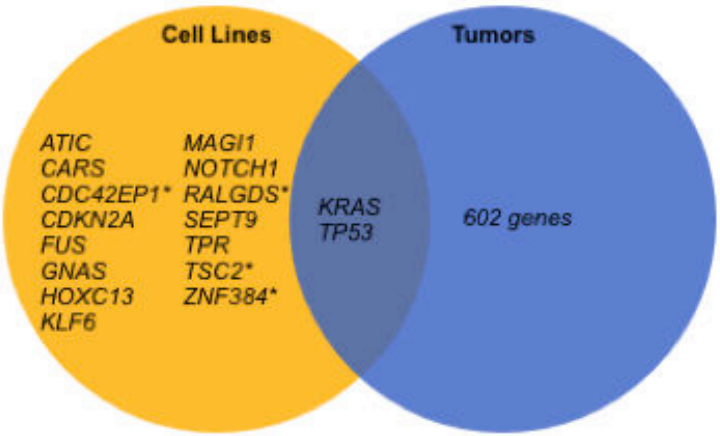
Preliminary Results: Cell Lines vs. PAAC tumors



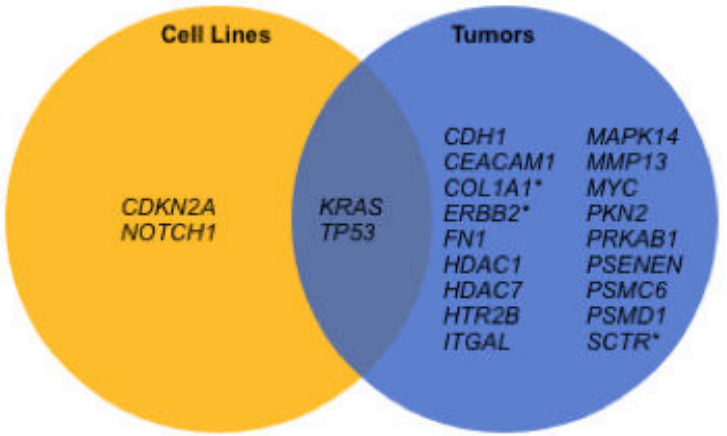
(a) all rare and deleterious mutations



(b) all rare, deleterious, and druggable mutations



(c) rare and deleterious mutations in highly-expressed genes



(d) rare, deleterious, and druggable mutations in highly-expressed genes

Trajectory

Standardize workflows for use by research oncologists

Improve workflows using new tools

Generate clinician-friendly reports from workflow

Common Galaxy Questions

How does Galaxy scale?

- to whatever computing resources are available

Can I modify workflows/pipelines?

- yes, either via Web or programmatically

Can I add my own tools and visualizations?

- yes, there are frameworks for both

Concluding Thoughts

Galaxy is a very useful platform for high-throughout genomics

- accessible, reproducible, collaborative
- tools, workflows, visualizations
- public, local, cloud

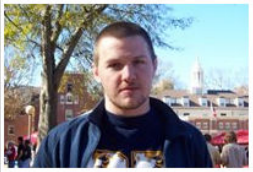
Use Galaxy to go from tumor sequence reads to:

- rare, deleterious (RD) variants (exome)
- gene fusions and gene expression (RNA-seq)
- drugs associated with genes that have RD variants (integrative)

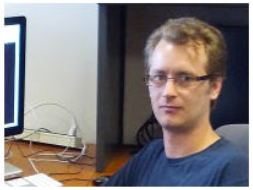
Galaxy


Enis Afgan
IRB


Guru Ananda
Penn State


Dannon Baker
Emory


Dan Blankenberg
Penn State


Dave Bouvier
Penn State


Dave Clements
Emory


Nate Coraor
Penn State


Carl Eberhard
Emory


Jeremy Goecks
Emory

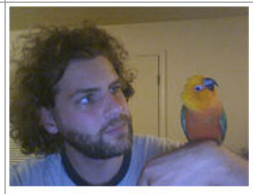

Sam Guerler
Emory


Jennifer Hillman Jackson
Penn State


Greg von Kuster
Penn State


Ross Lazarus
BakerIDI


Anton Nekrutenko
Penn State


James Taylor
Emory

EMORY | WINSHIP CANCER INSTITUTE

A Cancer Center Designated by the National Cancer Institute


Mike Rossi

JOHNS HOPKINS UNIVERSITY

PENNSTATE 1855

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

genome.gov
National Human Genome Research Institute
National Institutes of Health

National Science Foundation
WHERE DISCOVERIES BEGIN

Thanks!
Questions?

Galaxy

http://galaxyproject.org http://usegalaxy.org

Postdoc and software engineer positions available in
Interactive Genomics Lab @ THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

http://jeremygoecks.com
@jgoecks jgoecks@gwu.edu