



# Create Genome Browsers for Collaborative Eukaryotic Genome Annotations

Luke Sargent<sup>1</sup>, Yating Liu<sup>2</sup>, Wilson Leung<sup>2</sup>, Sarah C.R. Elgin<sup>2</sup>, Jeremy Goecks<sup>1</sup>



PAG XXVII





# Challenge: Genomics is Hard

“Genomics is a **‘four-headed beast’**; considering the computational demands across the lifecycle of a dataset—acquisition, storage, distribution, and analysis—**genomics is either on par with or the most demanding of the Big Data domains**”





# Challenge: Cross-Discipline Skill Requirements

## Answering an experimental question requires:

- Hardware resources
  - compute, storage, network
- Software resources
  - tools, platform
- An analytic pipeline
  - workflow of tools
- System administration
  - skills to keep it all running

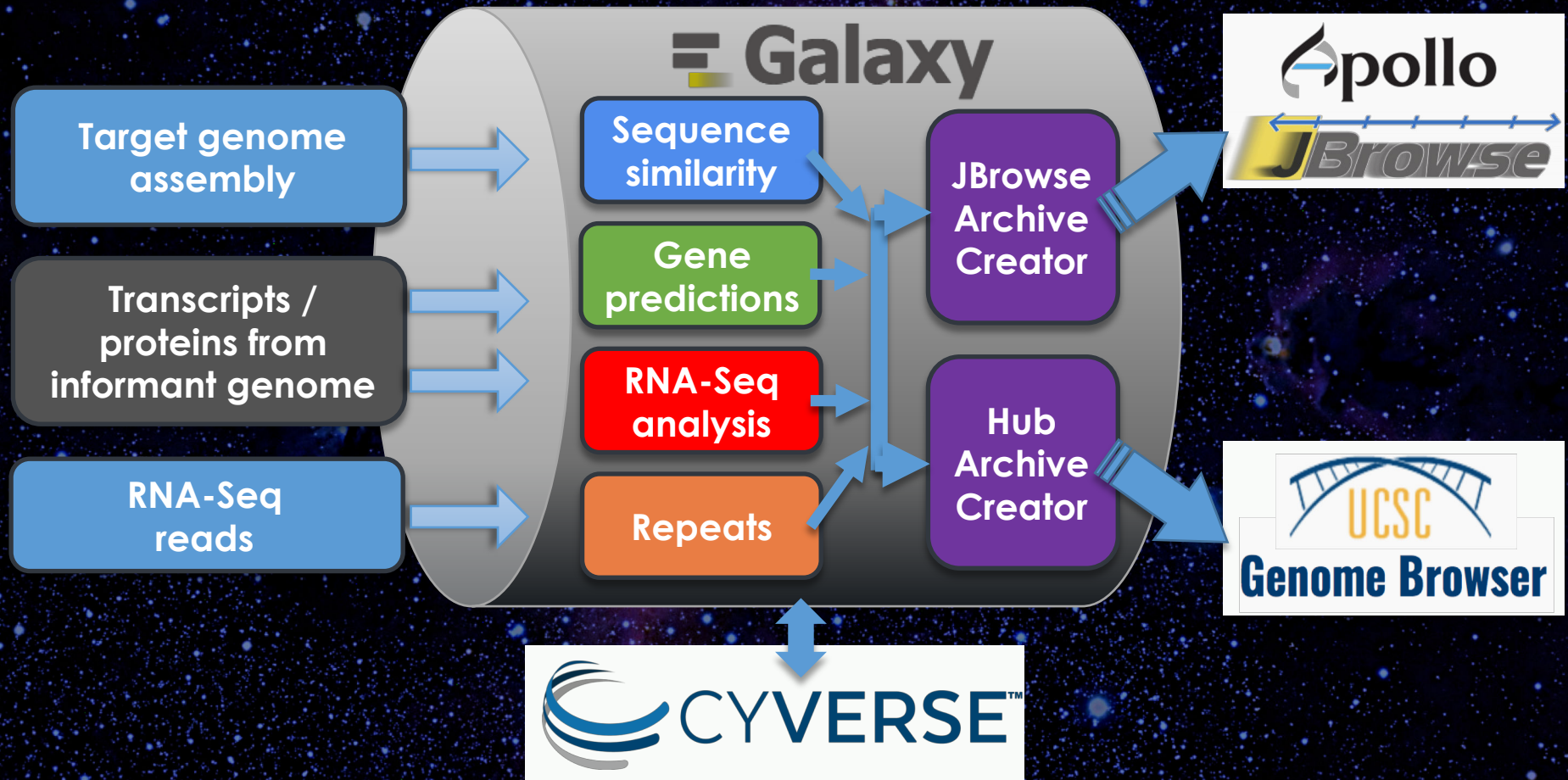


Obligatory xkcd: <https://xkcd.com/1739/>



# G-OnRamp:

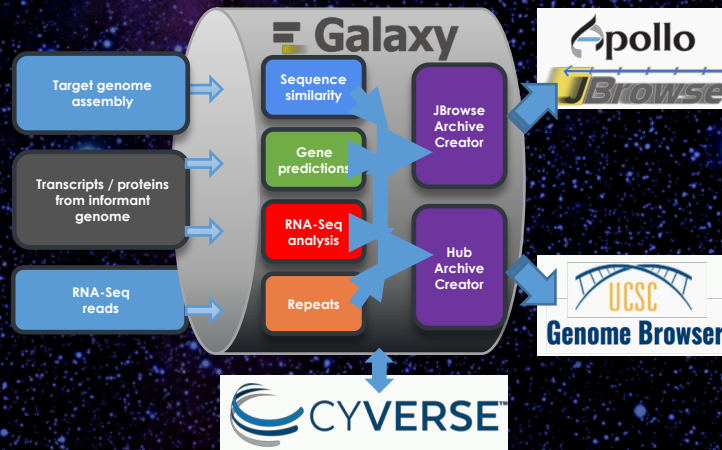
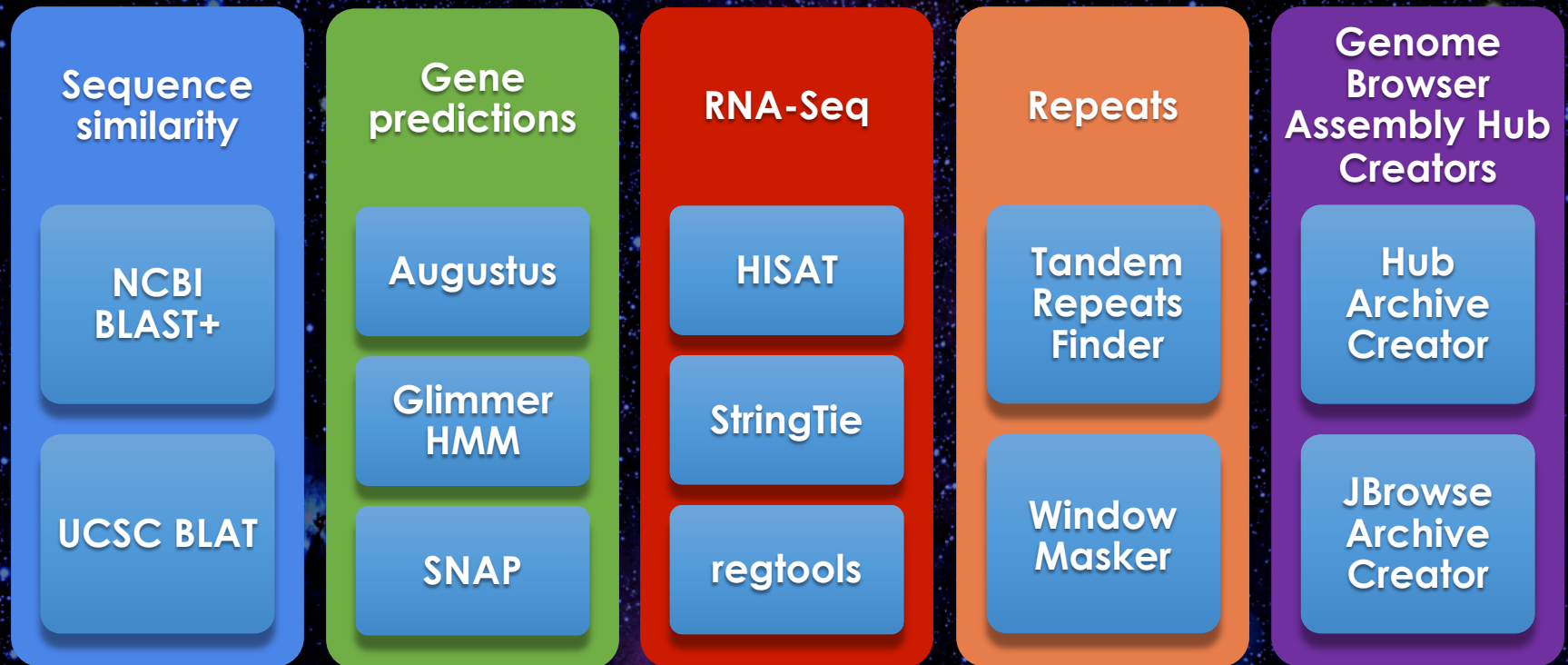
## Galaxy + Genomics Education Partnership (GEP)



- An **integrated**, **web-based**, and **scalable** environment
- Enables **interactive annotation of any eukaryotic genome**



# G-OnRamp: Tools Do the Work!





# G-OnRamp: Galaxy-Based Analytical Platform

The screenshot displays the Galaxy web interface with the 'G-OnRamp workflow for JBrowse/Apollo' selected. The interface includes a top navigation bar with tabs for 'Analyze Data', 'Workflow', 'Visualize', 'Shared Data', 'Admin', 'Help', and 'User'. A left sidebar lists various tool categories like 'Get Data', 'Send Data', 'Collection Operations', etc. The main workspace shows the workflow configuration with five steps: 1. Target genome (dbia3.fa), 2. Informant (dbia3.fa), 3. Informant protein sequences (dmel-hits-translation-r6.11.fa), 4. RNA-Seq: Forward reads (D.biamipes\_RNAseq\_whole\_adult\_females\_s\_1\_2\_sequence.dotchromosome.contig16.fastqsanger), and 5. RNA-Seq: Reverse reads (D.biamipes\_RNAseq\_whole\_adult\_females\_s\_1\_1\_sequence.dotchromosome.contig16.fastqsanger). A red arrow points to the 'Run workflow' button. Overlaid on the workflow steps are four callout boxes: 'Run workflow' (red), 'Target genome assembly' (blue), 'Transcripts / proteins from informant genome' (grey), and 'RNA-Seq reads' (blue). The right sidebar shows the 'History' tab with a list of datasets, including '5: dmel-mrna-chrom4.gb.txt' and '4: dmel-hits-translation-r6.11.fa'.

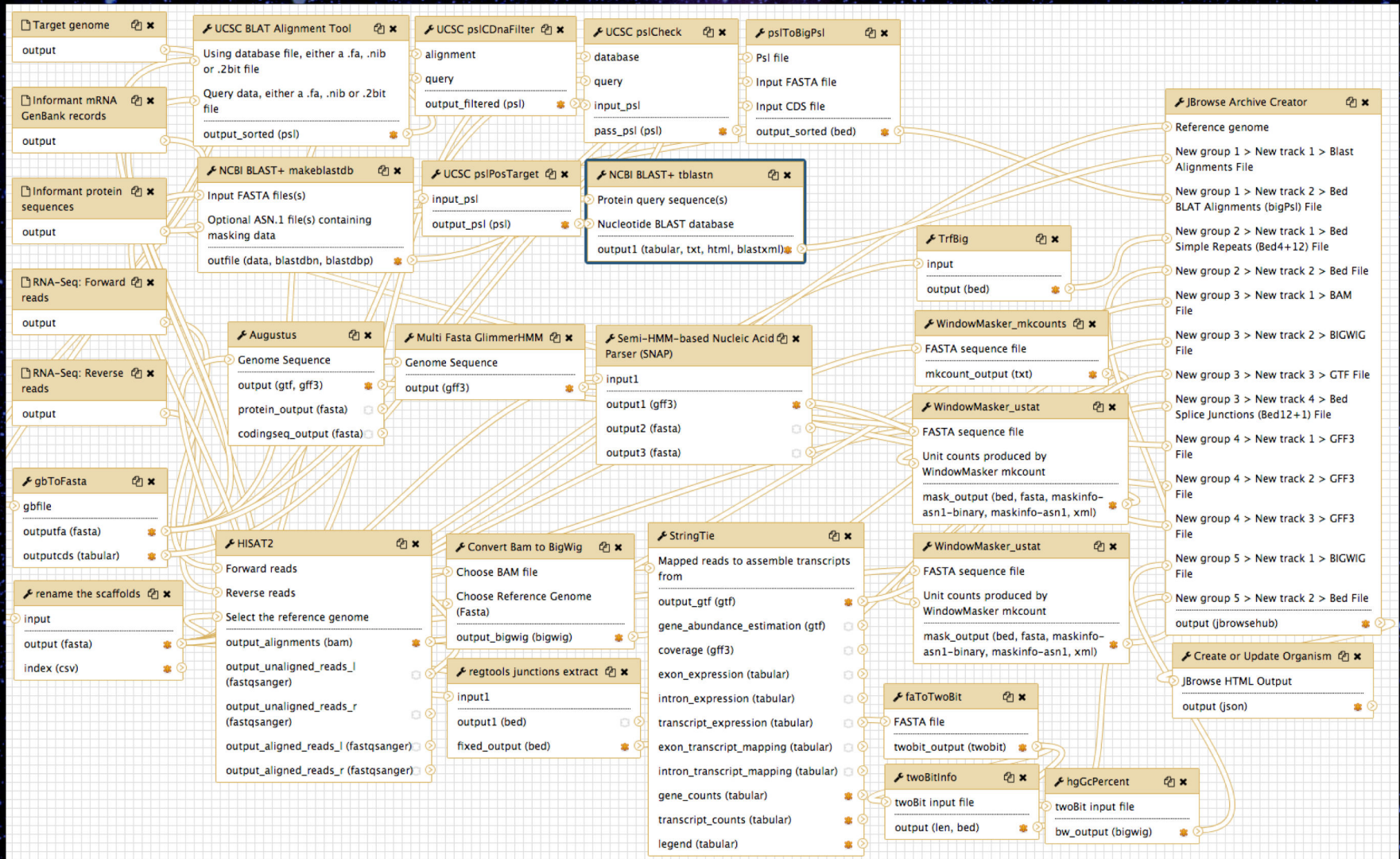
**Specify Inputs, Tune Parameters, Run!**

**Galaxy base provides:**

- Web UI for **accessibility**
- Programmatic API for **automation** and **scalability**
- Histories with specified parameters and intermediates for **transparency** and **reproducibility**



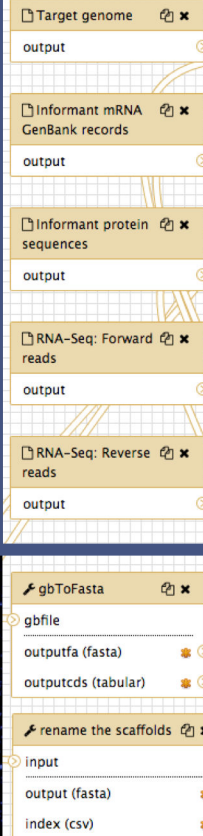
# G-OnRamp: Workflows, the Analytical Pipelines



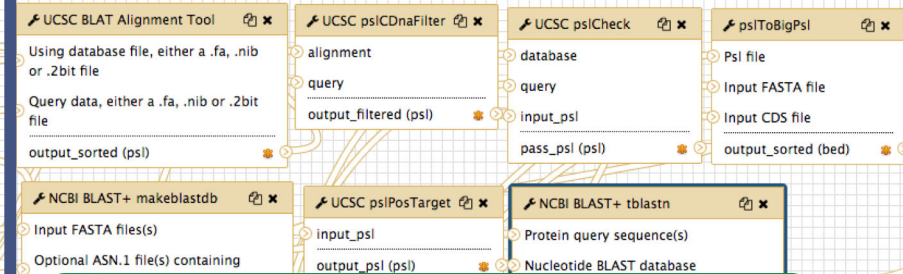


# G-OnRamp: Workflows By Sub-workflow

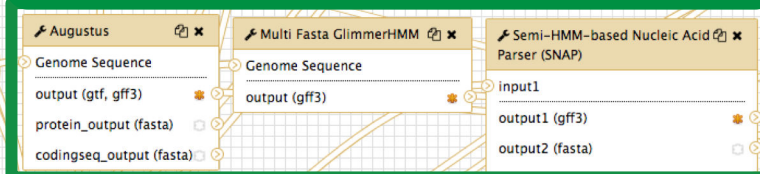
## Input data



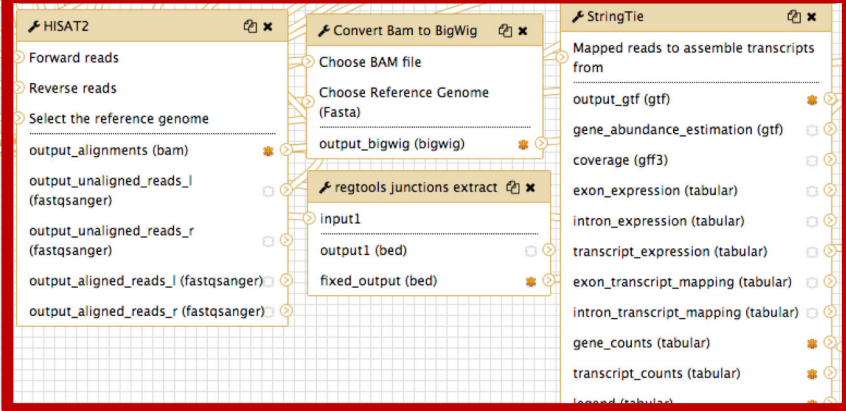
## Sequence similarity



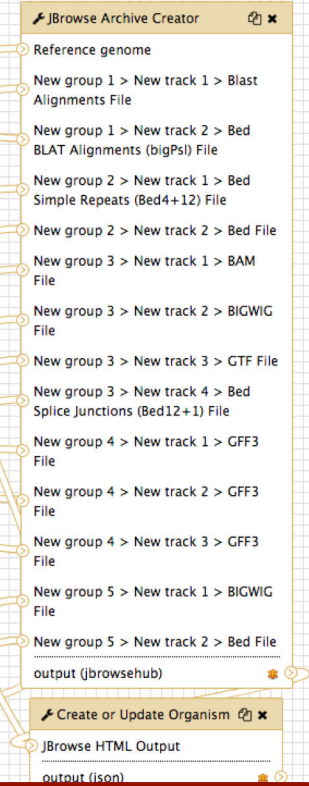
## Gene predictions



## RNA-Seq



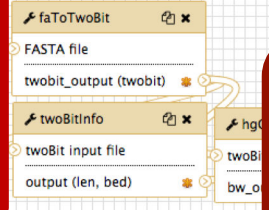
## JBrowse Archive Creator



## Repeats



## Create or Update Organism





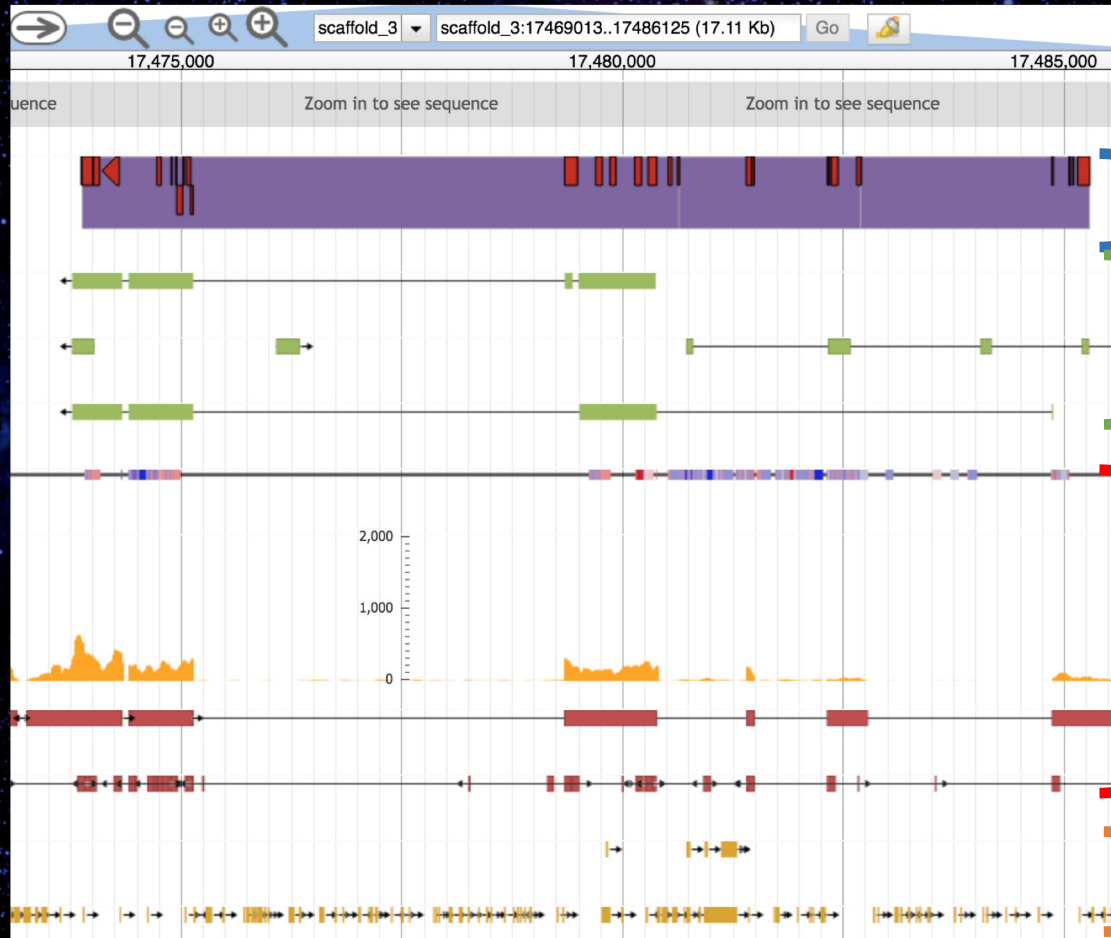
# G-OnRamp: A JBrowse Genome Browser Factory

Sequence  
Similarity

Gene  
Predictions

RNA-Seq  
Analysis

Repeats



• UCSC BLAT

• Augustus  
• GlimmerHMM  
• SNAP

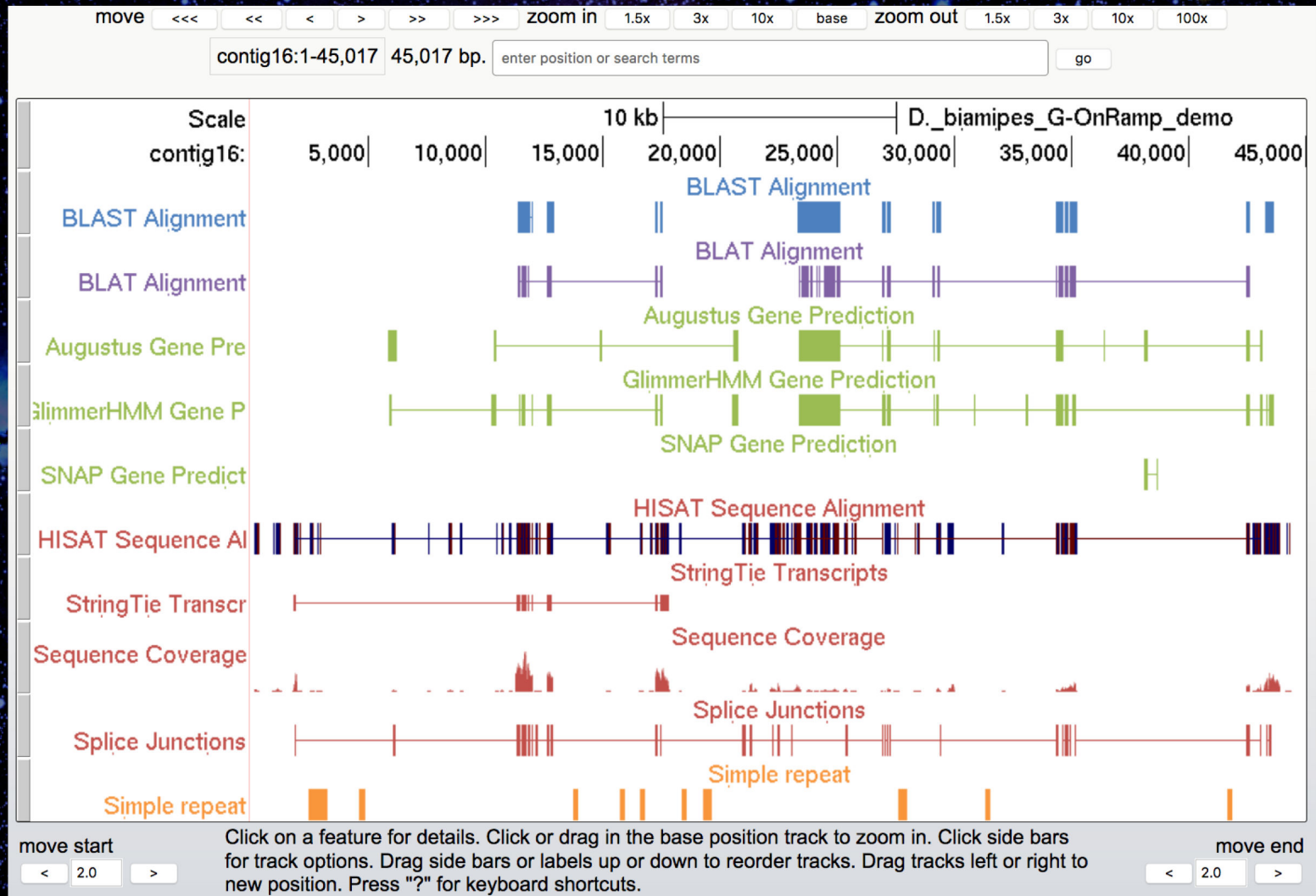
• HISAT  
• StringTie  
• regtools

• TRF  
• WindowMasker

<https://github.com/GMOD/jbrowse>



# G-OnRamp: A UCSC Genome Browser Factory



<https://genome.ucsc.edu>



# G-OnRamp: Collaborative Annotation Platform

The screenshot displays the Apollo genome annotation editor interface. On the left, the 'Available Tracks' panel lists various tracks including BLAT Alignment, Gene Prediction (Augustus, GlimmerHMM, SNAP), Homology (Blast Alignment), Mapping and Sequencing (GAP, GC Percent), RNA-Seq (Assembled Transcripts, Splice Junction, Sequence Alignment, Sequence Coverage), and Variation and Repeats (Simple Repeats, Window Masker). The main workspace shows a genomic track for 'drosophila\_demo' with a scale from 0 to 45,000. A specific region, 'scaffold\_16:8991..14090 (5.1 Kb)', is highlighted. Below this, a 'User-created Annotations' track shows a blue box labeled 'scaffold\_16\_FBpp0071086-00001'. The bottom track shows BLAT Alignments for transcripts NM\_001103380 and NM\_001103381. On the right, a sidebar shows the 'Annotations' tab with a table listing annotations for 'drosophila\_demo'.

Name	Annotations	Ref Sequences
drosophila_demo	0	70

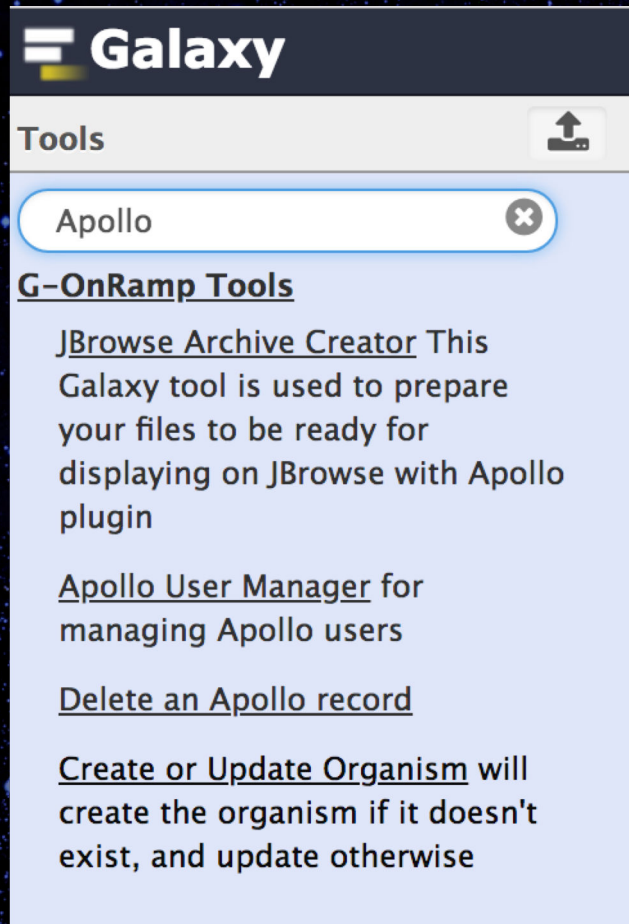
Details section shows: Name: drosophila\_demo

**Apollo**: an **instantaneous, collaborative** genome **annotation editor** that provides a custom track for **user-created annotations** on JBrowse hub archives

<https://github.com/GMOD/apollo>



# G-OnRamp: Tight Apollo Integration



## Bundled Tools for Apollo Administration:

- JBrowse Archive Creator (JAC)
- Apollo User Manager\*
  - **Batch management of user accounts**
- Create or Update Organism\*
  - **Create Apollo instance from JBrowse**
- Delete an Apollo Record\*



# G-OnRamp: The Genomics Education Partnership

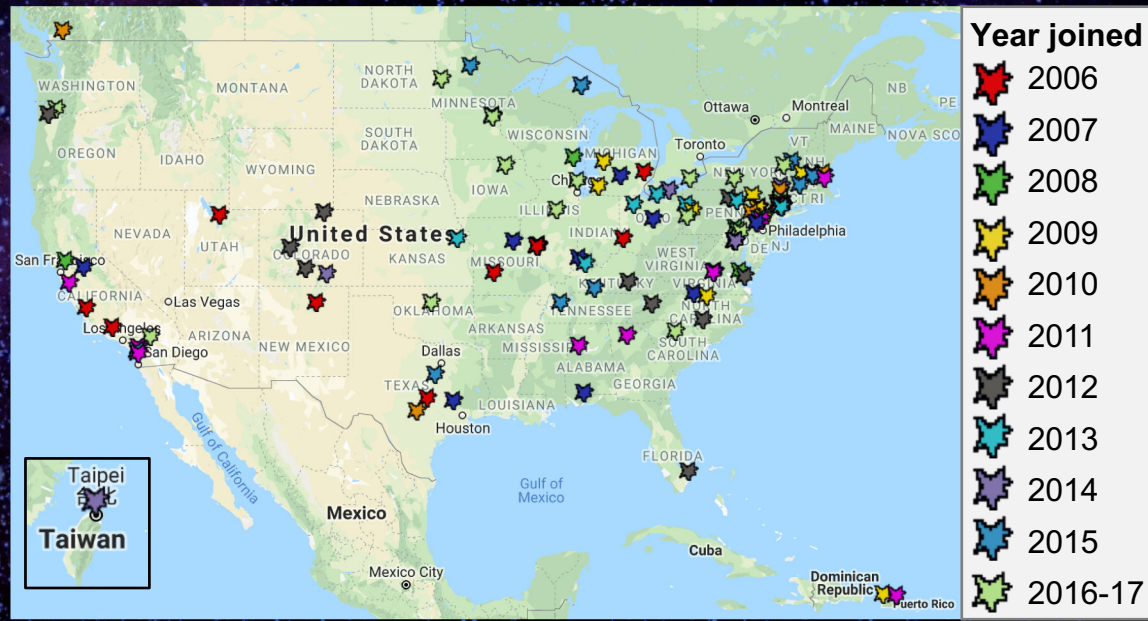
## Goals:

- Integrate genomics into the undergraduate biology curriculum
- Integrate research thinking into the academic year curriculum



## Partnership Scope:

- >100 faculty
- >1000 undergraduates participate annually
- G-OnRamp enables a diversity of projects creating manually-curated gene models





# G-OnRamp: Progress and Results

- 6 workshops from 2016-2018
- 65 participants from 40+ institutions
- > 20 genome browsers created from participant submissions
  - Assembly sizes: **70Mb - 3.9Gb**
  - Number of scaffolds: **54 - 402,501**
- Genome Browsers hosted on the CyVerse Data Store
  - <http://g-onramp.org>  
→ “View Genome Browser”





# G-OnRamp: Long-Term Hub Storage on CyVerse

The screenshot shows the Galaxy web interface. The top navigation bar includes 'Galaxy', 'Analyze Data', 'Workflow', 'Visualize', 'Shared Data', 'Admin', and 'Help'. On the left sidebar, under 'Tools', 'CyVerse' is selected. Below it, 'G-OnRamp Tools' is listed with a description: 'CyVerse interaction via iRODS This Galaxy tool is used to upload files to a user's Data Store on CyVerse'. Under 'Workflows', 'All workflows' is listed. The main panel shows the 'CyVerse interaction via iRODS' tool configuration. It includes a title bar with 'Options', a description, a 'Data to upload' section with a file selection button and a dropdown menu showing '27: JBrowse Archi...', a 'Subdirectory to install to - descriptive is good' text input field, a 'Place Timestamp in directory name for uniqueness? (ex: aName\_20180604T212729)' section with two radio buttons ('Timestamp Directory' and 'Don't Timestamp Directory'), 'CyVerse user name' and 'CyVerse user password' text input fields, and an 'Execute' button at the bottom.

**Galaxy** Analyze Data Workflow Visualize Shared Data Admin Help

**Tools**

CyVerse

**G-OnRamp Tools**

CyVerse interaction via iRODS This Galaxy tool is used to upload files to a user's Data Store on CyVerse

**Workflows**

- All workflows

**CyVerse interaction via iRODS**

This Galaxy tool is used to upload files to a user's Data Store on CyVerse (Galaxy Version 0.1.0)

**Data to upload**

27: JBrowse Archi...

**Subdirectory to install to - descriptive is good**

**Place Timestamp in directory name for uniqueness? (ex: aName\_20180604T212729)**

☒ Timestamp Directory ☐ Don't Timestamp Directory

**CyVerse user name**

**CyVerse user password**

**Execute**

## CyVerse Upload Tool

- Requires a (free) CyVerse account

## Advantages of External Storage:

- **No Cost:** CyVerse offers 100GB free (more on request)
- **Flexible:** Accessible from anywhere, by anyone
- **Simple:** View hubs without running a server

## Disadvantages:

- **Privacy:** Accessible from anywhere, by anyone
- **Just the Hub:** No workflow intermediates
- **No Analytics:** Cannot run analysis tools or workflows



# G-OnRamp: Getting Started



**Virtual Machine for Local  
Deployment**

**Detailed instructions at <http://g-onramp.org>:**

Get started with G-OnRamp

G-OnRamp Ubuntu  
Virtual Machine Image



CloudLaunch  
Deployment



G-OnRamp Training  
Materials





# G-OnRamp: Getting Started

<http://galaxy1.cloudve.org>

Create user from **Login or Register** menu:

- **Shared Data** -> **Histories** -> "Apollo Demo"
- **AND/OR**
- **Shared Data** -> **Data Libraries** -> Import "Intro to G-OnRamp"
- **Shared Data** -> **Workflows** -> Workflow of choice

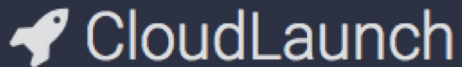
The screenshot shows the G-OnRamp Galaxy web interface. The top navigation bar includes 'Galaxy', 'Analyze Data', 'Workflow', 'Visualize', 'Shared Data', 'Help', 'Login or Register', and 'Using 0 bytes'. The 'Shared Data' menu is open, showing options: 'Data Libraries', 'Histories', 'Workflows', 'Visualizations', and 'Pages'. The 'Login or Register' button is highlighted with an orange box. The main content area features the G-OnRamp logo and a section titled 'G-OnRampKickStart launched Galaxy release 18.05' with several blue buttons: 'G-OnRamp Home »', 'Learn G-OnRamp »', 'View G-OnRamp Workshop Assembly Hubs »', 'Configuring Galaxy »', and 'Installing Tools »'. A blue arrow points to the 'Learn G-OnRamp »' button. The left sidebar contains a 'Tools' section with a search bar and various tool categories like 'Get Data', 'Collection Operations', 'Text Manipulation', etc. The right sidebar shows a 'History' section with a search bar and a message: 'This history is empty. You can load your own data or get data from an external source'.



# G-OnRamp: Production Deployments



**Amazon AMI for EC2  
Deployment**

The CloudLaunch logo, featuring a white rocket icon to the left of the word "CloudLaunch" in a white sans-serif font, all on a dark grey rectangular background.

**[launch.usegalaxy.org](http://launch.usegalaxy.org)**



**Ansible for Flexible  
Deployment**

**Detailed instructions at <http://g-onramp.org>:**

Get started with G-OnRamp

G-OnRamp Ubuntu  
Virtual Machine Image



CloudLaunch  
Deployment



G-OnRamp Training  
Materials





# CloudLaunch: Deploy G-OnRamp Easily

## Appliance Catalog

An online depot to discover and launch pre-configured software for a variety of clouds.

🔍 Search for an appliance



### Genomics Virtual Lab (GVL)

A versatile genomics workbench with Galaxy, RStudio and Jupyter.  
USE THIS FOR LATEST GALAXY.



### Galaxy CloudMan

Pre-configured Galaxy instance on a scalable cluster-in-the-cloud.  
DEPRECATED - USE THE GVL INSTEAD.



### G-OnRamp

An integrated, web-based environment for the interactive annotation of any eukaryotic genome.

### G-OnRamp

G-OnRamp is an integrated, web-based, and scalable environment that enables biologists to utilize large genomic datasets for the interactive annotation of any eukaryotic genome. It also serves as a platform to introduce undergraduates to "big data," to train them in one type of analysis (genome annotation) that is based on using large datasets. G-OnRamp is a collaboration between the Galaxy Project (<https://galaxyproject.org/>) and the Genomics Education Partnership (<https://gep.wustl.edu/>). For more information, documentation, and tutorials, visit <http://g-onramp.org>.

Maintainer: G-OnRamp team

Support URL: <http://g-onramp.org>

Added: Dec 7, 2018



# CloudLaunch: Log in

## Login

Please Sign In

Before you can launch any appliances via this server, you need to log in. Logging in will allow you to launch and monitor appliances you have launched in the past as well as store your credentials for convenience.

Login via one of the sites listed below:



Login with Facebook



Login with Github



Login with Google



Login with Twitter



# CloudLaunch: Log in

Fill out the form below to launch the selected appliance. ☒

1 Select Target Cloud

2 Select Appliance Settings

3 Launch!

Which version of this appliance would you like to launch?

1.0

On which cloud would you like to launch your appliance?

Amazon US East 1 - N. Virginia

For detailed instructions on how to obtain credentials for this cloud, [click here](#).

What type of credentials do you want to use?



Temporary Credentials



Saved Credentials

Select saved credentials

gxyoutreach

NEXT >



# CloudLaunch: Configure Instance

## Launching G-OnRamp appliance



Fill out the form below to launch the selected appliance. ☒

1 Select Target Cloud 2 Select Appliance Settings 3 Launch!

Provide a name for your deployment

**g-onramp-deployment**

A deployment name helps you identify your appliance. The name must be at most 63 characters long and can consist of lowercase letters, numbers, and dashes.

What type of virtual hardware would you like to use?

**c5.4xlarge**



16 VCPUs



32.0 GB RAM



0 GB Disk



☐ Advanced cloud launch options

[< PREVIOUS](#)









**LAUNCH**



# CloudLaunch: Ready, Set, Annotate

## My Appliances

This page is a list of active appliances you have launched. 

Name	Created	Status	Access address	Actions
<b>g-onramp-deployment</b> Appliance: G-OnRamp      Version: 1.0 Cloud: amazon-us-east-n-virginia Credentials: gxyoutreach	a few seconds ago	 PROGRESSING Waiting for application to become ready at <a href="http://34.197.88.248/">http://34.197.88.248/</a>		 
				
<b>g-onramp-deployment</b> Appliance: G-OnRamp      Version: 1.0 Cloud: amazon-us-east-n-virginia Credentials: gxyoutreach	5 minutes ago	 RUNNING	<a href="http://34.197.88.248/">http://34.197.88.248/</a> 	 



# G-OnRamp: Components in Concert





# G-OnRamp: Progress and Results

## Read the Pre-publication Manuscript

G-OnRamp: A Galaxy-based platform for creating genome browsers for collaborative genome annotation

- <http://g-onramp.org/2018-preprint>

The screenshot shows the bioRxiv preprint server interface. At the top left is the Cold Spring Harbor Laboratory logo. The bioRxiv logo is prominently displayed with the tagline 'THE PREPRINT SERVER FOR BIOLOGY'. Navigation links include HOME, ABOUT, SUBMIT, ALERTS / RSS, and CHANNELS. A search bar is located on the right. The main content area features the title 'G-OnRamp: A Galaxy-based platform for creating genome browsers for collaborative genome annotation' by Yating Liu, Luke Sargent, Wilson Leung, Sarah C.R. Elgin, and Jeremy Goecks. The DOI is https://doi.org/10.1101/499558. A note states: 'This article is a preprint and has not been peer-reviewed [what does this mean?]'.

Navigation and interaction options include 'New Results', 'Comment on this paper', 'Previous', 'Next', 'Download PDF', 'Email', 'Share', 'Citation Tools', and a 'G+' button. The 'Subject Area' section shows 'Genomics' as the selected category. The 'Abstract' section is active, displaying the summary: 'Summary: G-OnRamp provides a user-friendly, web-based platform for collaborative, end-to-end annotation of eukaryotic genomes using UCSC Assembly Hubs and JBrowse/Apollo'.

**Abstract**

Summary: G-OnRamp provides a user-friendly, web-based platform for collaborative, end-to-end annotation of eukaryotic genomes using UCSC Assembly Hubs and JBrowse/Apollo



# G-OnRamp: The Team, Acknowledgements

## The Team:



Yating Liu <sup>1</sup>



Wilson Leung <sup>1</sup>



Sarah Elgin <sup>1</sup>



Jeremy Goecks <sup>2</sup>



Luke Sargent <sup>2</sup>

1. Washington University in St. Louis

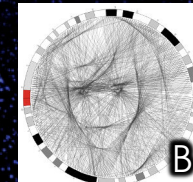
2. Oregon Health & Science University

## Thanks to:

- The Galaxy Core Team
- The Galaxy Genome Annotation Community
- GalaxyKickStart developers
- the GMOD team

## Bonus Thanks:

- Enis Afgan <sup>A</sup> (for CloudLaunch support)
- Helena Rasche <sup>B</sup> (Apollo/Galaxy tooling)
- Nathan Dunn <sup>C</sup> (Apollo user management)
- GEP members (beta workshop participants)
- You! (for listening)



Funding: NIH (BD2K grant 1R25GM119157 to SCRE)



**G-OnRamp:** Questions, More Information

# Questions?

**More Information:**

<http://g-onramp.org>

Documentation, generated genome browsers, VM downloads, more

**Preprint Manuscript**

<http://g-onramp.org/2018-preprint>

**Live G-OnRamp Demo**

<http://galaxy1.cloudve.org>

**Ansible Github Repo**

<https://github.com/goeckslab/gonrampkickstart>