G-OnRamp: Create Genome Browsers for Collaborative Eukaryotic Genome Annotations

Luke Sargent[1], Yating Liu[2], Wilson Leung[2], Jeremy Goecks[1], Sarah C.R. Elgin[2]

[1]Oregon Health & Science University, [2]Washington University in St. Louis

G-OnRamp Project Objectives: Create an easy-to-use web environment for anyone to utilize large genomics datasets for annotating any eukaryotic genome, and provide educators with a platform to train undergraduate students on "big data" biomedical analyses using a problem-based approach.

Abstract

G-OnRamp (http://g-onramp.org) is a collaboration between two successful and long-running projects — the Genomics Education Partnership (GEP; http://gep.wustl.edu) and the Galaxy Project (http://galaxyproject.org). G-OnRamp provides researchers with an integrated, web-based, scalable environment for interactive annotation of eukaryotic genomes using large genomic datasets. It also provides educators with a platform to help undergraduates develop "big data" science skills through eukaryotic genome annotation.

GEP is a consortium of faculty from over 100 colleges/universities that provides Course-based Undergraduate Research Experiences (CUREs) in bioinformatics/genomics for students at all levels. GEP faculty currently use the gene annotation of multiple Drosophila and parasitoid wasps species to introduce genomics and research thinking to undergraduates. Galaxy is a popular open-source, web-based scientific gateway for accessible, reproducible, and transparent analyses of large biomedical datasets. G-OnRamp extends Galaxy with tools and workflows that create UCSC Assembly Hubs and Apollo/JBrowse genome browsers with evidence tracks for sequence similarity to proteins and transcripts from an informant genome, ab initio gene predictions, RNA-Seq data, and repeats. G-OnRamp also provides tools for creating Apollo instances from JBrowse genome browsers and for managing Apollo user accounts, thereby enabling multiple users to collaborate on a genome annotation project in research and educational settings. Users can optionally upload UCSC Assembly Hubs and JBrowse genome browsers to the CyVerse Data Store for statically-hosted, external visualization. Users can deploy G-OnRamp on the Cloud (Amazon EC2) using CloudLaunch, or locally using a virtual appliance. For researchers, G-OnRamp provides a dedicated, customizable analysis platform for public or private deployments. For educators, G-OnRamp provides a platform to design CUREs based on their favorite eukaryotic species (e.g., parasitoid wasps) and research questions (e.g., evolution of a chromosomal domain, genes involved in a pathway).

G-OnRamp training workshops have attracted 65 participants from 40+ institutions across the world, generating 18 genome browsers for genomes that range from 70MB to 3.9GB in size, with 13 genomes featuring RNA seq data (https://bit.ly/2AqDLq8). Future plans for G-OnRamp include developing additional tools for validating and integrating annotations, plus adding new evidence tracks (e.g., ChIP-seq). Supported by NIH 1R25GM119157.
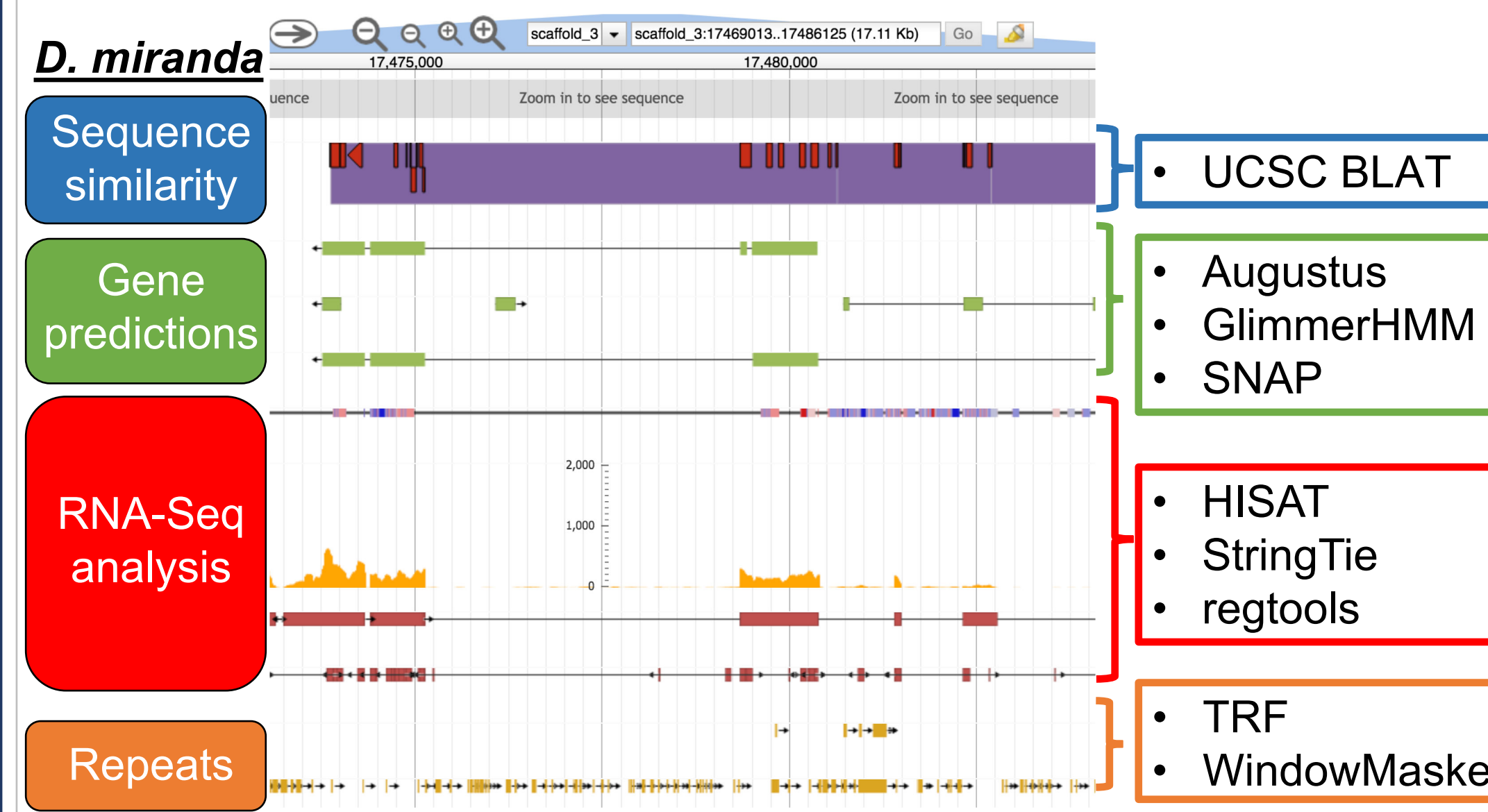
G-OnRamp | http://g-onramp.org
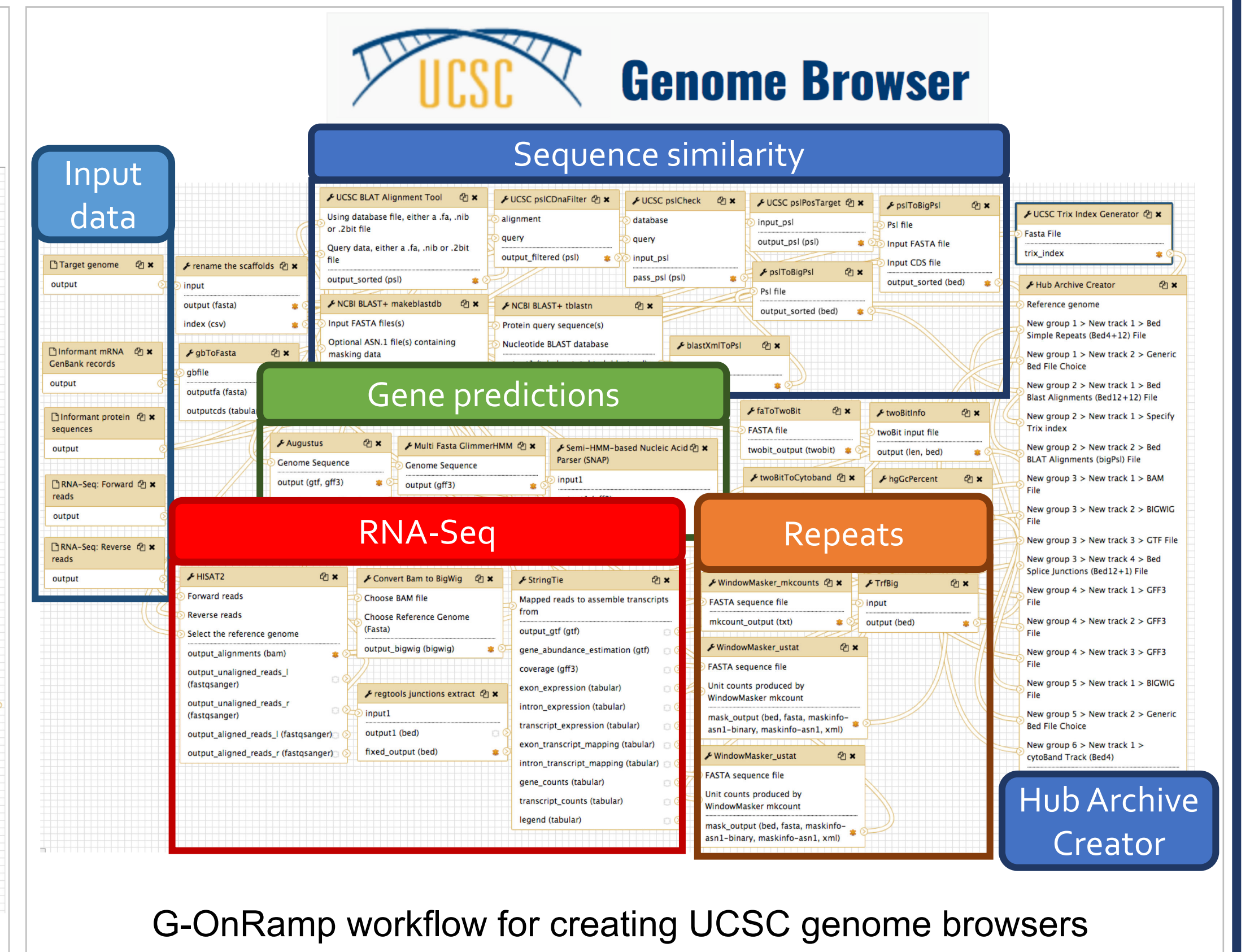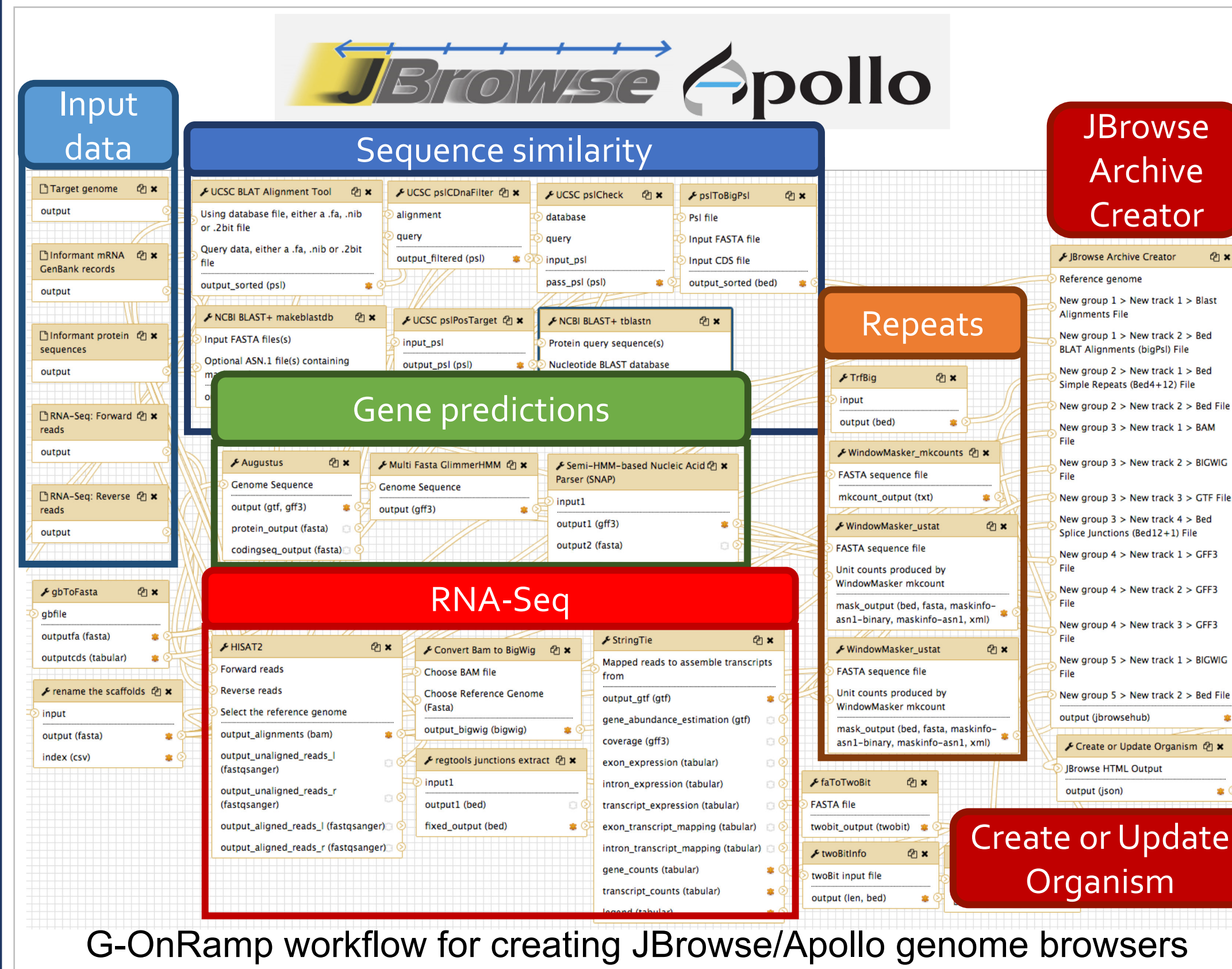
- Collaboration between the **GEP** and **Galaxy**
- Produces **UCSC Assembly Hubs** and **JBrowse/Apollo genome browsers** for any eukaryotes
- User can upload hubs to **CyVerse** for free, long-term storage

https://galaxyproject.org
open, web-based platform for bio-informatics analyses

- **Accessible**: does not require programming experience
- **Reproducible**: repeat analyses that contain multiple steps
- **Transparent**: share and publish workflows and results

Genomics Education Partnership (GEP; http://gep.wustl.edu)

- Introduces **genomics and bioinformatics** into the undergraduate curriculum
- **>100** faculty from **>100** affiliated schools
- **>1000** undergraduates participate annually
- Maintains up-to-date curriculum
- Engages students in genomics research through gene annotation, phylogenetic comparisons
- Founding member of the **Genomics Education Alliance** (GEA; https://qubeshub.org/community/groups/gea)

Genomics Education Alliance

- NSF RCN in Undergraduate Biology Education
- Unites existing genomics education networks/faculty, aiming to make genomics accessible to students from any college or university by maintaining a common, accessible platform
- Identifies and curates common tools, curricular and assessment materials, and training strategies to enable CUREs in genomics / bioinformatics; materials will be posted on the QUBES platform

G-OnRamp Training Workshops

- 6 workshops from 2016-2018
- 65 participants from 40+ institutions
- Over 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"

Acknowledgements

Contacts

Sarah C.R. Elgin
selgin@wustl.edu

Jeremy Goecks
goecksj@ohsu.edu

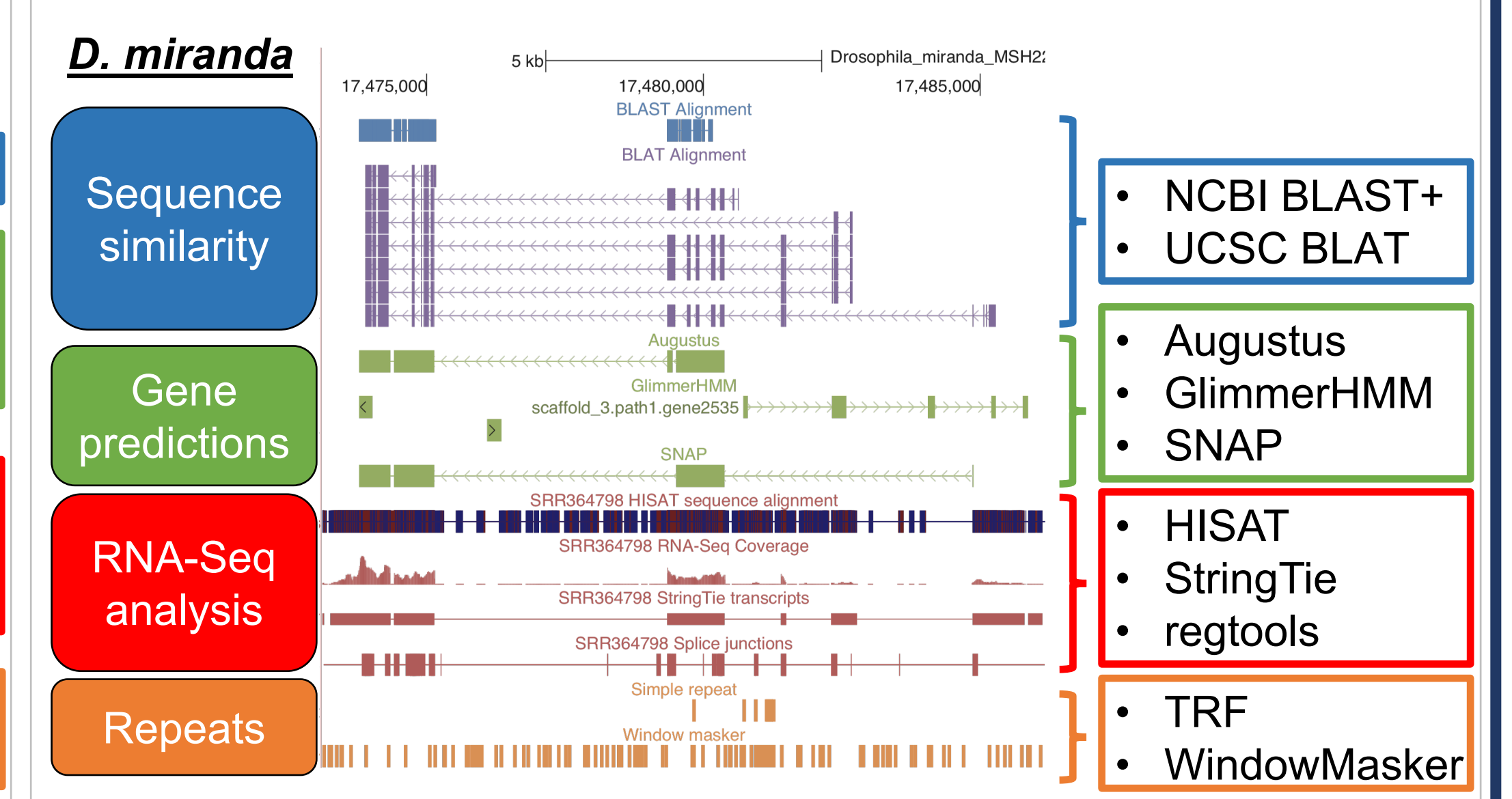## G-OnRamp Architecture

G-OnRamp workflows

- Extends Galaxy with tools and workflows that create UCSC Assembly Hubs and JBrowse/Apollo genome browsers
- Four sub-workflows: sequence similarity, gene predictions, RNA-Seq and repeats
- Convert and validate outputs from sub-workflows to file formats that are compatible with genome browsers

G-OnRamp workflow for creating JBrowse/Apollo genome browsers

G-OnRamp workflow for creating UCSC genome browsers

### Use the JBrowse/Apollo browser to visualize multiple genomic datasets

D. miranda

- Sequence similarity — • UCSC BLAT
- Gene predictions — • Augustus • GlimmerHMM • SNAP
- RNA-Seq analysis — • HISAT • StringTie • regtools
- Repeats — • TRF • WindowMasker

Use the UCSC genome browser to visualize multiple genomic datasets

D. miranda

- Sequence similarity — • NCBI BLAST+ • UCSC BLAT
- Gene predictions — • Augustus • GlimmerHMM • SNAP
- RNA-Seq analysis — • HISAT • StringTie • regtools
- Repeats — • TRF • WindowMasker

Collaborative Genome Annotation with Apollo

Apollo http://genomearchitect.github.io

- A JBrowse plug-in that supports collaborative genome annotation
- Automatic synchronization of annotation updates
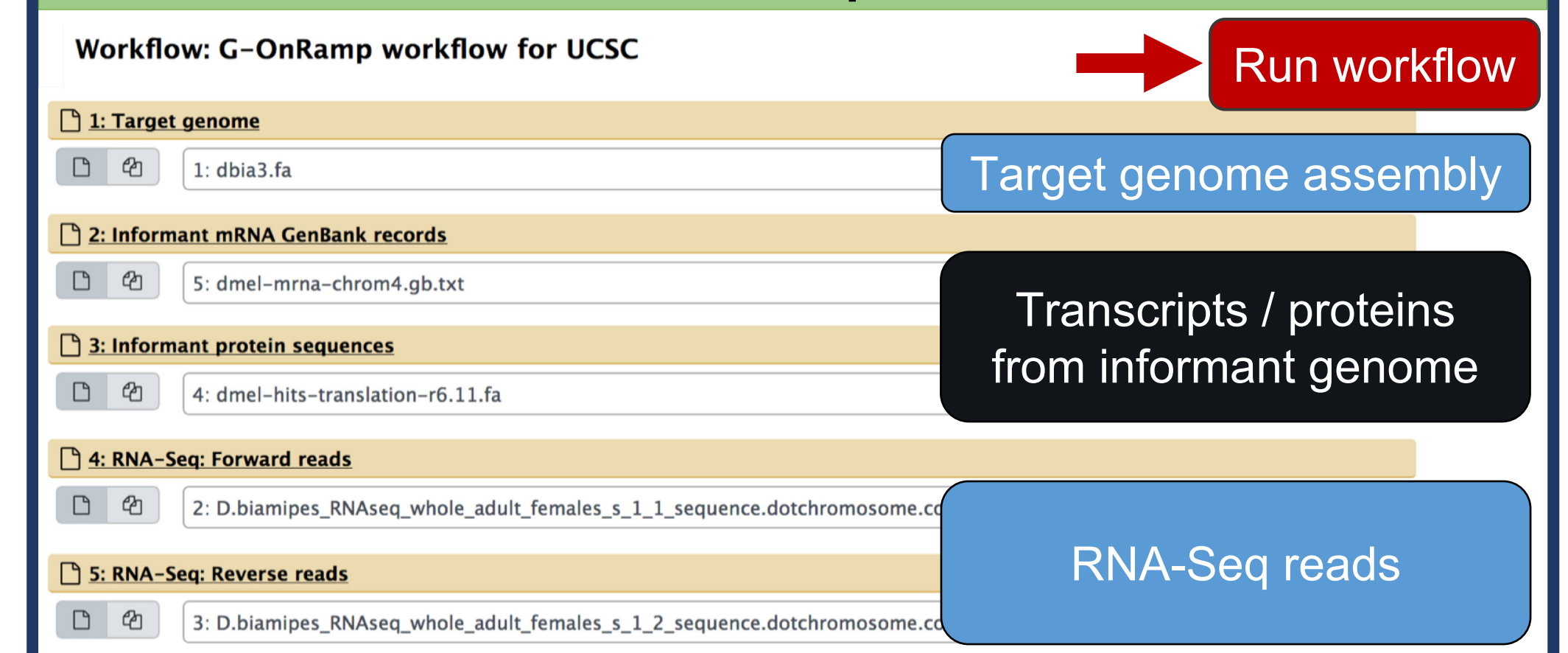- Annotating genes, transposable elements, and non-coding RNAs
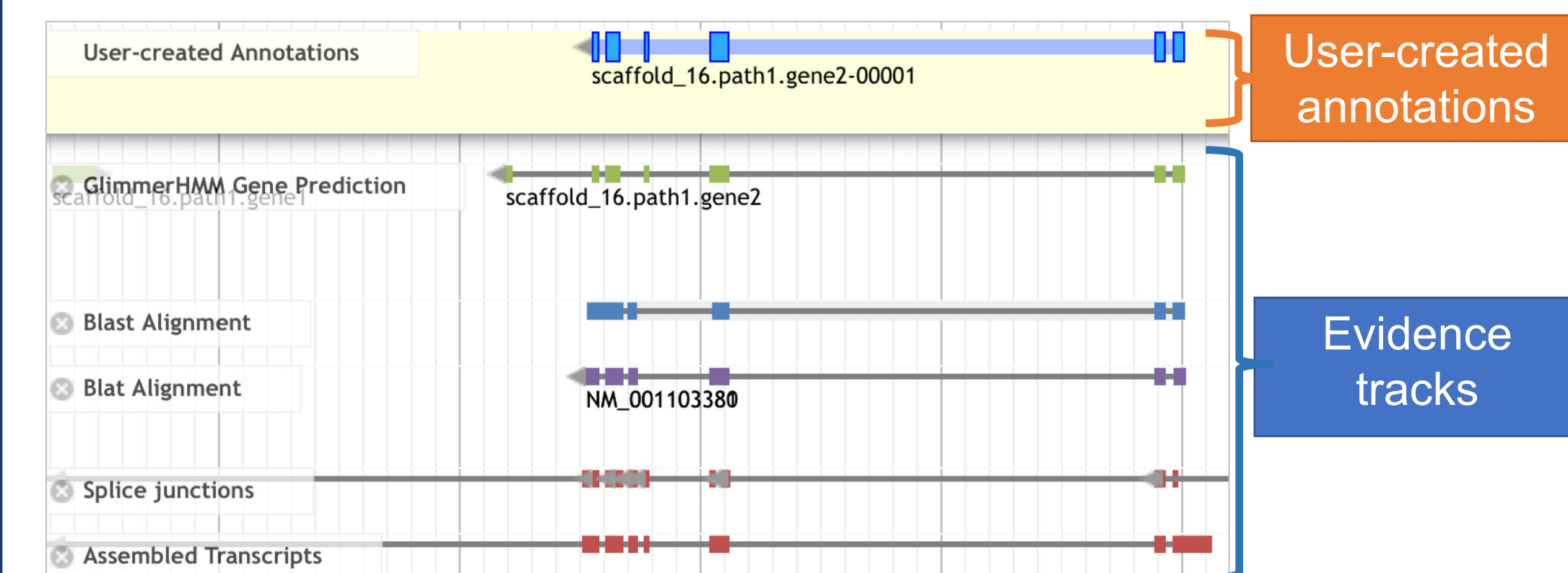
Integrations between Galaxy and Apollo

- Use **Create or Update Organism** tool to add or update assembly hubs on Apollo
- Use **Apollo User Manager** to manage multiple Apollo users
  - Use Apollo for collaborative genome annotation in class

G-OnRamp Products

http://g-onramp.org/deployments

Install a G-OnRamp virtual machine image locally

- Use for testing and training due to performance limitations

Launch G-OnRamp on the cloud

- Amazon Web Services
- CloudLaunch
  - https://launch.usegalaxy.org

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

bioRχiv

G-OnRamp Tools per Sub-Workflow

Sequence similarity	Gene predictions	RNA-Seq	Repeats	Create Genome Browsers
NCBI BLAST+	Augustus	HISAT	Tandem Repeats Finder	Hub Archive Creator
UCSC BLAT	Glimmer HMM	StringTie	Window Masker	JBrowse Archive Creator
	SNAP	regtools		

Create Genome Browsers by Running the G-OnRamp Workflow

Workflow: G-OnRamp workflow for UCSC — **Run workflow**

- Target genome assembly
- Transcripts / proteins from informant genome
- RNA-Seq reads

Related PAG Posters and Talks

- **PO0033:** The **Genomics Education Alliance**: Working Together to Facilitate Undergraduate Research in Genomics
- **PE0034:** The **Genomics Education Partnership**: Collaborating With Science Partners to Engage Undergraduates in Genomics Research
- **PE1178:** **Apollo**: Genome Annotation Editing and Publishing for Research Groups
- **W914:** The **Genomics Education Alliance**: Working Together to Facilitate Undergraduate Research in Genomics
  - Sun, Jan 13 @ 9:25 AM, Town and Country Meeting House
- **W477-82:** **Galaxy**: An Open Platform for Data Analysis & Integration
  - Tues, Jan 15 @ 4:00 PM, California Room