G-OnRamp

Using genome browsers constructed by G-OnRamp to provide students with a Course-based Undergraduate Research Experience in genome annotation

Wilson Leung

Washington University in St. Louis

http://g-onramp.org

01/2020

Creating genome browsers requires **substantial bioinformatics expertise**

Manage multiple data formats

```
$ faToTwoBit newGenome.fa newGenome.2bit
```

FASTA

twoBit

Manage input and output of multiple tools

```
$ twoBitInfo newGenome.2bit stdout | \
    sort -k 2,2nr > newGenome.chrom.sizes
```

Set up genome database

```
INSERT INTO defaultDb (genome, name)
VALUES ("Ganaspis species 1", "ganSpe1");
```

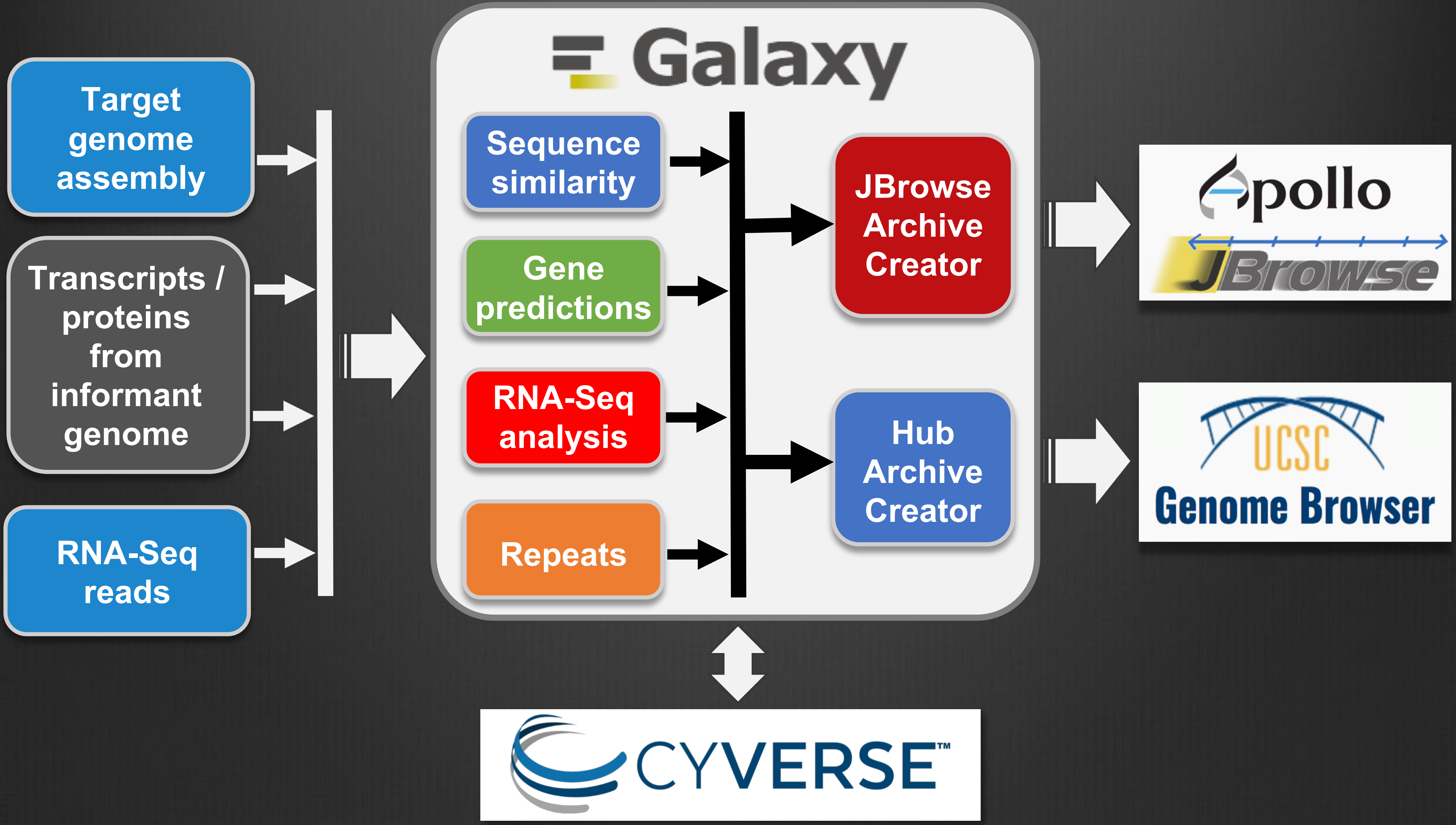
Step 0: Download and install bioinformatics tools

GEP + Galaxy = G-OnRamp

Goals:

- Enable researchers and biology faculty to create **genome browsers** for their favorite **eukaryotic species**
- Enable faculty to use the genome browsers created by G-OnRamp to **engage students in genomics research**

Use G-OnRamp to create genome browsers for **eukaryotic genomes**



Target genome assembly

Transcripts / proteins from informant genome

RNA-Seq reads

Galaxy

Sequence similarity

Gene predictions

RNA-Seq analysis

Repeats

JBrowse Archive Creator

Hub Archive Creator

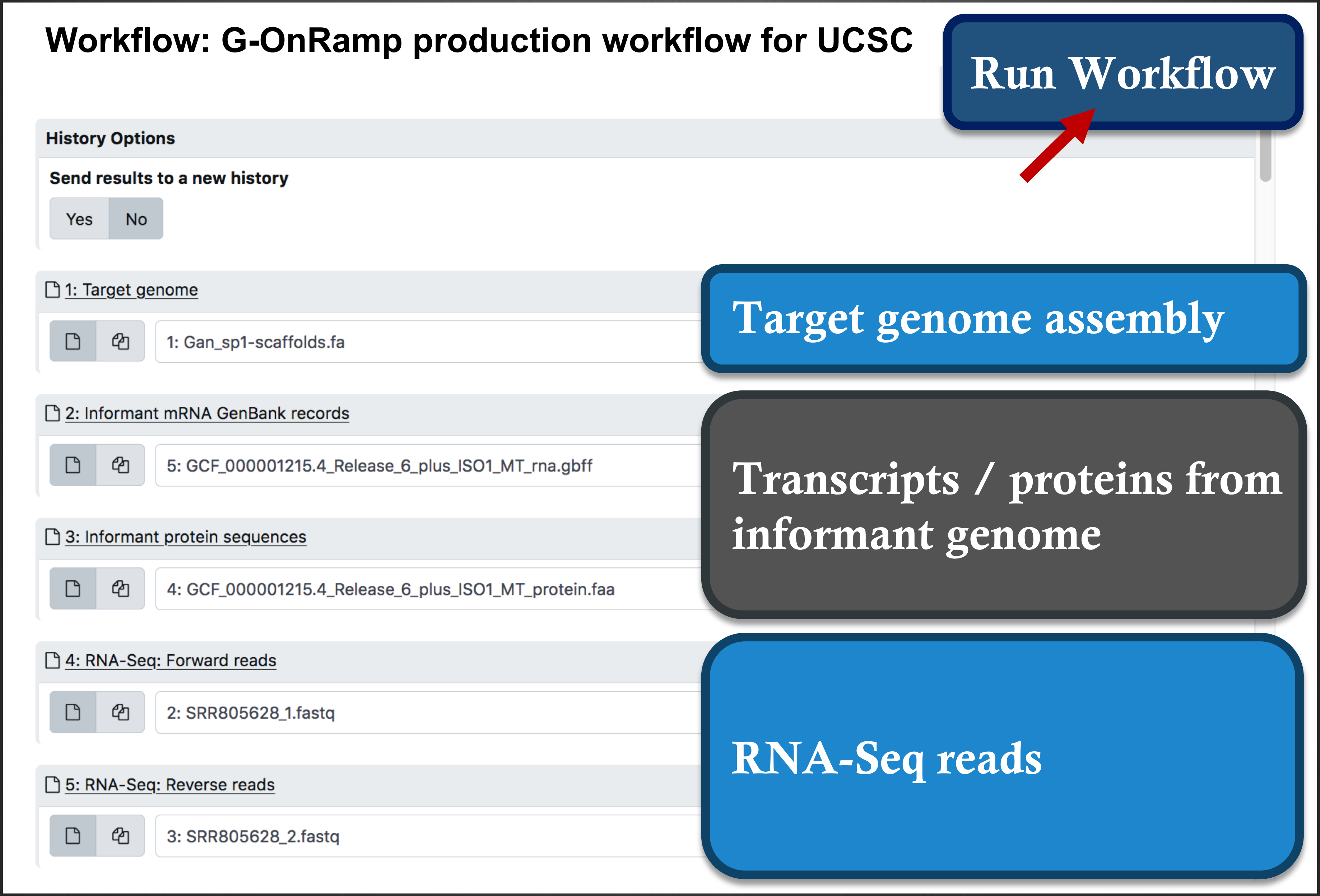
Apollo

JBrowse

UCSC Genome Browser

CYVERSE™

It is easy to get started with G-OnRamp



Workflow: G-OnRamp production workflow for UCSC

Run Workflow

History Options

Send results to a new history

Yes [✓ No]

1: Target genome

1: Gan_sp1-scaffolds.fa

2: Informant mRNA GenBank records

5: GCF_000001215.4_Release_6_plus_ISO1_MT_rna.gbff

3: Informant protein sequences

4: GCF_000001215.4_Release_6_plus_ISO1_MT_protein.faa

4: RNA-Seq: Forward reads

2: SRR805628_1.fastq

5: RNA-Seq: Reverse reads

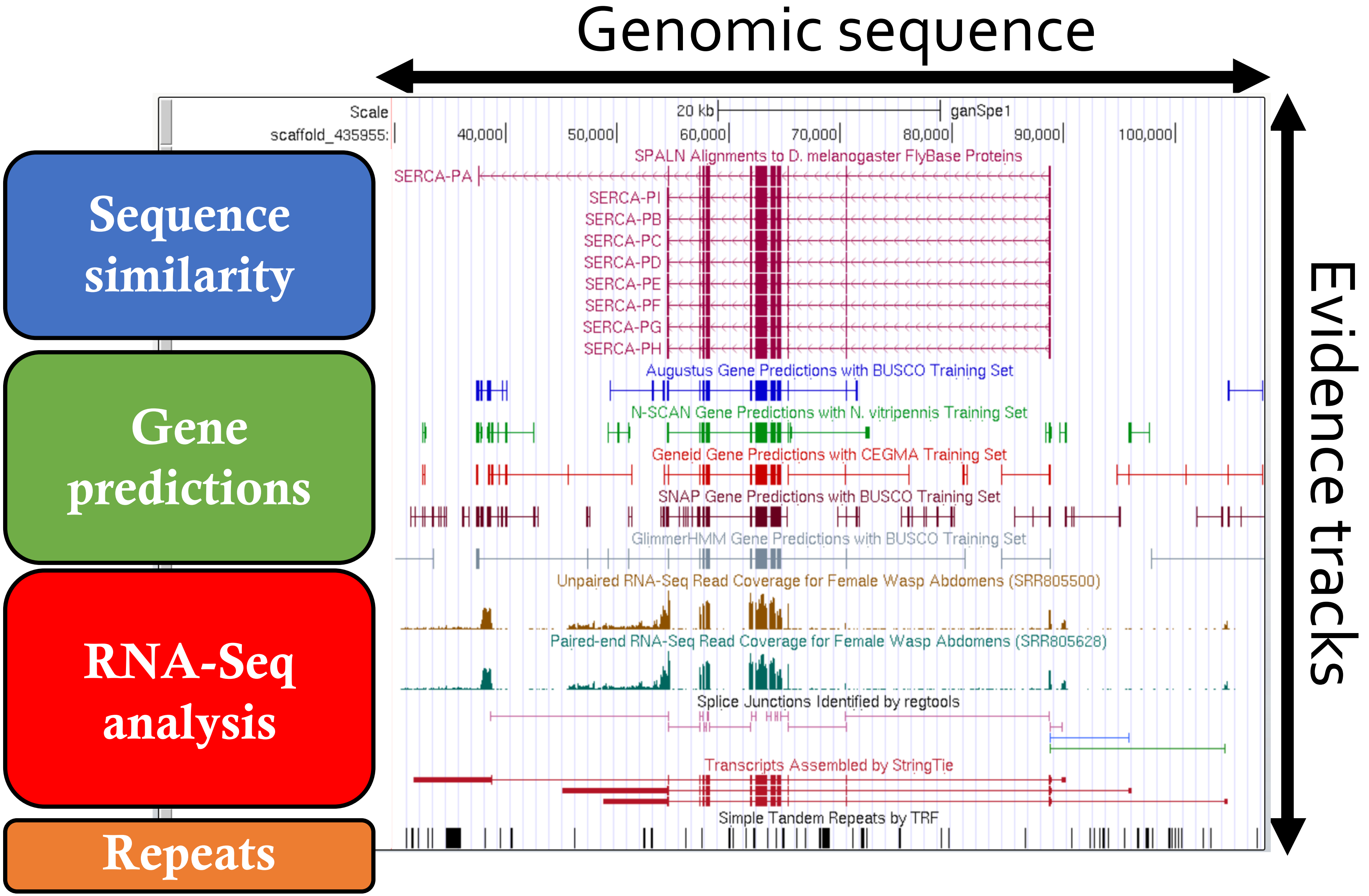
3: SRR805628_2.fastq

Target genome assembly

Transcripts / proteins from informant genome

RNA-Seq reads

UCSC Assembly Hub for *Ganaspis species 1*



Genomic sequence

Evidence tracks

Sequence similarity

Gene predictions

RNA-Seq analysis

Repeats

SPALN Alignments to D. melanogaster FlyBase Proteins

Augustus Gene Predictions with BUSCO Training Set

N-SCAN Gene Predictions with N. vitripennis Training Set

Geneid Gene Predictions with CEGMA Training Set

SNAP Gene Predictions with BUSCO Training Set

GlimmerHMM Gene Predictions with BUSCO Training Set

Unpaired RNA-Seq Read Coverage for Female Wasp Abdomens (SRR805500)

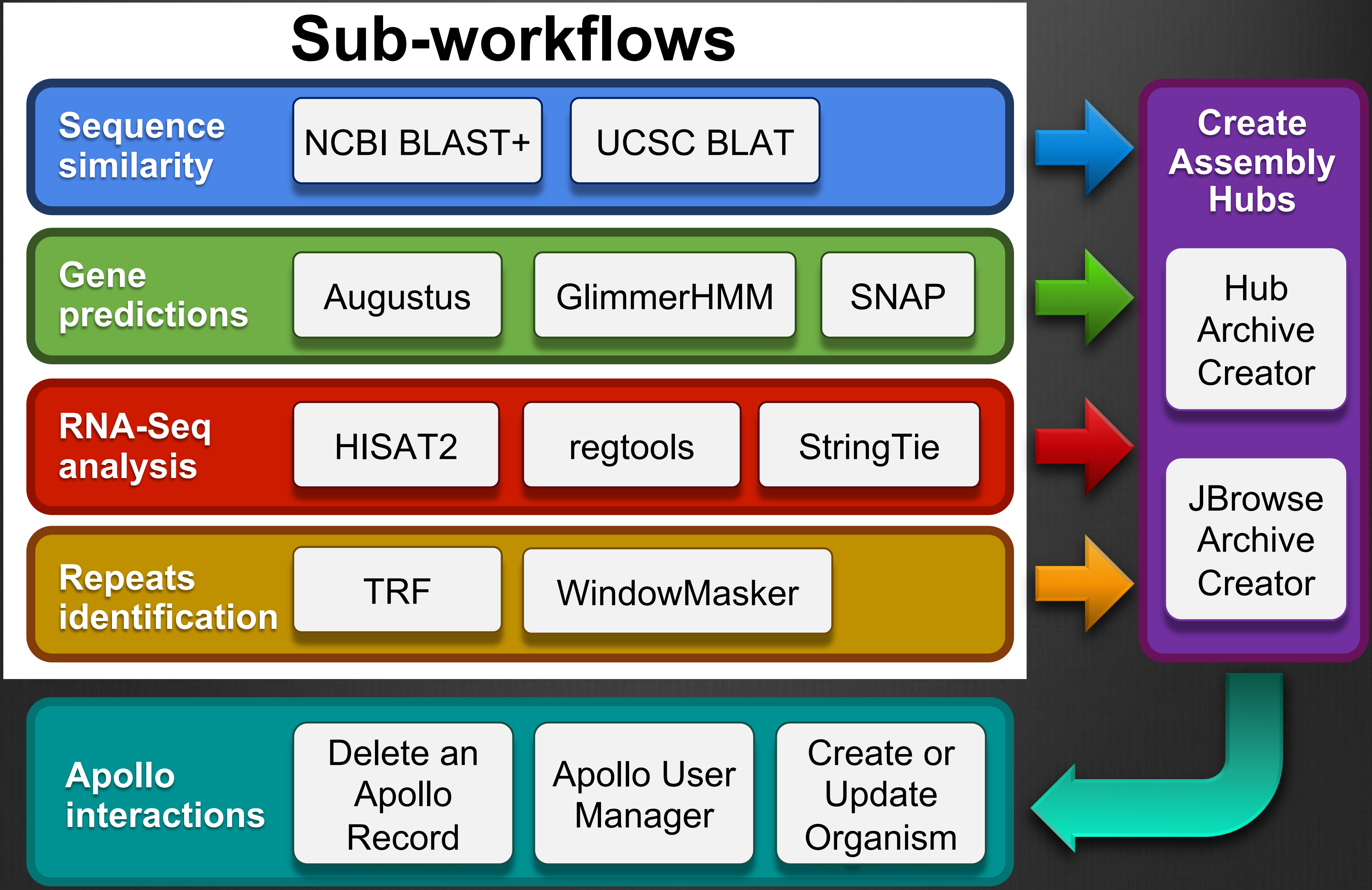
Paired-end RNA-Seq Read Coverage for Female Wasp Abdomens (SRR805628)

Splice Junctions Identified by regtools

Transcripts Assembled by StringTie

Simple Tandem Repeats by TRF

G-OnRamp is composed of **sub-workflows**



Sub-workflows

Sequence similarity: NCBI BLAST+, UCSC BLAT

Gene predictions: Augustus, GlimmerHMM, SNAP

RNA-Seq analysis: HISAT2, regtools, StringTie

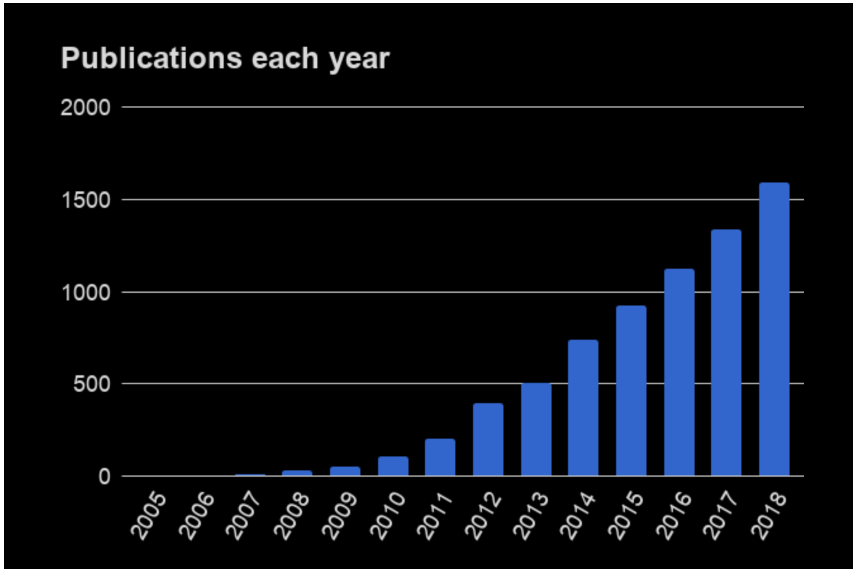
Repeats identification: TRF, WindowMasker

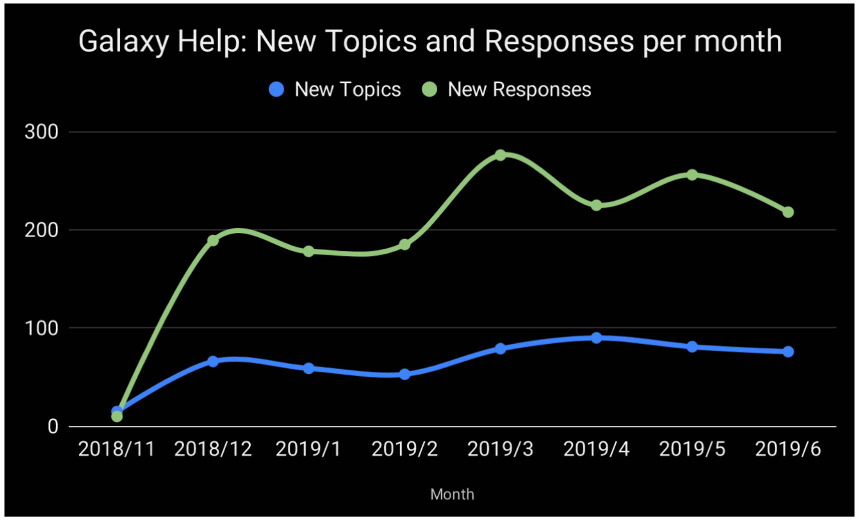
Create Assembly Hubs: Hub Archive Creator, JBrowse Archive Creator

Apollo interactions: Delete an Apollo Record, Apollo User Manager, Create or Update Organism

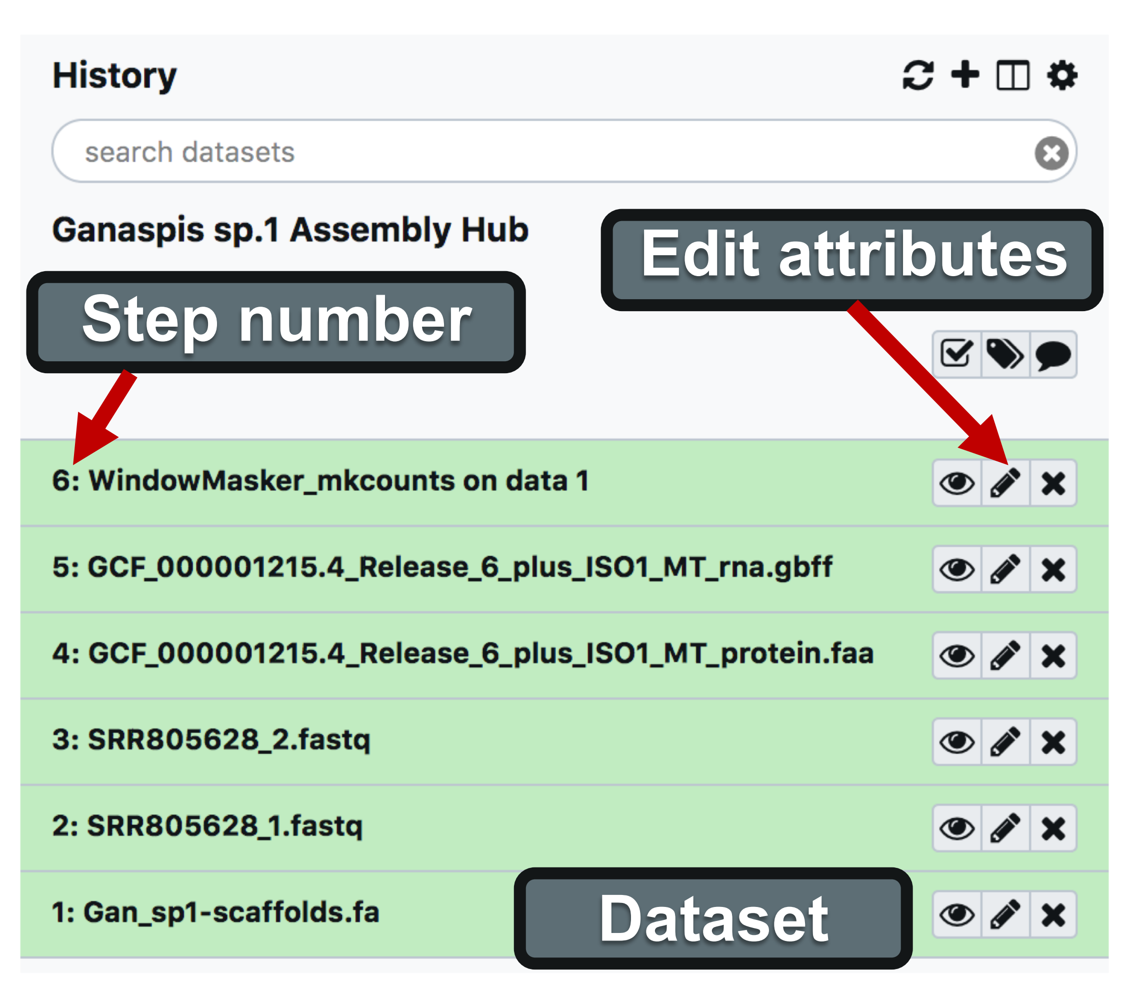
Galaxy PROJECT an open, web-based platform for bioinformatics analyses

- **Accessible**
  - Does not require programming experience
- **Reproducible**
  - Easily repeat analyses that contain multiple steps
- **Transparent**
  - Share and publish workflows and results


Publications each year


Galaxy Help: New Topics and Responses per month

Galaxy automatically keeps track of **each step of the analysis** (History)



History

search datasets

Ganaspis sp.1 Assembly Hub

6: WindowMasker_mkcounts on data 1

5: GCF_000001215.4_Release_6_plus_ISO1_MT_rna.gbff

4: GCF_000001215.4_Release_6_plus_ISO1_MT_protein.faa

3: SRR805628_2.fastq

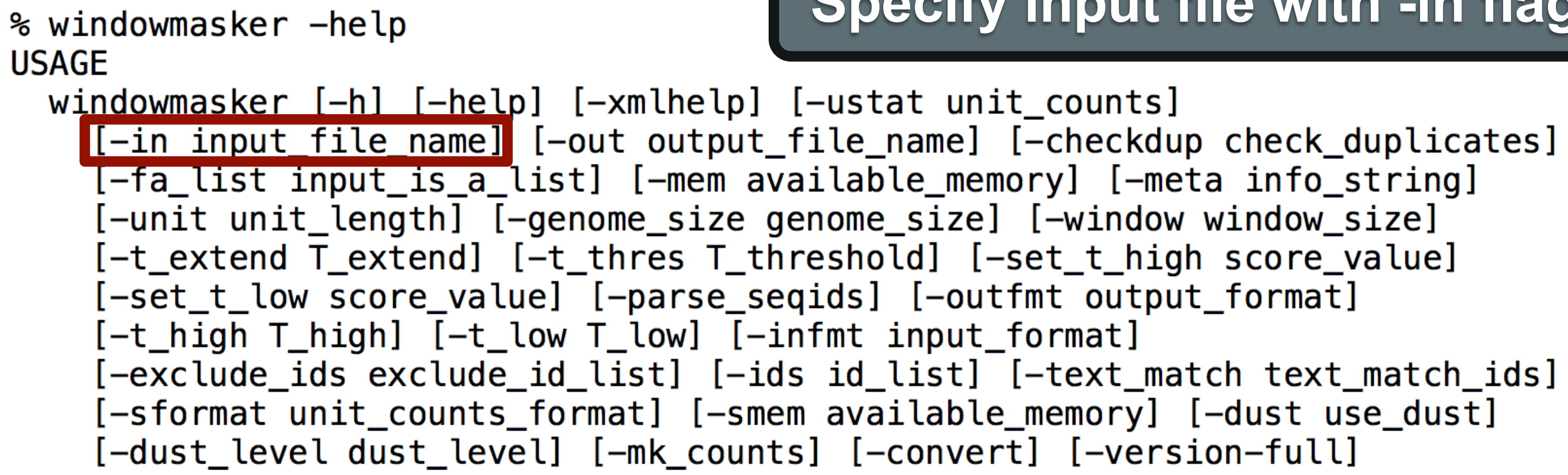
2: SRR805628_1.fastq

1: Gan_sp1-scaffolds.fa

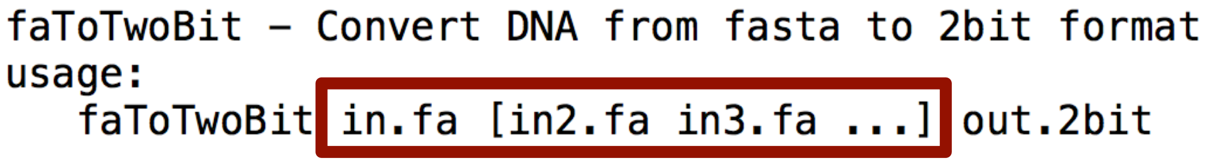
- Keep track of the metadata associated with each Dataset
- Examine details of each step of the analysis
- Easily repeat each step of the analysis

Bioinformatics tools often have **different user interfaces**

Specify input file with -in flag



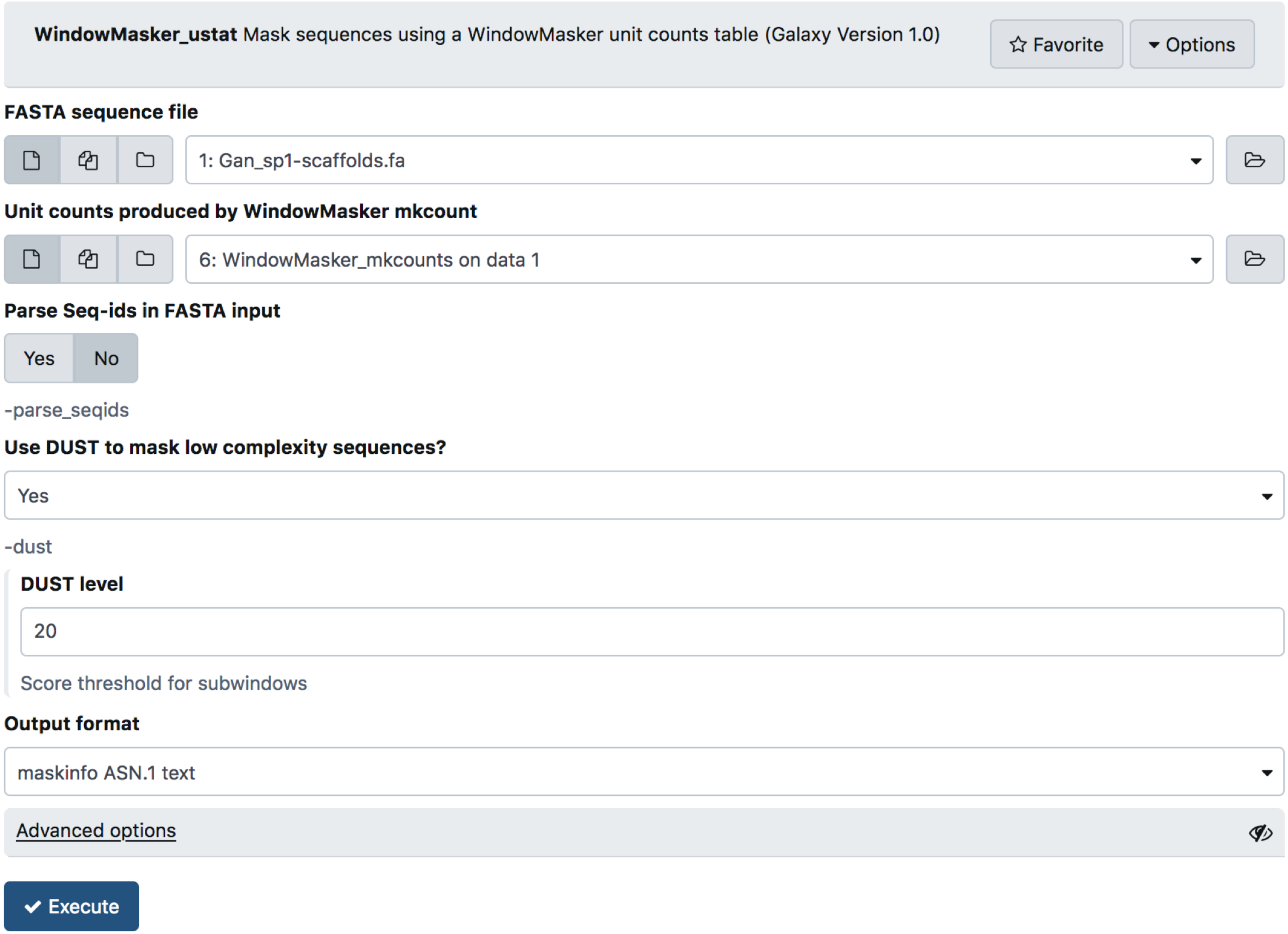
```
% windowmasker -help
USAGE
  windowmasker [-h] [-help] [-xmlhelp] [-ustat unit_counts]
    [-in input_file_name] [-out output_file_name] [-checkdup check_duplicates]
    [-fa_list input_is_a_list] [-mem available_memory] [-meta info_string]
    [-unit unit_length] [-genome_size genome_size] [-window window_size]
    [-t_extend T_extend] [-t_thres T_threshold] [-set_t_high score_value]
    [-set_t_low score_value] [-parse_seqids] [-outfmt output_format]
    [-t_high T_high] [-t_low T_low] [-infmt input_format]
    [-exclude_ids exclude_id_list] [-ids id_list] [-text_match text_match_ids]
    [-sformat unit_counts_format] [-smem available_memory] [-dust use_dust]
    [-dust_level dust_level] [-mk_counts] [-convert] [-version-full]
```



```
faToTwoBit - Convert DNA from fasta to 2bit format
usage:
   faToTwoBit in.fa [in2.fa in3.fa ...] out.2bit
```

Specify one or more input files without a flag

Galaxy provides a **standardized interface** for specifying inputs, parameters, and outputs



WindowMasker_ustat Mask sequences using a WindowMasker unit counts table (Galaxy Version 1.0)

FASTA sequence file

1: Gan_sp1-scaffolds.fa

Unit counts produced by WindowMasker mkcount

6: WindowMasker_mkcounts on data 1

Parse Seq-ids in FASTA input

Yes No

-parse_seqids

Use DUST to mask low complexity sequences?

Yes

-dust

DUST level

20

Score threshold for subwindows

Output format

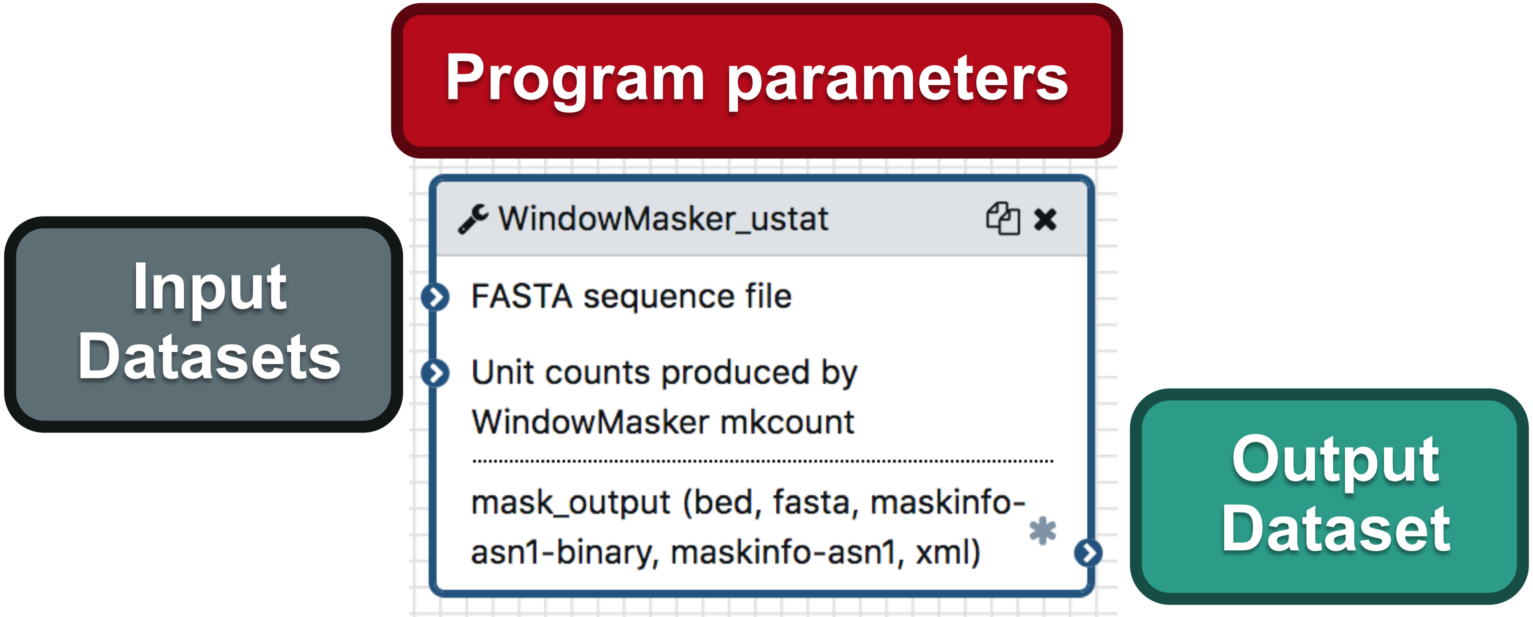
maskinfo ASN.1 text

Input Datasets

Program parameters

Output Datasets

Create a sub-workflow by adding tools and specifying the **input and output datasets**

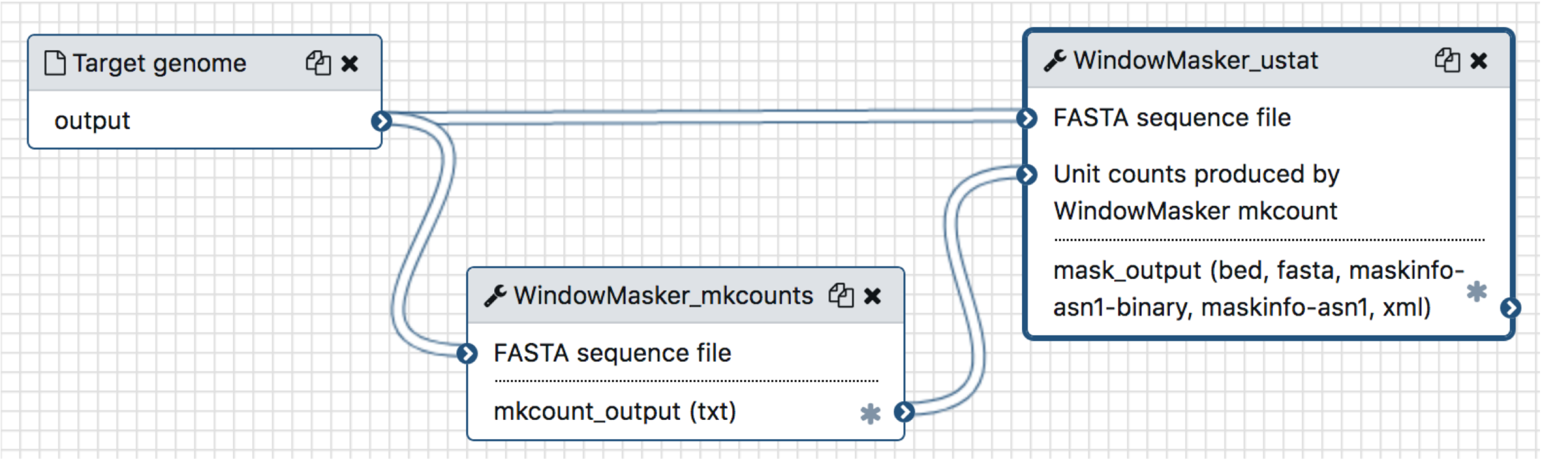


Program parameters

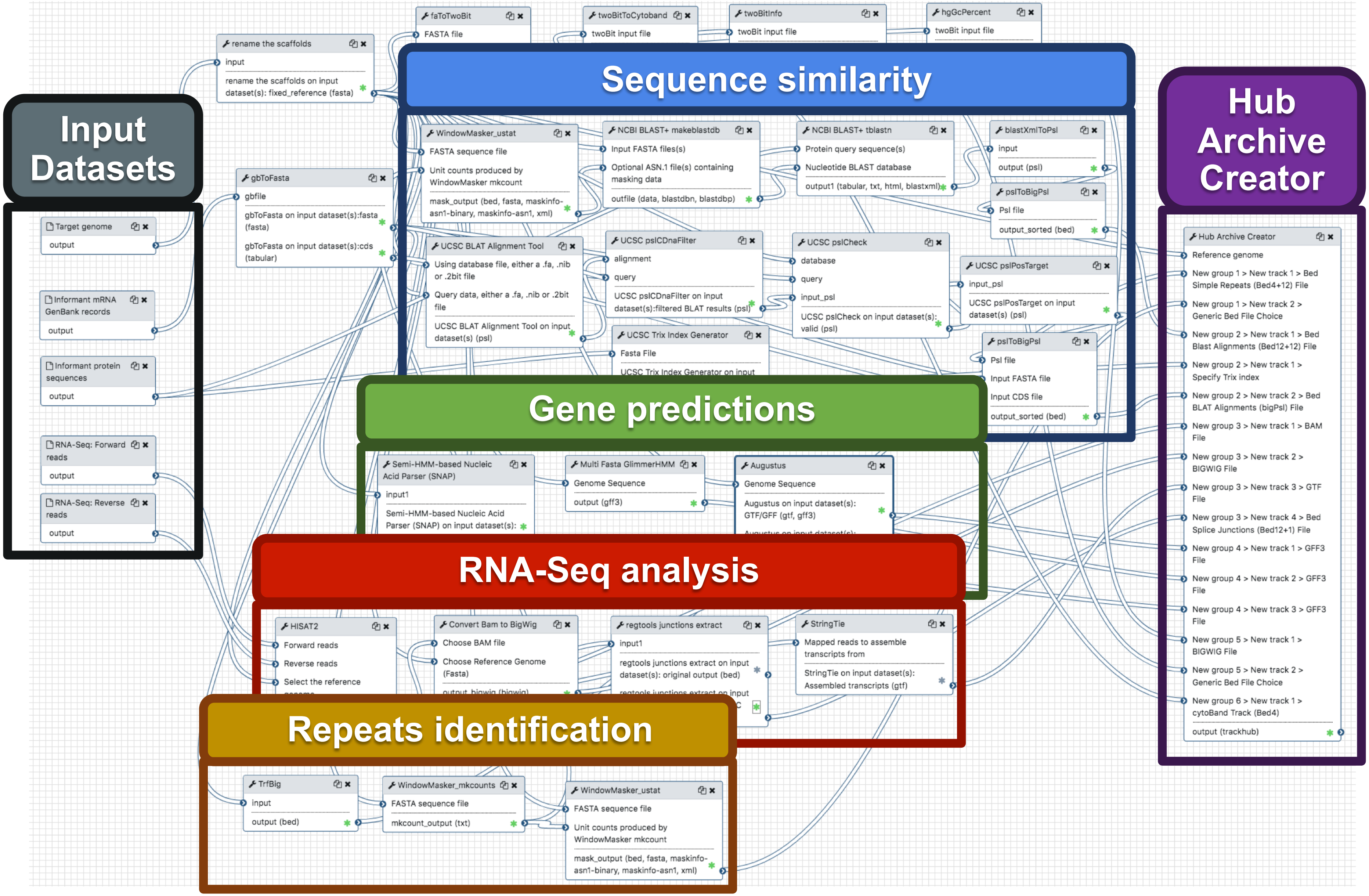
Input Datasets

Output Dataset

WindowMasker_ustat
FASTA sequence file
Unit counts produced by WindowMasker mkcount
mask_output (bed, fasta, maskinfo-asn1-binary, maskinfo-asn1, xml)



Combine sub-workflows to create the large workflow



Input Datasets
Target genome
output
Informant mRNA GenBank records
output
Informant protein sequences
output
RNA-Seq: Forward reads
output
RNA-Seq: Reverse reads
output
rename the scaffolds
input
rename the scaffolds on input dataset(s): fixed_reference (fasta)
gbToFasta
gbfile
gbToFasta on input dataset(s):fasta (fasta)
gbToFasta on input dataset(s):cds (tabular)
faToTwoBit
FASTA file
twoBitToCytoband
twoBit input file
twoBitInfo
twoBit input file
hgGcPercent
twoBit input file
Sequence similarity
WindowMasker_ustat
FASTA sequence file
Unit counts produced by WindowMasker mkcount
mask_output (bed, fasta, maskinfo-asn1-binary, maskinfo-asn1, xml)
NCBI BLAST+ makeblastdb
Input FASTA files(s)
Optional ASN.1 file(s) containing masking data
outfile (data, blastdbn, blastdbp)
NCBI BLAST+ tblastn
Protein query sequence(s)
Nucleotide BLAST database
output1 (tabular, txt, html, blastxml)
blastXmlToPsl
input
output (psl)
pslToBigPsl
Psl file
output_sorted (bed)
UCSC BLAT Alignment Tool
Using database file, either a .fa, .nib or .2bit file
Query data, either a .fa, .nib or .2bit file
UCSC BLAT Alignment Tool on input dataset(s) (psl)
UCSC pslCDnaFilter
alignment
query
UCSC pslCDnaFilter on input dataset(s):filtered BLAT results (psl)
UCSC pslCheck
database
query
input_psl
UCSC pslCheck on input dataset(s): valid (psl)
UCSC pslPosTarget
input_psl
UCSC pslPosTarget on input dataset(s) (psl)
UCSC Trix Index Generator
Fasta File
UCSC Trix Index Generator on input
pslToBigPsl
Psl file
Input FASTA file
Input CDS file
output_sorted (bed)
Hub Archive Creator
Hub Archive Creator
Reference genome
New group 1 > New track 1 > Bed Simple Repeats (Bed4+12) File
New group 1 > New track 2 > Generic Bed File Choice
New group 2 > New track 1 > Bed Blast Alignments (Bed12+12) File
New group 2 > New track 1 > Specify Trix index
New group 2 > New track 2 > Bed BLAT Alignments (bigPsl) File
New group 3 > New track 1 > BAM File
New group 3 > New track 2 > BIGWIG File
New group 3 > New track 3 > GTF File
New group 3 > New track 4 > Bed Splice Junctions (Bed12+1) File
New group 4 > New track 1 > GFF3 File
New group 4 > New track 2 > GFF3 File
New group 4 > New track 3 > GFF3 File
New group 5 > New track 1 > BIGWIG File
New group 5 > New track 2 > Generic Bed File Choice
New group 6 > New track 1 > cytoBand Track (Bed4)
output (trackhub)
Gene predictions
Semi-HMM-based Nucleic Acid Parser (SNAP)
input1
Semi-HMM-based Nucleic Acid Parser (SNAP) on input dataset(s):
Multi Fasta GlimmerHMM
Genome Sequence
output (gff3)
Augustus
Genome Sequence
Augustus on input dataset(s): GTF/GFF (gtf, gff3)
RNA-Seq analysis
HISAT2
Forward reads
Reverse reads
Select the reference
Convert Bam to BigWig
Choose BAM file
Choose Reference Genome (Fasta)
output_bigwig (bigwig)
regtools junctions extract
input1
regtools junctions extract on input dataset(s): original output (bed)
regtools junctions extract on input
StringTie
Mapped reads to assemble transcripts from
StringTie on input dataset(s): Assembled transcripts (gtf)
Repeats identification
TrfBig
input
output (bed)
WindowMasker_mkcounts
FASTA sequence file
mkcount_output (txt)
WindowMasker_ustat
FASTA sequence file
Unit counts produced by WindowMasker mkcount
mask_output (bed, fasta, maskinfo-asn1-binary, maskinfo-asn1, xml)

G-OnRamp training materials (http://g-onramp.org/training)

http://g-onramp.org/genome-browsers

- 6 G-OnRamp workshops from 2016–2018:
  - 65 participants
  - 40+ institutions
  - Half are from Primarily Undergraduate Institutions (PUIs)
- Created genome browsers for 18 species
  - Available through the CyVerse Data Store

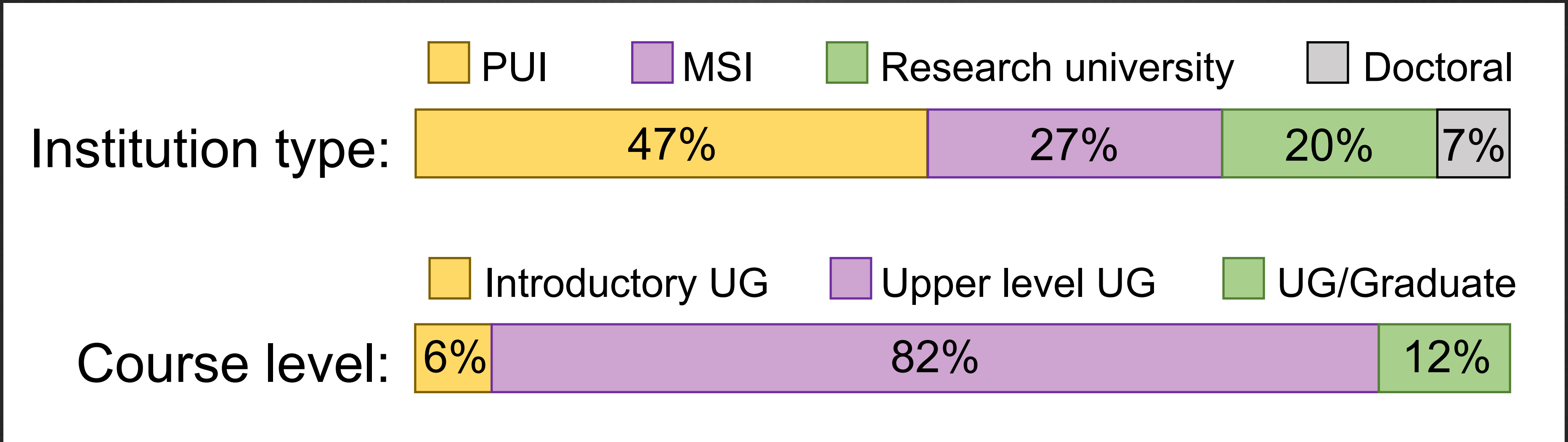
Genomics Education Partnership (http://gep.wustl.edu)

- Integration of **genomics** and research thinking into the **undergraduate biology curriculum**
- Creation of **student-scientist partnerships**
- **Publication** of research in genomics and in science education

Comparative annotation of **four parasitoid wasp species**

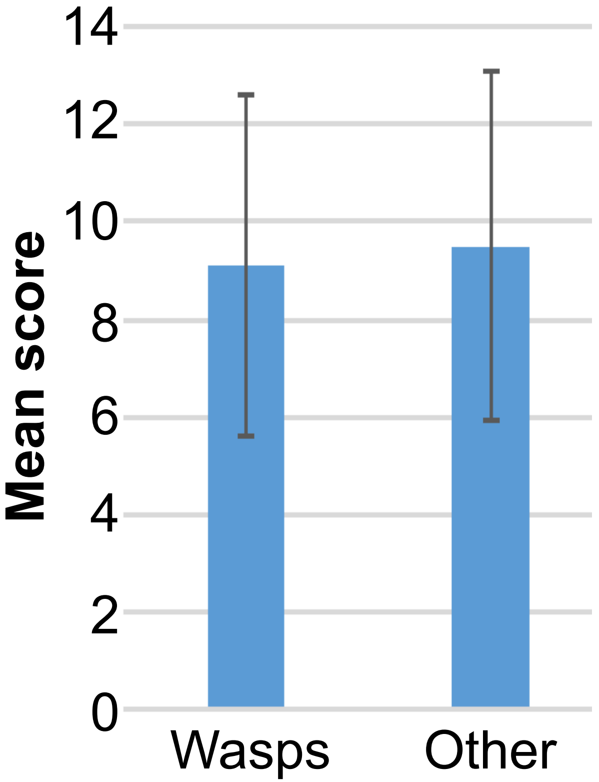
- Goal: understand how **venom proteins** from parasitoid wasps manipulate the signal transduction pathways and second messenger system of their hosts
  - Dr. Nathan T. Mortimer (Illinois State University)

- Engaged >200 GEP students from 15 institutions:

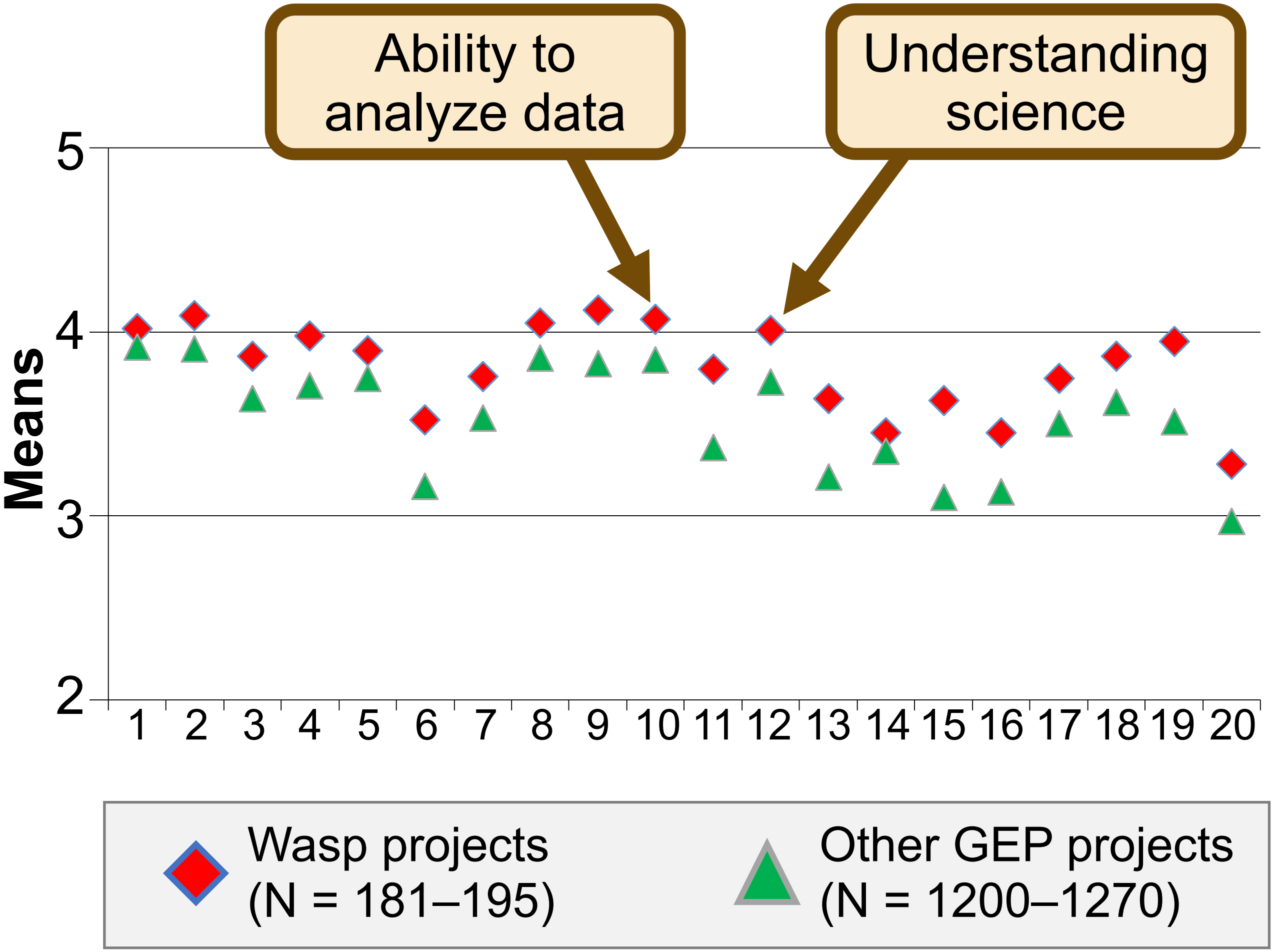


The Genome Browsers produced by G-OnRamp **work well in the classroom**

Mean annotation post-test scores



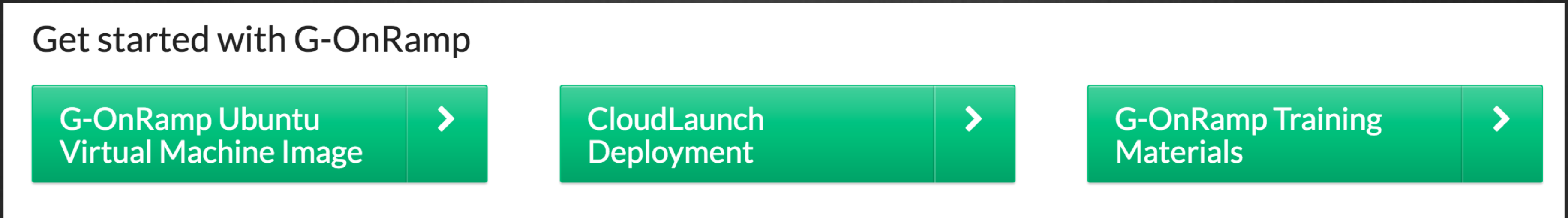
Responses to SURE survey questions


Ability to analyze data

Understanding science

G-OnRamp deployment options (http://g-onramp.org/deployments)

- G-OnRamp virtual appliance
  - Suitable for local testing and training
  - Freely available
- G-OnRamp on Amazon Web Services (AWS)
  - Production analysis of whole genome assemblies
  - CloudLaunch: https://launch.usegalaxy.org



Get started with G-OnRamp

G-OnRamp Ubuntu Virtual Machine Image ›
CloudLaunch Deployment ›
G-OnRamp Training Materials ›

Summary

- G-OnRamp provides a web-based platform for creating genome browsers for eukaryotic genomes
- Faculty have successfully used the genome browsers created by G-OnRamp to engage students in genomics CUREs

GEP is recruiting Science Partners

http://gep.wustl.edu/contact_us

Visit the G-OnRamp poster PO0137

Related posters and sessions

GEP

Genomics Education Partnership

Poster PE0138

Galaxy PROJECT

Posters PE0141 and PE0142

Galaxy Session:
1/14 @4:00pm (California)

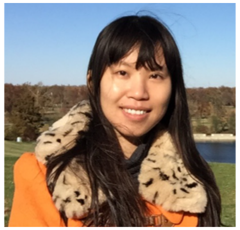
Acknowledgements

Washington University in St. Louis

Oregon Health & Science University

Sarah C. R. Elgin

Jeremy Goecks



Yating Liu

Luke Sargent

Illinois State University
Illinois' first public university

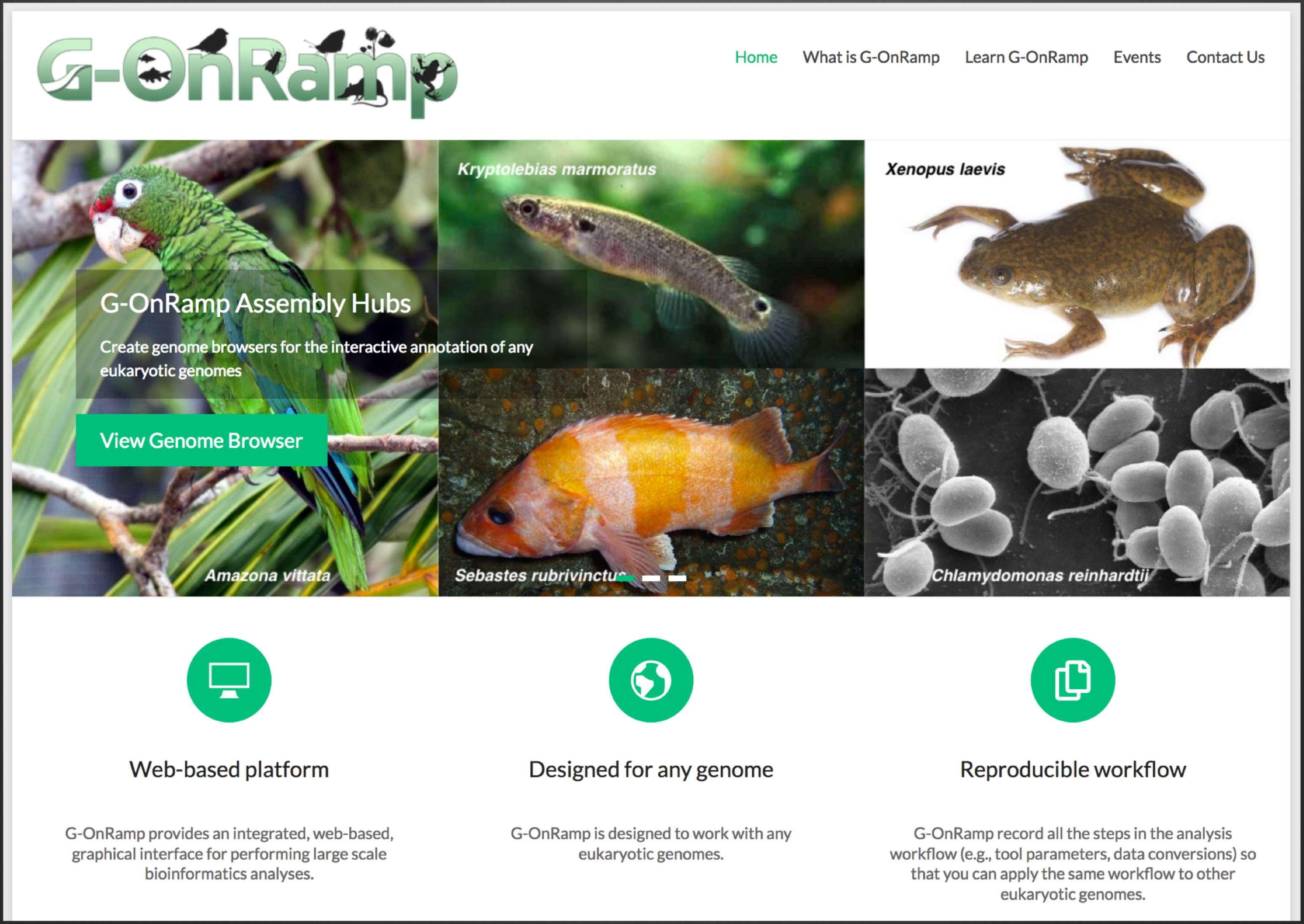
Grinnell College

Nathan T. Mortimer

David Lopatto

Supported by NIH BD2K grant # 1R25GM119157

Questions?



Home What is G-OnRamp Learn G-OnRamp Events Contact Us

G-OnRamp Assembly Hubs

Create genome browsers for the interactive annotation of any eukaryotic genomes

View Genome Browser

Web-based platform

G-OnRamp provides an integrated, web-based, graphical interface for performing large scale bioinformatics analyses.

Designed for any genome

G-OnRamp is designed to work with any eukaryotic genomes.

Reproducible workflow

G-OnRamp record all the steps in the analysis workflow (e.g., tool parameters, data conversions) so that you can apply the same workflow to other eukaryotic genomes.

http://g-onramp.org