

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

Yating Liu¹, Luke Sargent², Wilson Leung¹, Sarah C.R. Elgin¹, Jeremy Goecks^{2,*}

¹Washington University in St. Louis, St. Louis, MO 63130, USA

²Oregon Health & Science University, Portland, OR 97239, USA

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 genome browsers with evidence tracks derived from sequence alignments, *ab initio* gene predictors, RNA-Seq data, and repeat finders. Researchers can use the G-OnRamp output to visualize large genomics datasets, and can utilize the output to drive collaborative genome annotation projects in both research and educational settings.

Availability and Implementation: The G-OnRamp virtual machine images and tutorials are freely available on the G-OnRamp web site (<http://g-onramp.org/deployments>). The G-OnRamp source code is freely available under an Academic Free License version 3.0 through the goeckslab GitHub repository (<https://github.com/goeckslab>).

Contact: goecksj@ohsu.edu

Supplementary information:

1. Introduction

As the cost continues to decrease, many more eukaryotic genomes are being sequenced (Cheng *et al.*, 2018; Koepfli *et al.*, 2015; i5K Consortium, 2013). The first task after genome sequencing is often annotating the genome with the locations of functional elements such as genes (exons, introns, splice sites) and promoters. High quality genome annotations are critical for understanding organism function and evolution. Genome browsers facilitate annotation, enabling researchers to visually synthesize results from multiple experiments. Genome annotation requires the integration of multiple lines of evidence, both experimental (*e.g.*, RNA-Seq) and computational (*e.g.*, gene predictions, sequence alignments) (Yandell and Ence, 2012). However, generating the evidence tracks and visualizations needed to analyze eukaryotic genomes remains technically challenging and time consuming.

Many analysis workflows for genome annotation are already available (Haas *et al.*, 2008; Hoff *et al.*, 2016; Holt and Yandell, 2011), but these workflows are often difficult for biologists with limited bioinformatics expertise to use. Key challenges include configuring needed tools and software dependencies, learning to use command-line tools, running tools on a computing

cluster, converting the results for visualization [e.g., using the UCSC Genome Browser (Kent *et al.*, 2002) or JBrowse (Buels *et al.*, 2016)], maintaining web servers with the visualization platform to facilitate collaboration with other researchers, and keeping track of analysis steps and program parameters in order to re-use the workflow on other genome assemblies.

To address these challenges, the Genomics Education Partnership (GEP; <http://gep.wustl.edu>) has collaborated with Galaxy (<https://galaxyproject.org/>) to develop G-OnRamp, a scalable user-friendly web-based platform for eukaryotic genome annotation. G-OnRamp has computational workflows that combine more than 25 community and custom analysis tools to create evidence tracks leading to complete UCSC Assembly Hubs (Raney *et al.*, 2014) and JBrowse/Apollo genome browsers. The output produced by G-OnRamp includes evidence tracks for homologous protein and transcript sequence alignments, *ab initio* gene predictions, transcriptional activity (based on RNA-Seq) with full transcripts and splice junctions, and repeats. G-OnRamp can be deployed on the Amazon Elastic Compute Cloud (EC2) via CloudLaunch (Afgan, Lonie, *et al.*, 2018), or locally via a virtual appliance. Genome browsers created with G-OnRamp can be exported to the CyVerse Data Store (Merchant *et al.*, 2016), where they can be used for annotation without the need to maintain a web server. G-OnRamp can convert a JBrowse genome browser into an Apollo instance (Lee *et al.*, 2013) to enable real-time, collaborative genome annotations. Training materials and documentations for G-OnRamp are available on the G-OnRamp web site (<http://g-onramp.org/training>).

2. Features

Galaxy is an open, web-based platform for accessible, reproducible, and transparent analyses of large biological datasets that is used by thousands of scientists throughout the world (Afgan, Baker, *et al.*, 2018). G-OnRamp extends Galaxy by providing the data analyses and conversions needed for constructing genome browsers for annotation.

G-OnRamp encapsulates the steps required to construct UCSC Assembly Hubs and JBrowse genome browsers into Galaxy workflows. Users need only to specify the input datasets—a genome assembly of the target genome, transcript and protein sequences from a closely-related informant genome, and RNA-Seq data from the target genome—and then run the workflow to create the genome browser. G-OnRamp can be used to generate a genome browser for annotation of almost any eukaryotic genome.

2.1 Primary tools in each sub-workflow

The G-OnRamp workflow consists of four sub-workflows: homologous sequence alignments, *ab initio* gene predictions, RNA-Seq analysis, and repeat identification. Each sub-workflow is composed of multiple bioinformatics tools. The outputs from these sub-workflows provide the input data for novel Galaxy tools, the Hub Archive Creator and the JBrowse Archive Creator, which construct the UCSC Assembly Hub and the JBrowse genome browser for the target genome (Figure 1). The key components of each sub-workflow are described below; more details are available at <http://g-onramp.org>.

Homologous sequence similarity: This sub-workflow uses NCBI tblastn (Gertz *et al.*, 2006; Camacho *et al.*, 2009) and BLAT (Kent, 2002) to align proteins and transcripts from an informant genome to the target genome. BLAT transcript alignments are filtered to determine the locations of the putative orthologs. The “UCSC Trix Index Generator” tool creates an index that enables users to search for protein and transcript matches by name.

RNA-Seq analysis: This sub-workflow uses HISAT2 (Kim *et al.*, 2015) to align RNA-Seq reads to the target genome. RNA-Seq read alignments are aggregated by the “Convert BAM to BigWig” tool to create a read coverage track. The “junctions extract” subprogram in regtools (<https://github.com/griffithlab/regtools>) is used to identify putative splice junctions. StringTie (Pertea *et al.*, 2015) is used to construct putative transcripts from the aligned RNA-Seq reads.

Ab initio gene predictions: This sub-workflow consists of three gene predictors: Augustus (Stanke *et al.*, 2006), GlimmerHMM (Majoros *et al.*, 2004), and SNAP (Korf, 2004). Species-specific gene prediction parameters can be specified when the workflow is run.

Repeats identification: Tandem Repeats Finder (Benson, 1999) and the TrfBig utility developed by the UCSC Genome Bioinformatics Group are used to identify tandem repeats. WindowMasker (Morgulis *et al.*, 2006) is used to identify transposon remnants and simple repeats.

2.2 Tools for constructing genome browsers

We have developed custom Galaxy tools, the Hub Archive Creator (HAC) and the JBrowse Archive Creator (JAC), that organize and format output data from the annotation sub-workflows so it is suitable for constructing new UCSC and JBrowse genome browsers. Users can customize names and colors of each evidence track, and assign related evidence tracks to the same group. Users can also incorporate custom evidence tracks (in GFF3, BED, GTF, BAM, and BigWig formats) into the genome browsers.

2.3 Collaborative annotation with Apollo

The “Create or Update Organism” tool creates an Apollo instance from the JAC output. The “Apollo User Manager” tool provides batch management by administrators of Apollo user accounts and roles. These G-OnRamp tools are based on tools developed by the Galaxy Genome Annotation project (<https://github.com/galaxy-genome-annotation/galaxy-tools>).

2.4 Data storage and visualization

The “CyVerse interaction via iRODS” tool facilitates the transfer of G-OnRamp output to the CyVerse Data Store for long-term storage and visualization using iRODS (<https://irods.org>). Genome browsers that have been produced by G-OnRamp are available through the CyVerse Data Store at https://de.cyverse.org/anon-files/iplant/home/shared/G-OnRamp_hubs/index.html.

3. Conclusions

G-OnRamp provides a scalable, user-friendly system for individual or collaborative eukaryotic genome annotation. G-OnRamp integrates the popular Galaxy platform, more than 25 community and custom bioinformatics tools, and the popular UCSC and JBrowse/Apollo genome browsers/annotation platforms into a powerful annotation platform. In addition to annotating novel genomes, G-OnRamp genome browsers can be used in comparative genomics studies (*e.g.*, evolution of genes and pathways, understanding organism-environment interactions, etc.). G-OnRamp is useful in both research and educational settings.

Funding information

This work was supported by the National Institutes of Health [1R25 GM119157 to SCRE]. GEP is supported by the National Science Foundation [IUSE, #1431407 to SCRE] and Washington University in St. Louis. Galaxy is supported by the National Human Genome Research Institute, National Institutes of Health [HG006620, HG005133, HG004909 and HG005542]; NSF [DBI 0543285, 0850103 and 1661497]; and Oregon Health & Science University.

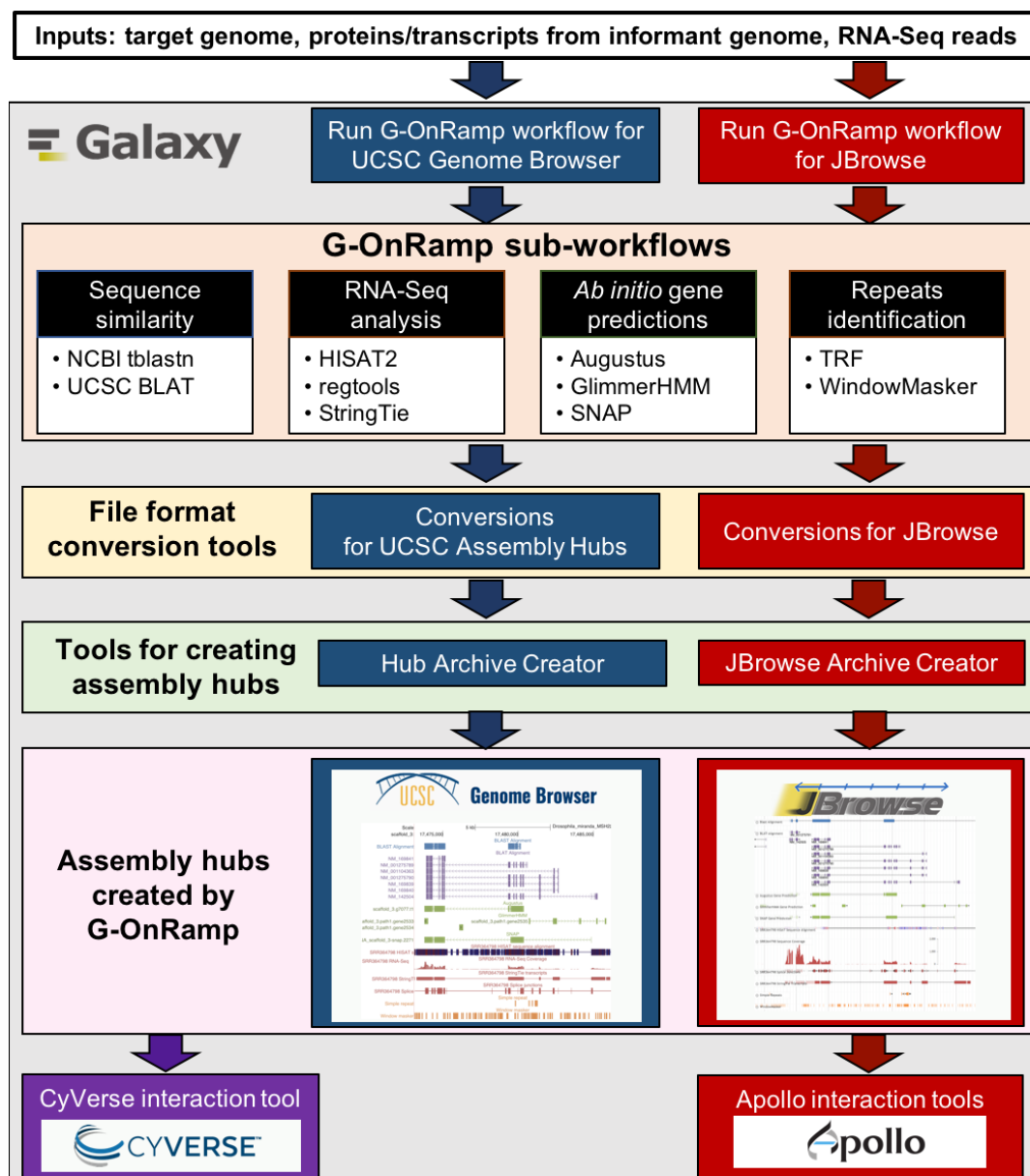


Figure 1: **Overview of the G-OnRamp workflows.** A researcher provides the target genome assembly for annotation, transcript and protein sequences from an informant genome, and RNA-Seq data from the target organism as input datasets to G-OnRamp. These datasets are processed by the sequence similarity, RNA-Seq analysis, *ab initio* gene predictions, and repeats identification sub-workflows to create UCSC Assembly Hubs and JBrowse genome browsers. The Apollo interaction tools can convert the JBrowse genome browser into an Apollo instance for collaborative annotation. The CyVerse interaction tool can transfer the G-OnRamp output to the CyVerse Data Store for long-term storage and visualization.

References

- Afgan,E., Lonie,A., *et al.* (2018) CloudLaunch: Discover and deploy cloud applications. *Future Generation Computer Systems*.
- Afgan,E., Baker,D., *et al.* (2018) The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res.*, **46**, W537–W544.

- Benson,G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.*, **27**, 573–580.
- Buels,R. *et al.* (2016) JBrowse: a dynamic web platform for genome visualization and analysis. *Genome Biol.*, **17**, 66.
- Camacho,C. *et al.* (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.
- Cheng,S. *et al.* (2018) 10KP: A phylodiverse genome sequencing plan. *Gigascience*, **7**, 1–9.
- Gertz,E.M. *et al.* (2006) Composition-based statistics and translated nucleotide searches: improving the TBLASTN module of BLAST. *BMC Biol.*, **4**, 41.
- Haas,B.J. *et al.* (2008) Automated eukaryotic gene structure annotation using EVIDENCEModeler and the Program to Assemble Spliced Alignments. *Genome Biol.*, **9**, R7.
- Hoff,K.J. *et al.* (2016) BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics*, **32**, 767–769.
- Holt,C. and Yandell,M. (2011) MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics*, **12**, 491.
- i5K Consortium (2013) The i5K Initiative: advancing arthropod genomics for knowledge, human health, agriculture, and the environment. *J. Hered.*, **104**, 595–600.
- Kent,W.J. (2002) BLAT — the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.
- Kent,W.J. *et al.* (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.
- Kim,D. *et al.* (2015) HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods*, **12**, 357–360.
- Koepfli,K.-P. *et al.* (2015) The Genome 10K Project: a way forward. *Annu Rev Anim Biosci*, **3**, 57–111.
- Korf,I. (2004) Gene finding in novel genomes. *BMC Bioinformatics*, **5**, 59.
- Lee,E. *et al.* (2013) Web Apollo: a web-based genomic annotation editing platform. *Genome Biol.*, **14**, R93.
- Majoros,W.H. *et al.* (2004) TigrScan and GlimmerHMM: two open source *ab initio* eukaryotic gene-finders. *Bioinformatics*, **20**, 2878–2879.
- Merchant,N. *et al.* (2016) The iPlant Collaborative: Cyberinfrastructure for Enabling Data to Discovery for the Life Sciences. *PLoS Biol.*, **14**, e1002342.

- Morgulis,A. *et al.* (2006) WindowMasker: window-based masker for sequenced genomes. *Bioinformatics*, **22**, 134–141.
- Pertea,M. *et al.* (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.*, **33**, 290–295.
- Raney,B.J. *et al.* (2014) Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser. *Bioinformatics*, **30**, 1003–1005.
- Stanke,M. *et al.* (2006) Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics*, **7**, 62.
- Yandell,M. and Ence,D. (2012) A beginner’s guide to eukaryotic genome annotation. *Nat. Rev. Genet.*, **13**, 329–342.