Engaging and Training Undergraduates in Big Data Analysis through Genome Annotation

Wilson Leung[1], Remi Marenco[2], Yating Liu[1], Jeremy Goecks[2], and Sarah C.R. Elgin[1]

[1]Washington University in St. Louis, [2]George Washington University

Project objectives: Create an integrated, web-based, and scalable environment (**G-OnRamp**) that enables biologists to utilize large genomics datasets in the annotation of **any eukaryotic genome**, and provide educators with a platform to train **undergraduate students** on "big data" biomedical analyses.

Abstract

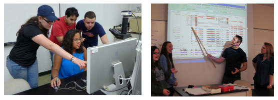
As data science becomes increasingly important in biomedicine, it is critical to introduce students to "big data" early in their studies, to prepare them for jobs in industry and for graduate education. To meet the needs of introductory data science training, we are developing **G-OnRamp**, a suite of software and training materials that enables anyone new to big data analysis (*e.g.*, undergraduates) to develop data science skills through eukaryotic genome annotation.

Genome annotation—identifying functional regions of a genome—requires the use of diverse datasets and many algorithmic tools. Annotators must interpret potentially contradictory lines of evidence in order to produce gene models that are best supported by the available evidence. The Genomics Education Partnership (GEP; http://gep.wustl.edu) is a consortium of over 100 colleges and universities that provide classroom undergraduate research experiences in bioinformatics / genomics for students at all levels. The GEP is currently focused on the annotation of multiple *Drosophila* species. G-OnRamp will enable GEP faculty to diversify, using any eukaryote with a sequenced genome that fits their particular pedagogical and research interests.

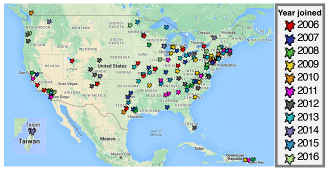
G-OnRamp is a Galaxy workflow that creates a genome browser for a new genome assembly. Galaxy (http://galaxyproject.org/, https://usegalaxy.org) is an open-source, web-based scientific gateway for accessible, reproducible, and transparent analyses of large biomedical datasets that is used throughout the world. G-OnRamp extends Galaxy with (a) analysis workflows that create a graphical genome browser for annotation, including evidence from sequence homology, gene predictions, and RNA-seq, and (b) a stand-alone virtual machine to ensure wide availability. Future versions of G-OnRamp will include (i) interactive visual analytics; (ii) collaborative genome annotation; and (iii) a public server for broad usage. Concomitant with the development of the G-OnRamp software, we are also developing training materials that can be used by educators in an instructional setting and by individual researchers.

Genomics Education Partnership (GEP)

(http://gep.wustl.edu)



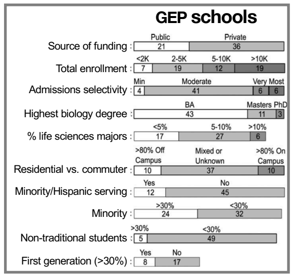
Photos by GEP faculty Michael Rubin (University of Puerto Rico – Cayey) and Heather Eisler (University of the Cumberlands)



- **>100** faculty from **>100** affiliated schools
- **>1000** undergraduates participate annually

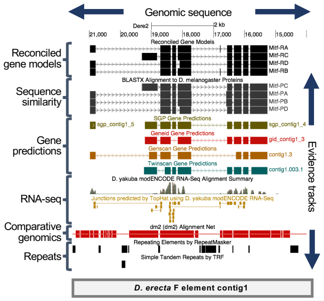
GEP goals:

- Introduce **genomics and bioinformatics** into the undergraduate curriculum
- Engage students in **genomics research**

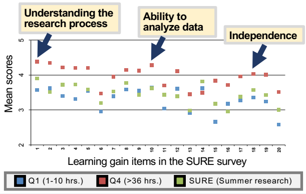


Shaffer CD *et al.* 2014, CBE Life Sci Educ. 13(1):111-30

Students evaluate evidence tracks on the UCSC Genome Browser to create optimal gene models



- Results produced by GEP students are assembled for domain analysis and **scientific publications**
  - Leung *et al.* 2015, G3. 5(5):719-40
- Students report **substantial learning gains**
  - Gains enhanced by increased time investment; Q4 > Q1 below



Use Galaxy to address GEP challenges

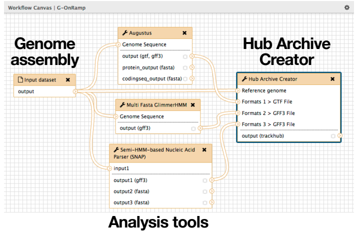
GEP challenges	Galaxy features
Requires expertise (*e.g.*, familiarity with Linux) to configure and run bioinformatics tools	Provides a web-based user interface to configure and run tools
Difficult to reproduce analysis results	Galaxy History describes the entire analysis workflow, including tool parameters and tool versions
Difficult to share workflows and results	Can make Histories, Datasets, and Workflows publicly available or share with individual Galaxy users
Difficult to incorporate additional analyses and tools	Can use the Workflow Canvas to modify existing workflows and add new tools from the Galaxy Tool Shed
GEP projects are currently limited to the analysis of different *Drosophila* species	Can extract a Workflow from History and run the Workflow on other genome assemblies

GEP + Galaxy = G-OnRamp

G-OnRamp architecture:

- Extends Galaxy with tools and workflows for genome annotation
- Combines multiple tools into reproducible sub-workflows
- Uses **Hub Archive Creator (HAC)** to create UCSC Assembly Hubs
- Displays genome browsers using the servers maintained by UCSC

Gene prediction sub-workflow



Types of evidence tracks:

- Sequence similarity (tblastn search against protein sequences from informant species)
- Gene predictions (GlimmerHMM, Augustus, and SNAP)
- RNA-Seq (HISAT2, read coverage, splice junctions, and StringTie)
- Repeats (TRF)

GitHub repository: https://github.com/goeckslab/hub-archive-creator

Collaborate with GEP faculty to improve the design and to develop training materials for G-OnRamp

GEP faculty identify challenges with creating genome browsers:

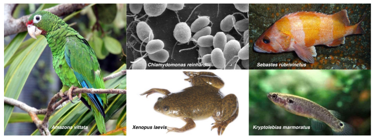
- Set up compute and storage infrastructure; install and configure bioinformatics tools
- **Optimize parameters** for each species (*e.g.*, gene prediction parameters, repeat library)
- **Validate and convert results** into file formats compatible with genome browsers
- Apply analysis workflow to a new version of the assembly or another species
- Set up and maintain a local instance of the genome browser

GEP faculty are serving as beta users of G-OnRamp:

- Ensure G-OnRamp is accessible to a broad audience
- Ensure G-OnRamp **meets real educational needs**
- Provide continuous feedback to help guide the development of G-OnRamp
- Help test and revise **curriculum and training materials**

Created UCSC Assembly Hubs for the G-OnRamp beta testers workshop (July 26-28, 2016)

- 10 participants from 9 institutions
- **Five genome assemblies:**
  - *Amazona vittata*, *Chlamydomonas reinhardtii*, *Kryptolebias marmoratus*, *Sebastes rubrivinctus*, *Xenopus laevis*
  - Assembly sizes: **111Mb - 2.8Gb**
  - Number of scaffolds: **54 - 402,501**
  - Four genomes with RNA-Seq data



Photos by Tom MacKenzie (*A. vittata*), Dartmouth Electron Microscope Facility (*C. reinhardtii*), Chad King (*S. rubrivinctus*), Brian Gratwicke (*X. laevis*), and Jean-Paul Cicéron (*K. marmoratus*)

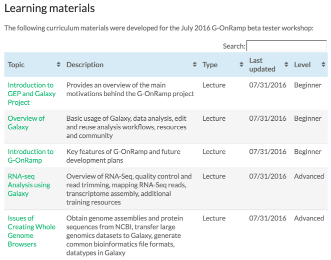
Develop training materials for G-OnRamp

Target audiences:

- Research scientists
- College faculty / undergraduate students

Curriculum materials:

- Overview of Galaxy
- Overview of the bioinformatics tools used by G-OnRamp
- Written walkthroughs on how to use and customize G-OnRamp
- Screencasts and interactive tours



Learning materials

The following curriculum materials were developed for the July 2016 G-OnRamp beta tester workshop:

Search:

Topic	Description	Type	Last updated	Level
Introduction to GEP and Galaxy Project	Provides an overview of the main motivations behind the G-OnRamp project	Lecture	07/31/2016	Beginner
Overview of Galaxy	Basic usage of Galaxy, data analysis, edit and reuse analysis workflows, resources and community	Lecture	07/31/2016	Beginner
Introduction to G-OnRamp	Key features of G-OnRamp and future development plans	Lecture	07/31/2016	Beginner
RNA-seq Analysis using Galaxy	Overview of RNA-Seq, quality control and read trimming, mapping RNA-Seq reads, transcriptome assembly, additional training resources	Lecture	07/31/2016	Advanced
Issues of Creating Whole Genome Browsers	Obtain genome assemblies and protein sequences from NCBI, transfer large genomics datasets to Galaxy, generate common bioinformatics file formats, datatypes in Galaxy	Lecture	07/31/2016	Advanced

Future plans

- Develop a sub-workflow for **identifying transposons**:
  - Reduce false positives in gene predictions and improve workflow performance
- Develop a sub-workflow for creating **species-specific gene prediction parameters**
- Extend the G-OnRamp Workflow to analyze other functional genomic data:
  - Data from ChIP-seq, DNase-seq / ATAC-seq, and Bisulfite sequencing
- Integrate with existing **collaborative annotation platforms** (*e.g.*, WebApollo, CoGE)
- Integrate with GEP annotation tools designed for teaching (*e.g.*, Gene Model Checker)
- Provide multiple methods to use and install G-OnRamp:
  - Public server, local installation, cloud deployment (Amazon EC2), and virtual machines
- Host G-OnRamp training workshops for educators and research scientists:

G-OnRamp workshops: June 20-22 and July 25-27, 2017

Acknowledgements

G-OnRamp is supported by NIH BD2K Grant 1R25 GM119157. GEP is supported by NSF grant #1431407 and WUSTL. Galaxy is supported by NIH grant HG006620-04 and GWU.

Washington University in St.Louis

THE GEORGE WASHINGTON UNIVERSITY WASHINGTON, DC

NIH National Institutes of Health

NSF

Contacts

Sarah C.R. Elgin
selgin@wustl.edu

Jeremy Goecks
jgoecks@gwu.edu