**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**

- Genomics has become increasingly important in education. 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 biotechnology training, we are developing G-OnRamp, a suite of software and training materials that enable 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 integrate potentially contradictory lines of evidence in order to produce gene models that are best supported by the available evidence. The Genomics Education Partnership (GEP; www.genepi.org) is a consortium of over 100 colleges and universities that provides classroom undergraduate research experiences in bioinformatics and 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) is an open-source, web-based scientific gateway for accessible, reproducible, and transparent data analysis and sharing that is used throughout the world. G-OnRamp extends Galaxy with (i) analysis workflows that create a graphical genome browser for annotation, including evidence from sequence homology, gene predictions, and RNA-seq, and (ii) an online virtual machine to ensure wider availability. Future versions of G-OnRamp will include (i) interactive visual analytics, (ii) collaborative annotation platforms, (iii) automated download of training datasets, and (iv) new workflows developed by GEP faculty to address particular annotation challenges. In addition, we are also developing training materials that can be used by educators in an institutional setting and by individual researchers.

**Use Galaxy to address GEP challenges**

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

**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

**Types of evidence tracks:**
- Sequence similarity (blastn search against protein sequences from informant species)
- Gene predictions (GlimmerHMM, Augustus, and SNAP)
- RNA-Seq (HSAT2, read coverage, splice junctions, and StringTie)
- Repeats (TRF)

**Develop training materials for G-OnRamp**

- **Learning materials:**
  - **Gene Model Checker**
  - **Integrate with GEP annotation tools designed for teaching (e.g., Gene Model Checker)**

- **Contact:**
  - Sarah C.R. Elgin
  - jgoecks@wustl.edu

**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., WebApolo, 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 HD02K Grant 1R25 GM119157. GEP is supported by NSF grant #1401477 and W35TL. Galaxy is supported by NIH grant HG00660-04 and GWU.

**Contacts**

- Sarah C.R. Elgin
  - jgoecks@wustl.edu
- Jeremy Goecks
  - jgoecks@wustl.edu