1-2019

Generating genome browsers to facilitate undergraduate-driven collaborative genome annotation

Luke Sargent

Yating Liu
Washington University in St. Louis

Wilson Leung
Washington University in St. Louis, wleung@wustl.edu

Nathan Mortimer

David Lopatto

See next page for additional authors

Recommended Citation

Sargent, Luke; Liu, Yating; Leung, Wilson; Mortimer, Nathan; Lopatto, David; Goecks, Jeremy; and Elgin, Sarah C.R., "Generating genome browsers to facilitate undergraduate-driven collaborative genome annotation" (2019). *Biology Faculty Publications & Presentations*. 241.
https://openscholarship.wustl.edu/bio_facpubs/241

Authors

Luke Sargent, Yating Liu, Wilson Leung, Nathan Mortimer, David Lopatto, Jeremy Goecks, and Sarah C.R. Elgin

G-OnRamp: Generating genome browsers to facilitate undergraduate-driven collaborative genome annotation

Luke Sargent[1], Yating Liu[2], Wilson Leung[2], Nathan T. Mortimer[3], David Lopatto[4], Jeremy Goecks[1], Sarah C. R. Elgin[2*]

[1] Department of Biomedical Engineering, Oregon Health & Science University, Portland, Oregon, United States of America

[2] Department of Biology, Washington University in St. Louis, St. Louis, Missouri, United States of America

[3] School of Biological Sciences, Illinois State University, Normal, Illinois, United States of America

[4] Department of Psychology, Grinnell College, Grinnell, Iowa, United States of America

* Corresponding author
E-mail: selgin@wustl.edu (SCRE)

Short title: G-OnRamp creates genome browsers to facilitate collaborative annotation

Abstract

Scientists are sequencing new genomes at an increasing rate with the goal of associating genome contents with phenotypic traits. After a new genome is sequenced and assembled, structural gene annotation is often the first step in analysis. Despite advances in computational gene prediction algorithms, most eukaryotic genomes still benefit from manual gene annotation. Undergraduates can become skilled annotators, and in the process learn both about genes/genomes and about how to utilize large datasets. Data visualizations provided by a genome browser are essential for manual gene annotation, enabling annotators to quickly evaluate multiple lines of evidence (*e.g.*, sequence similarity, RNA-Seq, gene predictions, repeats). However, creating genome browsers requires extensive computational skills; lack of the expertise required remains a major barrier for many biomedical researchers and educators.

To address these challenges, the Genomics Education Partnership (GEP; https://gep.wustl.edu/) has partnered with the Galaxy Project (https://galaxyproject.org) to develop G-OnRamp (http://g-onramp.org), a web-based platform for creating UCSC Assembly Hubs and JBrowse genome browsers. G-OnRamp can also convert a JBrowse instance into an Apollo instance for collaborative genome annotations in research and educational settings. G-OnRamp enables researchers to easily visualize their experimental results, educators to create Course-based Undergraduate Research Experiences (CUREs) centered on genome annotation, and students to participate in genomics research.

Development of G-OnRamp was guided by extensive user feedback from in-person workshops. Sixty-five researchers and educators from over 40 institutions participated in these workshops, which produced over 20 genome browsers now available for research and education. For example, genome browsers for four parasitoid wasp species were used in a CURE engaging 142 students taught by 13 faculty members — producing a total of 192 gene models. G-OnRamp can be deployed on a personal computer or on cloud computing platforms, and the genome browsers produced can be transferred to the CyVerse Data Store for long-term access.

Introduction

The need for G-OnRamp

A considerable effort has been made over the last two decades to improve undergraduate science education by engaging students in the process of science, as well as acquainting them with the resulting knowledge base. For the life sciences these efforts were perhaps best enunciated by the AAAS report *Vision and Change in Undergraduate Biology Education* [1]. One of the strategies found to be effective in engaging large numbers of undergraduates in doing science is the CURE, or Course-based Undergraduate Research Experience ([2]; see [3] and [4] for examples). Within computational biology, a number of groups have found that genome annotation is a research problem that can be adapted to this purpose.

With the decreasing cost and wide availability of genome sequencing [5], the bottleneck for utilizing genomics datasets to address scientific questions is shifting from the ability to produce data to the ability to analyze and interpret data. Genome annotation—labeling functional regions of the genome such as gene boundaries, exons, and introns—benefits from a combination of computational and manual curation of data. With appropriate tools and training, undergraduates can make a significant contribution to a community annotation project, where scientists work together to annotate an entire genome. Gene annotation builds on what students are learning about gene structure, while requiring them to grapple with multiple lines of evidence to establish defendable gene models. Student annotation projects thus are mutually beneficial for researchers and for students, enabling unique science and providing a multi-faceted learning experience for students [6, 7, 8, 9, 10].

However, despite the improvements in tool accessibility and quality, there remain technical barriers that must be overcome to perform genome annotation. Many biology researchers and educators lack detailed knowledge of informatics and computational tools. When these scientists acquire the genome assembly of their favorite organism, a major barrier is the need to use multiple bioinformatics tools to analyze the genome assembly and visualize the results in a genome browser — the display tool central to community annotation. There are several good options, but most either require substantial computer skills and bioinformatics expertise to use, or have compute and storage limits that restrict the size/complexity of genome assemblies that can be analyzed using the platform [11, 12, 13, 14, 15].

We developed G-OnRamp to address these concerns. G-OnRamp is a collaboration between the Galaxy project (https://galaxyproject.org/), an open-source, web-based computational workbench for analyzing large biological datasets [16], and the Genomics Education Partnership (GEP; http://gep.wustl.edu/) [8, 17]. Among G-OnRamp's principal goals is lowering technical barriers to enable biologists to construct either a UCSC Assembly Hub [18] or a JBrowse/Apollo genome browser [19]. G-OnRamp accomplishes this by providing a collection of tools, workflows and services pre-configured and ready to process data and enable annotation. Students, educators and researchers can bypass most of the system administration tasks involved in generating a genome browser and focus on using the genome browser to address scientific questions. Our assessment results in the classroom demonstrate that the genome browsers produced by G-OnRamp are effective tools for engaging undergraduates in research and in enabling their contributions to the scientific literature in genomics.

Results

Overview of the components

Genome annotation needs for the Genomics Education Partnership. The GEP is a consortium of faculty members from over 100 educational institutions, which annually introduces more than 1300 undergraduates to genomics research through engagement in collaborative annotation projects (Fig 1A). The GEP core organization provides technical infrastructure as well as identifying research questions that would benefit from high quality gene annotations, particularly those where utilizing comparisons across multiple species can provide insights. By engaging the talents of "massively parallel undergraduates," one can gather data (high quality annotations of hundreds of genes) that could not be obtained otherwise, given the high labor costs. To ensure that the gene annotations are high quality, each gene is annotated by at least two students working independently, and the results are reconciled by experienced students (Fig 1B).

These collaborative genome annotation projects can be performed by students using either a genome browser or a genome annotation editor such as Apollo. Pedagogically, there are advantages to requiring students to initially examine the evidence tracks on a genome browser, using the data to determine the precise exon coordinates for their gene model, and recording the results in an Excel worksheet or other table. These models can then be imported into the genome browser as custom tracks, and used as evidence in the final reconciliation. Currently, the GEP uses a hybrid approach, whereby students in GEP courses use a UCSC Genome Browser to construct the initial gene models, while experienced students use the Apollo annotation editor for finale reconciliation. See Fig 2 for an example of a typical error in a gene model submitted by a GEP student, viewed in Apollo for reconciliation. Overall, we see complete agreement in 60% – 80% of the gene models submitted, depending on the difficulty of the project.

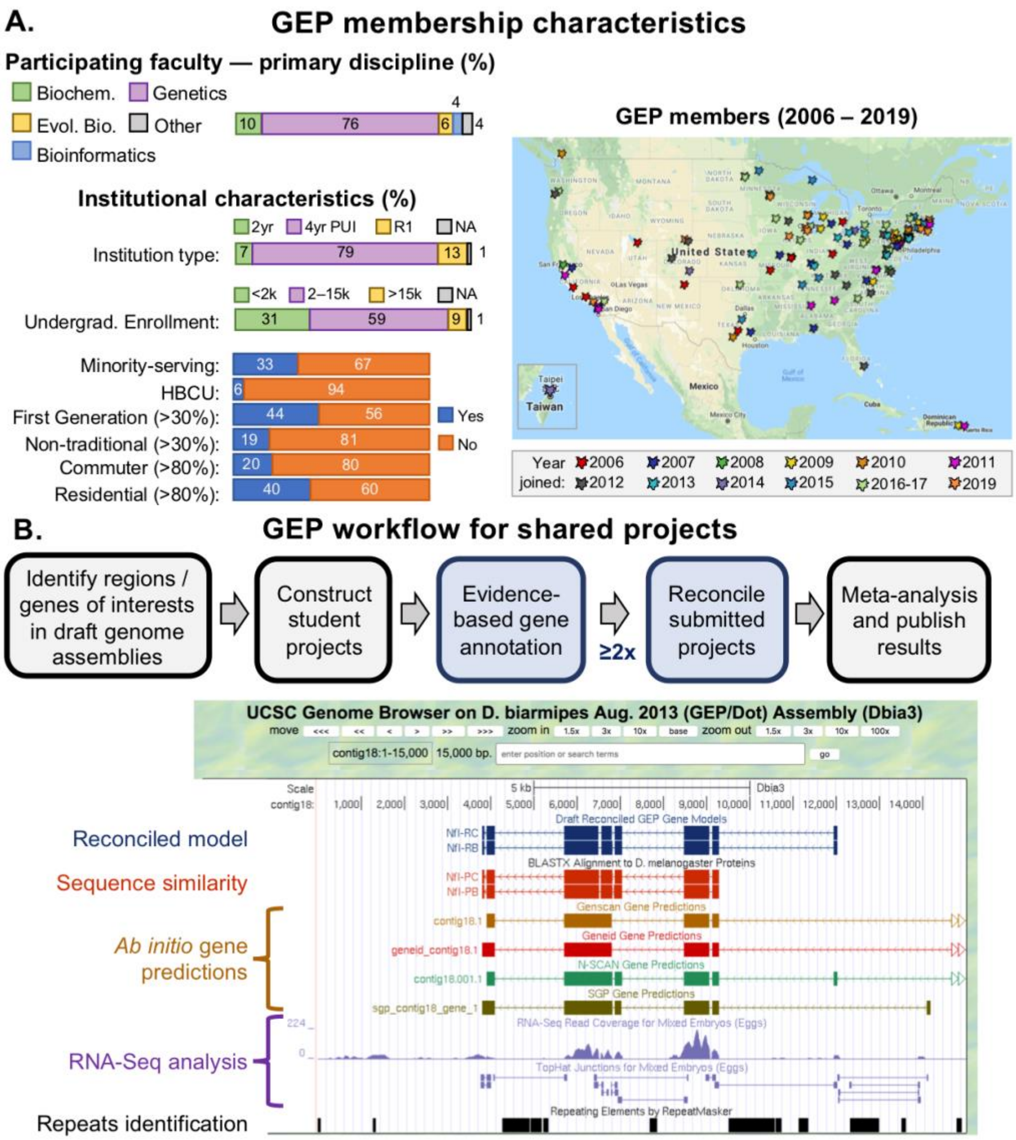
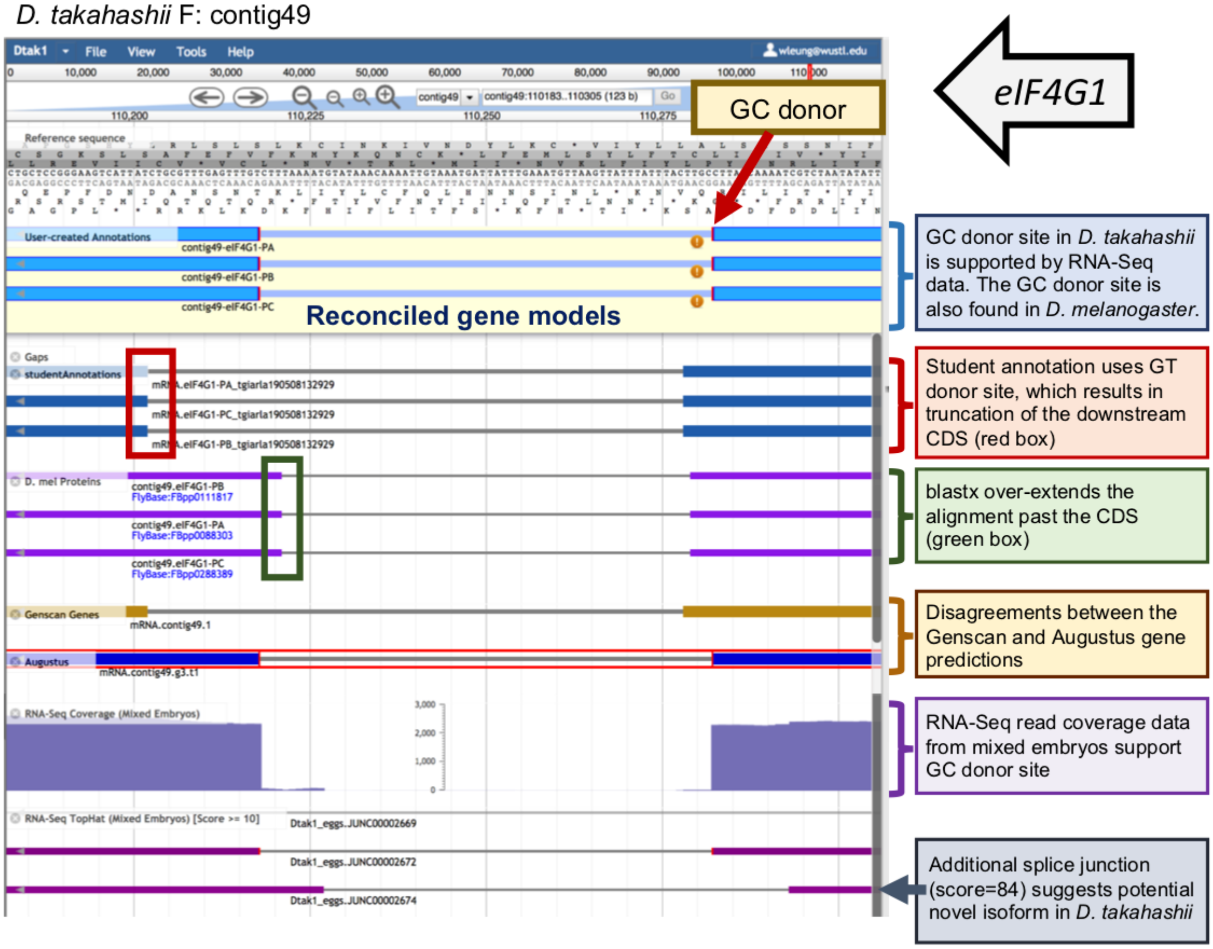


Fig 1. Overview of the Genomics Education Partnership. A. Membership characteristics: participating faculty primarily teach genetics (although other disciplines are represented), and most often teach at Primarily Undergraduate Institutions (PUIs) across the USA; faculty at community colleges and R1 research universities also participate. The geographical distribution of member schools and year of joining GEP are shown on the map. The member schools serve a diverse undergraduate student body, with 33% Minority-Serving Schools, including six HBCUs (Historically Black Colleges and Universities); 44% of the schools have 30% or more first-generation students, 11% have 30% or more non-traditional students (over 25 yrs of age), and 20% are commuter schools, with over 80% of the students commuting. See the Current GEP Members page (http://gep.wustl.edu/community/current_members) for a complete list of participating faculty with their schools. B. Students in the GEP work together to produce high-quality annotation of a genome region or a collection of genes of interest identified by a Lead Scientist. "Student projects" are provided as genome browser pages (see lower portion of the figure) with from one to seven potential genes (and other features of interest) for annotation. Browser tracks show available evidence for a gene, including gene conservation (Sequence similarity track and additional BLAST searches), presence of large open reading frames and other appropriate signals (*ab initio* gene predictions), and evidence of gene expression (RNA-seq data, Top-Hat analysis results, etc.). Students work from these multiple lines of evidence, some of which may initially appear contradictory, to generate a gene model that they can defend. In the case shown, the sequence similarity search (BLAST) failed to identify putative upstream exons, whose presence is supported by RNA-seq data and Top-Hat analysis. Students take responsibility for the workflow steps shown in light blue, while the Lead Scientist's research group is responsible for the steps shown in grey. Pre-/post course assessment has shown the effectiveness of such a collaborative annotation project both for supporting student learning about genes and genomes and in providing a research experience [17, 20, 21].



GC donor site in *D. takahashii* is supported by RNA-Seq data. The GC donor site is also found in *D. melanogaster*.

Student annotation uses GT donor site, which results in truncation of the downstream CDS (red box)

blastx over-extends the alignment past the CDS (green box)

Disagreements between the Genscan and Augustus gene predictions

RNA-Seq read coverage data from mixed embryos support GC donor site

Additional splice junction (score=84) suggests potential novel isoform in *D. takahashii*

Fig 2. Apollo overview. After uploading data to Apollo via G-OnRamp's "Create or Update Organism" tool, a user can choose which tracks to display with computational and experimental evidence, including submitted annotations from students, and begin to create her own gene model in a user-created annotations panel. Pictured is the Apollo interface showing provided sample data and computed lines of evidence, in addition to student annotation data and the final reconciled gene models (shown in the annotations panel). The genome browser image illustrates a typical error by one student annotator at an intron/exon boundary, and the reconciled model generated by an experienced student annotator. Based on RNA-Seq data and the use of the non-canonical GC donor site in the informant species (*Drosophila melanogaster*), the reconciled gene model for the *D. takahashii* ortholog of *eIF4G1* uses a non-canonical GC splice donor site instead of the GT donor site proposed by the student annotator.

GEP faculty have worked collaboratively to generate and maintain curricula to introduce students to the appropriate computer-based tools and to the scientific questions under study [8, 20]; all such materials are available on the GEP website under a "creative commons" license. Students who contribute documented gene models, and participate in reading and critiquing the final manuscript, are co-authors on the resulting scientific publication (*e.g.*, [22], [23]). G-OnRamp was conceived by the GEP as a component of the technical infrastructure, simplifying the process of generating genome browsers. This capability should allow biology faculty to diversify the research questions under study, exploiting newly sequenced genomes as they become available.

G-OnRamp overview. G-OnRamp is a Galaxy-based analysis platform providing a collection of tools and services that enable collaborative genome annotation in an efficient, user-friendly, and web-based environment (http://www.g-onramp.org; [24]). Galaxy is used across the world by thousands of scientists, and one of its key features is a web-based user interface that anyone can use for complex biological analyses regardless of their computational knowledge. G-OnRamp is configured with tools for sequence similarity searches, gene predictions, RNA-Seq data analysis, and repeat analysis (Fig 3). These tools are combined into multi-step workflows that process a target genome assembly and create a UCSC Assembly Hub (which can be viewed at the official UCSC Genome Browser; http://genome.ucsc.edu) or a locally-bundled JBrowse instance. G-OnRamp also provides tools to import a JBrowse instance into Apollo to facilitate real-time collaborative genome annotation (https://genomearchitect.readthedocs.io/en/latest/; [10]). In a pedagogical example, an instructor can deploy G-OnRamp, upload the data, run a workflow to generate a JBrowse genome browser for visualization, and use the G-OnRamp Apollo interaction tools to convert the genome browser hub to Apollo for collaborative analysis by students.

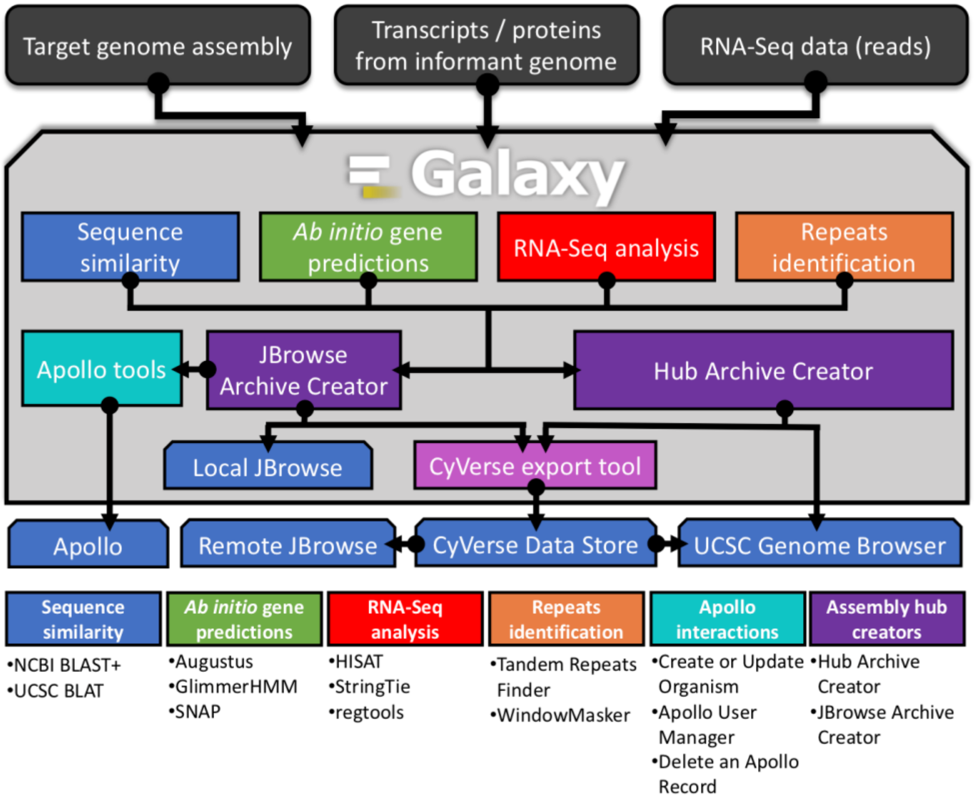


Fig 3. G-OnRamp overview. G-OnRamp is a Galaxy-based platform with analysis workflows that process a target genome assembly, transcripts and proteins from an informant genome, and RNA-Seq data from the target genome to create a genome browser for individual or collaborative annotation. Four sub-workflows (sequence similarity, *ab initio* gene predictions, RNA-Seq analysis, and repeats identification) run concurrently and generate the data for manual gene annotation. Data produced by the sub-workflows is used by the Hub Archive Creator (HAC) tool to create UCSC Assembly Hubs and by the JBrowse Archive Creator to create JBrowse genome browsers. The Apollo interaction tools convert JBrowse genome browsers into an Apollo instance to facilitate collaborative annotations. Genome browsers produced by G-OnRamp can be transferred to the CyVerse Data Store via the CyVerse export tool for long-term storage and visualization. The "Tool Suites" panel (below) lists the primary tools in each sub-workflow and the tools provided by G-OnRamp to create and manage Apollo instances. See [24] and http://g-onramp.org for further details.

Overview of genome annotation with Apollo: efficiency and crowd management. Apollo was included in G-OnRamp as it substantially increases the efficiency of gene annotation. Using Apollo, students can dynamically interact with evidence tracks, selecting the desired exons (by drag and drop) for assembly into a gene model. With effective permission management, annotation can be done separately (different students annotating different genes), iteratively (annotated genes being passed from one student to another) or simultaneously (students collaborate to annotate the same gene at the same time).

To aid permission-driven access control, G-OnRamp provides interaction tools (based on tools developed by the Galaxy community; [25]) for managing user accounts and genome assemblies in an Apollo instance. For example, a G-OnRamp administrator can use the “Create or Update Organism” tool to create a new Apollo instance or modify an existing Apollo instance. The Apollo User Manager tool provides fine-grained access controls; an administrator can control the read, write, and export permissions of individual users or groups of users. For example, instructors can use the Apollo User Manager to create accounts for a group of students enrolled in a course, and to limit their access to a subset of the genome assemblies in the Apollo instance.

Using G-OnRamp in research and education settings

G-OnRamp workshops and evaluation. To grow the community of users and better tailor G-OnRamp to their needs, we hosted two beta-testers workshops in 2017 and two “train the trainer” workshops in 2018 to introduce researchers and educators to the platform. The goal of these workshops was to familiarize members of the community with G-OnRamp and to solicit feedback. These workshops attracted 53 diverse participants from over 40 institutions across the world, demonstrating that G-OnRamp satisfies a need for both researchers and educators alike (Fig 4).

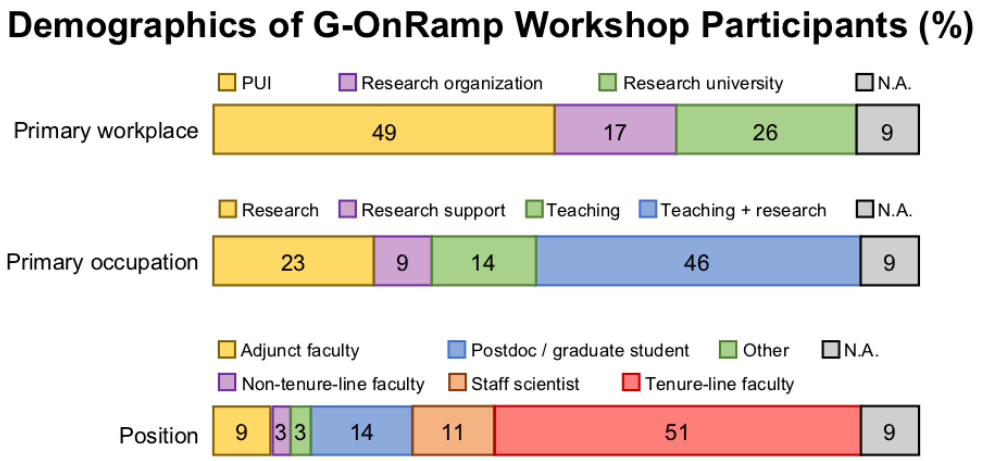


Fig 4. Demographics of G-OnRamp workshop participants. Of the 53 workshop participants eligible, 35 responded to the demographics questions (response rate = 66.0%). Many G-OnRamp workshop participants are tenure-line faculty members who work at primarily undergraduate institutions (PUIs), where they are involved in both teaching and research. Other participants focus mainly on research, either carrying out research or providing research support.

In addition to following a general training curriculum (available at http://g-onramp.org/training) on sample data, attendees were encouraged to bring their own genome assembly for processing and genome browser hub creation. Over 20 publicly-available genome browsers were created by workshop participants and the users that tested prototype G-OnRamp versions. Browsers generated during the 2017 and 2018 workshops demonstrate results obtained for genomes with assembly sizes ranging from 70Mb to 2.1Gb and with scaffold counts ranging from 53 to 271,888 (Table 1A). These genome browsers are hosted on the CyVerse Data Store [26] and are available via the "View Genome Browser" button on the G-OnRamp website (http://g-onramp.org/genome-browsers).

Table 1A. Publicly available genome browsers.

Target genome (common name)	Genome assembly file size	Number of scaffolds	Informant genome	Number of RNA-Seq samples	Genome Browser(s) created
Centrapalus pauciflorus (Vernonia)	1.2 GB	19,697	*Arabidopsis thaliana*	1	JBrowse
Spinus cucullatus (Red siskin)	1.1 GB	26,015	*Taeniopygia guttata*	0	JBrowse and UCSC Assembly Hub
Thlaspi arvense (Field pennycress)	539 MB	6,768	*Arabidopsis thaliana*	1	JBrowse and UCSC Assembly Hub
Xestospongia bocatorensis (Sponge)	70 MB	271,888	*Amphimedon queenslandica*	8	JBrowse
Tetrahymena thermophila (Ciliate)	155.6 MB	1,464	*Ichthyophthirius multifiliis*	1	UCSC Assembly Hub
Bemisia tabaci (Silverleaf whitefly)	690 MB	19,751	*Drosophila melanogaster*	2	UCSC Assembly Hub
Solenodon paradoxus (Haitian solenodon)	2.1 GB	40,372	*Erinaceus europaeus*	0	UCSC Assembly Hub
Ganaspis sp.1 (Parasitoid wasp)	500 MB	54,394	*Drosophila melanogaster*	1	UCSC Assembly Hub
Fragaria vesca (Wild strawberry)	240 MB	3,263	*Arabidopsis thaliana*	4	UCSC Assembly Hub
Chlamydomonas reinhardtii (Green algae)	113.3 MB	53	*Arabidopsis thaliana*	2	UCSC Assembly Hub
Solenodon paradoxus (Haitian solenodon)	2.1 GB	3,078	*Homo sapiens*	0	UCSC Assembly Hub
Taeniopygia guttata (Zebra finch)	1.26 GB	37,096	*Taeniopygia guttata*	0	UCSC Assembly Hub
Amazona ventralis (Hispaniolan parrot)	1.1 GB	18,948	*Gallus gallus*	0	UCSC Assembly Hub
Amazona vittata (Puerto Rican parrot)	1.2 GB	16,449	*Gallus gallus*	2	UCSC Assembly Hub
Schrenkiella parvula (Saltwater cress)	137 MB	1,457	*Arabidopsis thaliana*	4	JBrowse
Aiptasia pallida (Coral reef)	260 MB	5,065	*Nematostella vectensis*	2	JBrowse + Apollo
Thalassiosira pseudonana (Diatoms)	32.8 MB	64	*Arabidopsis thaliana*	2	JBrowse + Apollo

List of publicly available genome browsers generated with user-submitted data during the 2017-2018 workshops. These and additional G-OnRamp browsers generated by earlier prototypes with user-submitted data can be seen at http://g-onramp.org/genome-browsers.

G-OnRamp features. Feedback collected from participants after each workshop was used to determine priority areas for improvements in documentation, performance and scalability of the workflows, accessibility of the user interface, and quality-of-life improvements to extant tools. For example, the 1.1 release of G-OnRamp includes requested improvements to Galaxy's support for Augustus, a tool that performs comparative gene prediction [27], enabling users to limit the genomic range to search or to add extrinsic 'hints' for improved search specificity. Beyond this, the 1.1 release features the latest (as of this writing) versions of Galaxy (19.05), Apollo (2.4.1) and JBrowse (1.16.6). A more complete list of features is provided in Table 1B.

Table 1B. Feature: G-OnRamp provides…

Processing / Analysis:
The UCSC Hub Archive Creator, a tool to create genome browser archives for display with the UCSC browser
The JBrowse Archive Creator, a tool to create JBrowse genome browsers with Galaxy
An RNA-seq analysis subworkflow to process and visualize RNA-seq data
A BLAT alignment subworkflow to align transcript sequences from an informant genome to the target genome
Tools to identify repeats using WindowMasker within Galaxy
Input / Data Acceptance:
Default workflows that accept genome assemblies in fasta format, RNA-seq data in fastqsanger format, transcripts from informant genomes in GenBank or fasta formats, and proteins from informant genomes in fasta format
Added tools to facilitate the incorporation of results from additional gene predictors and RNA-Seq alignment tools (e.g., bigWig and BAM files) into the genome browsers produced by G-OnRamp
An extended Augustus tool Galaxy wrapper, exposing more functionality (*e.g.*, ability to specify search range or add extrinsic hints)*
An improved Hub Archive Creator (HAC) and the JBrowse Archive Creator (JAC) tools (e.g., bug fixes, added support for new track types and custom tracks)*
Annotation Support:
Tools and a workflows to create Apollo instances from JBrowse genome browsers, and to support collaborative genome annotation using Apollo
Improved role-based access control in Apollo to facilitate collaborative annotation in educational settings*
Reporting features for instructor roles in Apollo to enable faculty to monitor student annotation progress*
General Ease of Use:
The G-OnRamp website (http://g-onramp.org), which hosts documentation, training resources and previously processed data
A CyVerse interaction tool to facilitate the data import and export between G-OnRamp and the CyVerse Data Store
JBrowse improvements to display tblastn alignments that span larger genomic regions*
Optimized search index strategies for feature names and descriptions in JBrowse to reduce the number of index files (*e.g.*, Tabix-indexed GFF3 files)*
The ability to look up gene predictions, and the BLAST and BLAT alignments by name (*e.g.*, RefSeq accession numbers) and by description
Links to external database records (*e.g.*, at NCBI, FlyBase) for the tblastn and BLAT alignment tracks
Improved organization, grouping, and labeling of evidence tracks on UCSC Assembly Hubs
Comprehensive training materials based on feedback from the participants of the G-OnRamp beta testers workshops
Deployment:
Automated local and cloud (Amazon EC2) deployments of G-OnRamp with GalaxyKickStart — an Ansible playbook for deploying production Galaxy servers
A G-OnRamp image deployable via CloudLaunch (https://launch.usegalaxy.org) to enable users with limited technical expertise to run G-OnRamp on the cloud (Amazon EC2)
A G-OnRamp image deployable via the Amazon Web Services EC2 console

List of major features developed for the G-OnRamp platform, and improvements made to various software components. Feature and improvement development was driven predominantly by user feedback, most of which was gathered from attendees of our biannual G-OnRamp workshops. While improvements were made throughout the cycle of G-OnRamp development, feedback from these events was a valuable aid to prioritization.

*** Features or improvements that were developed for component services of G-OnRamp which are now generally available for those services.**

Based on the results from an anonymous survey of G-OnRamp workshop participants, we find that the overall response by users has been very good. Both researchers and educators reported that G-OnRamp has facilitated their work (Fig 5). A majority of the respondents found G-OnRamp useful in their research and/or teaching, and planned to continue to use it, including setting up new student research courses.

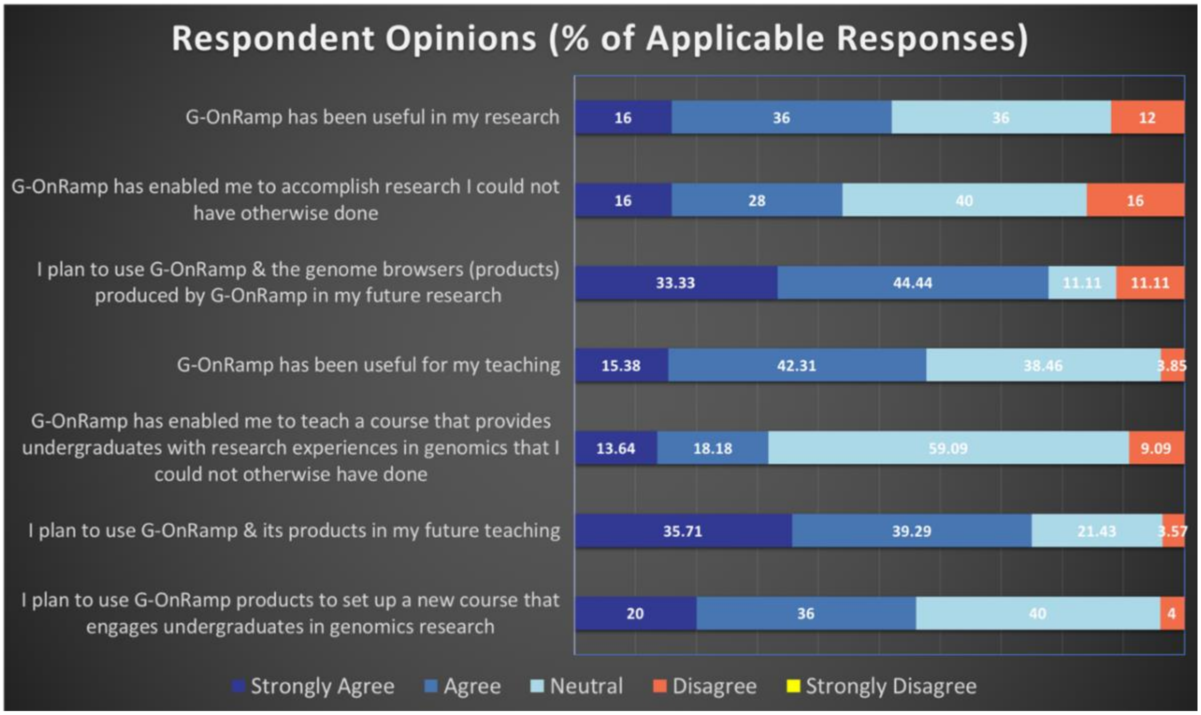


Fig 5. Survey responses on the utility of G-OnRamp. An anonymous survey asked respondents (N = 35 of 53 eligible) to check "strongly agree,", "agree," "neutral," "disagree," or "strongly disagree." Participants ranged from those whose primary occupation is teaching to those managing a research support service (see Fig 4). Consequently from 20% to 38% of the participants checked "not applicable" for any given statement; these responses were removed before percentages were calculated. Overall, participants reported that G-OnRamp facilitates both research and teaching.

G-OnRamp in a CURE: Examining lipid synthesis pathways in parasitoid wasps. As discussed above, many bioinformatics educators have found that a genome annotation project is a good way to introduce students to genomics while providing a research experience. This can be implemented as a one-semester CURE, or as a shorter unit to provide students with an introduction to research.

Many genomics projects that can benefit from careful manual annotation will be focused on a limited set of genes. Because these genes of interest are commonly defined by a shared functional annotation or membership in a specific pathway, they are likely to be dispersed throughout the genome. In the case study presented here, the project is focused on the evolution of lipid synthesis pathways in parasitoid wasps, and so the genes of interest are defined based on their predicted functions rather than their genomic locations. This case was used to test the acceptability and utility of G-OnRamp products in the undergraduate lab.

Fig 6A illustrates the workflow underlying the creation of student annotation projects, in which the approximate locations of the genes of interest are identified in the newly sequenced genomes and assigned as student projects. Fig 6B outlines the approach taken by the student annotator, which is predicated on sequence similarity between the gene of interest in the target genome and genes from an informant genome. The difficulty of the student project primarily depends on the result of the homology search.

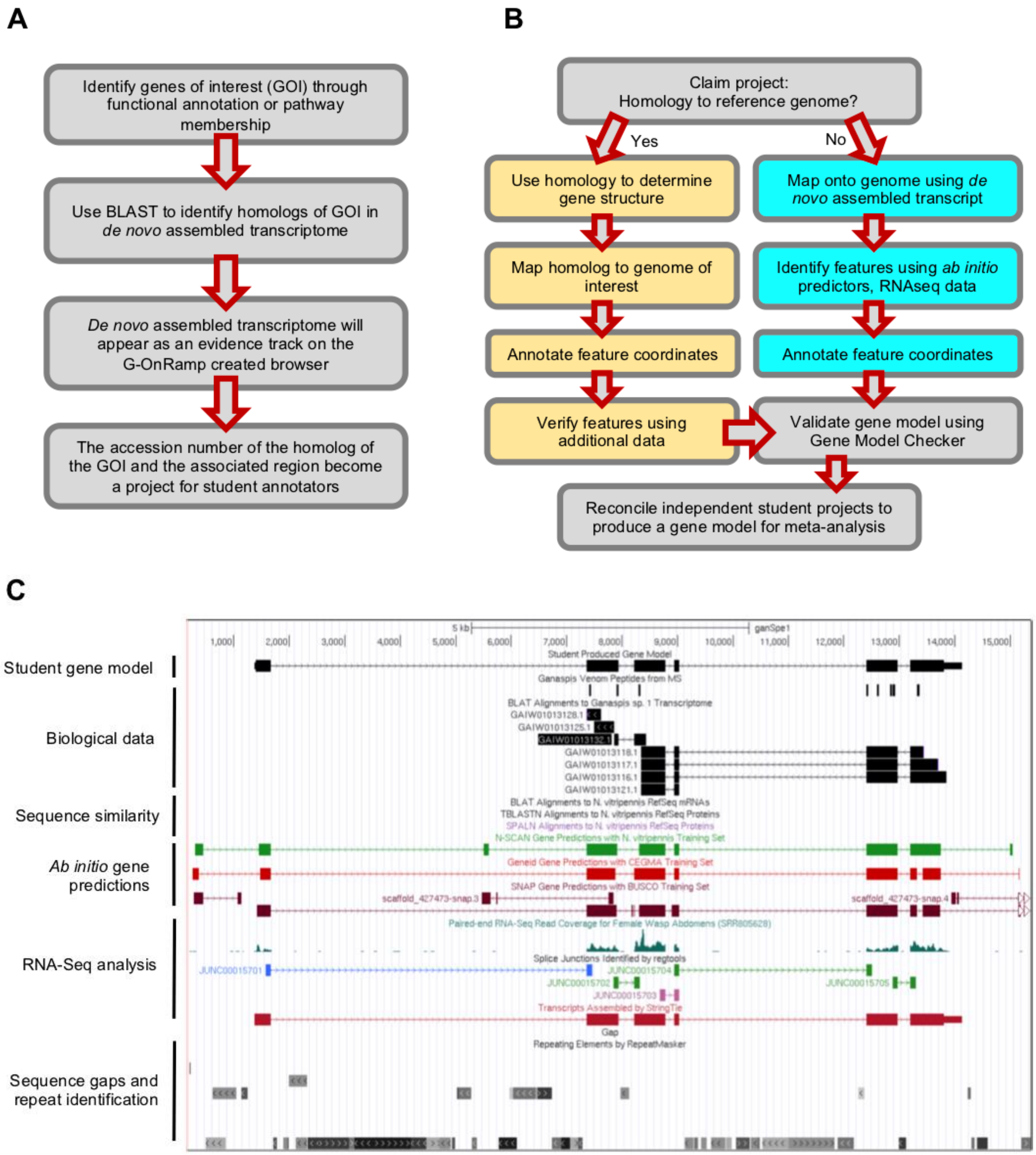

A
Identify genes of interest (GOI) through functional annotation or pathway membership
Use BLAST to identify homologs of GOI in *de novo* assembled transcriptome
De novo assembled transcriptome will appear as an evidence track on the G-OnRamp created browser
The accession number of the homolog of the GOI and the associated region become a project for student annotators
B
Claim project: Homology to reference genome?
Yes
No
Use homology to determine gene structure
Map homolog to genome of interest
Annotate feature coordinates
Verify features using additional data
Map onto genome using *de novo* assembled transcript
Identify features using *ab initio* predictors, RNAseq data
Annotate feature coordinates
Validate gene model using Gene Model Checker
Reconcile independent student projects to produce a gene model for meta-analysis
C
Student gene model
Biological data
Sequence similarity
Ab initio gene predictions
RNA-Seq analysis
Sequence gaps and repeat identification
5 kb
ganSpe1
1,000
2,000
3,000
4,000
5,000
6,000
7,000
8,000
9,000
10,000
11,000
12,000
13,000
14,000
15,000
Student Produced Gene Model
Ganaspis Venom Peptides from MS
BLAT Alignments to Ganaspis sp. 1 Transcriptome
GAIW01013128.1
GAIW01013125.1
GAIW01013132.1
GAIW01013118.1
GAIW01013117.1
GAIW01013116.1
GAIW01013121.1
BLAT Alignments to N. vitripennis RefSeq mRNAs
TBLASTN Alignments to N. vitripennis RefSeq Proteins
SPALN Alignments to N. vitripennis RefSeq Proteins
N-SCAN Gene Predictions with N. vitripennis Training Set
Geneid Gene Predictions with CEGMA Training Set
SNAP Gene Predictions with BUSCO Training Set
scaffold_427473-snap.3
scaffold_427473-snap.4
Paired-end RNA-Seq Read Coverage for Female Wasp Abdomens (SRR805628)
Splice Junctions Identified by regtools
JUNC00015701
JUNC00015704
JUNC00015702
JUNC00015703
JUNC00015705
Transcripts Assembled by StringTie
Gap
Repeating Elements by RepeatMasker

Fig 6. Case study: Annotation using parasitoid wasp G-OnRamp browsers. A. The workflow for identifying genes of interest and creating student annotation projects based on G-OnRamp browsers. B. The student annotation workflow. Students are assigned a project and will then work through either of the two sub-workflows depending on homology of the gene of interest to the reference genome. Boxes in yellow define the sub-workflow for genes with homology to the reference genome; cyan boxes define the sub-workflow for genes lacking homology to the reference genome. C. An example student annotation of a gene with no homology to the reference genomes (*D. melanogaster* or *N. vitripennis*). Survey respondents identified lack of homology to an informant genome as one of the main challenges in annotating new species.

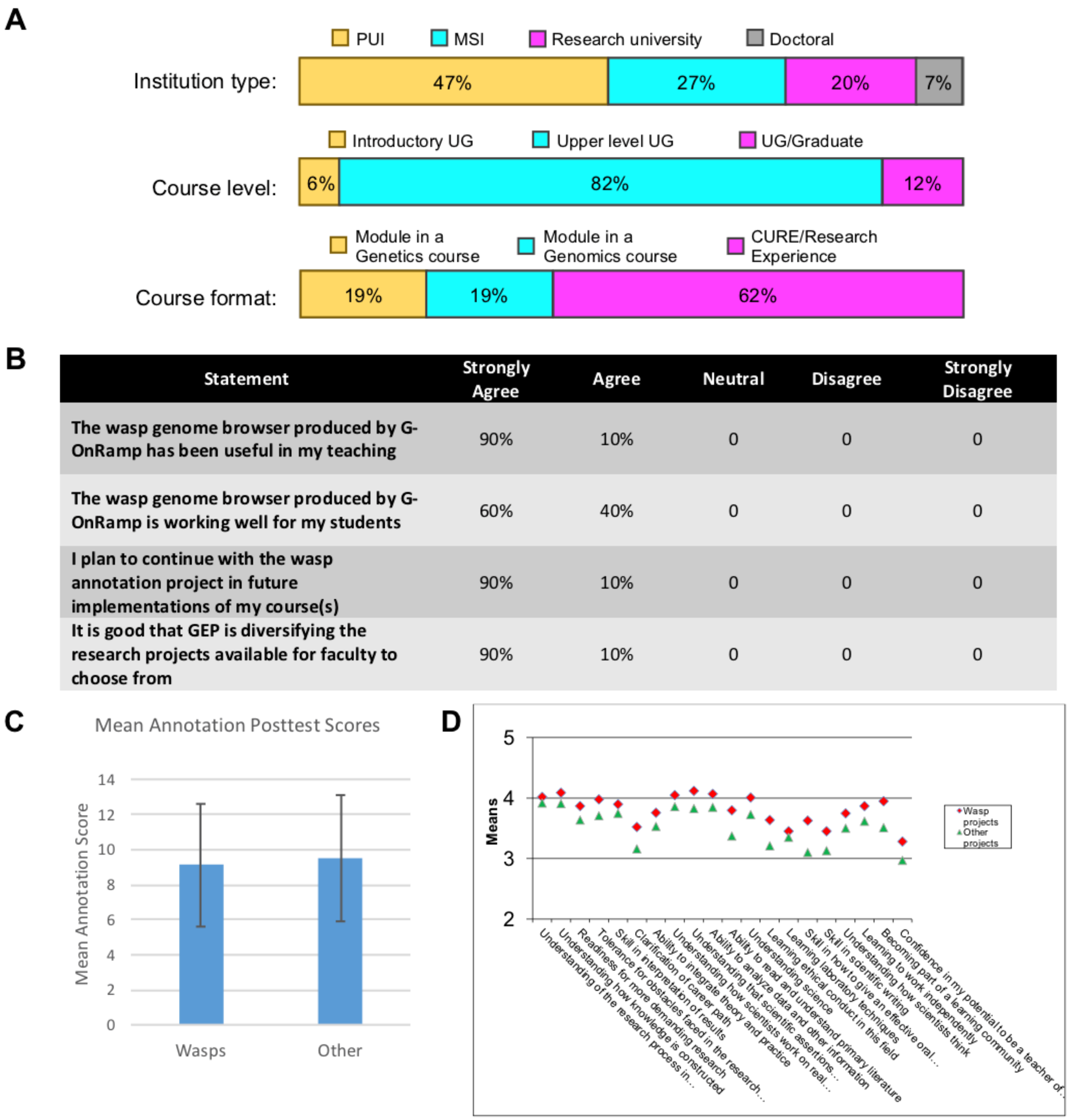
A gene that aligns to an ortholog in a well-studied informant species will not be very difficult for an undergraduate to annotate, while the absence of orthologs will create a challenge. If the gene of interest has significant similarity to a gene in the informant genome, then the student annotator would construct the most parsimonious gene model compared to its putative ortholog in the informant genome. Otherwise, the student annotator would use RNA-Seq data to construct the gene model. Instructors can pre-screen projects to select those at the appropriate level of difficulty for their students.

Fig 6C illustrates an example of a student annotation of a gene that has diverged from the informant genomes (*Nasonia vitripennis* and *Drosophila melanogaster*) such that homology data are not available. The student annotator has to construct a gene model based on other lines of evidence, such as proteomics data, RNA-Seq data (*e.g.*, read coverage, *de novo* transcriptome assembly), and *ab initio* gene predictions. The flexibility of the genome browsers produced by G-OnRamp, and the annotation workflow described above, have facilitated annotation in this case, and should make comparative genomics more accessible for use in the classroom, creating opportunities to study other newly sequenced genomes.

In this pilot implementation of a CURE project using genome browsers generated by G-OnRamp, 15 faculty from the GEP designed CUREs for their students based on the parasitoid wasp research project. These faculty members came from diverse schools (Fig 7A; a full list of faculty with their schools is given in the Acknowledgements). The courses ranged from freshman/sophomore level to those that provided graduate credit. The majority were structured as a research experience.

Responses from an anonymous survey show that most faculty found that the wasp genome browser produced by G-OnRamp worked well for their students, and was generally useful in teaching (Fig 7B). Faculty members who responded to the survey all planned to continue involving their students in the parasitoid wasp project the following year, and all applauded the effort by the GEP/Galaxy partnership to support genomics research broadly.

Direct assessment of the students engaged in a parasitoid wasp CURE was obtained by comparing the responses of this group to those of GEP students as a whole, looking at pooled data from 2017–2018 and 2018–2019. The results show no significant difference in student attainment as exhibited by post-course quiz scores (Fig 7C), indicating that the G-OnRamp-produced genome browsers and the wasp research project are as effective as the UCSC mirror *Drosophila* genome browsers and Muller F element research project in teaching the fundamentals of eukaryotic genes and genomes. Interestingly, there is a small increase in the responses to the SURE survey questions [28], which ask students to self-report perceived gains in the understanding of how science is done and their acquisition of research skills (Fig 7D). This suggests that G-OnRamp can increase student and faculty enthusiasm for genomics research by enabling a variety of projects. Eventually we hope to see multiple collaborative annotation projects that would allow all faculty to participate in a project according to their research interests.



Statement	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
The wasp genome browser produced by G-OnRamp has been useful in my teaching	90%	10%	0	0	0
The wasp genome browser produced by G-OnRamp is working well for my students	60%	40%	0	0	0
I plan to continue with the wasp annotation project in future implementations of my course(s)	90%	10%	0	0	0
It is good that GEP is diversifying the research projects available for faculty to choose from	90%	10%	0	0	0

Fig 7. Using G-OnRamp in a CURE. Classroom implementation with G-OnRamp browsers. A. Implementations of the parasitoid wasp project during 2017-2018 and 2018-2019 characterized by institution type (n = 15), course level (n = 16) and course format (n = 16). Abbreviations: PUI = Primarily undergraduate institution, MSI = Minority-serving institutions, UG = undergraduate, CURE = Course-based Undergraduate Research Experience. B. Results from a survey of faculty who have used a G-OnRamp-generated browser in a course. Participants were asked to respond on a 5-point Likert Scale with N.A. as an option; of the 14 faculty responding to this portion of the survey, the four checking "NA" for these questions were removed before calculating percentage responses, giving n = 10. Responses are shown by percentage of respondents. C. Mean annotation post-course test scores: The mean for the Wasp group is 9.1 (N = 173; SD = 3.6) and the mean for the other GEP students is 9.5 (N = 1185; SD = 3.5). The difference is not significant (bars represent the means; error bars represent one standard deviation). D. Responses to the SURE survey questions: the means for the wasp project students are in red (N ranges from 181 to 195, as some students did not answer all questions) and the means for the other GEP students (working in Drosophila) are in green (N ranges from 1200 to 1270). For some items the wasp group scores significantly higher than the comparison group; however, these results should be interpreted with caution, given the small sample size.

Using G-OnRamp on your own. Steps for acquiring and deploying G-OnRamp, like the platform itself, minimize technical complexity and accelerate data analysis activities. The two principal methods of deployment meet different user needs: 1.) a VirtualBox virtual appliance for small-scale local testing and training and 2.) an Amazon Machine Image (AMI) for cloud-based production deployments. Users can launch the G-OnRamp AMI on Amazon Web Services (AWS) via the CloudLaunch web application (https://launch.usegalaxy.org/; Table 2).

Table 2. Deployment options.

Deployment Option	URL	Notes	Documentation
Virtual Machine (VM) Image	https://ohsu.app.box.com/folder/60271031318	For local testing/training with G-OnRamp; not sufficiently performant for high-scale analysis. However, the VM can be used for smaller genomes, depending on the resources allocated to the VM.	https://wustl.box.com/s/9626q6n2mjnd3vuas26j20w419f5v0fc
AWS via CloudLaunch	https://launch.usegalaxy.org/catalog	For any level of analysis; instance resources configurable by the user. Select 'G-OnRamp' from the Appliance Catalog to launch on AWS without using the console	https://wustl.box.com/s/rg7xaezf22p75d8yardsooa2izbdlkd5
Amazon via AWS Marketplace	https://console.aws.amazon.com/ec2/	For any level of analysis; instance resources configurable by the user. When launching an instance, search for "G-OnRamp" from "Community AMIs"	https://wustl.box.com/s/agjynmu9endhknm37zvr6yfdcshrqa4j

Alternative G-OnRamp deployment methods, their strengths and weakness, and relevant documentation.

For more fine-grained control of the installation and launch of G-OnRamp, the scripts used to create the two principal deployment options are open-source and available on GitHub (https://github.com/goeckslab/gonrampkickstart). This option provides much greater control, but comes with additional complexity that requires technical expertise. For more complex deployment configurations within the AWS infrastructure, a G-OnRamp image can be found under “Community AMIs” when launching an Elastic Cloud Compute (EC2) instance.

Conclusion

The importance and efficacy of providing undergraduates with a research experience is widely accepted. While it is difficult to identify the impact of research *per se* [29], students engaged in a CURE are reported to be both retained in the sciences and to graduate within six years at a higher frequency than matched students who do not have this experience [30]. CUREs in bioinformatics have many advantages, both practical and pedagogical: infrastructure costs are low (only requires computers and Internet connectivity), and there is a large and growing pool of publicly available data, along with tools to manage and analyze that data (*e.g.*, Galaxy, CyVerse). Because no physical lab is required, access is 24/7, and there are no lab safety issues; this situation lends itself to peer instruction, an important multiplier. Perhaps most important, student mistakes are inexpensive in time and money, as the annotation process can be quickly reiterated, problems explored, and investigations taken to the next level.

Recognizing these advantages, a growing number of faculty groups have emerged over the last decade to organize CUREs that include collaborative genome annotation [8, 31, 32, 33]. Recently, several of these groups have come together to form a Genomics Education Alliance (GEA; https://qubeshub.org/community/groups/gea/), which seeks to support this effort by creating a common, well-maintained platform with common curriculum and tools [34]. G-OnRamp removes one bottleneck to CURE growth in bioinformatics by facilitating creation of the genome browsers needed for collaborative genome annotation projects. The G-OnRamp survey results and the parasitoid wasp pilot project have shown G-OnRamp to be a useful tool for researchers and educators alike.

Acknowledgements

We thank the GEP faculty members and their students who participated in the parasitoid wasp project during the last two years: Cindy Arrigo (New Jersey City University), Rebecca Burgess (Stevenson University), Thomas Giarla (Siena College), Rivka Glaser (Stevenson University), Shubha Govind (City College, CUNY), Adam Haberman (University of San Diego), Christopher Jones (Moravian College), Lisa Kadlec (Wilkes University), Adam Kleinschmit (University of Dubuque), Leocadia Paliulis (Bucknell University), Srebrenka Robic (Agnes Scott College), Michael Rubin (University of Puerto Rico at Cayey), Sheryl Smith (Arcadia University), Joyce Stamm (University of Evansville), and Melanie Van Stry (Lane College). Development of G-OnRamp was supported by NIH grant 1R25 GM119157 “A genome browser on-ramp to engage biologists with big data” awarded to SCRE; the work on parasitoid wasps is supported by NIH grants 1R35 GM133760 and 1R03 AG063314 to NTM.

References

1. American Association for the Advancement of Science (2011). Vision and Change in Undergraduate Biology Education: A Call to Action, Washington, DC: https://live-visionandchange.pantheonsite.io/wp-content/uploads/2013/11/aaas-VISchange-web1113.pdf (accessed 18 Sept 2019)

2. Auchincloss LC, Laursen SL, Branchaw JL, Eagan K, Graham M, Hanauer DI, et al. Assessment of Course-Based Undergraduate Research Experiences: A Meeting Report. LSE. 2014 Mar;13(1):29–40.

3. Jordan TC, Burnett SH, Carson S, Caruso SM, Clase K, DeJong RJ, et al. A Broadly Implementable Research Course in Phage Discovery and Genomics for First-Year Undergraduate Students. Losick R, editor. mBio. 2014 Feb 4;5(1):e01051-13.

4. Kowalski JR, Hoops GC, Johnson RJ. Implementation of a Collaborative Series of Classroom-Based Undergraduate Research Experiences Spanning Chemical Biology, Biochemistry, and Neurobiology. Hatfull GF, editor. LSE. 2016 Dec;15(4):ar55.

5. Wetterstrand, Kris, 2019 DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). https://www.genome.gov/27541954/dna-sequencing-costs-data/

6. Wang, Q., C. N. Arighi, B. L. King, S. W. Polson, J. Vincent et al., 2012 Community annotation and bioinformatics workforce development in concert--Little Skate Genome Annotation Workshops and Jamborees. Database 2012: bar064–bar064.

7. Staub, N. L., M. Poxleitner, A. Braley, H. Smith-Flores, C. M. Pribbenow et al., 2016 Scaling Up: Adapting a Phage-Hunting Course to Increase Participation of First-Year Students in Research (S. Elgin, Ed.). CBE—Life Sci. Educ. 15: ar13.

8. Elgin, S. C. R., C. Hauser, T. M. Holzen, C. Jones, A. Kleinschmit et al., 2017 The GEP: Crowd-Sourcing Big Data Analysis with Undergraduates. Trends Genet. 33: 81–85.

9. Hosmani, P. S., T. Shippy, S. Miller, J. B. Benoit, M. Munoz-Torres et al., 2019 A quick guide for student-driven community genome annotation. PLoS Comput. Biol. 15: e1006682.

10. Dunn, N. A., D. R. Unni, C. Diesh, M. Munoz-Torres, N. L. Harris et al., 2019 Apollo: Democratizing genome annotation. PLoS Comput. Biol. 15: e1006790.

11. Campbell, M. S., C. Holt, B. Moore, and M. Yandell, 2014 Genome Annotation and Curation Using MAKER and MAKER-P. Curr. Protoc. Bioinforma. 48: 4.11.1-39.

12. Hoff, K. J., S. Lange, A. Lomsadze, M. Borodovsky, and M. Stanke, 2016 BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS. Bioinforma. Oxf. Engl. 32: 767–769.

13. Humann, J. L., T. Lee, S. Ficklin, and D. Main, 2019 Structural and Functional Annotation of Eukaryotic Genomes with GenSAS, pp. 29–51 in Gene Prediction, edited by M. Kollmar. Springer New York, New York, NY.

14. Papanicolaou, A., 2019 Just Annotate My Genome. https://github.com/genomecuration/JAMg

15. Sallet, E., J. Gouzy, and T. Schiex, 2019 EuGene: An Automated Integrative Gene Finder for Eukaryotes and Prokaryotes, pp. 97–120 in Gene Prediction, edited by M. Kollmar. Springer New York, New York, NY.

16. Afgan, E., D. Baker, B. Batut, M. van den Beek, D. Bouvier et al., 2018 The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Res. 46: W537–W544.

17. Lopatto, D., C. Alvarez, D. Barnard, C. Chandrasekaran, H.-M. Chung et al., 2008 Genomics Education Partnership. Science 322: 684.

18. Raney, B. J., T. R. Dreszer, G. P. Barber, H. Clawson, P. A. Fujita et al., 2014 Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser. Bioinforma. Oxf. Engl. 30: 1003–1005.

19. Buels, R., E. Yao, C. M. Diesh, R. D. Hayes, M. Munoz-Torres et al., 2016 JBrowse: a dynamic web platform for genome visualization and analysis. Genome Biol. 17: 66.

20. Shaffer, C. D., C. Alvarez, C. Bailey, D. Barnard, S. Bhalla et al., 2010 The Genomics Education Partnership: Successful Integration of Research into Laboratory Classes at a Diverse Group of Undergraduate Institutions (B. Wakimoto, Ed.). CBE—Life Sci. Educ. 9: 55–69.

21. Shaffer, C. D., C. J. Alvarez, A. E. Bednarski, D. Dunbar, A. L. Goodman et al., 2014 A course-based research experience: how benefits change with increased investment in instructional time. CBE Life Sci. Educ. 13: 111–130.

22. Leung, W., C. D. Shaffer, L. K. Reed, S. T. Smith, W. Barshop et al., 2015 Drosophila Muller F Elements Maintain a Distinct Set of Genomic Properties Over 40 Million Years of Evolution. G3. 5: 719–740.

23. Leung, W., C. D. Shaffer, E. J. Chen, T. J. Quisenberry, K. Ko et al., 2017 Retrotransposons Are the Major Contributors to the Expansion of the *Drosophila ananassae* Muller F Element. G3. 7: 2439–2460.

24. Liu, Y., L. Sargent, W. Leung, S. C. R. Elgin, and J. Goecks, 2019 G-OnRamp: a Galaxy-based platform for collaborative annotation of eukaryotic genomes (J. Hancock, Ed.). Bioinformatics btz309.

25. Rasche, H., B. Grüning, N. Dunn, and A. Bretaudeau, 2018 GGA: Galaxy for genome annotation, teaching, and genomic databases [version 1; not peer reviewed]. F1000Research 7:1597.

26. Merchant, N., E. Lyons, S. Goff, M. Vaughn, D. Ware et al., 2016 The iPlant Collaborative: Cyberinfrastructure for Enabling Data to Discovery for the Life Sciences. PLoS Biol. 14: e1002342.

27. Nachtweide, S., and M. Stanke, 2019 Multi-Genome Annotation with AUGUSTUS, pp. 139–160 in Gene Prediction, edited by M. Kollmar. Springer New York, New York, NY.

28. Lopatto, D., 2007 Undergraduate Research Experiences Support Science Career Decisions and Active Learning (P. Williams, Ed.). CBE—Life Sci. Educ. 6: 297–306.

29. Committee on Strengthening Research Experiences for Undergraduate STEM Students, Board on Science Education, Division of Behavioral and Social Sciences and Education, Board on Life Sciences, Division on Earth and Life Studies et al., 2017 Undergraduate Research Experiences for STEM Students: Successes, Challenges, and Opportunities (J. Gentile, K. Brenner, & A. Stephens, Eds.). National Academies Press, Washington, D.C.

30. Rodenbusch, S. E., P. R. Hernandez, S. L. Simmons, and E. L. Dolan, 2016 Early Engagement in Course-Based Research Increases Graduation Rates and Completion of Science, Engineering, and Mathematics Degrees (J. Knight, Ed.). CBE—Life Sci. Educ. 15: ar20.

31. Buonaccorsi V, Peterson M, Lamendella G, Newman J, Trun N, Tobin T, et al. Vision and change through the genome consortium for active teaching using next-generation sequencing (GCAT-SEEK). CBE Life Sci Educ. 2014;13(1):1–2.

32. Rosenwald AG, Russell JS, Arora G. The genome solver website: a virtual space fostering high impact practices for undergraduate biology. J Microbiol Biol Educ. 2012;13(2):188–90.

33. Wiley, Emily A., Chalker, Douglas L. A community model for course-based student research that advances faculty scholarship. CUR Quarterly. 37(2):12–4.

34. Elgin, S. C. R., G. Bangera, V. P. Buonaccorsi, D. L. Chalker, E. Dinsdale et al., 2017. A Genomics Education Alliance. https://figshare.com/articles/A_Genomics_Education_Alliance/5197228