Cold Spring Harbor Laboratory
DNA LEARNING CENTER

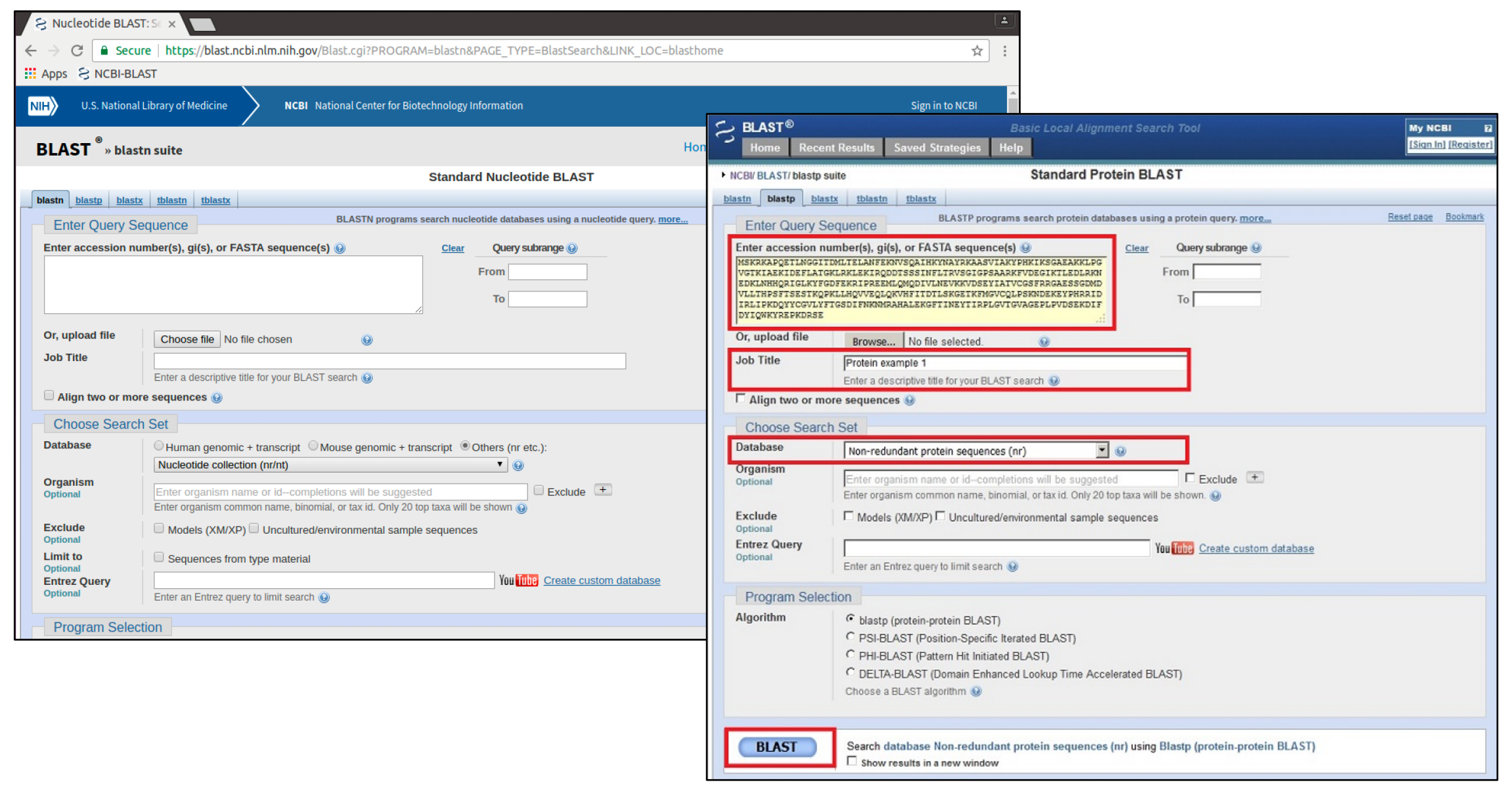
Making bioinformatics tools classroom-friendly

Jason Williams +, Wilson Leung ◆, Cornel Ghiban + and the Genomics Education Alliance
+DNA Learning Center, Cold Spring Harbor Laboratory, williams@cshl.edu; ◆ Washington University in St. Louis; Genomics Education Alliance

GEA
Genomics Education Alliance

Tackling BLAST

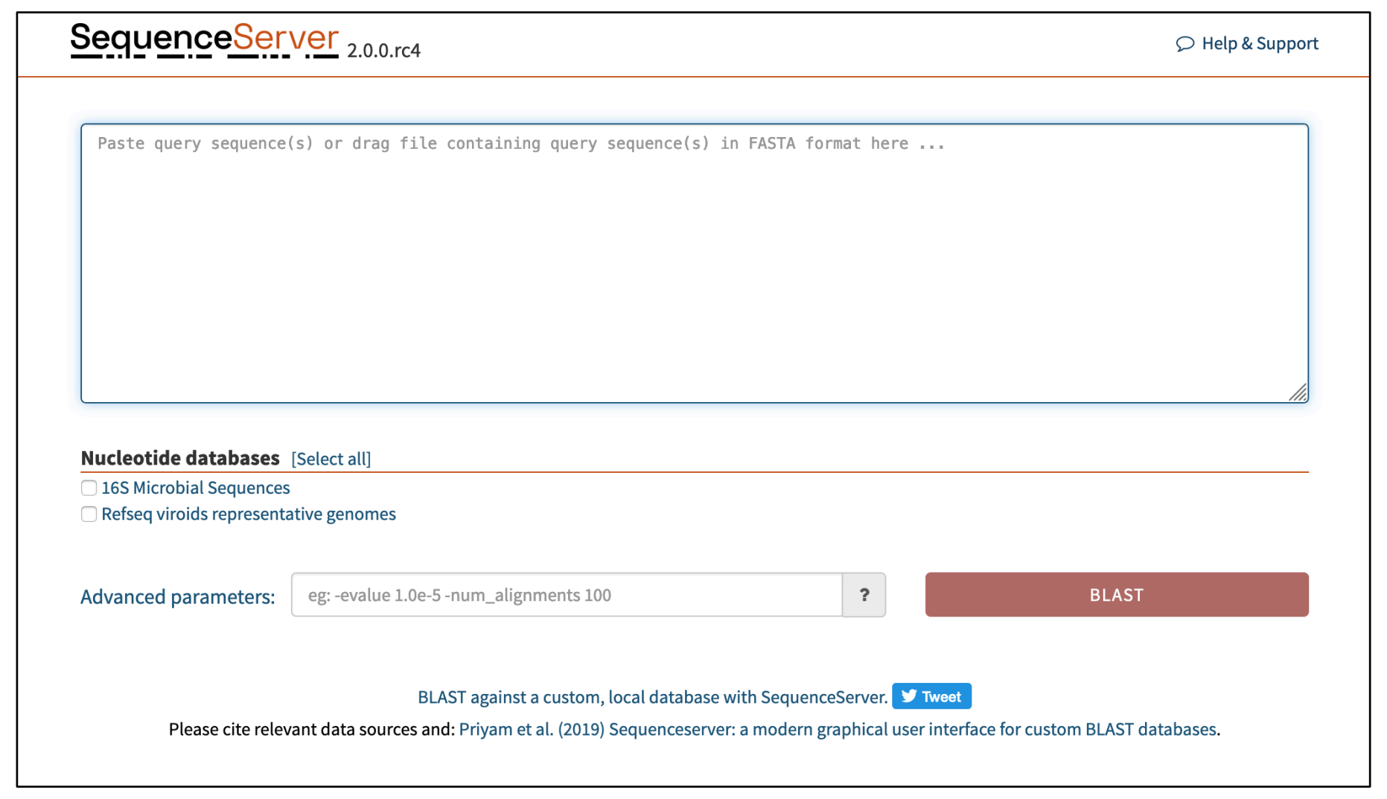
NCBI BLAST is probably the single most explored bioinformatics exercise in the undergraduate classroom. A familiar frustration for faculty is to plan exercises and labs using the NCBI website only to find that shortly before class, the website interface or the BLAST databases have changed (without warning); making carefully documented exercises out of date. This and other frustrations with bioinformatics tools were the reasons the *Genomics Education Alliance (GEA)* formed a Bioinformatics Infrastructure Committee to explore approaches for making bioinformatics tools more classroom friendly.



Although not unreasonable to expect faculty to update classroom exercises, the unpredictable nature of changes and updates to tools they don't control is one more burden to their effective use in the classroom.

GEA Solution: StarBLAST

Working with CyVerse and cyberinfrastructure students at the University of Arizona and the open source *SequenceServer* software, we developed a modular and efficient interface to BLAST. This version-controlled implementation (*StarBLAST*) allows faculty to launch BLAST servers on-demand, with a stable interface.



SequenceServer 2.0.0.rc4

Help & Support

Paste query sequence(s) or drag file containing query sequence(s) in FASTA format here ...

Nucleotide databases [Select all]

☐ 16S Microbial Sequences
☐ Refseq viroids representative genomes

Advanced parameters: eg: -evalue 1.0e-5 -num_alignments 100

BLAST

BLAST against a custom, local database with SequenceServer. Tweet

Please cite relevant data sources and: Priyam et al. (2019) Sequenceserver: a modern graphical user interface for custom BLAST databases.

***SequenceServer* running on CyVerse VICE**

How can bioinformatics tools and computational infrastructure be adapted to increase classroom use?

Faculty who want to teach genomics often need to maintain a variety of computational resources (software tools, websites, servers, etc.) for bioinformatics education. To overcome the challenge presented by the vast, constantly changing set of bioinformatics tools, we wanted to provide a proof-of-concept bioinformatics toolkit – providing a small, but well-maintained menu of software and computational tools (databases, cloud virtual machine images, etc.) that are optimized for bioinformatics instruction and implementation of CUREs. By limiting choice and providing straightforward descriptions, rather than expecting independent exploration and evaluation, we will appeal to faculty with high teaching loads (e.g. at PUIs, MSIs, CCs) who need help making choices efficiently, and have limited time to learn and introduce new materials. The goal is a stable and reliable framework from semester to semester that places the focus on learning genomics concepts, not maintaining bioinformatics tools.

CyVerse VICE – A Platform for Reproducible Software Deployment

CYVERSE

CyVerse (cyverse.org) — is funded by the National Science Foundation (DBI-0735191, DBI-1265383, and DBI-1743442) to develop computational resources for life science. CyVerse's web-based platforms include 1) data store and commons, 2) cloud and HPC computing, and 3) graphical interface to bioinformatics applications, including a visual and interactive computing environment for Jupyter Labs and RStudio applications. A key attribute to CyVerse infrastructure is that this platform supports the storage, sharing, and analyses of large datasets – enabling students to work with the same data and tools used by professional researchers.

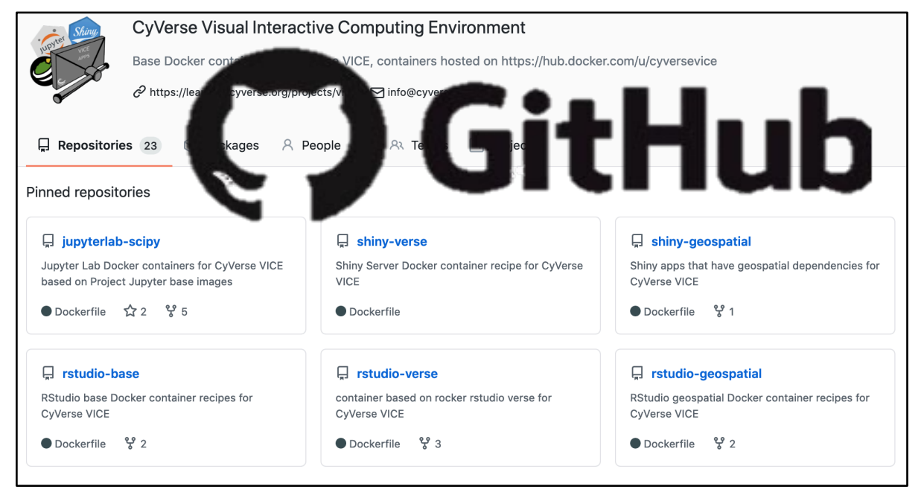
Deploying Apps on CyVerse

1. Identify Software

BIOCONDA python R

Identify your favorite bioinformatics software

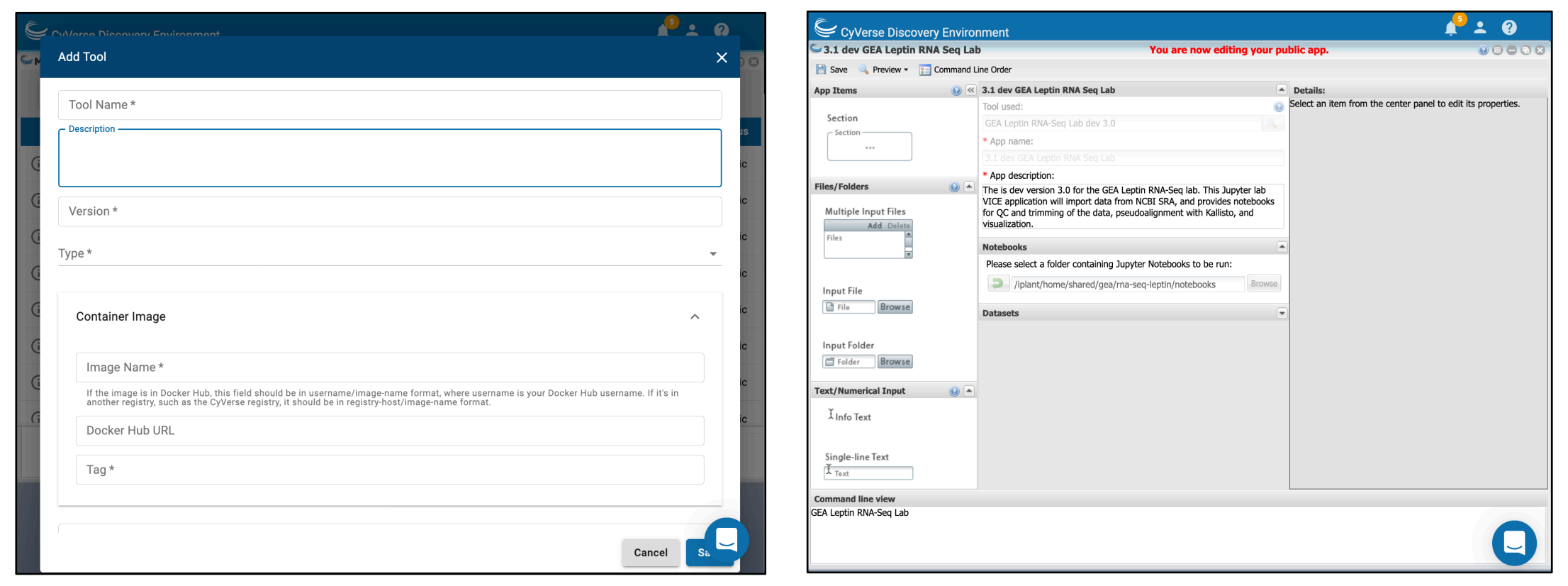
2. Add Software to Docker Container



GitHub docker

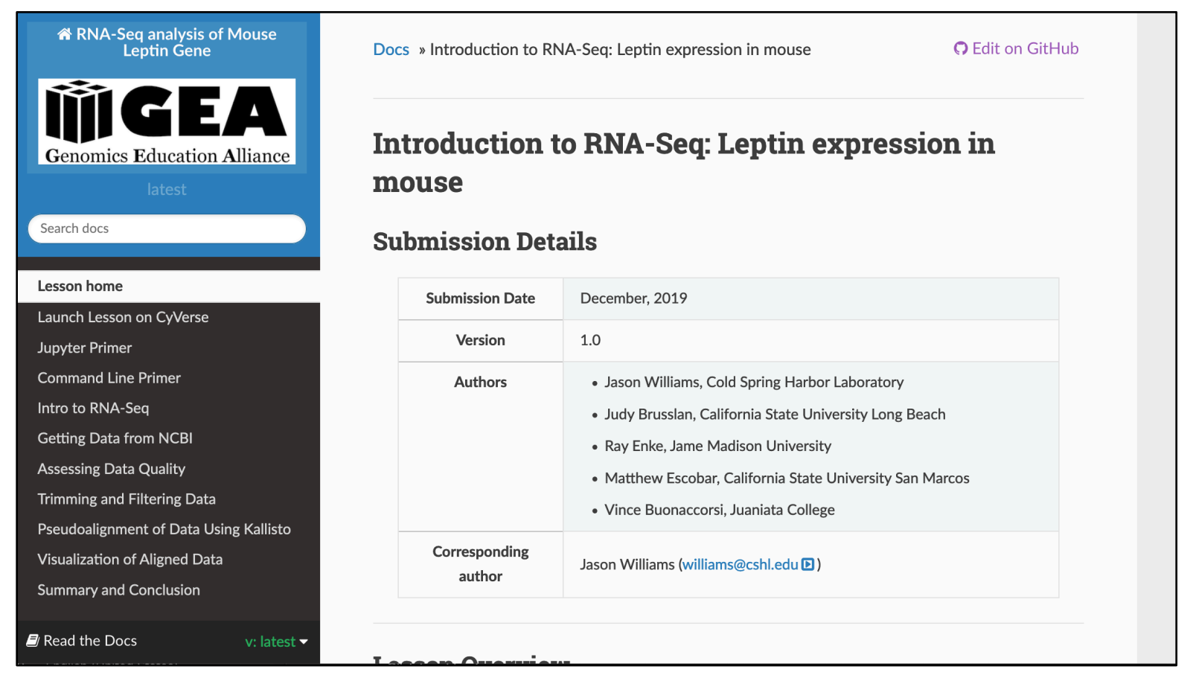
Use a CyVerse template to build Dockerfile

3. Push to CyVerse



Complete App publication; upload test data

4. Document and make it public



Build lesson with GEA ReadTheDocs template

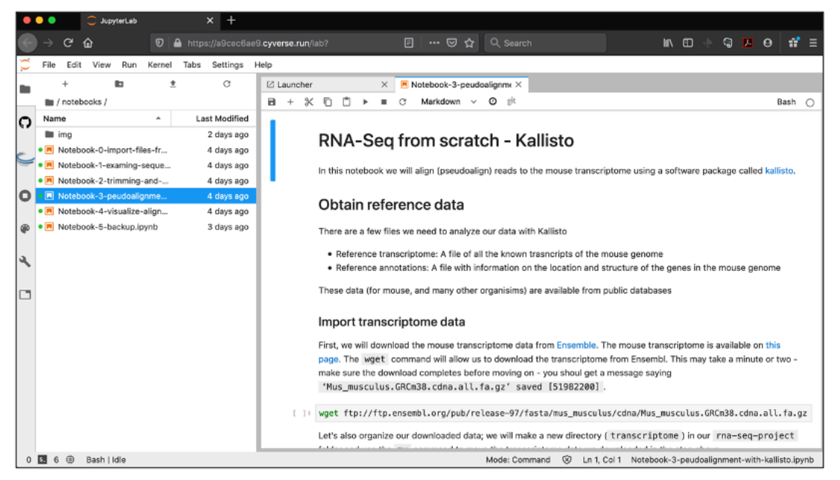
Interactive Apps – Friendly Coding

jupyter RStudio Shiny

Interactive apps such as Jupyter (jupyter.org), Rstudio (rstudio.com), and RShiny apps (shinyapps.io) are popular tools of data science that are increasingly used in the life science to present interactive code. Jupyter and RStudio both have "notebook" applications that allow you to mix command line code in a variety of languages (R, Python, Bash, Julia, etc.) with descriptive text, images, videos, and more. These features allow faculty and students to interact with code-based exercises without having deep expertise with command line computing. As such, these are an on-ramp to bioinformatics tools that guide learners as they gain experience.

The GEA has made use of Jupyter to deploy an RNA-Seq lesson that guides students though an introductory data analysis experiment.

Pilot lessons are available online:
gea.qubeshub.org/lessons



0 - Import from SRA | **1 - Sequence QC** | **2 - Sequencing Trimming** | **3 - Read Pseudoalignment** | **4 - Read Visualization**

jupyter jupyter jupyter jupyter jupyter

A set of 5 Jupyter notebooks for RNA-Seq analysis that goes from data import to visualization

The Genomics Education Alliance

The Genomics Education Alliance (GEA) is a newly NSF-funded research collaboration network funded by NSF (RCN-UBE #1827130). This alliance has brought together several well-known educators and educational projects. The work presented on this poster is the result of work of this network. Learn more and join at:
gea.qubeshub.org