CSH Cold Spring Harbor Laboratory DNA LEARNING CENTER



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

#### **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the website interface or the BLAST databases have changed (without warning); making carefully documented exercises out of date.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.

## 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 Rstudio optimized for bioinformatics instruction and implementation of CUREs. By limiting choice and providing straightforward (shinyapps.io) are popular tools of data science descriptions, rather than expecting independent exploration and evaluation, we will appeal to faculty with high teaching that are increasingly used in the life science to loads (e.g. at PUIs, MSIs, CCs) who need help making choices efficiently, and have limited time to learn and introduce present interactive code. Jupyter and RStudio both new materials. The goal is a stable and reliable framework from semester to semester that places the focus on learning have "notebook" applications that allow you to mix genomics concepts, not maintaining bioinformatics tools.

ucleotide BLAST: S ×			<b>(</b>		
C Secure https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE TY	VPE-BlastSearch&UNK_LOC-blasthome		* :		
S S NCBI-BLAST	r L-bioscscalengenike_coc-bioschome				
U.S. National Library of Medicine NCBI National Center for Biotechnology Information	n	Sign in to NC	ві		
ST <sup>®</sup> » blastn suite	Hon S BLAST	® Recent Results Saved Strategi	Basic Local Alignme ies Help	ent Search Tool	My NCBI 2 [Sign In] [Register]
Standard Nucle	eotide BLAST    NCBI/ BLAS	T/ blastp suite	Standard Prot	tein BLAST	
blastp blastx tblastn tblastx	blastn bla	stp <u>blastx tblastn</u> tblastx			
ter Query Sequence BLASTN programs search nucleotide data r accession number(s), gi(s), or FASTA sequence(s)	y subrange 🕢 Enter act VSKRKAP/ VSKRKAP/	Query Sequence ession number(s), gi(s), or FASTA sea stlnggitdmltelanfeknysgaihkynayra idbflatgkurklakirgddfsssinflrysg	quence(s) 🕑 KAASVIAKYPHKIKSGAEAKKLPG GIGPSAARKEVDEGIKILEDLARN	ases using a protein query. <u>more</u> <u>Clear</u> <u>Query subrange @</u> From	<u>Reset page</u> <u>Bookmark</u>
το	VLLTHPS	RIGLKYFGDFEKRIPREENLOMODIVLNEVKK TSESTKOPKLLHOVVEQLOKVHFITDTLSKGEI YYCGVLYFTGSDIFNKNMRAHALEKGFTINEYI EPKDRSE	TKFMGVCQLPSKNDEKEYPHRRID	То	
pload file Choose file No file chosen	Or, uploa	d file Browse No file select	ted. 😡		
Title	Job Title	Protein example 1			
Enter a descriptive title for your BLAST search 😡		Enter a descriptive title for yo	our BLAST search 🕖		
gn two or more sequences 😡	□ Align	wo or more sequences 🥹			
oose Search Set	Choos	e Search Set			
Human genomic + transcript OMouse genomic + transcript Others (nr	r etc.): Database	Non-redundant protein se	quences (nr)	0	
Ism Enter organism name or Id-completions will be suggested	Crganism     Optional	Enter organism name or id	d-completions will be suggested	d Exclude 🛨	
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown (			ame, binomial, or tax id. Only 20 to		
de General Models (XM/XP) Uncultured/environmental sample sequences	Optional	I Models (XIW/XP) I Unc	ultured/environmental sample se	equences	
to Sequences from type material	Entrez Q Optional	Enter an Entrez query to limit	t search 😣	You Tube Create custom database	
A Contract of the search and the sea	Tube Create custom database Progra	m Selection			
ogram Selection	Algorithm	lastp (protein-protein E			
		C PSI-BLAST (Position-S C PHI-BLAST (Pattern Hit C DELTA-BLAST (Domain Choose a BLAST algorithm	t Initiated BLAST) n Enhanced Lookup Time Accele	erated BLAST)	
	BLA	St Search database Non-re-		nr) using Blastp (protein-protein BLAST)	

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.

Sequence <mark>Ser</mark>	V <u>er</u> 2.0.0.rc4		
Paste query sequence	(s) or drag file containing query sequence(s) in FASTA	format here	
L			
Nucleotide databases			
<ul> <li>16S Microbial Sequences</li> <li>Refseq viroids representation</li> </ul>			
	0		
Advanced parameters:	eg: -evalue 1.0e-5 -num_alignments 100	?	BLAST
	BLAST against a custom, local database with Sequ	JenceServer. 🎔 Tweet	
Please cite relev	ant data sources and: Priyam et al. (2019) Sequenceserver: a mode		for custom BLAST databases.





	rea Discovery Environment
То	ool Name *
De	scription
Ve	ersion*
ype	5 ×
	e *
	Container Image
	Container Image Image Name *
	Container Image Image Name * If the image is in Docker Hub, this field should be in username/image-name format another registry, such as the CyVerse registry. It should be in registry-host/image-n

Complete App publication; upload test data

SequenceServer running on CyVerse VICE

# Making bioinformatics tools classroom-friendly

#### **CyVerse VICE – A Platform for Reproducible Software Deployment**

CyVerse (cyverse.org) — is funded by the National Science Foundation (DBIcommand line computing. As such, these are an 0735191, DBI-1265383, and DBI-1743442) to develop computational resources on-ramp to bioinformatics tools that guide learners for life science. CyVerse's web-based platforms include 1) data store and as they gain experience. commons, 2) cloud and HPC computing, and 3) graphical interface to bioinformatics applications, including a visual and interactive computing The GEA has made use of Jupyter to deploy an environment for Jupyter Labs and RStudio applications. A key attribute to RNA-Seq lesson that guides students though an CyVerse infrastructure is that this platform supports the storage, sharing, and introductory data analysis experiment. analyses of large datasets – enabling students to work with the same data and tools used by professional researchers. **Pilot lessons are available online:** 

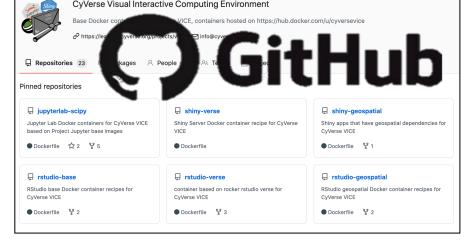
#### **Deploying Apps on CyVerse**

#### **1. Identify Software**

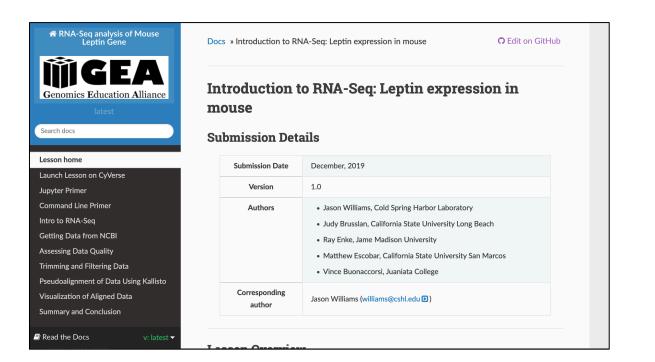
Identify your favorite bioinformatics software

#### **3. Push to CyVerse**

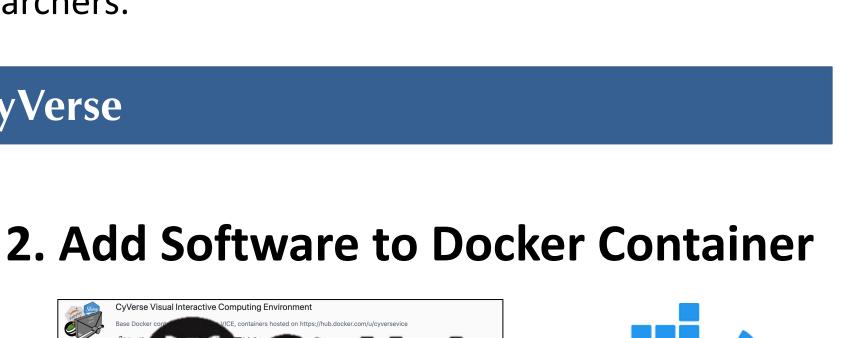
See Constant of Constant		Sal dev GEA Leptin	RNA Seq Lal	nment You are no	ow editing your p	ublic app.	般 🗆 🖨 🖓 🗙
Section		💾 Save 🔍 Preview 🔹	:: Command L	ine Order			
Section     Section       His/Folders     Section       Input File     Section       Input File     Recesse       C     Input File       C     Inp		App Items		3.1 dev GEA Leptin RNA Seq Lab			
Section   C   Files/Folders   Files/Folders   Multiple Input Files   Files   Files   Files   C   Input File   Input Folder   Input File   Input File   Input File   Input Folder   Input Folder				Tool used:	6	Select an item from the center panel to edit	ts properties.
<ul> <li>App dame:</li> <li>App dame:</li> <li>App dame:</li> <li>App description:</li> <li>App description:</li> <li>The is dev version 3.0 for the GEA Leptin RNA-Seq lab. This Jupyter lab</li> <li>App description:</li> <li>The is dev version 3.0 for the GEA Leptin RNA-Seq lab. This Jupyter lab</li> <li>App description:</li> <li>The is dev version 3.0 for the GEA Leptin RNA-Seq lab. This Jupyter lab</li> <li>App description:</li> <li>The is dev version 3.0 for the GEA Leptin RNA-Seq lab. This Jupyter lab</li> <li>App description:</li> <li>The is dev version 3.0 for the GEA Leptin RNA-Seq lab. This Jupyter lab</li> <li>App description:</li> <li>The is dev version 3.0 for the GEA Leptin RNA-Seq lab. This Jupyter lab</li> <li>App description:</li> <li>The is dev version 3.0 for the GEA Leptin RNA-Seq lab. This Jupyter lab</li> <li>Add The Plane Seq lab. This Jupyter lab. This Jupyter lab</li> <li>Add The Plane Seq lab. The Plane Seq lab. This Jupyter lab. The Plane Seq lab. The Plan</li></ul>				GEA Leptin RNA-Seq Lab dev 3.0			
Files/Folders Multiple Input Files Input File Input File Input File Text/Numerical Input Text/Numerical Input Text/Numerical Input				* App name:			
Piles/Folders     Imput File       C     Imput File       D     Trian       Browse     Datasets	ic						
C Multiple Input Files Multiple Input Files Multiple Input Files C Input File C In							
C     Input File       C     Input File       C     Input File       C     Input File       Datasets       C     Text/Numerical Input	ic	Multiple Input Files Add Delete	0	VICE application will import data from NCBI SRA, and for QC and trimming of the data, pseudoalignment wit	provides notebooks		
c     Input File     Please select a folder containing Jup/ter Notebooks to be run:       nput File     input File     input File       c     Input File     Datasets       c     Input Folder     input File       if rider     Browse     Datasets       c     Input Folder     Input Folder       Text/Numerical Input     image	IC			Notebooks	-		
C Input File Browse C Input Folder C Text/Numerical Input @ •	·	×		Please select a folder containing Jupyter Notebooks to	be run:		
C Input Folder C Text/Numerical Input	ic						
C Text/Numerical Input	~ <sup>IC</sup>	File Browse		Datasets		•	
C Text/Numerical Input	ic						
		Folder Browse					
			0				
L Info Text	ic	I Info Text					
Single-line Text							
Ĭ <sub>Test</sub>		1 Text					
Command line view GEA Leptin RNA-Seq Lab							

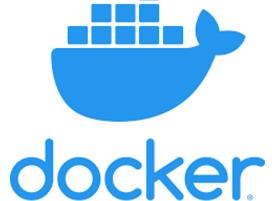


### 4. Document and make it public



Build lesson with GEA ReadTheDocs template





Use a CyVerse template to build Dockerfile



#### **Interactive Apps – Friendly Coding**



Interactive apps such as Jupyter (jupyter.org), (rstudio.com), and 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

# gea.qubeshub.org/lessons

4	→ C ŵ	▼ 🔒 https://a9	rcec6ae9.cyverse.run/lab?	🕫 🚥 🔽 🔍 Search	IN E 🔶 🤤 🖪 😖
đ	File Edit View Run	Kernel Tabs Sett			
	+ 10	± 0	C 🛛 🖄 Launcher	× 🗷 Notebook-3-peudoalignme ×	
	/ notebooks /		B + % E	) 🗂 ⊨ 🔳 ♂ Markdown ∨ Ø git	Bas
0	Name	<ul> <li>Last Mo</li> </ul>			
	<ul> <li>img</li> <li>Notebook-0-import-fil</li> </ul>		<sup>/s ago</sup> R	NA-Seq from scratch - Kal	listo
9	Notebook-1-examing-:			-	
	Notebook-2-trimming-		/s ago	this notebook we will align (pseudoalign) reads to the	e mouse transcriptome using a software package called kalliste
0			/s ago	btain reference data	
	Notebook-4-visualize-		/s ago		
æ	• Notebook-5-backup.ip	iyno 3 day	/s ago Th	ere are a few files we need to analyze our data with F	Callisto
2				Reference transcriptome: A file of all the known tra	
					e location and structure of the genes in the mouse genome
			Th	ese data (for mouse, and many other organisims) are	available from public databases
			In	port transcriptome data	
					rom Ensemble. The mouse transcriptome is available on this he transcriptome from Ensembl. This may take a minute or two
				ke sure the download completes before moving on -	
			9	<pre>lus_musculus.GRCm38.cdna.all.fa.gz' saved</pre>	[51982200].
			E 1: Mg	et ftp://ftp.ensembl.org/pub/release-97/fas	ta/mus_musculus/cdna/Mus_musculus.GRCn38.cdna.all.fa
			Le	's also organize our downloaded data; we will make a	a new directory ( transcriptome ) in our rna-seq-projec
	5. 6 🗇 Bash   Idle			Mode: Comma	nd 🛞 Ln 1, Col 1 Notebook-3-peudoalignment-with-kallisto
0					
		-			_
		l - Sequ	uence	2 - Sequencing	g 3 - Read
t f		- Sequ	uence	2 - Sequencing	-
		l - Sequ Q	uence C		g 3 - Read Pseudoalignment
t f		l - Sequ Q(	uence C		-
t f		- Sequ Q(	uence C		-
t f		L - Sequ Qu	uence C		-
t f	rom 1	Q	c		Pseudoalignment
t f	rom 1	L - Sequ Qa	c		-

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

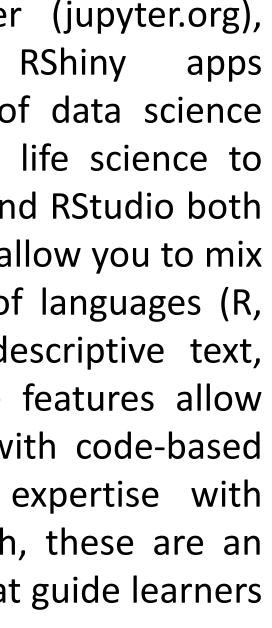
#### **The Genomics Education Alliance**

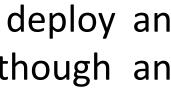
The Genomics Education Alliance (GEA) is a newly NSF-funded research collaboration network funded by NSF (RCN-UBE #1827130). This alliance brought together several has educators and educational projects. The work presented on this poster is the result of work of this network. Learn more and join at:

gea.qubeshub.org



Shiny,





alizatio upyter

